Reducing the credit gap in Mexico in an environment of uncertainty generated by the COVID-19 pandemic: A data science approach (machine learning)

Rodríguez-García, Jair Hissarly and Venegas-Martínez, Francisco

Instituto Politécnico Nacional, México, Instituto Politécnico Nacional, México

4 January 2021

Reducción de la brecha del crédito en México en un ambiente de incertidumbre generada por la pandemia COVID-19: Un enfoque de ciencia de datos (*machine learning*)

(Reducing the credit gap in Mexico in an environment of uncertainty generated by the COVID-19 pandemic: A data science approach (machine learning))

Jair Hissarly Rodríguez-García
Instituto Politécnico Nacional, México
jair.hissarly@hotmail.com

Francisco Venegas-Martínez
Instituto Politécnico Nacional, México
fvenegas1111@yahoo.com.mx

Resumen

El otorgamiento de microcréditos de forma eficiente y transparente a través de plataformas digitales a individuos que desarrollan actividades económicas y que buscan mantener su empleo y el de sus trabajadores y que no tienen acceso al sistema financiero convencional es, sin duda, un problema urgente por resolver en la crisis sanitaria por la que atraviesa actualmente México. La presente investigación desarrolla varios modelos y estrategias de riesgo de crédito que permiten promover la inclusión crediticia en México de manera justa y sostenible en un ambiente de incertidumbre generada por los estragos presentes y esperados por la pandemia COVID-19. Para ello se utiliza el enfoque de ciencia de datos de *machine learning,* particularmente, se emplean las herramientas: regresión del árbol de decisión, bosques aleatorios, función de base radial, *boosting*, *K-Nearest Neigbor* (KNN) y Redes Neuronales.

Clasificación JEL: G32, E44, G23, G53
Palabras clave: riesgo crédito, ciencia de datos, mercados de créditos, instituciones financieras, inclusión financiera.

Abstract

The efficient and transparent granting of microcredits through digital platforms to people who carry out economic activities and who seek to maintain their employment and that of their workers and who do not have access to the conventional financial system is, without a doubt, an urgent problem be solved in the health crisis that Mexico is going through. This research develops various credit risk models and strategies that allow promoting credit inclusion in Mexico in a fair and sustainable manner in an environment of uncertainty generated by the present and expected ravages of the COVID-19 pandemic. For this, the data science approach of machine learning is used, in particular, the used tools are: decision tree regression, random forests, radial basis function, *boosting*, *K-Nearest Neigbor* (KNN), and Neural Networks.

JEL classification: G32, E44, G23, G53
Keywords: credit risk, data science, credit markets, financial institutions, financial inclusion.

1. Introducción

De acuerdo con el Instituto Nacional de Estadística y Geografía (INEGI, 2018) en su Encuesta Nacional de Inclusión Financiera (ENIF) se destaca que de la población de 18 a 70 años, sólo un 68% tiene acceso a algún tipo de producto financiero, mientras que las personas que poseen acceso al crédito, sólo representan un tercio de la población, existiendo una marcada distinción entre los estados del norte y el sur del país. Asimismo, el 11 de febrero de 2020 la Organización Mundial de la Salud reconoce un nuevo virus, coronavirus de tipo 2 causante del síndrome respiratorio agudo severo (SARS-CoV-2), también conocido como COVID-19. En muy poco tiempo, el SARS-CoV-2 fue catalogado epidemia por los efectos en el continente europeo. Posteriormente, Estados Unidos de América de Norte América se vio vulnerado por el nuevo coronavirus. Para el 28 de febrero de 2020, el coronavirus arribó a México. Las autoridades sanitarias del país reportaron el primer caso positivo de coronavirus en la CDMX. Los alarmantes niveles de transmisibilidad del SARS-CoV-2 orillaron a las autoridades sanitarias a catalogarlo como pandemia el miércoles 11 de marzo de 2020 (OMS, 2020). El efecto de ello abundó en la afectación de México, al menos siete de cada 10 mexicanos consideran que la emergencia sanitaria afectó ya su economía familiar, del 71% de los mexicanos que reconoció estar afectado en su economía por la emergencia sanitaria, el 44% respondió que mucho y el 27.7% señaló que poco. (Expansión Política, 2020). A finales de julio se daba a conocer que el país contaba con más de 40,000 fallecimientos relacionadas con el virus, ubicándose en entre los primeros lugares de decesos en el mundo (Universidad Johns Hopkins). Para el tercio de la población que podría tener acceso a un crédito de acuerdo con la ENIF del INEGI, en las condiciones actuales es a todas luces insuficientes para cubrir las familias y microempresas afectadas, por lo cual es necesario aumentar la inclusión de microcréditos.

Desde hace muchos años se han enfatizado las ventajas que tiene un país cuando posee mayor acceso al crédito, (Beck *et al*., 2000) en su trabajo explican, a través de un modelo de panel de datos, los efectos del desarrollo financiero sobre la economía real. Particularmente, los autores destacan que el acceso al crédito acelera la productividad total de factores, repercutiendo positivamente en el crecimiento económico a largo plazo. Continuando con esta idea, (Perossa y Gigler, 2015) realizan un estudio, en donde analizan las implicaciones del microcrédito en diferentes países Latinoamericanos, los resultados de la incursión de este tipo de crédito fue relativamente exitoso en la mayoría de los países observados, ya que en general, se percibe una disminución de la pobreza en el aumento y expansión del microcrédito, siendo peculiarmente anómalo en México. En este sentido, los autores concluyen, de acuerdo con su metodología de análisis, que el microcrédito no repercutió en la disminución de la pobreza en México; además de mencionar que la disminución o aumento de la pobreza se vio colateralmente afectada por el dinamismo general de la economía. Este efecto negativo se ha potenciado en México con la pandemia COVID-19.

Por otra parte, (Esquivel, 2017) podría tener una sugerencia parcial con respecto a este fenómeno, ya que argumenta que las características del microcrédito al menos en México y en Perú es el aprovechamiento de la vulnerabilidad de los más desfavorecidos a través del otorgamiento de recursos para enfrentar una contingencia a una alta tasa de

interés. Es por ello, que es urgente el diseño e implementación de modelos y estrategias integrales de riesgo que permitan a esa población que ha sido excluida del sistema financiero, obtener acceso al crédito a través de los diferentes intermediarios financieros que existen en México; teniendo en cuenta el ajuste por riesgo de las condiciones *a priori* del crédito (tasa de interés, plazo y capacidad de pago)

La pregunta de investigación es ¿Cuál sería una estrategia integral óptima desde el punto de vista de riesgo de crédito para enfrentar la brecha financiera en México en el contexto de la pandemia del COVID-19? La hipótesis es: la inclusión financiera se puede realizar a través de la mitigación del riesgo de crédito, minimizando las posibles pérdidas dadas el riesgo, y esto únicamente se logra sí y solo sí se cumplen estas tres condiciones:

1.) Se obtiene el mejor desempeño de un modelo, es decir, que dado un conjunto de datos se llega a los máximos indicadores de separación (KS, GINI, ROC)
2.) El punto de corte (*cutoff*) es optimizado, es decir que hay "punto de corte único", que permite que se maximice la precisión por categoría, y que a su vez se minimice la proporción de falsos positivos (FP) de tal manera que se maximiza el costo beneficio, así como el costo de oportunidad de la institución financiera.
3.) Una vez obtenido el punto de corte "optimizado", todos los sujetos de crédito se les ajusta la capacidad de pago, así como la tasa de interés, de tal manera que se la posible pérdida se minimice sin que esto sugiera la exclusión de los posibles de sujetos de crédito.

El objetivo principal de esta investigación es el desarrollo de un modelo integral de "originación" de riesgo de crédito orientado a disminuir la brecha de inclusión financiera en México, optimizando el nivel de rentabilidad de las Instituciones Financieras. Otros objetivos específicos son:

1.) Elaborar *scores* de crédito basados en técnicas econométricas y *Machine Learning* que permitan perfilar mejor la probabilidad de impago de prospectos con y sin antecedentes crediticios.
2.) Generación de un modelo ajustado por riesgo que permita asignarle a cada individuo una tasa y una capacidad de pago en función de su probabilidad de default.
3.) Estimar el ingreso por interés, estimaciones preventivas de riesgo de crédito, costos financieros y pérdidas por incobrabilidad derivados de las operaciones concedidas por dichos modelos.
4.) Evaluación del comportamiento en segmentos de crédito que la "banca tradicional" no atiende.

Para lograr los objetivos planteados se utiliza ciencia de datos y en particular *Machine Learning* (ML). Esta metodología pertenece a la Inteligencia Artificial (IA) y permite que un algoritmo pueda aprender de los datos en lugar de aprender de la programación explícita. Conforme al algoritmo de un modelo se le introducen datos de entrenamiento, el modelo tiene usualmente un mejor desempeño basado en datos. Si se introducen datos en el modelo predictivo para que éste aprenda, el modelo proporcionará un pronóstico basado en los datos que entrenaron al modelo. Lo más importante de ML es que ésta permite que los

modelos se entrenen con conjuntos de datos antes de ser implementados. Usualmente la complejidad y tamaño de estos modelos, no permite percibir patrones. No obstante, después de que un modelo ha sido entrenado, se pueden identificar patrones a través del aprendizaje con datos y, posteriormente, utilizar el modelo en tiempo real para seguir aprendiendo de los datos. Esta investigación utiliza el enfoque de ciencia de datos de ML para identificar las variables que son candidatas para distinguir entre clientes potenciales con distintas características y comportamientos para incluirlas en el modelo en el proceso de entrenamiento, particularmente se utilizan las herramientas de regresión del árbol de decisión, bosques aleatorios, función de base radial, Boosting, K-Nearest Neigbor (KNN) y redes neuronales, entre otras.

El presente trabajo se encuentra organizado de la siguiente manera: la sección 2 trata con los antecedentes y la situación actual en el contexto de la emergencia sanitaria y la crisis económica en México; la sección 3 se realiza una discusión detallada sobre las características que tiene el microcrédito y la justificación del enfoque a dicho producto; y la sección 4 en adelante proporcionan el fundamento matemático y de ciencia de datos para la aplicación de las estrategias de riesgo de crédito necesaria para abatir la brecha financiera en un ambiente de incertidumbre generada por el COVID-19.

La metodología se establece a continuación. En primer lugar, se realizarán dos *Scorings* reactivos, en donde se utilizaron fundamentalmente bases de datos de créditos simples de una Institución Financiera que por confidencialidad se mantendrá en anonimato, adicionalmente, se complementó con Bases de Datos externas. Esta información contiene una población total de créditos otorgados desde enero 2014 hasta enero de 2019 cuya población es específicamente el segmento de Microcrédito Individual, dentro de esta población, se pueden distinguir en dos grupos muy peculiares: *Clientes Hit*[1] *y Clientes No Hit*[2].

Esta información contiene los días de atraso a cierre de mes por crédito, las características, atributos de los acreditados, algunas variables asociadas con el historial de crédito y condiciones generales de los contratos con los cuáles fueron concedidos los créditos. Todo el análisis se realizó a través del software R que es un *open source*, el cuál es muy flexible para hacer cada uno de los cálculos y procesos necesarios para el desarrollo de modelos. La metodología se divide en:

1.) **Análisis de Costo de un malo:** Consiste en generar un panorama muy general acerca de lo que cuesta a la institución el adquirir un crédito malo.

2.) **Análisis de Cosechas**: En este punto, el objetivo es encontrar el número de meses necesarios de vida para que un cliente se torne como malo o alcance una mora indeseable para la institución que ejecuta el modelo.

[1] Concentra una población de clientes cuya característica central es que han tenido experiencia crediticia previa en alguna institución de crédito y que actualmente mantienen el historial de crédito activo con cuentas que son válidas para generar score genérico.

[2] Este grupo tiene como principal característica que son clientes que no han tenido ninguna experiencia de crédito previa (en su mayoría), sin embargo, también contiene aquella población que no cuentan con el historial crediticio robusto para generar score genérico.

3.) **Definición de un Bueno o Malo**: Aquí se analiza a partir de qué estado, dada una Cadena de Markov, el crédito presenta una alta probabilidad de migración a un punto específico de absorbencia[3]. Esos umbrales, dadas las probabilidades de migración nos definirán que es un crédito "Bueno", "Malo" o "Indeterminado".

4.) **Validación de Datos**: En este punto se explica que información tiene cada variable, adicionalmente se realiza un análisis estadístico básico basado en medidas de tendencia central para determinar:
   a.) Outliers
   b.) Valores máximos y mínimos
   c.) Cantidad de *missing values*
   d.) La distancia de la media respecto de la mediana para analizar el grado de asimetría en la variable (En caso de ser continua o discreta)

5.) **Análisis con Ciencia de Datos**: El análisis que se realiza en esta parte del modelo, consiste en realizar comparaciones a través de gráficos bivariantes, donde estos muestran las funciones de densidad de cada variable separando "Buenos" y "Malos" con el fin de identificar las variables que son candidatas para incluirlas en el modelo. Los siguientes puntos están relacionados con ML.

6.) **Selección y tratamiento de variables**: En este punto, fundamentalmente se eligen las variables a través estadísticos de separación, así como el tratamiento que se les puede dar.
   a.) Tratamiento *WOE* (*Weight of Evidence*): Bajo esta metodología se mide la "fuerza" de un atributo o del grupo de atributos dada una variable separando cuentas "Buenas" de "Malas", al aplicar esta metodología nos evitamos problemas autocorrelación y puede tomar en cuenta no-linealidades. Se puede usar de igual manera para variables continuas, así como para categóricas. El estadístico que nos dará pauta para la selección de las variables es el *Information Value* (IV) el cuál es la fuerza total de la variable.
   b.) *Escalamiento:* Es un tratamiento basado en el valor de la observación comparado con el mínimo y el máximo de la variable y así acotar valores entre 0 y 1. Las variables categóricas se tratan como "*dummies*". Se puede usar la *SommersD* como estadístico de separación.
   c.) *Normalización*: Es un tratamiento basado en el valor de la observación con respecto al promedio y a la desviación estándar para evitar valores extremos. Las variables se tratan como "*dummies*". Las variables categóricas se tratan como "*dummies*". Se puede usar la *SommersD* como estadístico de separación.

7.) **Validación Cruzada**: Una vez seleccionadas las variables se someten a un *K-Fold Cross Validation*, la cuál es una técnica de iteraciones con diferentes muestras de *training* y de *testing*, para medir la eficiencia de los modelos a través del error, es decir, se ponen a competir diferentes metodologías estadísticas ya sean Econométricas o Algoritmos de *Machine Learning*. La importancia de este punto es

[3] Este es un punto irreversible en el tiempo, en donde el crédito ya no tiende a regresar a estados anteriores.

determinar cuál técnica estadística es la mejor para hacer el modelo, se selecciona a principalmente a través estadísticos como la desviación absoluta y el error promedio.

8.) **Entrenamiento del modelo**: Una vez realizada las pruebas iterativas y teniendo los resultados de cuál técnica estadística es la más adecuada para realizar el modelo, se procede a hacer el entrenamiento del modelo con dos submuestras: Una de *training* y una de *testing*. En este punto, dependiendo de la técnica estadística utilizada se adecuan los parámetros estadísticos y se rechazan las variables que por parámetros estadísticos de cada metodología se deben rechazar.

9.) **Validación del performance**: Una vez ya con la selección del mejor modelo, procedemos a medir el performance, es decir, la capacidad de predicción a través de diferentes estadísticos como lo son:
   a.) Prueba K-S (Kolmogrov – Smirnov)
   b.) Prueba GINI
   c.) Prueba ROC
   d.) Gráfico Bivariante
   e.) Information Value
   f.) Matriz de Confusión y sus derivaciones.

10.) **Determinación del punto de corte (*Cutoff*):** Este punto es central dentro de nuestro trabajo, ya que es el punto de conexión, en el cuál se decide que perfiles son aptos para la concesión de crédito y cuáles no. Tradicionalmente se suele rechazar de facto a aquellos clientes que están por debajo de esta métrica. En este punto se analiza la forma en que se puede hacer completamente eficiente un modelo estadístico, a través de la curva de máximo rendimiento y mínima pérdida dado un *cutoff*. Este punto es fundamental, ya que hasta la actualidad en la literatura y en la práctica no se ha profundizado, el cual se solventa en este documento a través de una fórmula matemática basada en métricas de la matriz de confusión.

11.) ***Pricing* de Crédito y Capacidad de Pago**: Una vez obtenido el *cutoff* y los resultados del modelo materializados a través de una probabilidad de cumplimiento, se realiza una propuesta a través de un modulo, basado en el riesgo de cada observación para asignarle una capacidad de pago, así como una tasa de interés ajustada por riesgo a nivel agregado.

2. Antecedentes y situación actual de inclusión financiera en México en el contexto de la pandemia COVID-19

En México, a partir del confinamiento social, iniciado en marzo de 2020, existe la necesidad de generar un escenario que permita el acceso a productos financieros útiles que satisfagan las necesidades de distintos entes económicos como: transacciones, pagos, ahorro, crédito y seguros; es decir un contexto de inclusión financiera. Esta última trae consigo beneficios económicos porque impulsa una menor desigualdad. En este sentido,

(Burgess y Pande, 2005) realizaron un estudio en India en donde analizaron el impacto de la expansión bancaria en zonas rurales no bancarizadas y encontraron que existe una asociación en la reducción de la pobreza. Esto se debió a que, al instalar más sucursales en dichas zonas, hubo una mayor movilización de ahorros que incentivó la acumulación de capital y la obtención de préstamos para inversiones productivas a largo plazo.

La evidencia más cercana que se tiene de este efecto en México, se encuentra a través de la experiencia de Compartamos Banco, en un estudio realizado por (Angelucci et al., 2015) que analizan los efectos de un producto de crédito grupal, enfocado principalmente a mujeres. Mediante el uso de encuestas, encontraron, al imputarle un análisis de panel de datos, que dicho producto, brindaba estabilidad de flujo de efectivo a los acreditados, permitiendo que se mantuvieran sus activos físicos y no los usaran en caso de insuficiencia de liquidez.

Otro punto que ha tomado fuerza con respecto al enfoque antes mencionado es que, por el contrario, el microcrédito ha traído más desigualdad y se ha caracterizado por un crédito promedio muy bajo y altas tasas de interés. De acuerdo con un estudio, realizado por la Corporación Financiera Internacional (IFC) *et al*. (2017), acerca de los factores que intervienen en la sensibilidad de la tasa de interés, se concluye que principalmente, ésta se ven afectada por el fondeo y los costos de operación ajustados por la cartera vencida. Un punto importante de este artículo se enfoca en que las Instituciones Microfinancieras (IMF) pueden mejorar su tasa de interés, una vez que se mejoran sus prácticas de otorgamiento de crédito. En este mismo informe, se estipula en gran medida que el crédito promedio se mantiene en niveles bajos debido al hecho de que su metodología predominante es a través del crédito grupal[4], ya que se limita por la capacidad de pago colectiva y no, la individual, proponiendo a su vez, orientar los esfuerzos a otorgar créditos individuales. Sin embargo, el tema de la inclusión no sólo se reduce a préstamos pequeños o el acceso geográfico, sino que, en la mayoría de las instituciones formales son rechazados por su probabilidad de default imputado por un algoritmo[5]. Esto lo demostró el Banco Interamericano de Desarrollo (BID) a través de un estudio realizado por Azevedo *et al*. (2019), en colaboración con Banco Familiar en Paraguay, en donde se generó un producto financiero que incluyó a sectores de la población con características peculiares que se conformaban por personas que primordialmente:

a) No hubiesen tenido una experiencia bancaria previa
b) Sector informal, que no pudiera comprobar ingresos estables
c) Personas que tuvieron experiencia negativa crediticia en su pasado

De manera que, el algoritmo formal de dicho banco seguía una regla estricta dura de riesgos y era que, se declinará[6] el crédito cuando su probabilidad de impago fuera mayor a 19.6%.

[4] Es una modalidad de crédito que se otorga a un grupo de personas, cuya responsabilidad es de forma colectiva. Generalmente se da en el segmento de Microcrédito.
[5] Entiéndase como el resultado particular de una regresión, análisis paramétrico o no paramétrico que usando ciertas variables determinan un probabilidad de cumplimiento o incumplimiento de pago de un crédito.
[6] Representa que el banco o la institución financiera se reserve el derecho de conceder un crédito.

En dicha investigación se decidió analizar los casos que estuvieran por encima de este umbral y que no calificaran inicialmente para una oferta de crédito, dándoles una oportunidad de acceso al crédito, a través de un proyecto llamado "Credicédula", dicho proyecto buscaba como objetivo ver el resultado paralelo cuando un crédito no era concedido.

Los resultados anteriores fueron muy interesantes debido a que encontraron que la probabilidad de impago[7] era 11% mayor en el largo plazo en aquellos que eran inicialmente elegibles por el algoritmo, mientras que los que no eran elegibles por dicha regla dura de riesgo y asociados a un escaso o nulo acceso previo al crédito, no sufrieron un impacto negativo en su calificación crediticia al final del estudio. Este estudio sugiere de manera implícita que la inclusión está ligada a la calidad crediticia "documentada" de los postulantes, ya que, en la visión tradicional de riesgos, los historiales crediticios, permiten "vislumbrar mejor" un perfil crediticio, condenando a la declinación a aquellos prospectos que no poseen un historial extenso y aceptando aquellos que sí tienen historia, bajo una regla dura de riesgos, sin embargo, dicho documento nos ayuda a entender que no es así, debido a sus resultados positivos y en contra de la lógica ortodoxa bancaria.

En adición, detectan un efecto particularmente a los usuarios de este servicio y es que, las personas que eran no elegibles en un principio y que el experimento, demostró que eran incluso mejores pagadores, fue que, en el largo plazo, dichos consumidores encontraron opciones más baratas de otras fuentes, debido, que al haber formado un historial crediticio con Credicédula, otras instituciones los lograron captar con mejores alternativas de costos. Este último punto es muy importante, debido a que existe una relación implícita de Riesgo-Precio que juega un papel fundamental en la inclusión financiera.

Existen muchos modelos que analizan los requerimientos mínimos para poder determinar las tasas de interés de los créditos bancarios. Sin embargo, la mayoría de estos modelos se sustentan en los estados financieros, debido a la dificultad en el acceso a información bancaria más precisa. Un ejemplo de ello, es el caso de (Rozo et al., 2014), donde la exposición de su trabajo muestra una metodología de *pricing* de crédito sobre los siguientes factores:

1.) Costo de fondeo
2.) Costos de originación y administración de cartera
3.) Costos por Riesgo de Crédito

De esta manera, su principal aportación es generar un pricing basado en dichos costos. Algo que llama la atención de esta investigación, es que, para determinar el riesgo de crédito (que es parte fundamental para asignar el precio), el default lo obtienen de la investigación de (Gómez-González et al., 2007), en donde adquieren matrices de transición bajo supuestos markovianos y se utiliza un modelo de duración para así obtener pronóstico del default en la cartera crediticia comercial. La esencia del documento mencionado es de vital importancia, pues muestra que a través de cadenas de Markov, en donde de manera

[7] Representa la probabilidad de que un sujeto de crédito en algún momento del tiempo llegue a un punto de no pago del crédito. Este último es diferente para cada banco.

genérica se supone total exogeneidad y dependen únicamente de su estado anterior, no es lo suficientemente consistente para pronosticar el default en el largo plazo. Dentro de este ejercicio, se establece la importancia de introducir variables explicativas como determinantes en la probabilidad de migración.

Continuando con la idea anterior, (Pulgar y Rojas, 2019), realizan un trabajo sobre los requerimientos mínimos en la tasa de interés, planteando una metodología para estimar una prima por riesgo de crédito sostenida en una tasa implícita basada en costos, que suponen factores esenciales: tasa libre de riesgo, premio por riesgo de crédito y un spread adicional dadas las imperfecciones del mercado. Este trabajo documenta el impacto cuando existen cambios de la normativa en el entorno financiero y su impacto en el pricing. De este documento, es importante destacar la relevancia que le dan al cambio en la normativa, algo no muy común a tomar en cuenta en los modelos, ya que en la mayoría de la literatura suponen constantes en los marcos regulatorios en los que se sostienen los modelos.

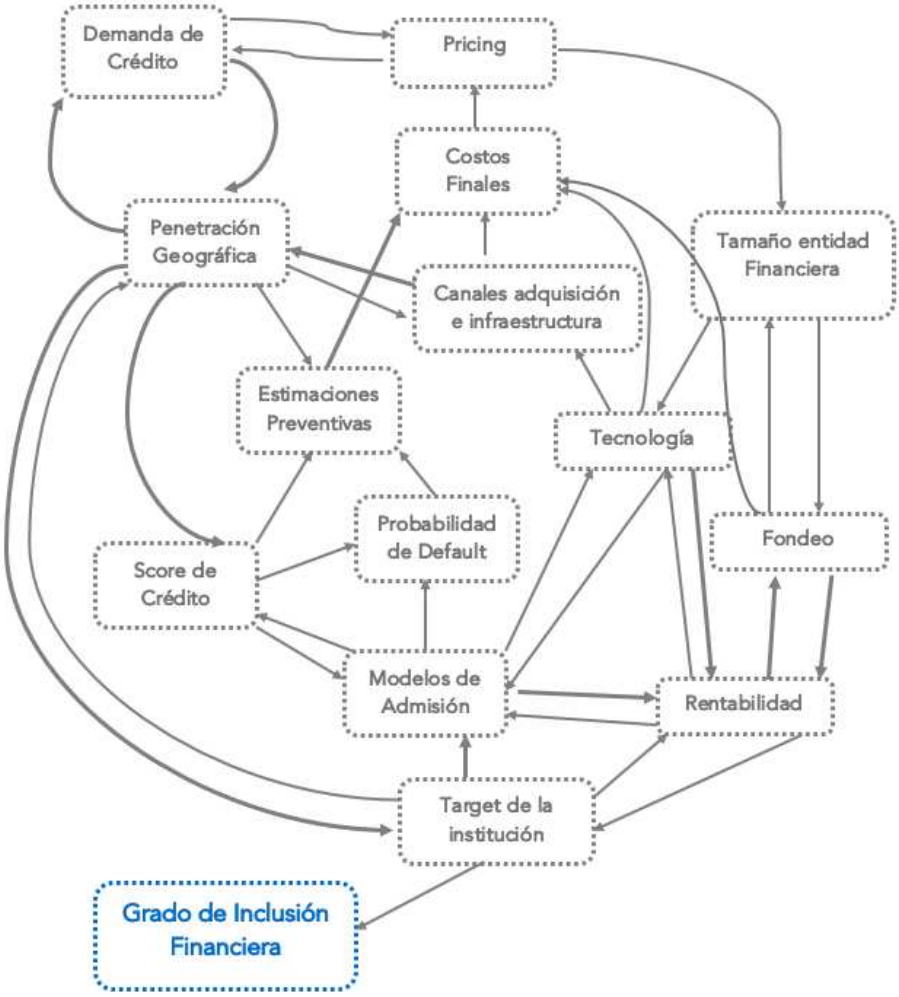
Adicionalmente, estos costos se ven altamente influenciados por la penetración geográfica y la tecnología que tenga cada institución financiera para ofrecer productos de crédito. En 2018 la Comisión Nacional Bancaria y de Valores (CNBV) a través del reporte de inclusión financiera, sostiene que no se han logrado obtener un margen de puntos de acceso tradicionales (sucursales y cajeros) en toda la República Mexicana dado que los costos asociados son altamente significativos.

Por otra parte, (Domínguez y Ortega ,2019), a través de su investigación sugieren que contar con una sucursal bancaria a menos de tres kilómetros de usuarios finales, genera un efecto positivo en la actividad económica. Esta idea va de la mano con un documento realizado por (Bruhn y Love, 2014) en el cual, a través de la experiencia con Banco Azteca, determinan que dicha institución, a través de la apertura de sucursales, incentivó a dueños de negocios informales a mantener sus negocios; lo que implicó un aumento del 1.4% del empleo y 7.6% en magnitud de negocios. Por ello, la inclusión financiera debe estar enfocada a sectores vulnerables, como el empleo informal. Sin embargo, como lo sugieren Azevedo, Lafortune *et al*. (2019), este tipo de personas son las menos adeptas, por reglas genéricas de riesgo, a tener acceso al crédito.

Dado lo anterior, para tratar de vislumbrar estas relaciones a través de un diagrama de relaciones que permitirá definir a profundidad este fenómeno. La principal conexión que se establece para que se dé la inclusión financiera es a través del *Target* de la Institución el cuál es afectado directamente por la Rentabilidad la cuál es alimentada y alimenta constantemente a los modelos de admisión, los cuáles a su vez, se interconectan con los *scores* de Crédito, los cuales tienen implicaciones directas en las Estimaciones Preventivas y en los costos Finales que también tienen incidencia en el *pricing*, el cual afecta de nuevo a la Rentabilidad, creando un *loop*[8] eterno entre la Rentabilidad y los componentes principales del Riesgo de crédito.

[8] Es un ciclo que se repite constantemente hasta que cumple con una función objetivo.

Gráfica 1. Relaciones inclusión de crédito



Fuente: Elaboración Propia

Siguiendo con las ideas anteriormente expuestas, es importante señalar que el presente trabajo principalmente se concentrará en perfeccionar el target de las instituciones a través de modelos de riesgo, ya que, cuando se obtiene un mejor nivel de rentabilidad, el apetito al riesgo suele ser mucho mayor, debido a que se tienen muchos más recursos para enfrentar costos, esto permitirá corregir y aumentar en gran medida el grado de inclusión financiera.

2.1 La Probabilidad de Incumplimiento como medida de exclusión del sistema financiero

La probabilidad de Incumplimiento una medida de calificación crediticia que se otorga internamente a un cliente o a un contrato con el objetivo de estimar su probabilidad de incumplimiento a un año vista".

Esta definición nos la da (BBVA Financial Report, 2010) sin embargo en el mismo documento, nos menciona que dicha probabilidad de incumplimiento se basa en *scorings* y estos se subdividen en:

a.) Reactivos: tienen como principal objetivo pronosticar la calidad crediticia de las solicitudes de crédito realizadas por los clientes y tratan de predecir la morosidad de los solicitantes en un tiempo determinado. Es decir que, se ejecutan solamente a la hora de originar por primera vez con el prospecto.
b.) "Comportamentales": Calcula la probabilidad de mora de una operación viva. Este modelo es dinámico, ya que se puede calcular n veces, en tanto la información de la operación cambie (saldos, retrasos con otras instituciones, etc.). Scorings de este tipo (Que más adelante examinaremos) y que son de uso público con el fin de Calificar la Cartera Bancaria se encuentran en los artículos 91,92,97, 99 y 112 de las Disposiciones de Carácter General aplicables a instituciones de crédito (CNBV, 2020).
c.) Proactivos: Son *scorings* que realizan evaluación del riesgo que permite adelantarse a una solicitud de crédito, estos usualmente se utilizan para clientes que ya están en la cartera y permiten perfilar productos de crédito más adecuados.
d.) Buró: Estos scores se alimentan principalmente de la información crediticia que ha presentado el cliente desde el inicio con operaciones de crédito. Este es un producto que principalmente lo ofrecen Sociedades de Información crediticia como Buró de Crédito y Círculo de Crédito.

Otra definición, BASILEA II especifica al *scoring* como:

> Una técnica estadística utilizada para predecir el comportamiento futuro. La calificación o puntuación asignada a cada individuo en particular, indica el probable comportamiento que podría exhibir a futuro. En el ámbito crediticio, es la Probabilidad de Incumplimiento (PD) expresada en puntos. Es el resultado final de aplicar ponderaciones a las distintas variables determinantes para el otorgamiento de un Crédito."

Uno de los grandes retos al hablar de Scorings, son las limitaciones que se tienen a la hora de ejecutarlos, ya que, en muchas de las ocasiones, la naturaleza de la información no es completamente fidedigna o no se consideran como calificables por información incompleta. Por ejemplo, no tener historiales crediticios "actualizados"; esta problemática es mencionada por el Congressional Research Service, en 2020, en Estados Unidos. De acuerdo con tal información, se categoriza como "Invisibles de crédito" y excluyen de facto a aquella población que no tiene registro en ninguna de las 3 oficinas más grandes dedicadas a recopilar información crediticia, o que tienen historial insuficiente de crédito para ser calificado. Se estima que este segmento, representa al menos el 11% de la población adulta en Estados Unidos, un equivalente a 26 millones de habitantes que no tienen acceso al servicio de crédito por este tipo de inconsistencias. Una de las propuestas de este artículo es usar información alterna a la hora de realizar scores de crédito, por ejemplo, el pago de servicios públicos o pagos de alquiler.

Con base en el concepto de exclusión financiera, al hacer uso de modelos de *scoring, Consumer Financial Protection Bureau's Office of Research (The CFPB Office of Research*, 2015) realiza un estudio en retrospectiva de esta población "invisible de crédito" [9]y demuestra que existe una clara diferencia de estos grupos cuando se realiza un análisis de características demográficas. Por ejemplo, cuando hacen el análisis a través de la edad, se dieron cuenta que las personas que eran más propensas a ser invisibles por tener insuficiencia de score o no haber tenido experiencia crediticia eran las que tenían edades muy tempranas (18- 24 años) o muy longevas (70+ años). A la par, al analizar a través del ingreso, se dieron cuenta que casi el 50% de la muestra que se usó, no tuvo acceso al crédito cuando presentó ingresos bajos. Así mismo, sugieren que hay una brecha racial, ya que cuando analizaron los grupos a través de diferentes tipos de raza, el acumulado de porcentaje de raza negra o hispana, superaba el 25% de no tener acceso al crédito, mientras que, la raza blanca o asiática rondaba cerca del 18%. De manera que, los bancos se niegan a ofertar un crédito porque las características que evalúan los sistemas score representan que son de alto riesgo, y que impiden que el solicitante tenga acceso al crédito.

En una investigación denominada "Geografías de la exclusión Financiera", realizada por (Leyshon & Thrift, 1995) se muestra claramente como las personas en Reino Unido que no tenían acceso a una cuenta de débito se destacaba cuando la localidad donde vivían, pertenecía a una zona rural; mientras que en las metrópoli, tenían mayor acceso. Pero, lo más destacable, es que la concesión de cuentas de débito con chequera, dependía de sistemas de scoring bancarios.

Por lo anterior, en los siguientes sub-secciones, se analizará, en primer lugar; los efectos sustanciales del COVID en la Banca, en segundo lugar; el efecto que tiene la probabilidad de incumplimiento, en este caso de los diferentes *scorings* "comportamentales" de la metodología de calificación de cartera (como medida que refleja el riesgo) en el aumento o disminución de los saldos de cartera.

2.2 Impacto COVID en la Estructura Bancaria en México

La pandemia generada por el virus SARS-COVID19 ha generado una serie de desequilibrios económicos que han estribado en una afectación directa a la gran mayoría de las industrias en México, esto debido a la paralización inmediata de la economía con el objetivo de contener dicho virus y no colapsar los sistemas de Salud. Los pronósticos económicos no son optimistas ni en el mejor de los escenarios, tal es así que (Preciado y Schaefer, 2020) realizan un pronóstico en la desaceleración económica del -6.6% tomando como referencia el Producto Interno Bruto (PIB) y realizan una serie de proyecciones donde estipulan que el tiempo mínimo de recuperación de las principales industrias (Alimentación, *Retail*, Restaurantes, Viajes y Hotelería) se dará hasta inicios de 2021. Tal es así que los bancos no son ajenos a este fenómeno y frente a esta situación el Gobierno Federal a través de la Comisión Nacional Bancaria de Valores (CNBV) emitió una serie de criterios contables con el fin de contrarrestar los posibles efectos en Cartera Vencida y sus

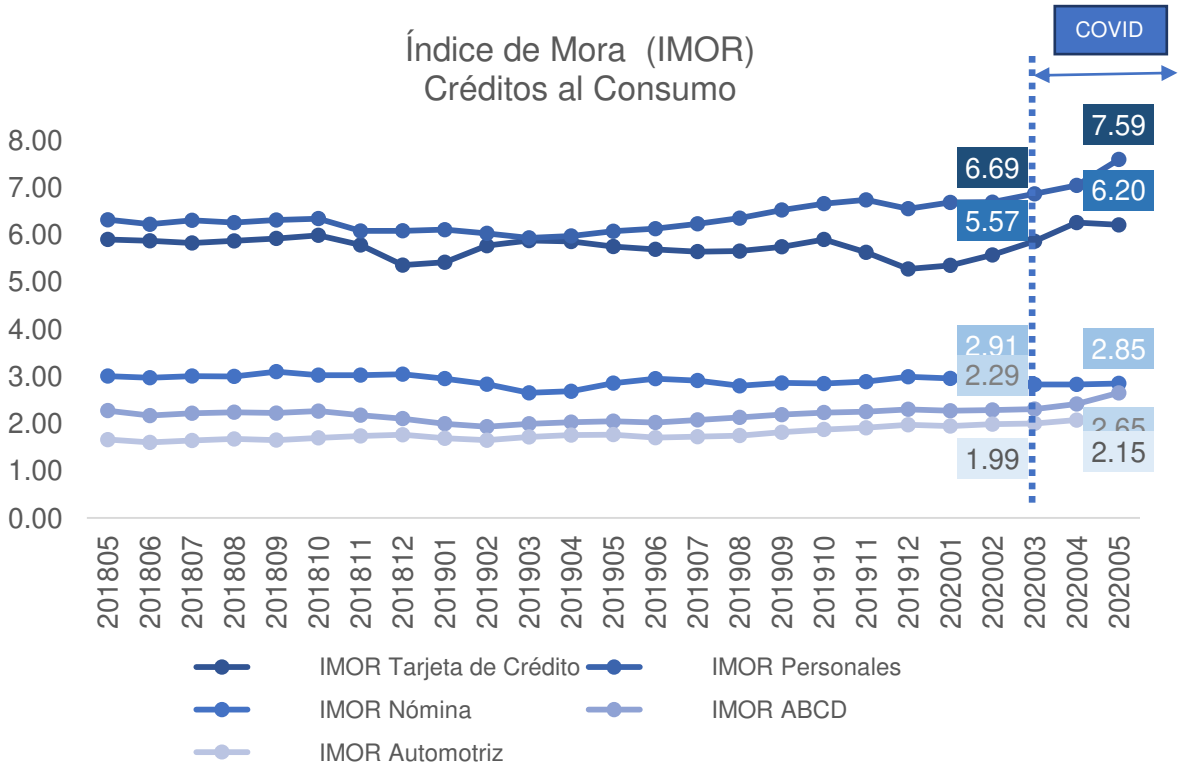
[9] Representa a aquella persona en búsqueda de algún tipo de Financiamiento, pero, que por sus características o su nulo historial de crédito se vuelve invisible o no atractivo ante las instituciones oferentes de crédito.

implicaciones en los Índices de Capitalización, dentro de esta serie de acciones las más relevantes son las siguientes:

1.) Permite la reestructuración[10] de la mayoría créditos[11] vigentes al 28 de Febrero del 2020.
2.) Se otorga un periodo máximo de gracia de hasta 180 días como extensión máxima del plazo original del contrato de crédito
3.) Permiten constituir Estimaciones Preventivas de Riesgo de Crédito de manera parcializada con el fin de no descapitalizar a las Instituciones de Crédito.
4.) Prohibición de la disminución o cancelación de líneas de crédito previamente autorizadas o pactadas.

Sin embargo, pese a todos los esfuerzos del sector público y privado, en los reportes de la CNBV al cierre del Mes de Mayo del 2020 se denota un incremento en la mora[12] de los principales productos de crédito que se presentan a continuación:

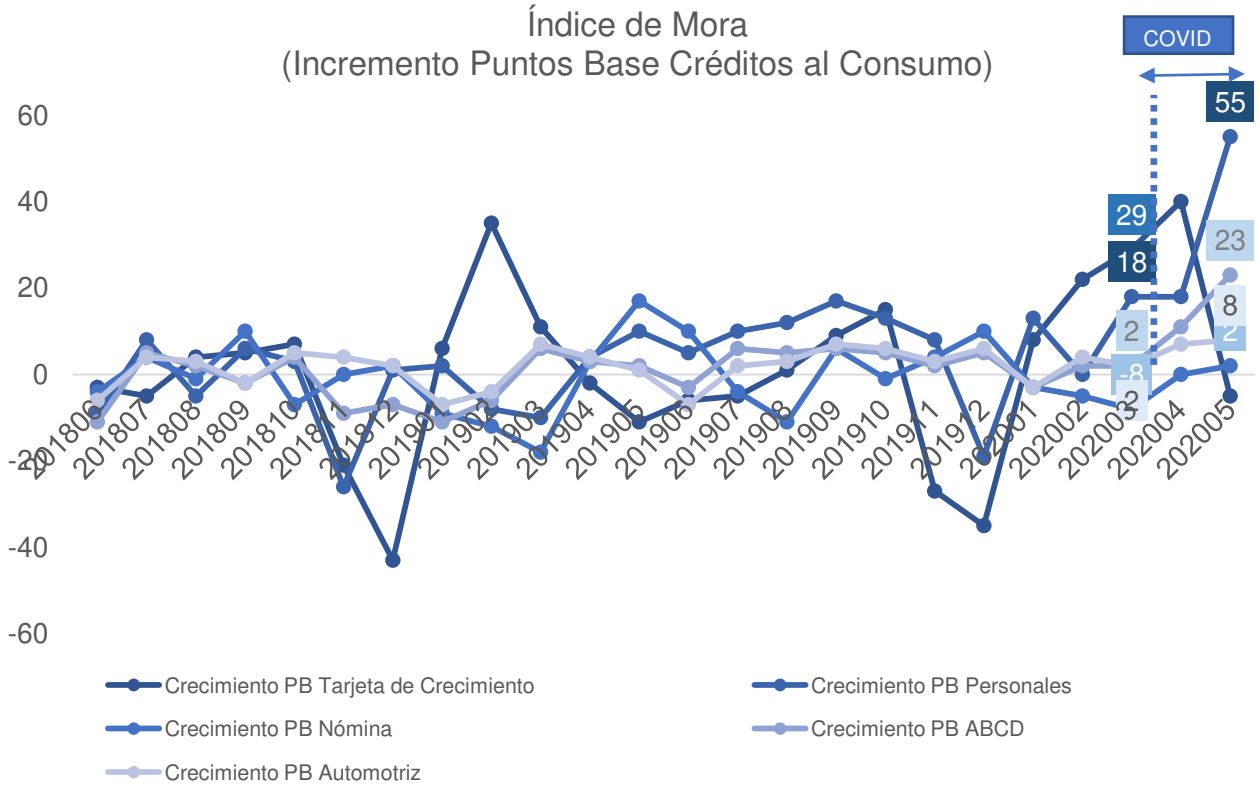
Gráfica 2.Índice de Mora Créditos al Consumo (SARS-COVID19)



[10] Modificación a cualquiera de las condiciones iniciales de cualquier contrato de crédito: Tasa, plazo, tipo de interés, pago.
[11] Se hace referencia específicamente a los Créditos de Consumo: Revolventes (Tarjetas de Crédito y Líneas de Crédito), No Revolventes (Automotriz, Personal, Nómina, ABCD, Microcrédito), Créditos Garantizados y Créditos Comerciales.
[12] Considérese como Cartera Vencida/ Cartera Total

Vemos que previo a los inicios duros de la pandemia SARS-COVID19 se notaba una estabilidad en el Índice de Mora (IMOR) de todos los créditos, sin embargo cuando se inició la paralización de actividades por parte del Gobierno Federal, se concibe un repunte anormal en la mora de los créditos. Esto se puede analizar más a profundidad cuando analizamos la variación mes a mes sobre puntos base:

Gráfica 3.Índice de Mora Créditos al Consumo 2 (SARS-COVID19)



Vemos que en efecto, hubo un repunte de puntos base que no había tenido antecedentes en los últimos dos años, por lo que esto tiene efectos en la concesión de créditos, obligando a las Instituciones de Crédito a realizar una selección más adversa, orillando aun más a la exclusión de crédito.

Paralelamente, de forma autónoma, otra de las medidas adoptadas por parte del Banco Central (Banco de México) para mantener la estabilidad de Cartera e impulsar el crecimiento de la misma, además de estimular el consumo, fue el recortar la tasa de interés de referencia 100 puntos base desde el 20 de Marzo, hasta el 14 de Mayo, este comportamiento no se había suscitado desde la crisis de 2008 en donde la explotó la burbuja relacionada con el mercado inmobiliario en Estados Unidos.

Gráfica 4.Tasa de Interés de Referencia (SARS-COVID19)

Entre otras medidas adoptadas por Banco de México (BANXICO) de acuerdo con información de (Clavellina Miller y Domínguez Rivas, 2020) encontramos lo siguiente:

1.) Banxico y la *Federal Reserve Board*[13] (FED) establecieron una línea de crédito temporal por 60 mil millones de dólares para fondear capital privado.
2.) Emitió recursos a Instituciones Bancarias para canalizar en crédito por 250 mil millones de pesos, por lo que disminuyó de manera importante el costo del fondeo para las Instituciones de Crédito.

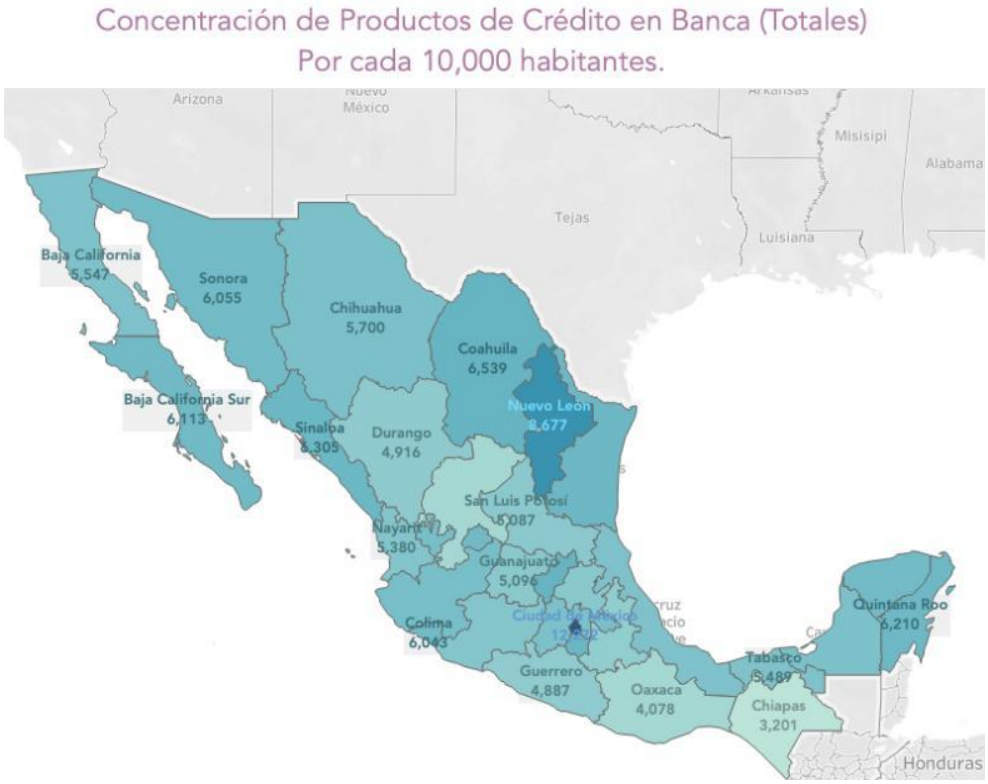
2.3 Análisis de la concentración de la cartera, productos de crédito y su probabilidad de incumplimiento en México

Es importante en esta investigación, resaltar, la constitución de la Cartera al Consumo[14] Mexicana a través de sus tipos de contratos, con el fin de puntualizar: 1.) los pormenores de la distribución del crédito y su destino en México para resaltar el poder de mercado meta de las Instituciones de Crédito y el destino de recursos que se le da al Crédito; 2.) El impacto

[13] Conjunto de Bancos Centrales de Estados Unidos.
[14] Constituida por créditos directos, incluyendo los de liquidez que no cuenten con garantía de inmuebles, denominados en moneda nacional, extranjera, en UDIs, o en VSM, así como los intereses que generen otorgados a personas físicas, derivados de operaciones de tarjeta de crédito, de créditos personales, de créditos para la adquisición de bienes de consumo duradero (conocidos como ABCD), que contempla entre otros al crédito automotriz y a las operaciones de arrendamiento financiero que sean celebradas con personas físicas; incluyendo aquellos créditos otorgados para tales efectos a los ex-empleados de las Instituciones.

financiero de tener probabilidades de incumplimiento altas a través de los modelos normativos materializados en Reservas y; 3.) El impacto que tienen dichas probabilidades de incumplimiento en el *Pricing* de Crédito, así como en el aumento o disminución de Saldos de Cartera en variables correlacionadas evidenciando una selección adversa más conservadora.

Gráfica 5. Concentración de contratos crédito Banca



Fuente: Elaboración propia con información de CNBV (2019)

De acuerdo con información de la CNBV al tercer trimestre de 2019 se encontró que la concentración de contratos de crédito por cada 10,000 habitantes se encuentra principalmente aparentemente en ciudades que poseen alto dinamismo económico, tal es el caso de la Ciudad de México, seguido por Nuevo León.

Al estimar el índice Herfindalh Hirschman[15] (IHH) con la distribución de contratos bruta y los contratos de crédito por cada 10,000 habitantes para tener un acercamiento de la concentración que se tiene registrado en 2019, y adicionalmente, lo comparamos con una distribución perfectamente equilibrada donde llega al punto más bajo del índice, se encontró encontramos lo siguiente:

[15] Véase anexo I

Cuadro 1. Concentración total de contratos por Estados en Bancos

Indicador	Resultado
IHH número de contratos totales	0.064577
IHH por cada 10 k hab 2019	0.03326
IHH ideal	0.03125

Fuente: Elaboración propia a partir de datos CNBV 2019

Se observa que aparentemente no existe tanta concentración de contratos a nivel general, si se toma el IHH por cada 10,000 habitantes. Sin embargo, a continuación, se replica este ejercicio para los cinco tipos de contrato que se tienen contabilizados en dicha base de datos.

Cuadro 2. Concentración por tipo contrato en Banca

Contrato	IHH Contratos Brutos	IHH Contratos 10K hab	Ideal[16]
Tarjeta Crédito	0.07536129	0.03654057	0.03125
Hipotecario	0.0643396	0.04050117	
Grupal	0.05944202	0.03697679	
Personal	0.05939986	0.03260165	
Nómina	0.05724558	0.03322917	
Automotriz	0.06512789	0.03598928	
ABCD	0.07669676	0.03417393	

Fuente: Elaboración propia a partir de datos CNBV 2019

Bajo esta perspectiva de analizar contratos por cada 10,000 habitantes, a primera vista, parece que no se encuentra concentrada en ninguno de los rubros, a excepción del Crédito Hipotecario, debido a que, en Nuevo León, Querétaro y Ciudad de México, ya que la media de contratos por cada 10,000 habitantes es de 169, mientras que dichos estados tienen 382, 327 y 352 respectivamente. Pese a estas aproximaciones, este indicador es muy lejano, debido a que la contabilidad por cada 10,000 habitantes no se debería realizar sobre el número de contratos, sino por el número de clientes únicos, ya que un cliente puede ser tenedor de más de un producto de crédito a la vez, por lo que, al analizar basado en número de contratos, asumimos ex ante que es un único contrato por cliente.

Dada la complejidad de este asunto, se revisó el trabajo realizado por Balmaseda y Necochea (2013) sobre conjuntos y curvas de nivel para aproximar la cantidad de número de clientes del sistema financiero, asumiendo una premisa muy importante, existen clientes compartidos. La desventaja de esta gran investigación fue que únicamente tomaron como referente las Instituciones de Banca Múltiple y basados en el producto de tarjeta de crédito. Aún así, llegan a un resultado convincente, aproximan que, para ese momento, el sistema bancario oscilaba los 47.6 millones de usuarios.

[16] Se considera a partir de dividir en partes iguales, el cual sugiere el punto menor del IHH.

De esta forma, se sabe que, en primera instancia que, al carecer de esta información, los datos que se presenten siempre tendrán sesgo si es que no se considera el factor de cliente único. Sin embargo, en la siguiente sección se analiza la estructura financiera sobre puntos muy particulares de interés para este trabajo. El enfoque principal será en lo que consideran las Disposiciones de Carácter General aplicables a Instituciones de Crédito de la CNBV (2020) como crédito al consumo. Cabe mencionar que, pese a que el Microcrédito Individual y Grupal, pertenecen a la categoría Consumo, no se encontró información oficial, de estas dos formas particulares de crédito. Sin embargo, a lo largo de esta sección se trata de encontrar una explicación acerca de este fenómeno. Luego entonces, enunciaremos de manera breve la serie de regulatorios utilizados para el análisis de la información.

Cuadro 3. Regulatorios banca múltiple

Tipo Cartera	Tipo Contrato	Reg1	Reg2	Reg3	Reg4	Reg5	Reg6
040	30 (Nómina)	A-R23	A-R62	R-R3	R-R20	A-R11	R-R11
Consumo	33 (Personal)	A-R23	A-R62	R-R3	R-R20	A-R11	R-R11
	31 (Automotriz)	A-R23	A-R62	R-R3	R-R20	A-R11	R-R11
	32 (ABCD)	A-R23	A-R62	R-R3	R-R20	A-R11	R-R11
	12 (Revolvente)	D-R5	E-R9	E-R10	H-R1		

La nomenclatura genérica de dichos reportes refiere a lo siguiente:

Cuadro 4. Nomenclatura regulatorios

Reporte Genérico	Detalle
A-R23	Saldo de Cartera por Entidad Federativa
A-R62	Tasa de Interés por Entidad Federativa
R-R3	Número de Créditos por Probabilidad de Incumplimiento[17]
R-R20	Probabilidad de Incumplimiento por Entidad Federativa
A-R11	Saldos de Cartera por Intervalo importe original Crédito
R-R11	Prob. De Incumplimiento por Intervalo importe original Crédito
D-R5	Saldo Total del crédito por Entidad Federativa
E-R9	Distribución de tarjetas por Pérdida Esperada
E-R10	Distribución de tarjetas por Probabilidad de Incumplimiento
H_R1	Tasa de interés por Pérdida Esperada

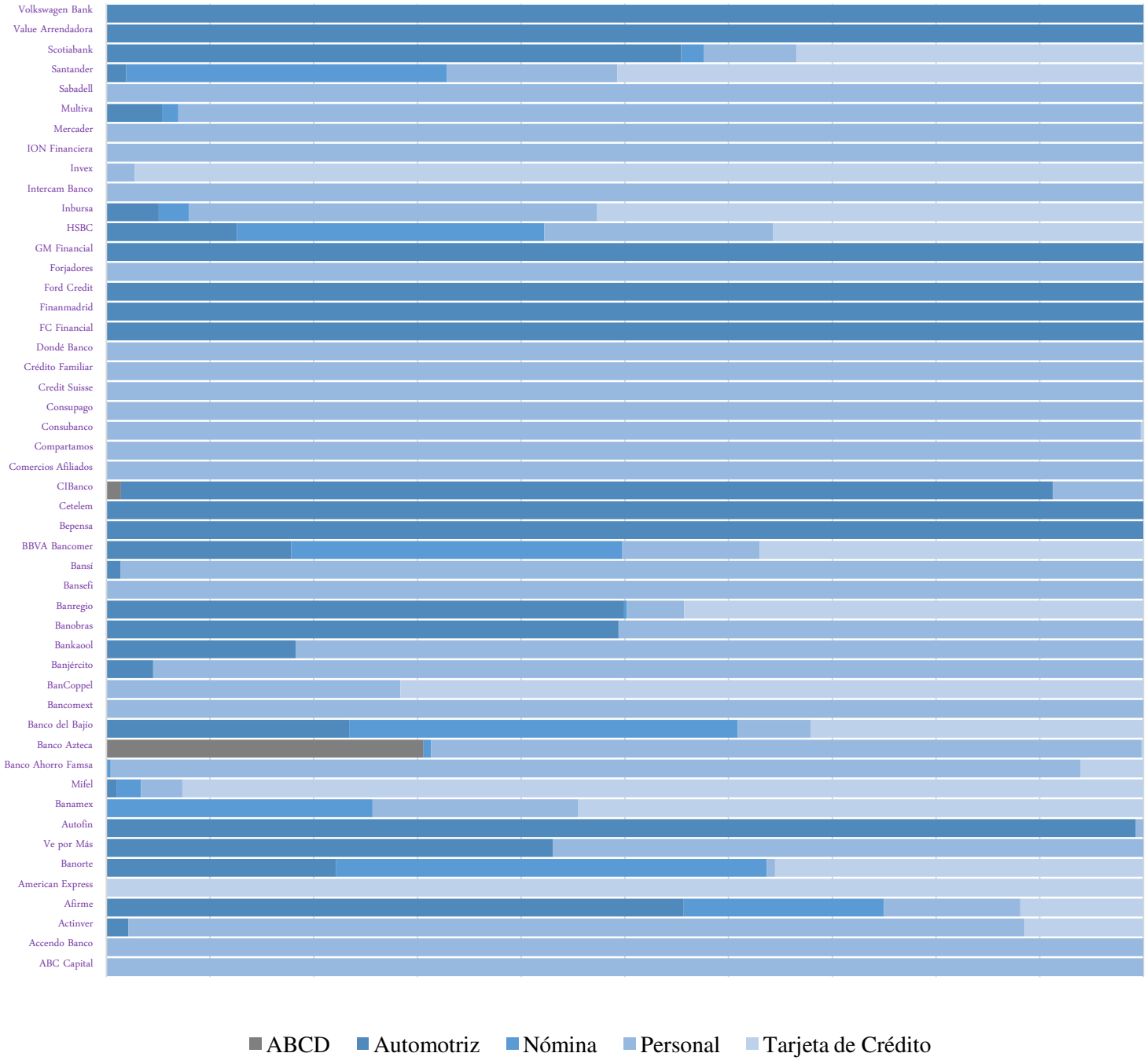
A diferencia de los créditos al consumo amortizables (Nómina, ABCD, Personal, Automotriz, Otros), los créditos al consumo revolventes no se encuentran estructurados con la misma información, es por ello que se consideran los reportes regulatorios cercanamente equivalentes.

[17] Esta probabilidad de incumplimiento deriva de la metodología de cálculo enunciada de los artículos 91 y 92 de las Disposiciones Aplicables a Instituciones de Crédito.

2.4 Estructura del sistema por tipo de contrato

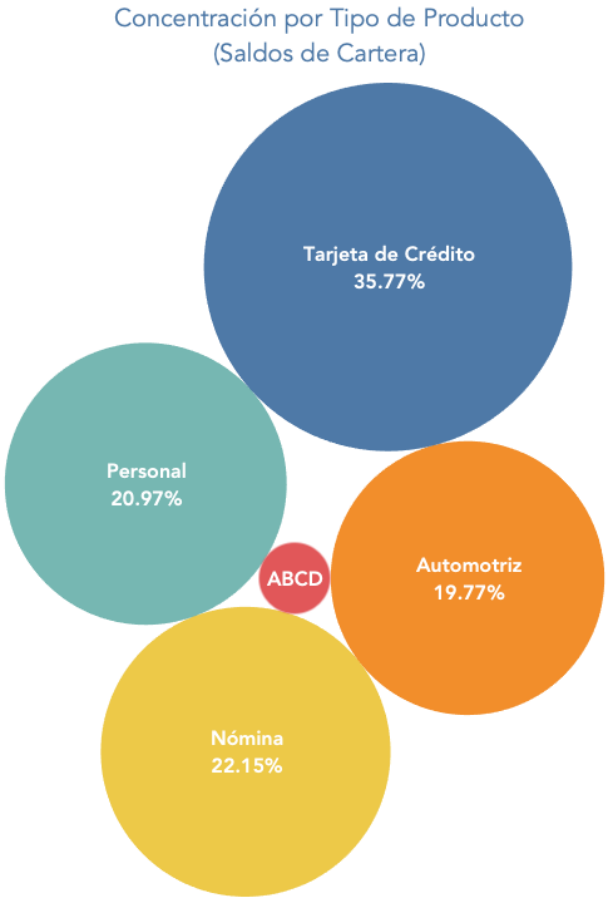
Al analizar de manera muy general el tipo de Target que tiene cada institución a través del tipo de contrato, se observa que muy pocas instituciones se encuentran diversificadas, ya que no hay ninguna que abarque los 5 contratos y hay muy pocas que cuentan con 4.

Gráfica 6. Distribución de Cartera Consumo
(Por Institución Financiera)



Observe que las Instituciones que cuentan con cuatro son BBVA Bancomer, Intercam y Afirme. En general el target por institución se enfoca primordialmente al crédito de nómina, posteriormente el Personal y, en tercer lugar, a las tarjetas de crédito, esto sólo es bajo la perspectiva de oferentes, por el contrario, a través de saldos, encontramos lo siguiente.

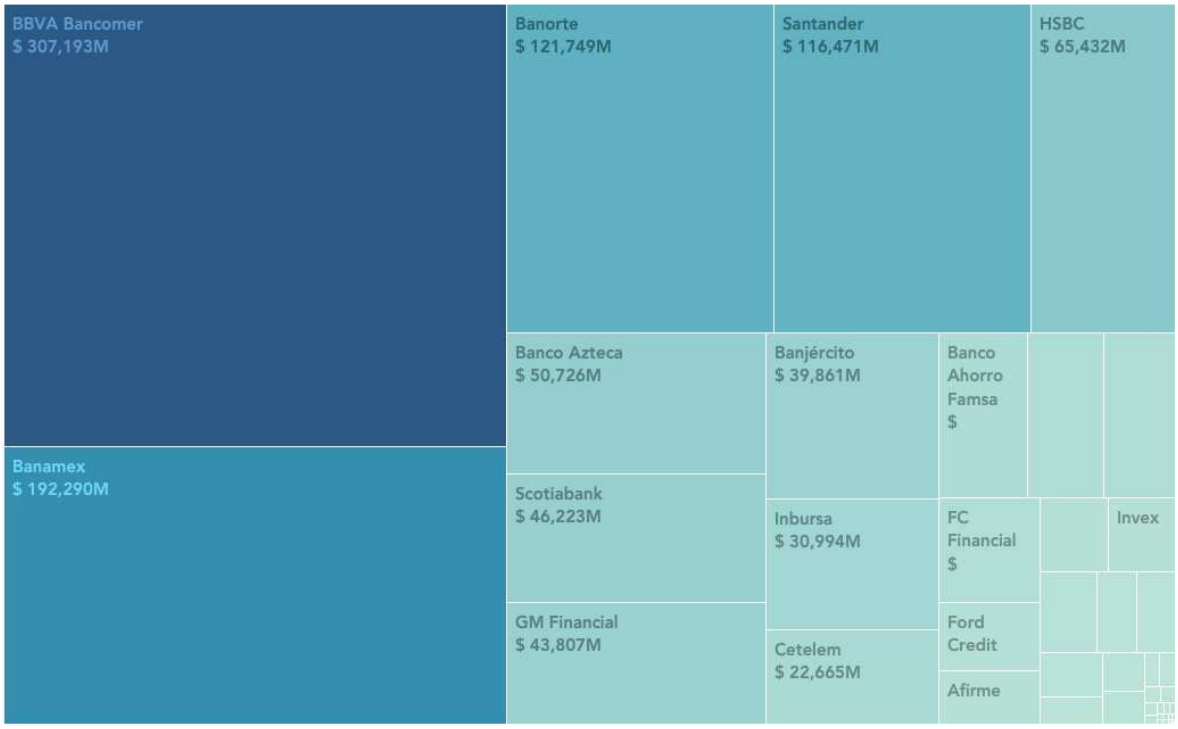
Gráfica 1. Distribución de Cartera por tipo de Producto



Se encontró que el producto que más concentra saldos de cartera en México a nivel sistémico es la Tarjeta de Crédito, con 35.77%, seguido del producto de nómina con 22.15%. Sin embargo, se observa que a la par, de los $1.1 billones de pesos de cartera al consumo, gran parte se encuentra concentrada en las siguientes instituciones:

De manera que al aplicar el IHH, se encontró que dicho indicador se ubica a 0.1290 puntos, lo que, para los 49 participantes que entraron en este análisis, claramente se ve concentrado en dos instituciones: BBVA Bancomer y Banamex a través de sus diferentes subsidiarias.

Gráfica 2. Cartera al Consumo: Distribución por Instituciones.



Al mismo tiempo se analiza el IHH por institución, tomando como base la distribución de cartera por Entidad Federativa, para determinar por tipo de producto la concentración que se tiene.

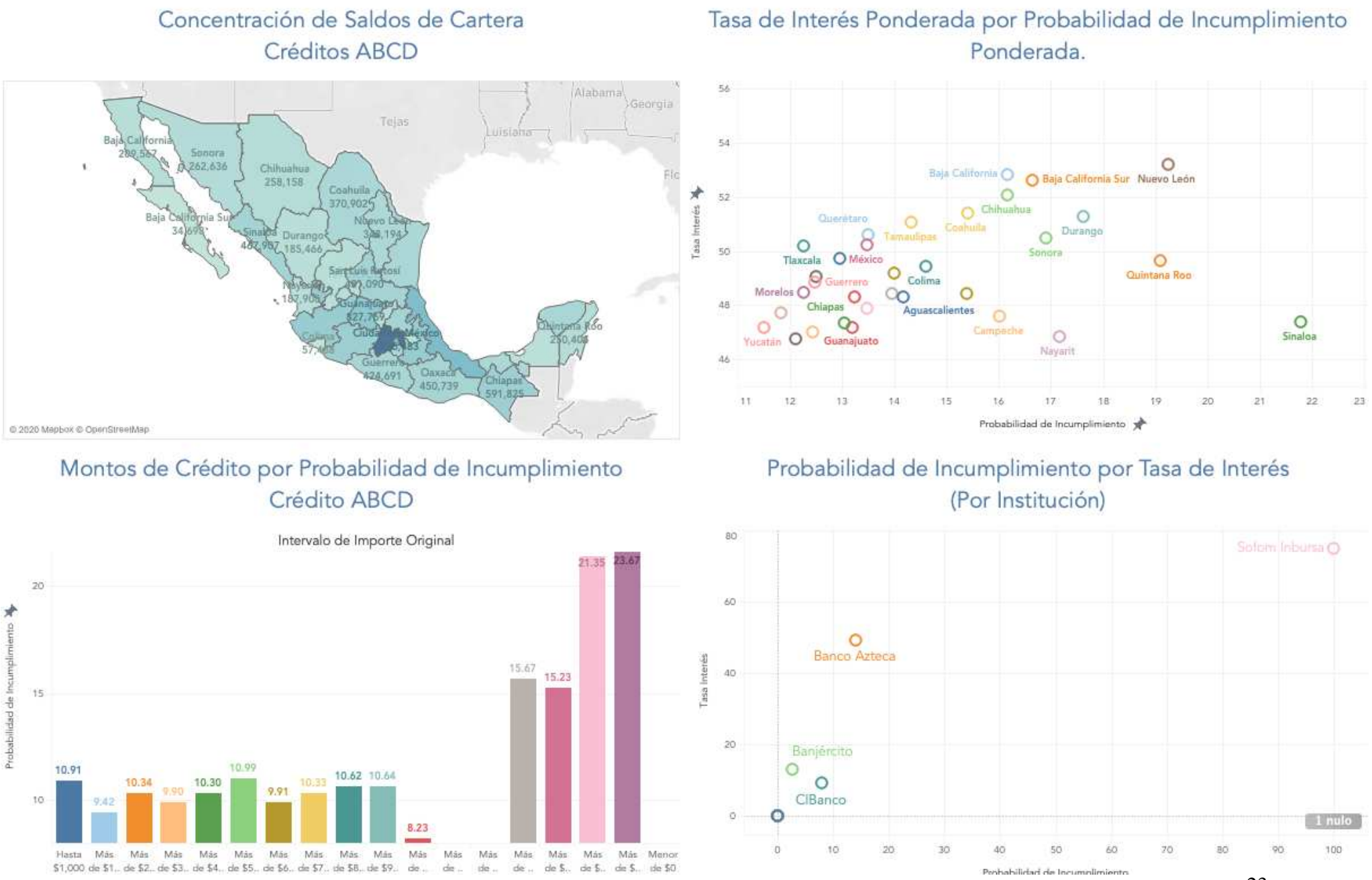
Cuadro 5. IHH por contrato y por Institución

Institución Financiera	ABCD	Automotriz	Nómina	Personal	TC
ABC Capital	0.00	0.00	0.00	0.38	0.00
Accendo Banco	0.00	0.00	0.00	0.41	0.00
Actinver	0.00	0.21	0.00	0.15	0.03
Afirme	0.00	0.34	0.20	0.19	0.03
American Express	0.00	0.00	0.00	0.00	0.02
Banorte	0.00	0.43	0.06	0.00	0.01
Ve por Más	0.00	0.09	0.00	0.00	0.00
Autofin	0.00	0.25	0.00	0.49	0.00
Banamex	0.00	0.25	0.06	0.09	0.01
Mifel	0.00	0.00	0.48	0.42	0.04
Banco Ahorro Famsa	0.00	0.00	0.35	0.06	0.01
Banco Azteca	0.06	1.00	0.33	0.06	0.02
Banco del Bajío	0.00	0.08	0.19	0.10	0.01

Bancomext	0.00	0.00	0.00	0.49	0.00
BanCoppel	0.00	0.00	0.05	0.05	0.01
Banjército	0.27	0.13	0.00	0.14	0.00
Bankaool	0.00	0.96	0.00	1.00	0.00
Banobras	0.00	0.34	0.00	0.19	0.00
Banregio	0.00	0.00	0.54	0.00	0.03
Bansefi	0.00	0.00	0.00	0.00	0.00
Bansí	0.00	0.00	0.00	0.00	0.00
BBVA Bancomer	0.00	0.00	0.07	0.00	0.01
Bepensa	0.00	0.00	0.00	0.00	0.00
Cetelem	0.00	0.00	0.00	0.00	0.00
CIBanco	0.37	0.00	0.00	0.00	0.00
Comercios Afiliados	0.00	0.00	0.00	0.00	0.00
Compartamos	0.00	0.00	0.00	0.00	0.00
Consubanco	0.00	0.00	0.00	0.00	0.08
GM Financial	0.00	0.00	0.00	0.00	0.00
HSBC	0.00	0.00	0.08	0.00	0.01
Inbursa	0.80	0.00	0.17	0.00	0.00
Intercam Banco	0.00	0.00	0.00	0.00	0.00
Invex	0.00	0.00	0.00	0.00	0.01
Scotiabank	0.00	0.00	0.00	0.00	0.00
Sistema					

Se puede observar que las instituciones que tienen un índice cercano a 1, son las que más concentran sus activos en una sola entidad federativa y los que están cercanos a cero son las que no están concentradas y los activos se encuentran diversificados en las diferentes entidades.

Gráfica 3.Créditos ABCD



2.5 Créditos Adquisición de Bienes de Consumo Duraderos (ABCD)

Las Disposiciones de Carácter General aplicables a Instituciones de Crédito (de ahora en adelante CUB de Bancos) a través del artículo 91 define a los tipos de crédito, según su tipo de contrato, al crédito ABCD, lo define como sigue:

"A los créditos que sean otorgados a personas físicas y cuyo destino sea la adquisición de bienes de consumo duradero, con excepción de los créditos cuyo destino sea la adquisición de vehículos automotrices particulares".

Es decir, que se contempla como un crédito ABCD, siempre y cuando tenga como respaldo algún tipo de producto de consumo no perecedero, por ejemplo, una licuadora, una lavadora, etcétera y sea otorgado a una persona física. Aquí el colateral[18] que mitiga de cierta forma el riesgo, es el bien que se adquiere a través del crédito.

Al aplicarle el IHH por institución, llega a un nivel de 0.99, lo que indica en términos generales, que está perfectamente concentrada, debido a que Banco Azteca, es quién tiene todo el mercado de este tipo de contrato (99.4%), sin embargo, existen otras instituciones que ofrecen este servicio, como por ejemplo Banjercito y CI Banco, pero en mucho menor cantidad. Lo que se observa, de acuerdo a la Gráfica 7, es que la cartera encuentra a la par una concentración por Entidad Federativa, particularmente en el centro del país, teniendo como los 5 principales estados de concentración los siguientes:

Cuadro 6. Las 5 principales Entidades Federativas con Crédito ABCD (Miles de Pesos)

Estado	Saldo Cartera	Tasa Interés	Probabilidad de Incumplimiento
México	$ 2,906,015.50	50.24%	13.48%
Ciudad de México	$ 1,136,483.33	49.16%	13.99%
Veracruz	$ 1,110,794.86	48.28%	13.23%
Guanajuato	$ 827,768.56	47.19%	13.20%
Puebla	$ 695,098.22	49.74%	12.94%

La probabilidad de incumplimiento (PI) de estos créditos se encuentra dada por el modelo establecido en el artículo 91 BIS 1 (CUB, 2020)

Cuadro 6. Probabilidad de Incumplimiento Créditos ABCD

Coeficiente	Valor	Var
β_0	-2.5456	Riesgo Natural (Beta autónoma)
β_1	2.2337	ATR
β_2	1.0526	%SDOIMP

[18] Es una figura dentro del contrato de crédito que funge como mitigante del Riesgo de Crédito.

β_3	0.4361	Alto
β_4	0.078	Medio
β_5	-0.5141	Bajo
β_6	-0.0152	Ant***[19]
β_7	-0.0672	Meses
PI Riesgo Natural	**7.27%**	
PI Escenario A	**1.62%**	
PI Escenario B	**99.61%**	

Respetando los anexos y reglas establecidas del mismo numeral, se determinan dos tipos de Escenarios:

A.) Escenario al corriente (A): Deuda sin ningún atraso, a punto de liquidar, con antigüedad de un año, con bajo riesgo de concentración de deuda con otras instituciones y con el máximo de meses transcurridos desde el último impago

B.) Escenario de máximo impago (B)(Sin llegar ATR >3): Tres atrasos, deuda recién contraída, cliente nuevo, alto riesgo de concentración de deuda y con el mínimo de meses transcurridos desde el último impago

Dichos escenarios, se resumen en la siguiente tabla.

Cuadro 7. Variables PI Créditos ABCD

Var	Escenario A	Escenario B
Riesgo Natural (Beta autónoma)	1	1
ATR	0	3
%SDOIMP	0.01	1
Alto	0	1
Medio	0	0
Bajo	1	0
Ant**	12	2
Meses	13	1
PI	1.62%	99.61%
Reservas por cada 10K[20]	$139.49	$8,566.58

Lo que se trata de explicar a través de dicho modelo, es el riesgo y costo intrínseco por regulación que conlleva tener el producto ABCD en cartera y que a primera instancia en el Escenario A, ideal desde el punto de vista de riesgos implica que tendremos una Pérdida Esperada[21] muy pequeña, que implica un menor aprovisionamiento de Reservas Crediticias, impactando directamente los Estados de Resultados de cada institución. Al menos a primera vista, vemos que el crédito ABCD no cuenta con una probabilidad de

[19] Tomamos como base la antigüedad de 12 meses. En adelante, retomaremos, la importancia de las variables sin tratamiento estadístico de imputación, ya que tiene implicaciones serias en los modelos que expresan probabilidades.

[20] Se considera como severidad de la pérdida dentro del intervalo donde ATR <3. Se considera sólo sobre cálculos de saldo insoluto, sin considerar, interés devengado no cobrado e IVA.

[21] Derivado del cálculo Pérdida Esperada= Probabilidad de Incumplimiento * Exposición * Severidad de la pérdida. Para más detalles, véase el artículo 91 BIS.

incumplimiento alta sí se encuentra al corriente; sin embargo, en el peor de los escenarios la PI, es del 99.61%.

Continuando con esta idea, aunado al premio por dicho riesgo, se observa, a través de la Gráfica 7, que sí la PI contra la tasa de interés a nivel sistémico por Entidad Federativa, determinamos que no hay una distinción marcada entre probabilidad de incumplimiento y tasa de interés, por ejemplo, observamos el caso de Tlaxcala que tiene una PI ponderada de 12.25%[22] con una tasa de interés ponderada del 50.18%, mientras que Guanajuato tiene una PI ponderada de 13.20% y una tasa de interés del 47.19%, es decir que, pese a que hay un incentivo de menor riesgo, la tasa de interés parece ir en dirección contraria.

Cuadro 8. Coeficiente correlación Pearson Tasa Interés Probabilidad de Incumplimiento ABCD

	Correlación	R2
Correlación Tasa de Interés	0.40710156	0.16573168

Aunque parece no concluyente el análisis gráfico, la correlación a través del método Pearson, sugiere que es débilmente positiva y que explica hasta el 16% a la tasa de interés. Otra relación que es de particular interés examinada en la gráfica 7, es que la probabilidad de incumplimiento por intervalos del importe original del crédito no tiene relación directa, es decir, se espera que entre menor sea el *exposure*, mayor, será la probabilidad de incumplimiento, y viceversa, lo que asegura una dirección sostenible de negocio. Sin embargo observamos lo siguiente:

Cuadro 9. PI por Intervalo original importe de Crédito

Intervalo de Importe Original	Probabilidad de Incumplimiento[23]
Hasta $1,000	10.90%
Más de $1,000 y hasta $2,000	9.41%
Más de $2,000 y hasta $3,000	10.33%
Más de $3,000 y hasta $4,000	9.89%
Más de $4,000 y hasta $5,000	10.30%
Más de $5,000 y hasta $6,000	10.98%
Más de $6,000 y hasta $7,000	9.91%
Más de $7,000 y hasta $8,000	10.33%
Más de $8,000 y hasta $9,000	10.61%
Más de $9,000 y hasta $10,000	10.63%
Más de $10,000 y hasta $25,000	8.22%

[22] Para el desglose de indicadores, consultar con autores.
[23] Esta PI ponderada, fue calculada sobre la distribución de saldos presentados a 2019, pero con la PI de 2017. Ya que la CNBV no tiene información actualizada de ese rubro.

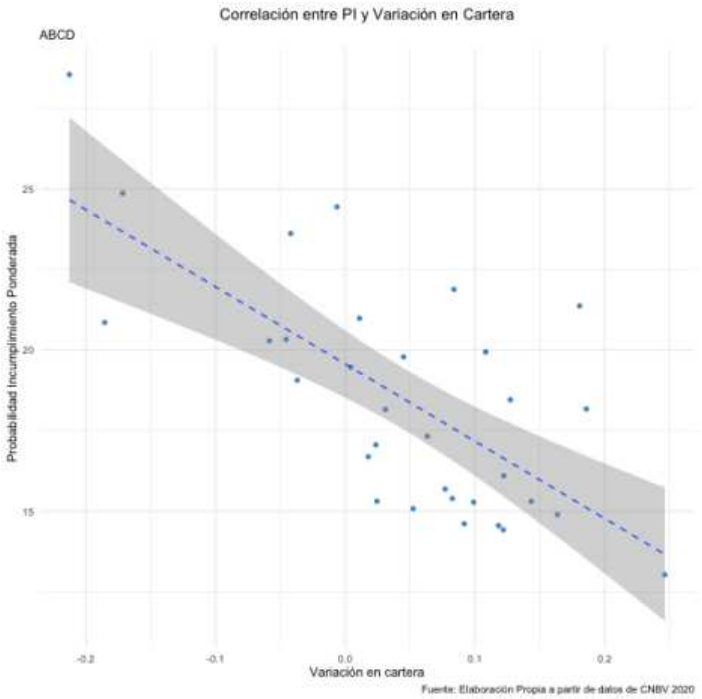
Más de $25,000 y hasta $50,000	5.77%
Más de $50,000 y hasta $75,000	7.81%
Más de $75,000 y hasta $100,000	15.66%
Más de $100,000 y hasta $250,000	15.23%
Más de $250,000 y hasta $500,000	21.35%
Más de $500,000	23.66%

Esto significa que, al menos con la información provista con la CNBV, no parece tener parsimonia la relación Riesgo-Rentabilidad. Así mismo, derivado de la gráfica 7, vemos que quién tiene el *princing* más alto para este producto, pero también la PI más alta es Banco Azteca.

Un efecto más que estudiamos, es la variación que tienen los saldos de cartera de un período con otro (2017-2018) con respecto a la probabilidad de incumplimiento de 2017 por Entidad Federativa. Partimos de la premisa en donde a mayor riesgo, los bancos dejarán de colocar dicho producto, y enfocarán sus esfuerzos a colocar créditos en donde la probabilidad de incumplimiento es baja.

A continuación aplicamos el *test* Shapiro-Wilk para ver que se cumple con condición de normalidad, de manera que el *p*-value, nos indica que podemos usar el coeficiente de correlación de Pearson.

Gráfica 4. Correlación PI y Variación Cartera ABCD (Saldos)



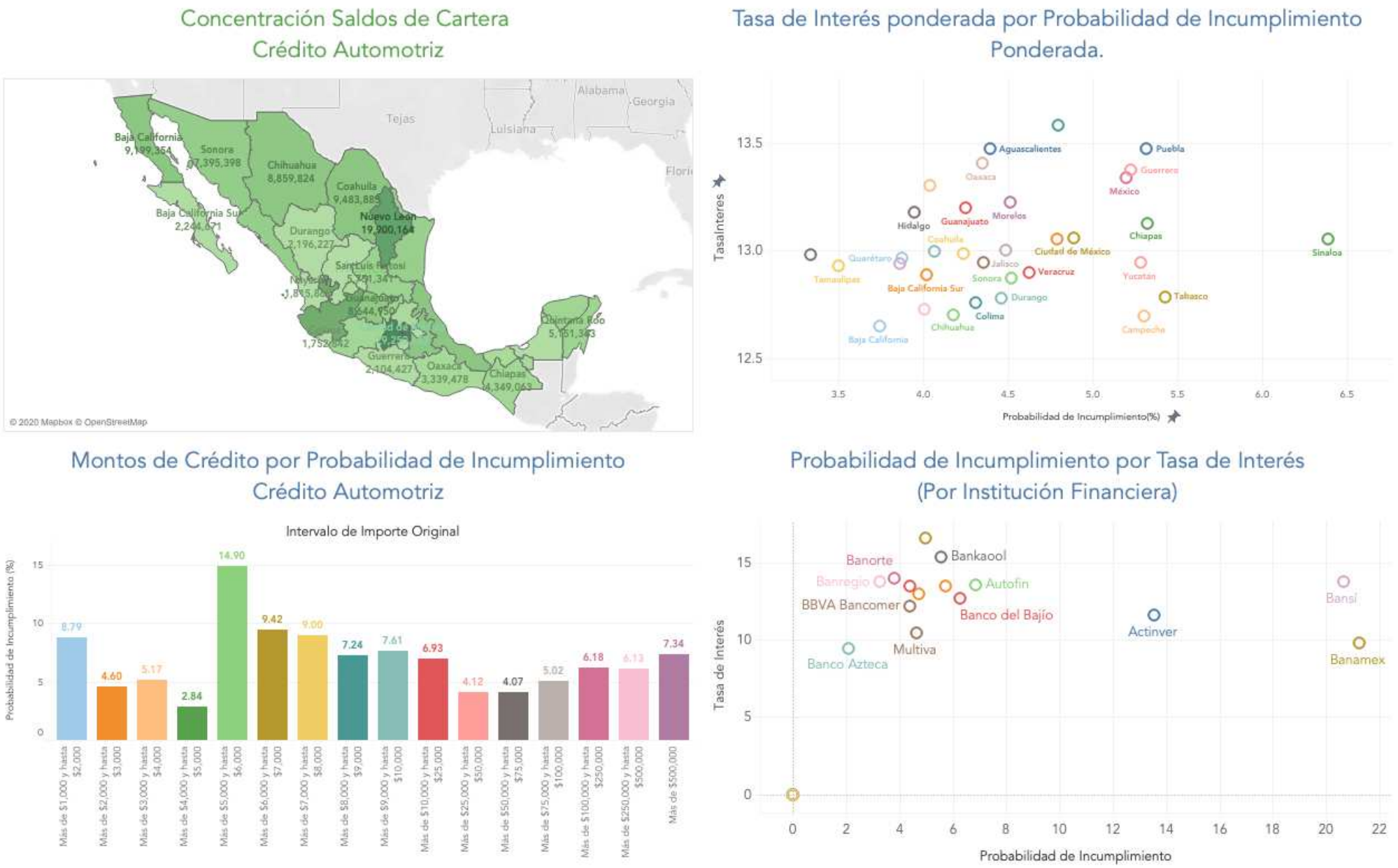
Cuadro 10. Correlaciones ABCD PI y Variación Saldos Cartera

Test Shapiro-Wilk

Test Shapiro	p-Value	W		
Probabilidad Incumplimiento	0.06054405	0.93674269		
Variación	0.1672812	0.95228276		
	Correlación	R^2	p-value	Rho
Correlación Pearson PI-Variación	-0.6889029	0.47458723	0.00001304	
Correlación Parcial sin IHH	-0.6371337	0.4059394	0.00011606	
Correlación Spearman		0.37475222	0.00026324	-0.6121701

Encontramos que existe una correlación fuerte y negativa (-.6889) entre la variación de saldos de cartera destinadas a las Entidades Federativas y la probabilidad de incumplimiento, tal es así que explica en 47% dicho efecto, aún si le hacemos la correlación a través de la correlación parcial, quitando el IHH, la variable es explicativa al 40% y continúa siendo significativa. Por lo que concluimos que en crédito ABCD la probabilidad de incumplimiento de un año anterior, incide en la colocación del próximo año. De manera que, la probabilidad de incumplimiento es factor de inclusión en ABCD, ya que por modelo de riesgo se comienzan a excluir Entidades con PI alta.

Figura 5. Crédito Automotriz



2.6 Crédito Automotriz

La CUB de Bancos (2020) define a través del mismo artículo 91 BIS, al crédito automotriz como:

"A los créditos que sean otorgados a personas físicas y cuyo destino sea la adquisición de vehículos automotrices particulares."

Es importante siempre mencionar la definición, ya que fácilmente se podría confundir con un crédito de arrendamiento automotriz, ya que los dos intrínsecamente están asociados al mismo bien, sin embargo la característica que los diferencia, es que este crédito se encuentra respaldado en la adquisición de un auto y el otro en arrendarlo, lo que nos lleva a tener diferentes niveles de riesgo.

Al aplicarle el IHH a nivel Institución Financiera, este indicador llega a ser de 0.1369 lo que revela que está altamente concentrado por Institución. A continuación, enunciamos las 5 instituciones que mayor participación tienen en la otorgación de créditos automotrices y concentran 75.44% de todo el portafolio.

Cuadro 11. Concentración de Crédito Automotriz por Institución (Miles de Pesos)

Banco	Saldo Cartera	Distribución
BBVA Bancomer	$ 54,756,251.00	23.77%
General Motors Financial	$ 43,806,974.62	19.02%
Banorte	$ 26,924,568.68	11.69%
Scotiabank	$ 25,626,792.71	11.13%
Cetelem (BNP Paribas)	$ 22,665,234.15	9.84%
Total 5 Instituciones Financieras	$ 173,779,821.15	75.44%
Total (31 Participantes).	$ 230,349,998.82	

A través de la Gráfica 8 observamos que dicho producto de crédito se encuentra concentrado en las ciudades de más dinamismo industrial, ya que la Entidad Federativa que lidera este tipo de Crédito es Ciudad de México, Estado de México, seguido de Nuevo León.

Cuadro 12. Concentración de Crédito por Entidad Federativa AUTO (Miles de Pesos)

Estado	Saldo Cartera	Tasa Interés	Probabilidad de Incumplimiento
Ciudad de México	$ 29,250,153.00	13.05%	4.88%
México	$ 27,539,467.00	13.33%	5.19%
Nuevo León	$ 19,900,164.00	12.94%	4.35%
Jalisco	$ 16,274,162.00	12.99%	4.49%
Coahuila	$ 9,483,883.00	12.98%	4.23%

Observamos que al menos en esta distribución, existe una correlación entre tasa de interés y probabilidad de incumplimiento, ya, al estimar el coeficiente correlación de Pearson, da como resultado lo siguiente:

	Correlación	R^2
Correlación Tasa de Interés	0.23497141	0.05521156

Lo que nos arroja la correlación de Pearson, es que existe una muy débil, pero positiva relación entre la tasa de interés ponderada y la probabilidad de incumplimiento y que explica sólo en 5.5% esta relación. Dicha probabilidad de incumplimiento viene dada por el modelo del artículo 91 BIS para productos definidos como "A"

Cuadro 13. Modelo probabilidad de incumplimiento ABCD CNBV AUTO

Coeficiente	Valor	VAR
β_0	-2.0471	Riesgo Natural (Beta autónoma)
β_1	1.0837	ATR
β_2	-0.7863	%Pago
β_3	0.5473	Alto
β_4	0.0587	Medio
β_5	-0.606	Bajo
β_6	-0.1559	Meses
Riesgo Natural	**11.43%**	
PI Escenario A	**0.42%**	
Pi Escenario B	**82.56%**	

Respetando los anexos y reglas establecidas del mismo numeral, establecemos dos tipos de Escenarios:

A.) Escenario al corriente (A): Deuda sin ningún atraso, con cuatro pagos exigibles completamente pagados en los últimos 4 meses, con endeudamiento bajo y con antigüedad de más de 54 meses con otras instituciones e internamente con una antigüedad mayor a 13 meses.

B.) Escenario de máximo impago (B) (Sin llegar ATR >3): Tres atrasos, deuda, un solo pago completamente cubierto en los últimos 4 meses, cliente nuevo, alto riesgo de concentración de deuda (Mayor a 65%) y con el mínimo de meses transcurridos desde el último impago

Cuadro 14. Escenarios PI AUTO

Var	Escenario A	Escenario B
Riesgo Natural (Beta autónoma)	1	1
ATR	0	3
%Pago	1	0.25
Alto	0	1
Medio	0	0
Bajo	1	0
Meses	13	0
PI	0.42%	82.56%
Reservas por cada 10K[24]	$ 36.21	$ 7,100.12

De manera que observamos que el aprovisionamiento por este producto es nulo en el escenario A, debido a que por cada $10,000.00, se debería reservar el equivalente a 0.36%, y 71% en caso de un problema mayor (escenario B), esto representa un costo menor que en crédito ABCD. Esto se debe a que hay un riesgo menor asumido debido a que la garantía que encara este producto es el auto, por lo que el riesgo se ve mitigado al estar colateralizado.

Por otra parte, al analizar los montos de exposición, vemos que la distribución que se comporta de la siguiente manera:

Tabla 15. PI Intervalo importe original del crédito AUTO

Intervalo de Importe Original	PI[25]
Hasta $1,000	46.05%
Más de $1,000 y hasta $2,000	8.79%
Más de $2,000 y hasta $3,000	4.60%
Más de $3,000 y hasta $4,000	5.17%
Más de $4,000 y hasta $5,000	2.84%
Más de $5,000 y hasta $6,000	14.90%
Más de $6,000 y hasta $7,000	9.42%
Más de $7,000 y hasta $8,000	9.00%
Más de $8,000 y hasta $9,000	7.24%
Más de $9,000 y hasta $10,000	7.61%
Más de $10,000 y hasta $25,000	6.93%
Más de $25,000 y hasta $50,000	4.12%
Más de $50,000 y hasta $75,000	4.07%
Más de $75,000 y hasta $100,000	5.02%

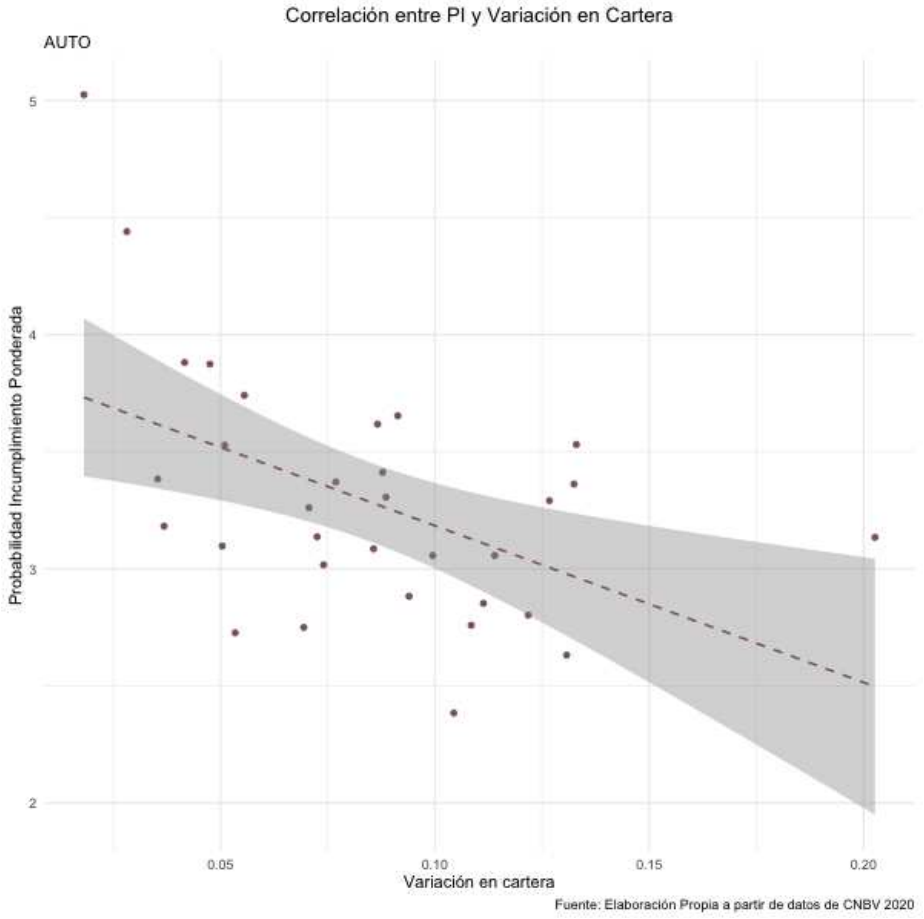
[24] Se considera como severidad de la pérdida dentro del intervalo donde ATR <=3. Se considera sólo sobre cálculos de saldo insoluto, sin considerar, interés devengado no cobrado e IVA.
[25] Esta PI ponderada, fue calculada sobre la distribución de saldos presentados a 2019, pero con la PI de 2017. Ya que la CNBV no tiene información actualizada de ese rubro.

Más de $100,000 y hasta $250,000	6.18%
Más de $250,000 y hasta $500,000	6.13%
Más de $500,000	7.34%

No observamos parsimonia entre el *exposure* y la Probabilidad de Incumplimiento, por lo que determinamos que los montos no están definidos de acuerdo con el riesgo que conlleva. A estas cifras no podemos darle el 100% de certeza, ya que los datos no corresponden a la temporalidad y asumimos que se mantienen las distribuciones a lo largo del tiempo, lo hace evidente un sesgo en el análisis.

Por último, medimos el efecto de la variación de cartera que se tienen de los saldos entre año 2017 y 2018, comparando con la probabilidad de incumplimiento del año 2017, esto con motivo de darle soporte a nuestro análisis en donde suponemos que las instituciones dejan de colocar en Entidades Federativas por la probabilidad de riesgo común que tiene cada una.

Figura 6. Correlación PI y Variación Cartera (Saldos Totales)



Test Shapiro-Wilk

Test Shapiro	P-Value	W
Probabilidad Incumplimiento	0.0189453	0.91850886
Variación	0.2429753	0.95806584

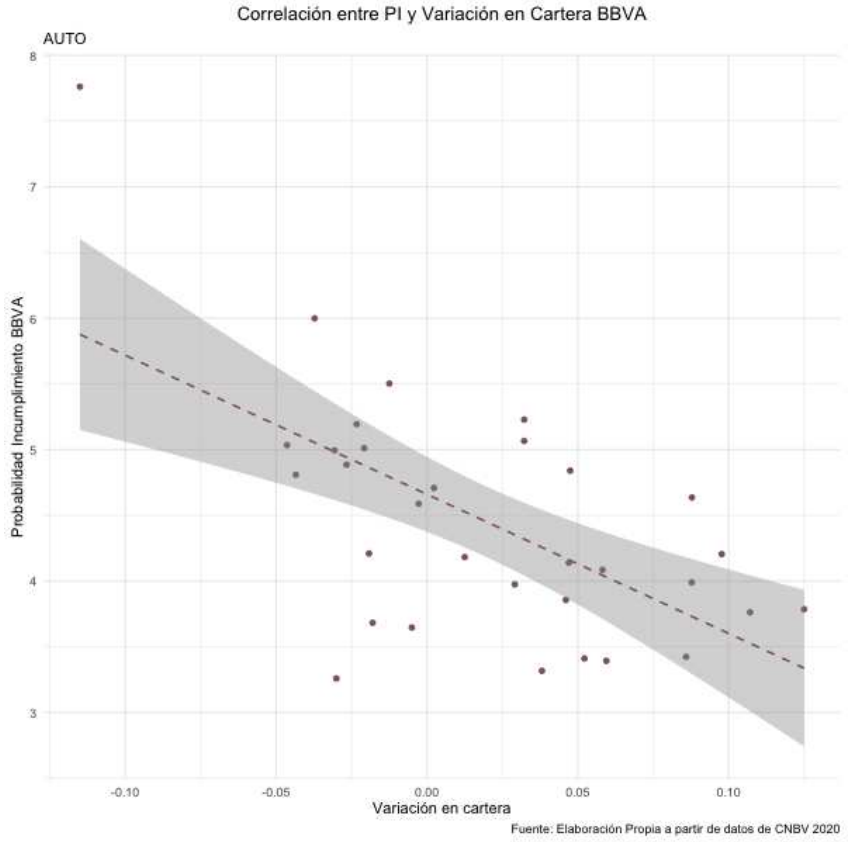
Correlación	R2	Pvalue	Rho

Correlación Pearson PI-Variación	-0.4764559	0.22701021	0.005838	
Correlación Parcial sin IHH	-0.4220552	0.1781306	0.01802813	
Correlación Spearman		0.19414231	0.01227236	-0.4406158

Podemos deducir que, por el valor del *p-value* del *Test* Shapiro, de la variable PI, que no podemos usar la correlación de Pearson, por lo que hacemos uso de la correlación *Spearman*, del cual asumimos que hay una correlación medianamente débil y negativa entre el aumento del PI y la variación de los saldos de cartera. Vemos que la Rho en este caso es de -0.44, mientras que explica un 19.41% la variación de saldos en cartera. Por lo que determinamos que, en crédito Automotriz, existe una propensión a dejar de colocar ante aumentos de PI.

Cuando analizamos este mismo efecto con la institución con más peso en el sistema (BBVA Bancomer), encontramos que dicha relación se agudiza. Observamos que las Entidades Federativas con mayor PI en 2017, en 2018 sufrieron una variación negativa en los saldos de cartera.

Figura 7. Correlación PI y Variación Cartera AUTO (Saldos BBVA)



***Test* Shapiro-Wilk**

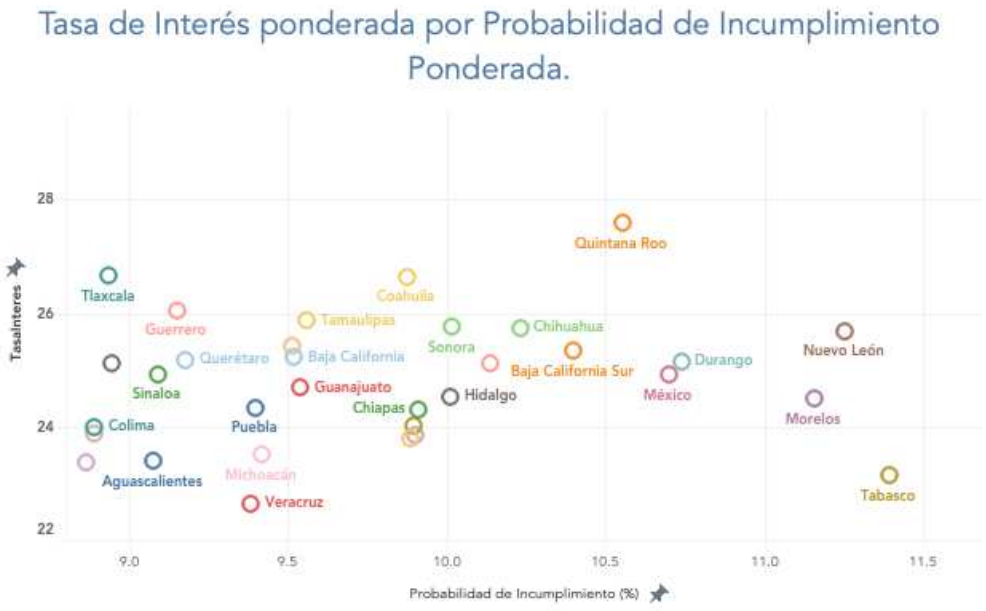
Test Shapiro	***p*-Value**	**W**
PI	0.003	0.889

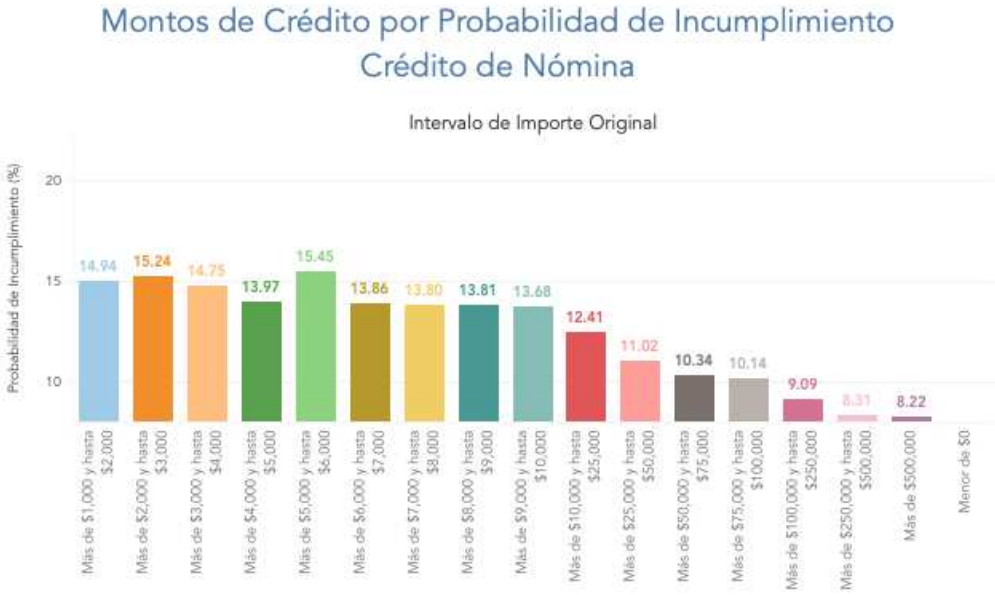
	Correlación	**R2**	***p*-value**	**Rho**
Correlación Pearson	-0.589	0.347	0.006	

Variación Car	0.497	0.970		Correlación Spearman		0.265	0.003	-0.514

Por el *p*-value mostrado en el *test* de Shapiro, descartamos el uso de la correlación Pearson, sin embargo, la correlación Spearman nos muestra que tiene una Rho de -0.514 lo que implica que hay una relación fuerte y negativa ante aumentos de la PI, con respecto de la variación de Saldos en cartera.

Gráfica 8. Crédito de nómina





2.7 Crédito de Nómina

La CUB de Bancos (2020) los define como al crédito de nómina como:

"Los créditos de liquidez que sean otorgados por la Institución que administra la cuenta de nómina del acreditado y que sean cobrados a través de dicha cuenta. No se considerará como crédito de "nómina" cuando la Institución no realice la cobranza de estos créditos a través de la cuenta de nómina del acreditado, por lo que estos deberán considerarse como 'Personales'"

Es decir, que sólo se podrá considerar como crédito de nómina siempre y cuando la nómina sea administrada por la misma institución que otorga el crédito. Esto es importante recalcarlo, ya que el nivel de riesgo es diferente, debido a que las gestiones de cobranza son de manera directa, debitando el pasivo contraído, incluso, antes de que el acreditado pueda tocar sus ingresos, lo que mitiga de manera abismal el riesgo. El riesgo que realmente presenta este tipo de créditos se fundamenta prácticamente sobre dos situaciones:

a.) Que el acreditado cambie de institución donde tiene la nómina y no se pueda debitar directamente.
b.) Que el acreditado pierda su empleo, por lo que presente insuficiencia de recursos en la cuenta de nómina.

El índice de concentración IHH de saldos de cartera por Institución Financiera presenta una estimación de 0.2418, es decir que se encuentra concentrada, incluso, en mayor medida que el crédito automotriz

Cuadro 16. Concentración de Crédito Nómina por Institución (Miles de Pesos)

Empresa	Total	Porcentaje
BBVA Bancomer	$ 98,100,674.79	37.79%
Banorte	$ 50,608,276.00	19.50%
Banamex	$ 49,298,747.71	18.99%
Santander	$ 36,012,988.09	13.87%
HSBC	$ 19,403,286.98	7.47%
Total 5 instituciones	$ 253,423,973.57	97.62%
Total Sistema (160. participantes)	$ 259,595,772.32	

Observamos que esta razón se da, debido a que las 5 instituciones con mayor saldo de cartera concentran el 98.18% de toda la cartera perteneciente a créditos de nómina, cuyo principal proveedor de este servicio es Bancomer con 38% de la cartera en total.

Cuadro 17. Concentración de Cartera por Entidad Federativa Crédito Nómina

Estado	Saldo Cartera	Tasa Interés	Probabilidad de Incumplimiento
México	35,822,913	24.92%	10.70%
Ciudad de México	33,833,481	24.03%	9.89%
Veracruz	16,836,995	22.66%	9.38%
Jalisco	13,747,145	23.85%	9.90%
Nuevo León	13,253,189	25.69%	11.25%

Observamos que la Entidad Federativa que concentra mayor saldo de cartera en créditos de nómina es el Estado de México, a la par observamos una leve relación entre tasa de interés y probabilidad de incumplimiento por estado, al estimar la correlación de Pearson obtenemos lo siguiente:

	Correlación	R^2
Correlación Tasa de Interés	0.15006015	0.02251805

Por lo que, determinamos que es una correlación bastante débil y que el premio al riesgo a nivel sistema no se ve reflejado. Esta probabilidad de incumplimiento al igual que en los demás productos deriva del artículo 91 BIS en la categoría "N".

Cuadro 18. Modelo probabilidad de incumplimiento Nómina CNBV

Coeficiente	Valor	VAR
β_0	**-2.4285**	Riesgo Natural (Beta autónoma)
β_1	**1.4298**	MAXATR
β_2	**0.4428**	Alto
β_3	**0.0616**	Medio
β_4	**-0.5044**	Bajo
β_5	**-0.047**	Meses
Riesgo Natural	**8.10%**	
Escenario A	**2.81%**	
Escenario B	**90.92%**	

Respetando las reglas del numeral, proponemos dos escenarios:

a.) Escenario corriente (A): Acreditado con pago al corriente en los últimos trece meses, ratio de endeudamiento menor a 0.40 respecto de la deuda total con otras instituciones a la fecha de corte y una antigüedad mayor a 29 meses dentro de la institución otorgante del crédito.

b.) Escenario de impago (B) (Sin llegar a ATR >3): Cliente con atraso de 3 periodos de facturación, un ratio de endeudamiento mayor a 0.40, una antigüedad menor a 29 meses.

Cuadro 19.Escenarios PI Nómina

VAR	Escenario A	Escenario B
Riesgo Natural (Beta autónoma)	1	1
MAXATR	0	3
Alto	0	1
Medio	0	0
Bajo	1	0
Meses	13	0
PI	2.81%	90.92%
Reservas por cada 10K[26]	$ 241.56	$ 7,818.97

Aquí observamos de primera instancia que el costo por cada $10,000.00 pesos por tener al corriente un crédito se eleva con respecto al de crédito ABCD ($139.40) y al Automotriz ($36.21). En un escenario de mayor problema, observamos que la PI es más baja que en ABCD y más alta que en crédito automotriz. Vemos que este tipo de crédito, el riesgo en términos generales es mayor, porque dicho crédito no tiene en colateral ningún bien físico, lo que ampara la operación de crédito es el contrato de nómina, pero como lo mencionamos anteriormente, existe el riesgo latente de cambiar de administración en la nómina, lo que pone en mayor riesgo de impago la operación.

Cuando analizamos los importes originales de crédito (en la Gráfica 10) versus la probabilidad de incumplimiento observamos que hay una tendencia más acentuada entre monto y PI. Por ahora, este ha sido el único que retoma esta parsimonia, no se puede hacer ningún análisis de correlación, debido a que la presentación del reporte R-R11 ya está definida dicha estructura, para ello detallamos el siguiente cuadro.

Cuadro 20. PI Intervalo importe original del crédito Nómina

Intervalo de Importe Original	PI[27]
Hasta $1,000	20.98%
Más de $1,000 y hasta $2,000	14.94%
Más de $2,000 y hasta $3,000	15.24%
Más de $3,000 y hasta $4,000	14.75%
Más de $4,000 y hasta $5,000	13.97%
Más de $5,000 y hasta $6,000	15.45%
Más de $6,000 y hasta $7,000	13.86%
Más de $7,000 y hasta $8,000	13.80%
Más de $8,000 y hasta $9,000	13.81%

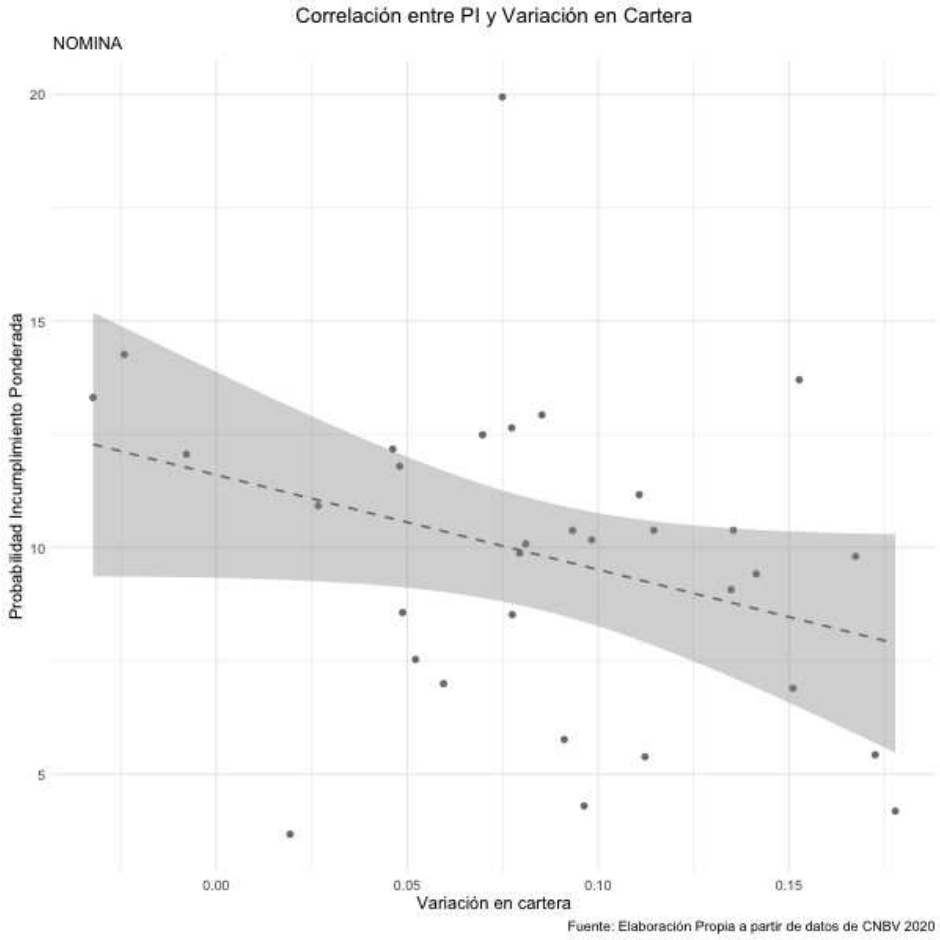
[26] Se considera como severidad de la pérdida dentro del intervalo donde ATR <=3. Se considera sólo sobre cálculos de saldo insoluto, sin considerar, interés devengado no cobrado e IVA.
[27] Esta PI ponderada, fue calculada sobre la distribución de saldos presentados a 2019, pero con la PI de 2017. Ya que la CNBV no tiene información actualizada de ese rubro.

Más de $9,000 y hasta $10,000	13.68%
Más de $10,000 y hasta $25,000	12.41%
Más de $25,000 y hasta $50,000	11.02%
Más de $50,000 y hasta $75,000	10.34%
Más de $75,000 y hasta $100,000	10.14%
Más de $100,000 y hasta $250,000	9.09%
Más de $250,000 y hasta $500,000	8.31%
Más de $500,000	8.22%

En donde observamos que es decreciente la tendencia, por otra parte, al igual que en las demás créditos, en esta información no se tiene la confianza al 100% debido a que está sostenida con cifras de diferentes temporalidades, esto se debe a que la CNBV no tiene actualizado dicho reporte.

Por último, medimos las variaciones existentes entre la Probabilidad de Incumplimiento de cierre de 2017, versus la variación de saldos de cartera de 2017 vs 2018, lo que nos indicará sí en términos generales las instituciones dejan de colocar crédito zonas donde el riesgo es mayor.

Gráfica 9. Correlación PI y Variación Cartera Nómina (Saldos Totales)

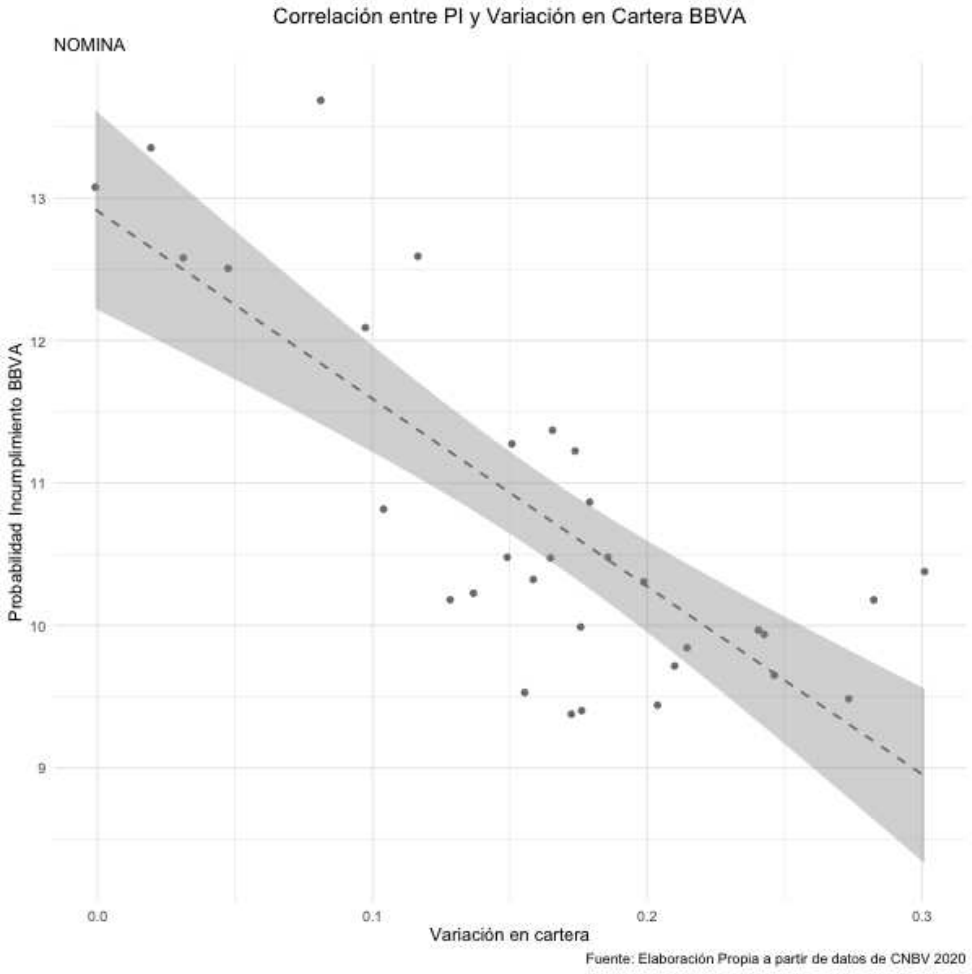


Test Shapiro-Wilk

Test Shapiro	**P-Value**	**W**		
Probabilidad Incumplimiento	0.25812383	0.95901687		
Variación	0.58275152	0.97289209		
	Correlación	**R^2**	**Pvalue**	**Rho**
Correlación Pearson PI-Variación	-0.3019025	0.09114513	0.09309	
Correlación Parcial sin IHH	-0.2070306	0.04286168	0.26378531	
Correlación Spearman		0.13090179	0.0425789	-0.3618035

Observamos que a nivel sistema, podemos utilizar la correlación de Pearson, ya que, por el Test de Shapiro, podemos asumir normalidad en las variables a utilizar. De manera que la correlación de Pearson nos muestra que está débilmente correlacionada la variable y explica sólo hasta en 9.11% esa relación y deja de tener sentido cuando le quitamos el factor IHH.
Cuando tomamos este análisis con la Institución Financiera de mayor concentración de cartera (BBVA Bancomer) observamos lo siguiente:

Gráfica 10. Correlación PI y Variación Cartera Nómina (Saldos BBVA)



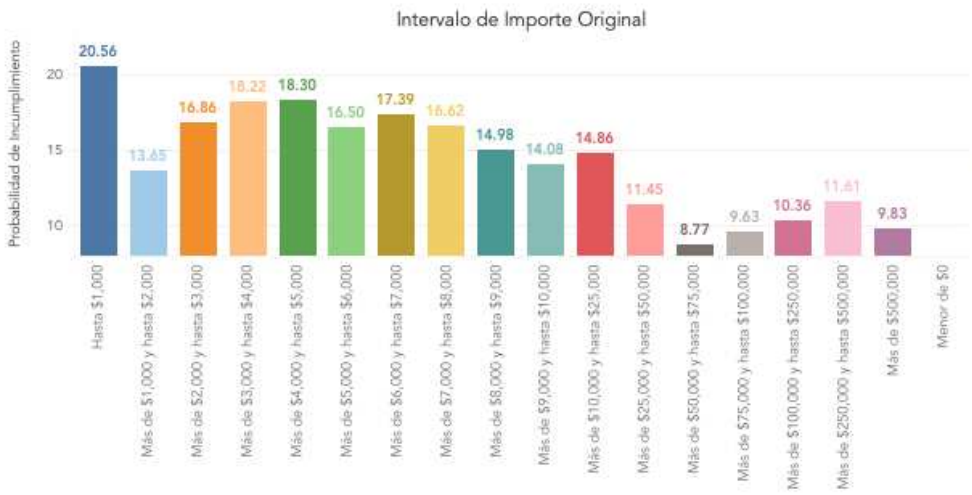
Test Shapiro-Wilk

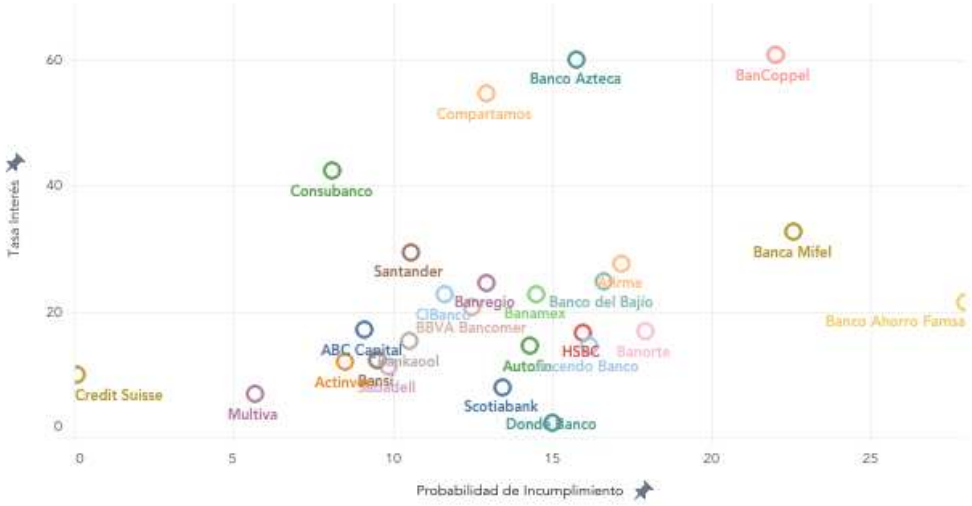
Test Shapiro	**p-Value**	**W**
PI	0.00144015	0.87402989
Variación	0.55652854	0.97200613

	Correlación	**R^2**	**p-value**	**Rho**
Correlación Pearson	-0.7775	0.6044	0.0931	
Correlación Spearman		0.4882	0.0000	-0.6987

Al aplicar el *test* de Shapiro, observamos que no podemos asumir normalidad en la PI, por lo que al aplicar la correlación Spearman observamos que la Rho es de -0.6987 lo que indica una relación fuerte y negativa entre la PI y la variación interanual de saldos de cartera y que explica hasta en un 48.82% esa relación. Por lo que determinamos que hay exclusión en Créditos de Nómina por localidad ante incrementos en PI.

Gráfica 11. Crédito Personal





2.8 Créditos personales

La CUB de bancos define como crédito personal:

"A los créditos que sean cobrados por la Institución por cualquier medio de pago distinto de la cuenta de nómina". Es decir que, este contrato es el que podría presentar aún mayor riesgo, debido a que no hay ningún tipo de colateral que ampare la operación a diferencia del ABCD, Automotriz o el de Nómina, aunque este último tuviera un colateral más laxo que los anteriores. Sin embargo, como lo vimos en la gráfica 4, es el producto más ofertado por el sistema. El IHH del sistema, tomando como base las instituciones da un 0.1156 lo que representa que se encuentra medianamente concentrado.

Cuadro 21. Concentración de Crédito Personal por Institución (Miles de Pesos)

Empresa	Total	Distribución
BBVA Bancomer	$ 40,664,232.92	16.65%
Banjército	$ 38,058,717.97	15.58%
Banamex	$ 37,899,324.46	15.51%
Banco Azteca	$ 34,745,778.05	14.22%
Santander Consumo	$ 19,008,022.21	7.78%
Total 5 Instituciones	$ 170,376,075.61	69.74%
Total Sistema (44 Participantes)	$ 244,287,859.42	

Sí revisamos la concentración por entidad federativa, nos encontramos que 0.0736 en puntos IHH lo que representa que está medianamente concentrado en saldos, a diferencia de lo visto en concentración por contratos brutos que era de 0.0593 (Tabla 2).

Cuadro 22. Concentración de Cartera por Entidad Federativa Crédito Personal

Estado	Saldo Cartera	Tasa Interés	Probabilidad de Incumplimiento
Ciudad de México	44,172,670	25.51	11.40
México	34,272,751	35.39	14.11
Jalisco	14,374,939	29.10	13.76
Veracruz	13,094,846	37.07	14.91
Nuevo León	11,960,788	25.55	15.69

Vemos que existe una leve relación entre la tasa de interés, y la PI, sin embargo al utilizar la correlación, obtenemos que hay una correlación negativa, pero con una significancia muy pequeña.

	Correlación	R^2
Correlación Tasa de Interés	-0.252728	0.06387142

Por lo que vemos que es muy débil, esta probabilidad de incumplimiento, deriva del Artículo 91 BIS

Cuadro 23.Modelo probabilidad de incumplimiento Personal CNBV

Coeficiente	Valor	VAR
β_0	**-1.2924**	Riesgo Natural (Beta autónoma)
β_1	**0.8074**	ATR
β_2	**-1.1984**	DEL
β_3	**0.3155**	MAXATR
β_4	**-0.8247**	%PAGO
β_5	**0.4404**	Alto
β_6	**0.0405**	Medio
β_7	**-0.4809**	Bajo
β_8	**-0.054**	Meses
Riesgo Natural	**21.54%**	
Escenario A	**1.10%**	
Escenario B	**3.56%**	
Escenario C	**90.97%**	

Se observa que el riesgo natural, es mucho mayor al de los demás productos, esto como ya mencionamos, se debe a que no se tiene ningún colateral que mitigue el riesgo de crédito, sin embargo hay una variable (DEL) en este modelo asume como una reducción del riesgo el que haya una cobranza delegada, es decir que, el pago realizado sea con cargo o descuento directo al salario de los acreditados a través de su empleador, lo que representa que "terciarizan" las acciones de cobranza al empleador, lo que estriba en una reducción del riesgo de crédito. Esto lo podemos ver más a fondo al hacer tres escenarios:

a.) Escenario al corriente con cobranza delegada (A): Cliente al corriente, sin ningún atraso en los últimos 13 meses, con cobranza delegada, con ratio deuda menor al 0.085 con respecto otras instituciones y con una antigüedad mayor a 28 meses.
b.) Escenario al corriente sin cobranza delegada (B): Cliente al corriente, sin ningún atraso en los últimos 13 meses, sin cobranza delegada, con ratio deuda menor al 0.085 con respecto otras instituciones y con una antigüedad mayor a 28 meses.
c.) Escenario de mayor problema (C) (Sin llegar a ATR>3): Cliente con atraso de 3 periodos consecutivos de facturación en los últimos 4 meses, sin cobranza delegada, sin antigüedad.

Lo que observamos de acuerdo a estos escenarios, es lo siguiente:

Cuadro 24. Escenarios PI Personal

VAR	Escenario A	Escenario B	Escenario C
Riesgo Natural (Beta autónoma)	1	1	1
ATR	0	0	3
DEL	1	0	0
MAXATR	0	0	3
%PAGO	1	1	0.25
Alto	0	0	1
Medio	0	0	0
Bajo	1	1	0
Meses	13	13	0
PI	1.10%	3.56%	90.97%
Reservas por cada 10K	$ 94.64	$ 305.91	$ 7,823.81

Vemos que hay una diferencia marcada entre que haya una cobranza delegada y el que no haya ese esquema, incluso cuando se está al corriente, esta variable hace que se reserve cerca de 3 veces más por lo que impacta de manera negativa los estados de resultados a nivel reservas. Es importante mencionar que el reservamiento incluso es menor si lo comparamos con el crédito de nómina en un escenario al corriente, por lo que podrían existir estrategias de swap de contrato (obvio, siempre y cuando existan convenios con las instituciones empleadoras) dentro de las instituciones con el objetivo de reservar menos.
Cuando observamos la distribución de importes originales del crédito vemos que sucede lo siguiente:

Cuadro25. PI Intervalo importe original del crédito Personal

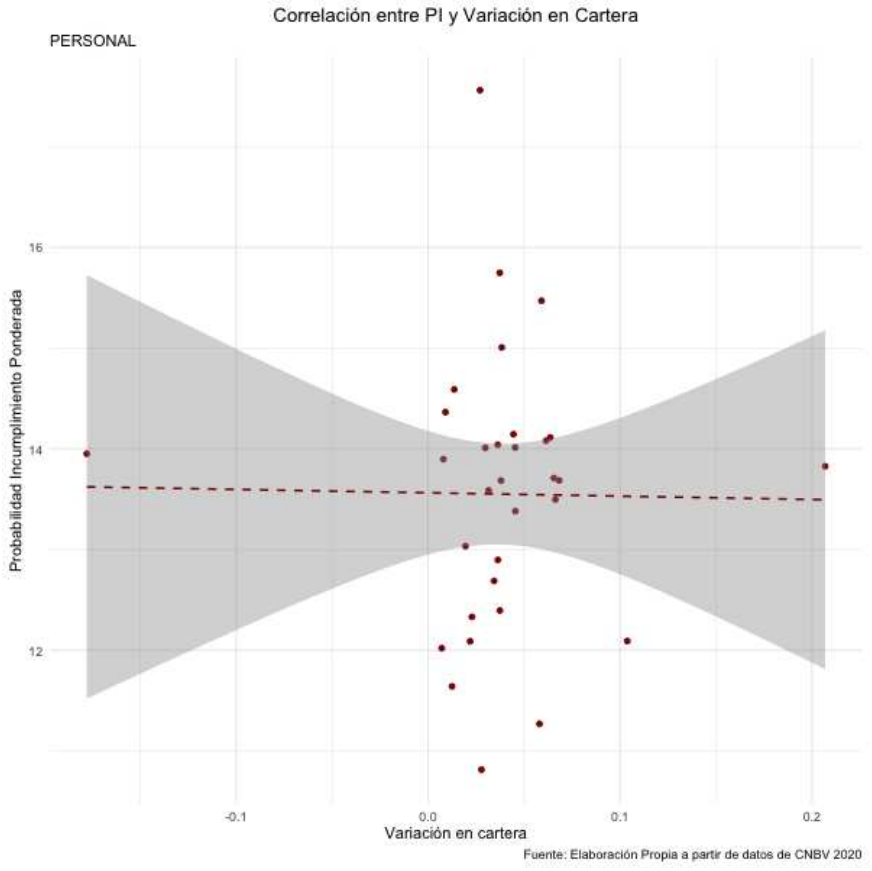
Intervalo de Importe Original	Probabilidad de [28]Incumplimiento
Hasta $1,000	20.56%
Más de $1,000 y hasta $2,000	13.65%
Más de $2,000 y hasta $3,000	16.86%
Más de $3,000 y hasta $4,000	18.22%
Más de $4,000 y hasta $5,000	18.30%
Más de $5,000 y hasta $6,000	16.50%
Más de $6,000 y hasta $7,000	17.39%
Más de $7,000 y hasta $8,000	16.62%
Más de $8,000 y hasta $9,000	14.98%
Más de $9,000 y hasta $10,000	14.08%
Más de $10,000 y hasta $25,000	14.86%
Más de $25,000 y hasta $50,000	11.45%
Más de $50,000 y hasta $75,000	8.77%

[28] Esta PI ponderada, fue calculada sobre la distribución de saldos presentados a 2019, pero con la PI de 2017. Ya que la CNBV no tiene información actualizada de ese rubro.

Más de $75,000 y hasta $100,000	9.63%
Más de $100,000 y hasta $250,000	10.36%
Más de $250,000 y hasta $500,000	11.61%
Más de $500,000	9.83%

De manera que no observamos parsimonia clara entre que exista un mayor riesgo, en montos pequeños, y un menor riesgo en montos grandes.
Lo que nos lleva a realizar un análisis entre la relación entre la probabilidad de incumplimiento y el nivel de variación en la cartera para ver sí una incide en la otra, para ello tomamos como base las probabilidades de incumplimiento ponderadas por el saldo total en cada rubro, versus la variación de saldos a nivel sistema por Entidad Federativa.

Figura 12. Correlación entre PI y variación de cartera Personal (Saldos Totales)



***Test* Shapiro-Wilk**

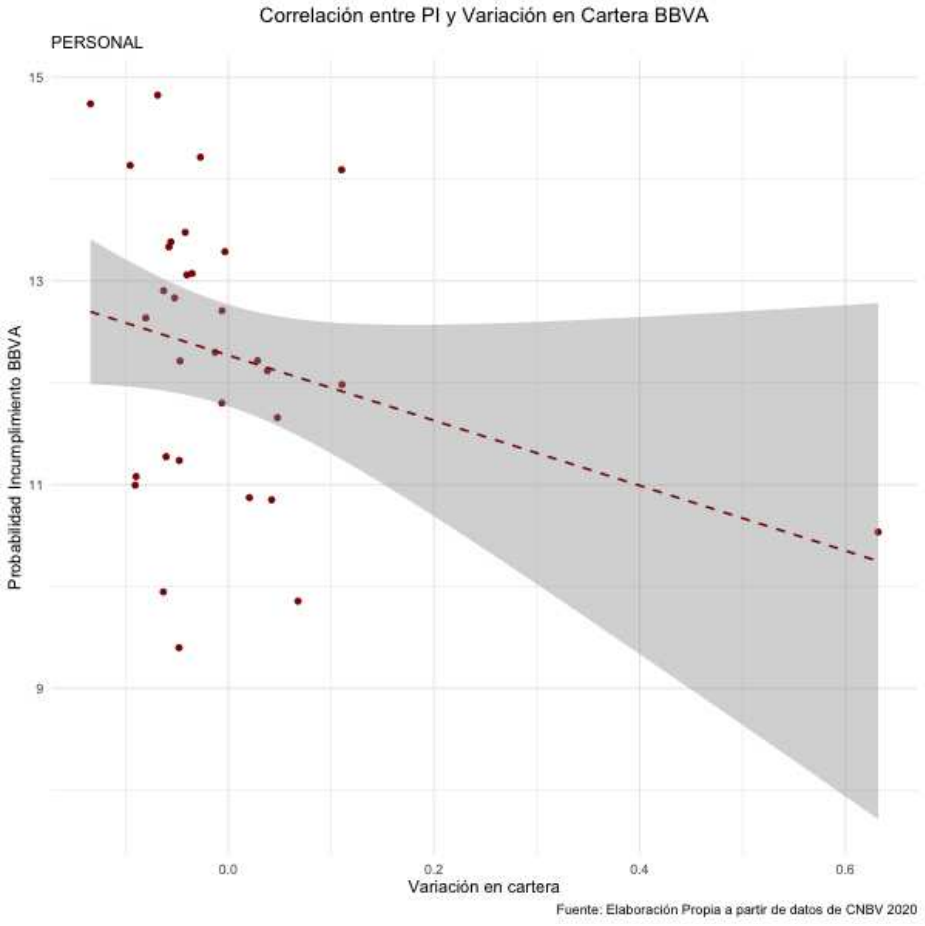
***Test* Shapiro**	**P-Value**	**W**
Probabilidad Incumplimiento	0.22393619	0.95679035
Variación	1.0195E-06	0.70460614

	Correlación	R2	Pvalue	Rho
Correlación Pearson PI-Variación	-0.0103684	0.0001075	0.9551	
Correlación Parcial sin IHH	-0.0613607	0.00376514	0.74297848	
Correlación Spearman		0.01008817	0.5830551	0.10043988

Primeramente, denotamos el hecho que no podemos usar la correlación de Pearson, ya que por el *test* de Shapiro, vemos que en la variación no se asume normalidad. Aun así, a nivel sistema, no existe ni gráficamente, pero tampoco a nivel estadístico, una ligera correlación entre la PI y la variación de saldos, como en productos anteriores, lo que llama la atención, ya que las correlaciones son muy débiles.

Cuando analizamos al principal proveedor de este servicio a nivel instituciones (BBVA Bancomer) encontramos lo siguiente:

Figura 13. Correlación entre PI y variación de cartera Personal (Saldos BBVA)



Test Shapiro-Wilk

Test Shapiro	P-Value	W
PI	0.818	0.981
Variación	0.000	0.580

	Correlación	R2	P-value	Rho
C Pearson	-0.2883	0.0831	0.9551	
C Spearman		0.0730	0.1346	-0.2702

Gráfica y estadísticamente podemos constatar el mismo fenómeno que a nivel sistema, esto se puede deber a que hubo un crecimiento anormal en una entidad (Durango), y ese *outlier* puede sesgar el análisis sin embargo en la gráfica 17 vemos aun así, sin Durango, no existe correlación marcada entre PI y variación de saldos de cartera.

2.9 Microcrédito

La CUB de Bancos (2020) define al microcrédito como sigue:

"Son créditos al consumo, los cuales podrán ser otorgados a personas físicas cuyos recursos estén destinados a financiar actividades de producción o comercialización de bienes o prestación de servicios, en los que la fuente principal de pago la constituyen los ingresos obtenidos por dichas actividades y cuyos montos y plazos serán bajo alguna de las modalidades siguientes:

1. Individual: cuando el crédito sea otorgado a un solo individuo y teniendo como límite máximo el monto equivalente en moneda nacional a 30,000 UDIS y un plazo máximo de tres años.

2. Grupal: cuando el crédito sea otorgado a grupos de individuos que avalen los adeudos o se constituyan como deudores solidarios entre sí y teniendo como límite máximo el monto equivalente en moneda nacional de 11,500 UDIS por cada integrante del grupo y un plazo máximo de un año.

Como lo mencionábamos, no existe información acerca de este grupo particular de créditos, siendo que pertenecen a la categoría de Consumo. Instituciones Financieras como Compartamos Banco, sólo tienen registrado en reportes abiertos de la CNBV un total de $99.2 millones de pesos bajo la modalidad de créditos personales, mientras que Compartamos a través de sus Estados Financieros reporta una cartera al consumo de $26,518 millones de pesos, lo que marca una sesgo importante para nuestro análisis, ya que no contamos con la información para ver la probabilidad de incumplimiento que tiene el microcrédito, con respecto a todo el sistema. Esto es importante de recalcar, porque de ahí, pueden derivar razones, por las cuáles, la banca tradicional no se apunta a otorgar este tipo de créditos.

De la misma manera, retomaremos el análisis de Calificación de Cartera, tratando de encontrar el riesgo intrínseco que conlleva este producto.

Cuadro 26. Probabilidad de Incumplimiento Microcrédito Individual

Coeficiente	Valor	VAR
β_0	**-1.2924**	Riesgo Natural (Beta autónoma)
β_1	**0.8074**	ATR
β_2	**0.3155**	MAXATR
β_3	**-0.8247**	%PAGO

β_4	**0.4404**	Alto
β_5	**0.0405**	Medio
β_6	**-0.4809**	Bajo
β_7	**-0.054**	Meses
β_8	**-0.0282**	CAP

Cabe mencionar que este el modelo de aprovisionamiento de cartera para Microcrédito individual es idéntico al crédito Personal, inclusive las betas, son iguales, a excepción de que tienen la característica de que le asignan una variable *dummy* (CAP) la cuál castiga al crédito si es que pertenece a la cartera de Microcrédito al tener signo negativo y no tiene la variable de cobranza delegada (DEL). Al igual que en los demás contratos, establecemos escenarios que nos ayudarán a determinar el precio que se tiene bajo ciertos supuestos:

a.) Escenario al corriente (A): Cliente al corriente, sin ningún atraso en los últimos 13 meses, pertenece a la cartera de microcrédito, con ratio deuda menor al 0.085 con respecto otras instituciones y con una antigüedad mayor a 28 meses.

b.) Escenario de mayor problema (B) (Sin llegar a ATR>3): Cliente con atraso de 3 periodos consecutivos de facturación en los últimos 4 meses, perteneciente a cartera de microcrédito, sin antigüedad.

Cuadro 27. Escenarios PI Microcrédito Individual

VAR	**Escenario A**	**Escenario B**
Riesgo Natural (Beta autónoma)	1	1
ATR	0	3
MAXATR	0	3
%PAGO	1	0
Alto	0	1
Medio	0	0
Bajo	1	0
Meses	13	0
CAP	1	1
PI	3.46%	92.33%
Reservas por cada 10K	$ 245.78	$ 6,555.65

Vemos que, bajo condiciones naturales, el microcrédito individual, es el que más causa Reservas por cada 10K, ya que su riesgo es más alto, por ahora, no profundizaremos sobre el riesgo en Microcrédito. Pasando a la calificación de cartera de Microcrédito Grupal encontramos el siguiente modelo:

Cuadro 28. Modelo Probabilidad Incumplimiento Microcrédito Grupal

Coeficiente	Valor	VAR
β_0	**0**	Riesgo Natural (Beta autónoma)
β_1	**0.6297**	ATR
β_2	**-4.1889**	%PAGO
β_3	**0.847**	Alto
β_4	**0.232**	Medio
β_5	**-1.079**	Bajo
Escenario A	**0.51%**	
Escenario B	**3.42%**	

De primera instancia vemos algo anormal respecto a los demás modelos y es que, la beta autónoma, toma valor de cero, esto es importante mencionarlo, ya que no es usual realizar regresiones forzando a la recta a pasar por el origen, por lo que, cambia la pendiente de la derivada de la regresión, se pierden grados de libertad, además que hacemos un *overfitting* de la regresión, lo que impacta en la modelación negativamente.

Retomando el riesgo intrínseco que requiere el Microcrédito Grupal, definimos escenarios que nos ayudaran a obtener el precio de intrínseco del riesgo, bajo supuestos similares a los demás tipos de contrato:

a.) Escenario al corriente y con experiencia (A): Grupo de acreditados con pago al corriente y sin atraso en los últimos 3 períodos de facturación, con experiencia de 5 créditos experiencia previa del acreditado que será evaluado en fecha de corte, cuyo grupo al que pertenece está formado por más de 15 personas.

b.) Escenario al corriente y sin experiencia (B): Grupo de acreditados con pago al corriente y sin atraso en los últimos 3 períodos de facturación, con experiencia de 2 créditos experiencia previa del acreditado que será evaluado en fecha de corte, cuyo grupo al que pertenece está formado por 10 personas.

c.) Escenario de impago(C) (Sin llegar a ATR >3): Grupo de acreditados con retraso en los últimos 3 períodos de facturación, con experiencia de 0 créditos experiencia previa del acreditado que será evaluado en fecha de corte, cuyo grupo al que pertenece está formado por menos de 10 personas.

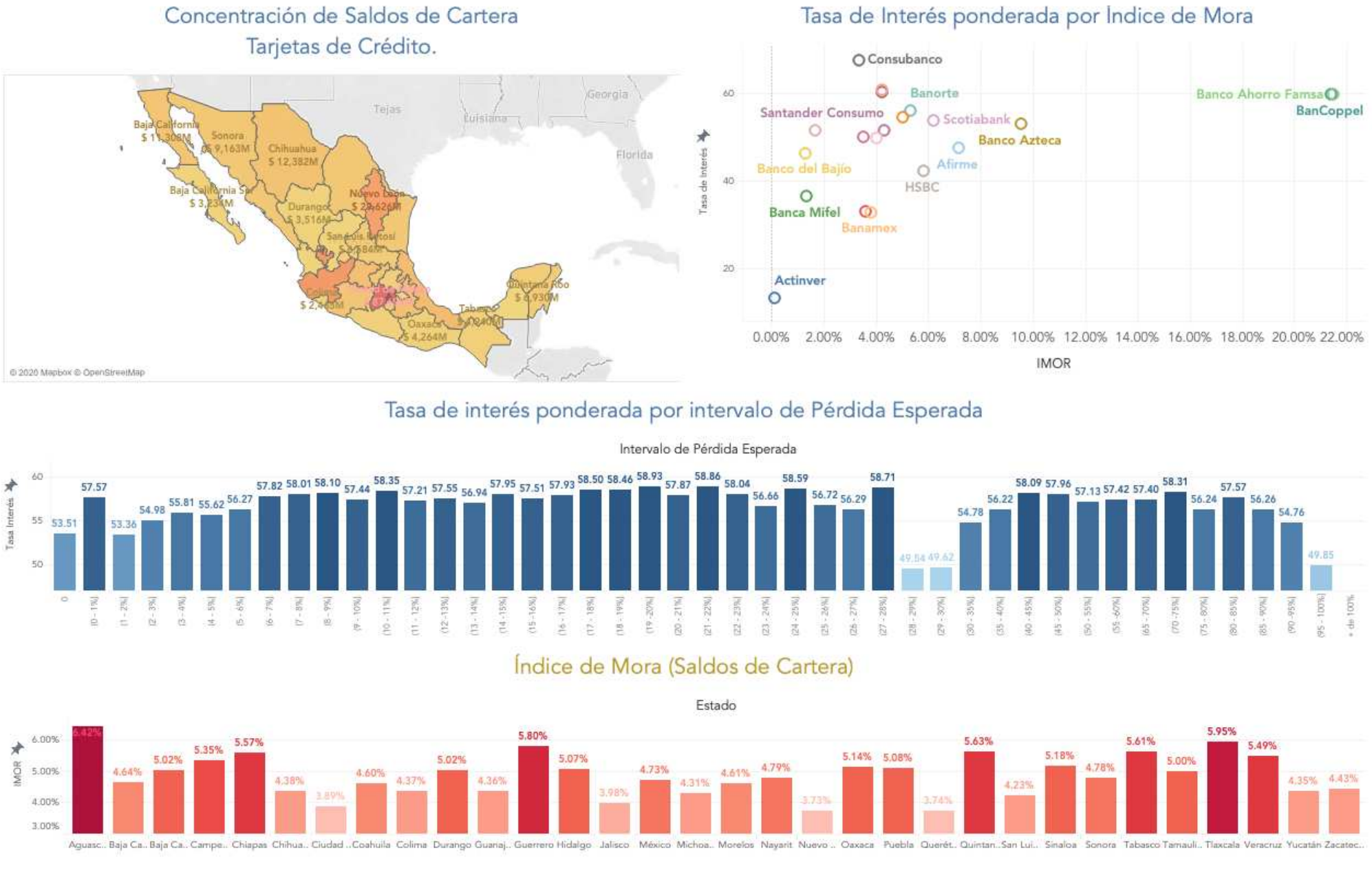
Cuadro 29. Escenarios PI Microcrédito Grupal

VAR	Escenario A	Escenario B	Escenario C
Riesgo Natural (Beta autónoma)	1	1	1
ATR	0	0	3
%PAGO	1	1	0
Alto	0	1	1
Medio	0	0	0
Bajo	1	0	0
PI	0.51%	3.42%	93.91%
Reservas por cada 10K	$ 40.00	$ 266.46	$ 7,325.16

Encontramos que en el escenario A, tiene una PI con respecto al sistema, sin embargo, este escenario es muy difícil de conseguir, por la cantidad de créditos previos que debe haber alcanzado un acreditado, esto a la par se ve potencialmente influido por la falta de beta autónoma, lo que nos pone en un escenario de muy bajo riesgo cuando se tiene un Microcrédito Grupal, sin embargo en un escenario más conservador (Escenario B), obtenemos que, la PI es más alta derivado de el cambio en las *dummies* de ciclos-deuda (Alto, Medio, Bajo), suponiendo que la mayoría de los créditos no tienen muchos ciclos en la institución financiera.

Por último revisaremos el último crédito al consumo que nos hace falta por explorar, este, no lo abordaremos de la misma manera, debido a que no existe información equiparable por la forma de operación, puesto que es un producto revolvente y es distinta la forma administración de riesgo, así como de cálculo de *exposure*.

Figura 14. Tarjeta de Crédito



2.10 Crédito de consumo revolvente

Las disposiciones de carácter general aplicables a instituciones de crédito definen revolvente como:

"Una característica contractual de la apertura de crédito, que da derecho al acreditado a realizar pagos, parciales o totales, de las disposiciones que previamente hubiere hecho, quedando facultado, mientras el contrato no concluya, para disponer en la forma pactada del saldo que resulte a su favor". Es decir, en este tipo de créditos no existe una tabla de amortización que ampare las operaciones de crédito, sino que es un crédito que se renueva de manera automática mientras el contrato no concluya. El IHH por institución financiera llega a ser de 0.1768 por lo que está altamente concentrada, podemos observar que en conjunto, las cinco instituciones con mayor saldo, concentran el 82.5% del mercado.

Cuadro 30. Concentración de Saldos TC por Institución (Miles de Pesos)

Institución	IMOR[29]	Tasa	Saldos Cartera	Dist.
BBVA Bancomer	4.20%	60.945268	$ 117,821,438,720.50	27.45%
Tarjetas Banamex	1.67%	51.5246466	$ 108,869,817,560.50	25.36%
Santander Consumo	4.31%	51.6764243	$ 60,550,866,114.57	14.11%
Banorte	5.33%	56.1250559	$ 44,087,978,815.84	10.27%
HSBC	5.80%	42.2089558	$ 24,162,550,547.96	5.63%
Total 5 inst.			$ 355,492,651,759.37	82.8%
Total sist. 21 part.			$ 429,232,583,169.35	

Algo que llama la atención en la gráfica 18 y la tabla 29 es que, no parece que existe una relación entre tasa de interés e IMOR. Al estimar la correlación de Pearson, encontramos que, existe una correlación positiva y que explica sólo en 5% a la tasa de interés.

	Correlación	R^2
Correlación Tasa de Interés	0.2244	0.0503

La concentración por Entidad Federativa al aplicar el IHH llega a un 0.08 lo que representa que está ligeramente concentrada.

Cuadro 31. Concentración de Cartera por Entidad Federativa Crédito TC

Estado	IMOR	Saldo Cartera	Distribución
Ciudad de México	3.89%	$ 87,692,899,265.23	21%
México	4.73%	$ 58,947,575,984.45	14%
Jalisco	3.98%	$ 33,806,978,768.59	8%
Nuevo León	3.73%	$ 29,626,152,084.00	7%
Veracruz	5.49%	$ 15,790,228,045.01	4%

[29] IMOR (índice de Mora) = Saldo de Cartera Vencida / Saldo Total

Se observa que las cinco entidades con los saldos más grandes concentran el 54% del saldo total de la república mexicana. Para el crédito revolvente no podremos hacer los análisis correspondientes de Probabilidad de Incumplimiento por Entidad Federativa, versus Probabilidad de cumplimiento, debido a que no existe un equivalente del reporte R-R11, así como del A-R11. Analizaremos el riesgo a través del índice de mora (IMOR) sin embargo no empataría con los análisis previamente realizados, pero consideramos que es un indicador equivalente de riesgo. Luego entonces, comenzaremos enunciando de manera breve como se conforma la probabilidad de incumplimiento en productos revolventes. Cabe mencionar, que por sus tipos de característica de contrato el producto revolvente es mucho más dinámico en la evaluación del riesgo, además de que es un *exposure* dinámico, ya que depende principalmente de la deuda total, versus el límite de crédito, estás relaciones se especifican en artículo 92 BIS de la CUB.

Cuadro 32. Modelo Probabilidad de Incumplimiento Revolvente

Coeficiente	Valor	VAR
β_0	**-2.1859**	Riesgo Natural (Beta autónoma)
β_1	**0.7864**	ACT
β_2	**0.3978**	HIST
β_3	**0.8731**	%USO
β_4	**-0.4112**	%Pago
β_5	**0.2912**	Alto
β_6	**-0.0294**	Medio
β_7	**-0.2618**	Bajo
β_8	**-0.1567**	GVeces1
β_9	**0.0238**	GVeces2
β_10	**0.1329**	GVeces3
β_11	**-0.0855**	BKATR
Riesgo Natural	**10.10%**	
Escenario A	**1.59%**	
Escenario B	**93.49%**	

Realizamos dos escenarios, para revisar cuál es el riesgo intrínseco por tener este producto en la cartera.

a.) Escenario al corriente (A). Tarjeta de Crédito al corriente sin ningún tipo de impago dentro y fuera de la institución en los últimos 13 meses, el porcentaje de uso es al 1% con respecto al límite de crédito, con el ratio a 100% de pago realizado- saldo a pagar de la deuda exigible, con un límite de crédito de $40,001.00 con antigüedad mayor a 42 meses dentro de la institución, con un monto a pagar menor a $640.00.

b.) Escenario de mayor problema (B) (Sin llegar a ACT>=4): Tarjeta de crédito con impago de 4 meses de facturación, con el porcentaje de uso al 100% respecto al límite de crédito, el ratio a 100% de pago realizado- saldo a pagar de la deuda

exigible, con un límite de crédito de $14,000.0 y un monto a pagar que supera en 2.2 veces a la deuda de la institución con respecto de la deuda con otras entidades.

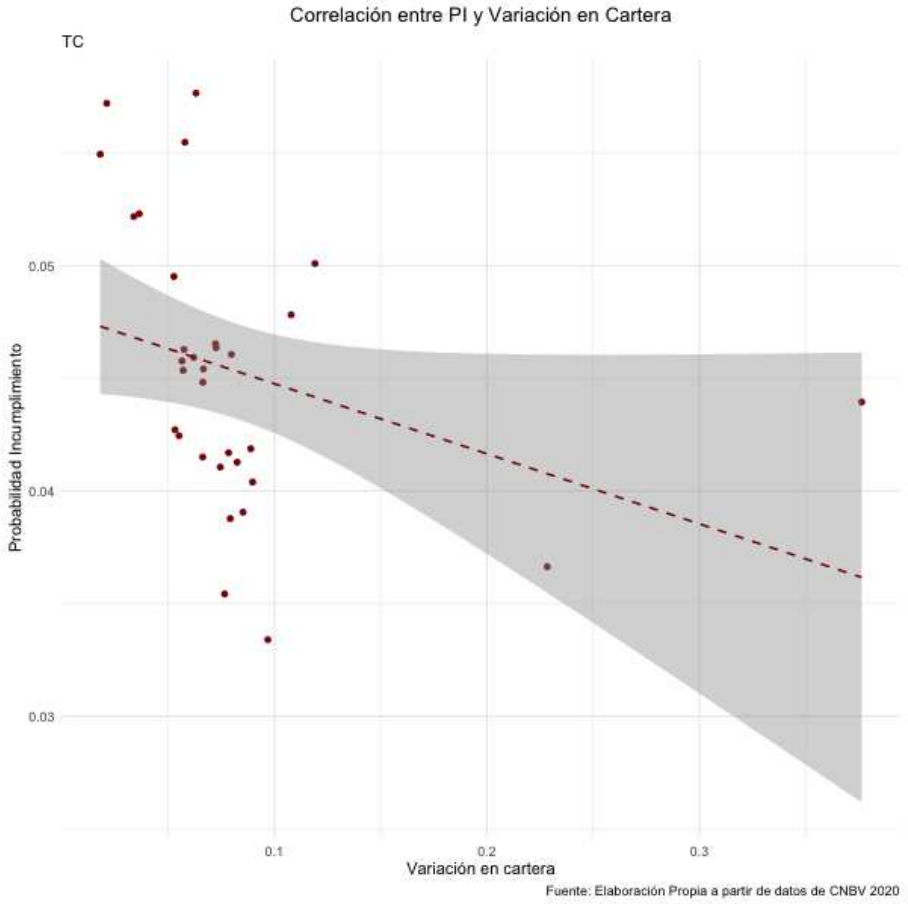
Cuadro 33. Escenarios PI TC

VAR	Escenario A	Escenario B
Riesgo Natural (Beta autónoma)	1	1
ACT	0	4
HIST	0	4
%USO	0.01	1
%Pago	1	0
Alto	0	1
Medio	0	0
Bajo	1	0
GVeces1	1	0
GVeces2	0	0
GVeces3	0	1
BKATR	13	0
PI	1.60%	97.91%
Reservas por cada 10K	$ 123.28	$ 7,539.26

Lo que determinamos, bajo estos supuestos es que, el costo de un cliente que está al corriente, por cada $10,000 de *exposure*, produce una reserva de $123.28, con una PI de 1.60%, mientras que en el peor escenario, tenemos una PI de 97.91% produciendo una reserva de $7,539.26.

Una vez analizado desde el punto de vista de probabilidad de incumplimiento, pasaremos a realizar el análisis gráfico de la interrelación entre la variación de la cartera versus el índice de mora, cabe mencionar que este lo hicimos en la periodicidad de 2018-2019, tomando como base el IMOR de 2018 para observar los efectos provocados en 2019.

Figura 15.Correlación entre IMOR y variación de cartera TC (Saldos Totales)

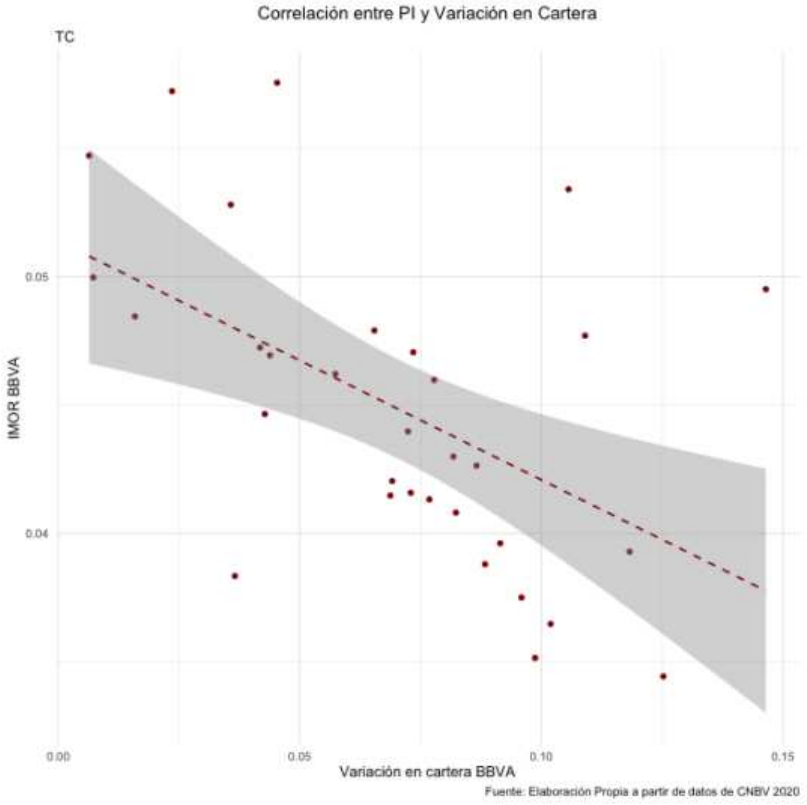


***Test* Shapiro-Wilk**

***Test* Shapiro**	**P-Value**	**W**		
Probabilidad Incumplimiento	0.5359148	0.97129571		
Variación	0.87168368	0.98264227		
	Correlación	**R2**	**Pvalue**	**Rho**
Correlación Pearson PI-Variación	-0.521033	0.27147542	0.9551	
Correlación Spearman		0.27247788	0.00249918	-0.5219941

Podemos observar a pesar de que existen *outliers* dentro de la muestra, existe gráficamente una relación entre el IMOR y la variación de cartera por entidad federativa. Por el *test* de Shapiro, nos da pauta a hacer uso de la correlación de Pearson, el cual nos da una estimación de -0.5210, con una significancia de 27.14% lo que nos indica que hay una relación fuerte y negativa ante índices de mora altos, esto refuerza nuestra concepción de que hay exclusión geográfica ante índices de riesgo altos.

Figura 16. Correlación entre IMOR y variación de cartera TC (Saldos BBVA)



***Test* Shapiro-Wilk**

***Test* Shapiro**	**P-Value**	**W**		
Probabilidad Incumplimiento	0.50821919	0.97031786		
Variación	2.7858E-08	0.589624		
	Correlación	**R2**	**Pvalue**	**Rho**
Correlación Pearson PI-Variación	-0.3244501	0.10526785	0.9551	
Correlación Spearman		0.3096384	0.00113458	-0.5564516

Gráficamente vemos que hay una relación entre IMOR y los aumentos de cartera que tuvo BBVA Bancomer en el período 2018-2019 en sentido inverso, ya que las Entidades Federativas que tuvieron un Índice de Mora más alto tuvieron un incremento menor en saldos de cartera que los que tuvieron un menor índice de mora, tal es así que la correlación Spearman (la que podemos utilizar, ya que por el *test* de Shapiro, no asumimos normalidad en la variable variación) es de -0.5564 que explica la relación en un 30.9%.

Cuadro 34.Resumen Indicadores

		Consumo No Revolvente						Revolvente
		ABCD	Automotriz	Nómina	Personal	Microcred I	Microcred G	TCrédito
	Cartera Total	$ 15,600,285,120.	$ 230,349,998,820	$ 259,595,772,035	$ 244,287,859,420.	S/I	S/I	$ 429,232,583,169
	Cartera Vencida	$ 780,909,966.	$ 4,699,266,540.	$ 6,892,347,384	$ 15,663,450,390	S/I	S/I	$ 17,612,076,763
Indicadores de concentración	IHH por Entidad	0.0645	0.05941	0.0594	0.0736	S/I	S/I	0.086241711
	IHH Por institución	0.9994	0.1369	0.2418	0.1156	S/I	S/I	0.1785
	Cantidad instituciones	4	32	16	43	S/I	S/I	21
	IHH Ideal por Entidad	0.03125	0.03125	0.03125	0.03125	0.03125	0.03125	0.03125
	IHH Ideal por Institución	0.25	0.03125	0.0625	0.023255814	S/I	S/I	0.047619048
	Concentración 5 instituciones mayor saldo	100%	75%	98%	69.74%	S/I	S/I	82.80%
	Concentración 5 Entidades con mayor saldo	42.80%	44.70%	43.97%	48.54%	S/I	S/I	54.19%
	Veces IHH entidad/IHH ideal Entidad	2.06	1.90	1.90	2.36	S/I	S/I	2.76
	Veces IHH institución/IHH institución ideal	4.00	4.38	3.87	4.97	S/I	S/I	3.75
Indicadores de Riesgo	IMOR	5.006%	2.040%	2.655%	6.412%	S/I	S/I	4.103%
	PI Ponderada	14.09%	4.62%	9.90%	14.00%	S/I	S/I	NA
	Tasa de Interés ponderada	49.11%	13.05%	24.75%	32.27%	S/I	S/I	54.61%
	PI Escenario Corriente	1.62%	0.42%	2.81%	1.10%	3.46%	0.51%	1.60%
	PI Escenario Impago	99.61%	82.56%	90.92%	90.97%	92.33%	93.91%	97.91%
	Riesgo Natural (Beta autónoma)	7.27%	11.43%	8.10%	21.54%	21.54%	NA	10.10%
	Reservas Escenario Corriente cada 10K	$139.49	$30.31	$191.00	$78.13	$245.78	$40.00	$123.28
	Reservas Escenario Impago cada 10K	$8,566.58	$5,944.28	$6,182.44	$6,459.19	$6,555.65	$7,325.16	$7,539.26
	Variación Cartera Sistema (2017-2018)	1.97%	9.20%	9.02%	1.89%	S/I	S/I	0.066399524
Test estadísticos sistema	*Test* Shapiro Wilk Variación	0.060544049	0.018945304	0.258123831	0.223936187	S/I	S/I	0.535914797
	Test Shapiro Wilk PI	0.167281204	0.242975298	0.582751524	1.01948E-06	S/I	S/I	0.871683683
	Corr Variación Cartera- PI	-0.68890292	-0.440615836	-0.301902525	0.100439883	S/I	S/I	-0.521994135
	R Cuadrado	0.474587233	0.194142315	0.091145134	0.01008817	S/I	S/I	0.272477877
	Correlación Tasa de Interés PI	0.4071	0.234971411	0.150060152	-0.252727964	S/I	S/I	NA
estadísticos Institución	Institución	Banco Azteca	BBVA Bancomer	BBVA Bancomer	BBVA Bancomer	S/I	S/I	BBVA Bancomer
	Test Shapiro Wilk Variación	0.060544049	0.003290155	0.001440147	0.817595419	S/I	S/I	0.508219189
	Test Shapiro Wilk PI	0.167281204	0.496568202	0.556528545	2.13708E-08	S/I	S/I	2.78578E-08
	Corr Variación Cartera- PI	-0.68890292	-0.514296188	-0.698680352	-0.27016129	S/I	S/I	-0.556451613
	R Cuadrado	0.474587233	0.264500569	0.488154234	0.072987123	S/I	S/I	0.309638398

2.11 Reflexiones finales de sección

A manera de resumen, después de la literatura debatida en los primeros apartados, así como los análisis mostrados en las secciones dedicadas al análisis de los productos de crédito de Consumo podemos sugerir lo siguiente:

a.) Los solicitantes de crédito son abiertamente excluidos del sistema financiero debido a que poseen características sociodemográficas que vista *scoring* de crédito determinan que hay una alta probabilidad de impago y que las instituciones financieras no están dispuestas a asumir.
b.) Los solicitantes de crédito que no cuentan con un historial de crédito consolidado tienen de la misma forma, gran probabilidad de ser candidatos rechazados, al no tener suficiente información crediticia que sustente la operación.
c.) La información que a menudo se utiliza para realizar los diferentes tipos de scoring a menudo carecen de datos alternos que pueden ayudar a hacer más eficiente y con un *accuraccy* superior, lo que agudiza la exclusión financiera.
d.) La cartera de crédito al Consumo en México, a nivel Entidad Federativa se encuentra concentrada en grandes metrópolis, que tienen como común denominador: Ciudad de México, Estado de México, Jalisco, Nuevo León y Veracruz, quedando fuera, las otras 27 entidades federativas.
e.) La concentración de saldos de cartera por institución es muy aguda y hasta cierto punto en productos como ABCD es preponderantemente monopólica.
f.) La tasa de interés ponderada por producto de crédito no corresponde en la gran mayoría a la probabilidad de incumplimiento ponderada. Esto implica que no se pide el premio al riesgo de forma adecuada y en ciertos segmentos pueden no ser rentables las operaciones.
g.) No existe información de la Cartera de Consumo Microcrédito, lo que hace difícil, estimar el riesgo promedio, lo que sí sabemos, es que es el más costoso de reservar en un "Estado Corriente".
h.) Existe un modelo "comportamental" dentro de la CUB de Bancos (Microcrédito Grupal), cuyo intercepto fue forzado a cero, que implica errores estadísticos de impacto que afectan el cálculo de PI. De la misma manera, para todos los modelos de los artículos mencionados de la CUB, tenemos el uso excesivo de variables dummy, así como variables de intervalo abierto (es decir, que no limitan los valores posibles de la variable), que a nivel tratamiento de datos, tienen implicaciones en la predicción de los modelos. Esto lo analizaremos más a fondo en las siguientes secciones.
i.) Vista sistema, en la gran mayoría de los contratos de crédito no se ve una correlación aguda entre probabilidad de incumplimiento ponderada y variación de cartera de Entidades Federativas, esto se puede deber a la competencia por el mercado, es decir, que el riesgo que no asume una institución financiera, lo asume otra.
j.) Vista Institución con mayor grado de participación de mercado por producto, nos dimos cuenta de que, la relación entre probabilidad de incumplimiento y variación de saldos de cartera se cumple, en más de .50 (en valores absolutos), a excepción del crédito personal, es decir, que las instituciones financieras, no están dispuestas a

asumir mayores riesgos en tanto a localidad se refiere, reduciendo su colocación y afectando los saldos de cartera.

k.) Los importes originalmente concedidos, no tienen parsimonia con el default, es decir, que no se controla el *exposure* ante probabilidades de incumplimiento altas. Cabe mencionar, que en este análisis, particularmente se ve sesgado porque asumimos la distribución de 2019 con la PI de 2017 a falta de datos por parte de la CNBV.

3. Características del microcrédito

Como ya hemos visto en la sección 2, más allá de las implicaciones de reservas y falta de datos en México, se carece de una definición clara acerca de lo que es microcrédito, ya que inicialmente, este en primera instancia no debería ser catalogado como crédito al consumo, porque su destino no está orientado a adquirir cualquier bien o servicio con fines de ser consumidos sin usufructo alguno, sino que por el contrario al igual créditos Comerciales están destinados a la adquisición de materia prima, capital de trabajo o activo fijo, es decir que se usarán con fines productivos. Es por ello, que inicialmente nos abocaremos al microcrédito, ya que es un producto que sigue sin ser atendido en México, además que a nivel financiero sigue siendo un término confuso, ya que habitualmente se mimetiza con el crédito empresarial enfocado a pequeñas empresas, pero hay factores que explicaremos a continuación que establecen esa gran diferencia y son las razones por las cuáles no se puede analizar el riesgo de la misma forma, como sugiere la CUB de Bancos (2020) a través de su modelo de Probabilidad de Incumplimiento para Microcréditos individuales, cuyas betas son iguales al crédito Personal, haciendo la distinción a través dos *dummies* (DEL y CAP)[30]. Continuando esta idea de entender el microcrédito, en su trabajo de investigación Chiapa (2009) establece dos definiciones que son centrales en este trabajo, citándolo a continuación:

"La distinción entre microfinanzas y microcrédito es importante. Ambos términos se refieren a transacciones pequeñas, pero el microcrédito se refiere únicamente a la parte relacionada con los préstamos dentro del universo de servicios financieros que se ofrecen a los pobres. Las microfinanzas, por su parte, se refieren a una amplia gama de servicios financieros ofrecidos a los más pobres y a sus microempresas entre los que se incluyen no sólo los microcréditos, sino también los microahorros, las transferencias de remesas, los microseguros, etc."

Por lo que en primera instancia, al igual que Corporación Financiera Internacional (IFC) *et al*. (2017), sostienen que el microcrédito está enfocado a la base de la pirámide, otra definición que empata con las vistas anteriormente es la de *Microcredit Summit* (2020):

"El microcrédito son programas que otorgan préstamos a personas muy pobres para proyectos de autoempleo para generar ingresos, lo que les permite cuidarse a sí mismos y a sus familias."

[30] Estas relaciones fueron revisadas en la sección 1.

Esto representa que el microcrédito, son aquellas personas que se establecen de manera informal, que tienen un grado agudo de pobreza y que tienen actividades de autoempleo como lo son: propietarios de pollerías, carpinteros, talabarteros, plomeros, taxistas, vendedores de frutas, etcétera y que adicionalmente no cuentan con los requisitos para ser acreedores en alguna institución formal bancaria. Esto es una condición clave para nuestro modelo, ya que el INEGI en 2019, estima que de la población informal es de 56.2% lo que implicaría que el total de esta población (31.2 Millones de Mexicanos) tendrían dificultades para tener acceso a un crédito, si partimos de la idea de que, por modelos de score, estas personas son vulnerables a ser rechazadas dadas sus características. Esto lo mencionamos, ya que muchas de los atributos que engloba el microcrédito, tienen implicaciones en la evaluación crediticia de la banca tradicional, y a su vez, explica la poca oferta de este tipo de productos financieros, debido a que la mayoría de microempresas que postulan se establecen en carácter informal y generalmente carecen de documentos que son indispensables en la concesión de un crédito, como lo son: comprobantes ingresos, garantías hipotecarias, historiales de crédito robustos, etcétera, por lo que, la información suministrada en un *scoring* de crédito dadas estas características, la mayoría optarían por una decisión automatizada de rechazar la operación, al tener un alta probabilidad de impago.

3. 1 El problema de la selección adversa y el microcrédito

La explicación detrás de la aversión al querer incursionar al microcrédito se debe en gran medida al riesgo de impago que pueda tener un prospecto. La teoría nos indica que se debe elegir la combinación de activos que resulte más beneficiosa, dado un portafolio, tomando como función objetivo la minimización de riesgos presentes. Es por ello que nos remontamos a Markowitz en 1952 que plantea un modelo de conducta racional del decisor para la selección de carteras, es decir que para él inversor, una cartera será eficiente siempre y cuando proporcione la máxima rentabilidad posible para un riesgo dado.

En investigaciones recientes Allen y Gale (2005) establecen a través de su modelo CVH supuestos de competencia perfecta, en donde suponen que los mercados competitivos van a ser el riesgo de quiebra que enfrentarían los bancos, dado que, cuanto mayor ser la competencia, el banco tomará mayores riesgos, lo que implica que sí existe un grave incumplimiento, el banco se verá obligado a salir del mercado, mientras que en un mercado monopólico, la institución al tener control sobre el mercado (valor de privilegio) tomaría menos riesgos. Sin embargo en la continuación de su modelo de Allen y Gale, (Nicolo et al., 2006) a través del modelo BDN (2006) se dan cuenta de que existen estás relaciones, pero adicionan que los bancos compiten más a través de la tasa de interés en los depósitos, por lo que influirá a que compitan más en el mercado de crédito, ya que al conseguir tasas más bajas en los depósitos, las tasas a las cuáles se ofertan los créditos deberían disminuir. Esto implica que las tasas de los créditos más bajas aumentarían la rentabilidad de los bancos, al hacer operaciones por volumen y, por lo tanto, hacen menos probable la bancarrota. Además, por la misma razón por la que los bancos eligen un mayor riesgo cuando aumentan las tasas de los depósito, ya que los prestamistas optan por estar más

seguros cuando las tasas de los créditos disminuyen. Ambos efectos, estriban en reducir el riesgo de quiebran en los bancos.

Bajo estos supuestos Castillo y León (2019) llevan a la práctica la teoría del CVH en Perú, poniendo el ejemplo con cajas municipales, lo que encontraron, a través de análisis de panel de datos fue que entre menos competencia hay en cajas municipales, el valor de privilegio, aumenta, lo que provoca una disminución en el riesgo, al igual que, corroboraron la tesis del modelo BDN a través de un z-score, donde concluyeron que a mayor tasa de interés, implicaba un mayor apetito al riesgo. Cuando nosotros en la sección 1 realizamos un ejercicio similar al tratar de ver los efectos del IHH en las correlaciones de incrementos en saldos de cartera contra PI, encontrábamos que no existía gran diferencia en el análisis, esto sucedió porque medimos el impacto en un punto específico, no de manera dinámica como lo proponen los autores.

Sin embargo, esto no es suficiente explicación a la falta de interés en apostar al microcrédito, Valencia (2010) explica que el racionamiento del crédito se da a partir de una contracción en la oferta de recursos, por lo que el crecimiento económico y las expectativas de la economía serán determinantes en apetito al riesgo de las entidades financieras, ya que ante quiebres económicos, impulsaría a asumir menos riesgo. Otra respuesta a este punto, lo podemos encontrar en la investigación de (Mei et al., 2019) donde analizan el racionamiento del crédito en empresas PyME y sostienen que la razón principal por la que es más difícil para las PYME al tratar de obtener un crédito es porque antes de aprobar la concesión de un crédito, las PYME no pueden transmitir sus niveles de riesgo, ya que el tamaño de sus activos iniciales generalmente está por debajo del valor de la garantía, y en su gran mayoría causan mayores costos a los bancos, ya que el monto de su préstamo generalmente es inferior al monto mínimo del préstamo. Esto sugiere de igual manera, que el microcrédito tiene esta misma cualidad, adicionalmente, que las PyME tienen un mayor grado de asimetría de información con los bancos. Este autor propone que como una medida que se pudiera optar para resolver este problema, es que se utilice el Big Data con el fin de poder analizar y predecir ll comportamiento crediticio de los prospectos y así reducir el grado de asimetría.

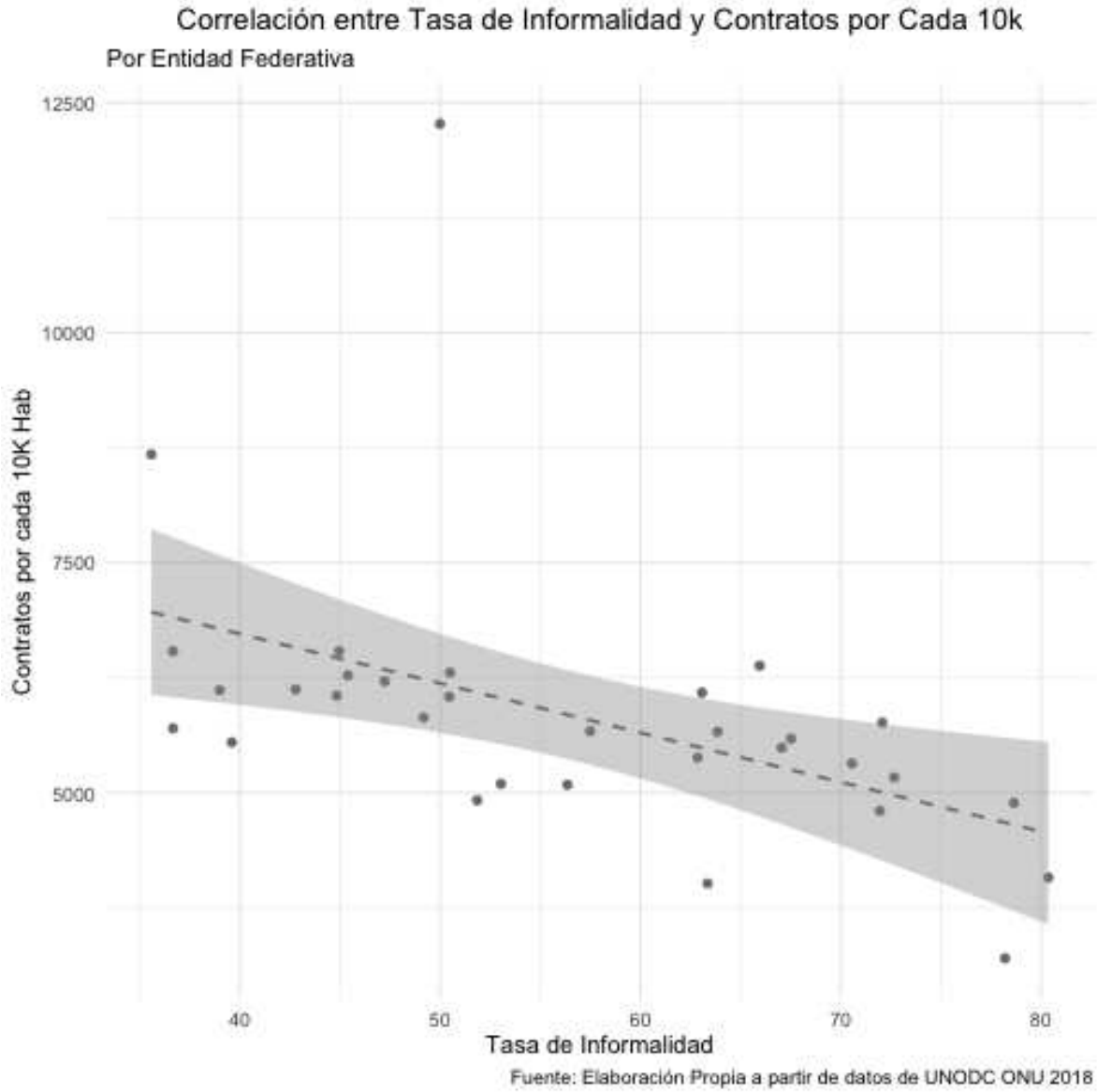
3.2 ¿Porqué el Microcrédito?

Como bien lo mencionamos, existen ciertas características que los modelos de *scoring* toman en cuenta para la asignación de una probabilidad de impago, la cual es fundamental para conceder o rechazar créditos. Comentábamos que la diferencia fundamental del microcrédito, es la informalidad, bajo este argumento, la OIT (2016) en un estudio reveló que el principal obstáculo (35%) por lo que una pequeña o mediana empresa informal no lograba crecer, era por la falta de acceso a servicios financieros. En México, es un tema

preocupante, ya que la cifra de empleo informal por Entidad Federativa es muy acentuada, por ejemplo, observamos que en la Gráfica 21 que, la informalidad se ve marcada en los estados del sur, mientras que en el norte, la cifras son menores. Si esto, lo relacionamos con el número de contratos totales de créditos de la banca por cada 10,000 habitantes, encontramos lo siguiente:

Gráfica 21. Tasa de Informalidad en México

Gráfica 17. Correlación entre Informalidad y Contratos Crédito



Observamos que hay una clara relación, casi sincrónica entre la cantidad de contratos de crédito de consumo por cada 10,000K y la tasa de informalidad por Entidad Federativa, el único *outlier* que se llega a apreciar es CDMX y Nuevo León con más de 8,000 contratos por cada 10k habitantes. Estas relaciones las podemos apreciar en la tabla 36, donde observamos que la correlación es fuerte y negativa, es decir, que entre mayor sea la tasa de informalidad, menor cuantía de contratos de crédito al consumo por cada 10k habitantes, con una significancia de hasta el 44.04% (excluyendo a CDMX).

Cuadro 35.Contratos de Crédito y Tasa de Informalidad por Entidad Federativa

Entidad federativa Nacional	Tasa de Informalidad Laboral (TIL1)	Contratos Banca 10K Hab	Entidad federativa Nacional	Tasa de Informalidad Laboral (TIL1)	Contratos Banca 10K Hab
Aguascalientes	42.78%	6,121.29	*Morelos*	65.95%	6,379.41
Baja California	39.59%	5,546.75	*Nayarit*	62.85%	5,381.05
Baja California Sur	38.99%	6,113.48	*Nuevo León*	35.58%	8,675.88
Campeche	63.85%	5,662.83	*Oaxaca*	80.37%	4,077.52
Coahuila	36.66%	6,537.98	*Puebla*	71.95%	4,803.14
Colima	50.45%	6,042.41	*Querétaro*	44.96%	6,537.50
Chiapas	78.20%	3,200.67	*Quintana Roo*	47.23%	6,210.24
Chihuahua	36.66%	5,699.80	*San Luis Potosí*	56.35%	5,086.41
Ciudad de México	49.99%	12,271.65	*Sinaloa*	50.49%	6,305.31
Durango	51.85%	4,916.31	*Sonora*	44.83%	6,054.92
Guanajuato	53.03%	5,096.18	*Tabasco*	67.05%	5,487.79
Guerrero	78.65%	4,887.85	*Tamaulipas*	45.39%	6,275.14
Hidalgo	72.67%	5,165.75	*Tlaxcala*	72.07%	5,762.09
Jalisco	49.17%	5,817.68	*Veracruz*	67.52%	5,586.57
México	57.49%	5,668.48	*Yucatán*	63.07%	6,087.49
Michoacán	70.56%	5,316.18	*Zacatecas*	63.35%	4,012.21
	Correlación	**-0.48248**		R2	**0.232795**
	Corr. sin CDMX	**-0.66364**		R2 Sin CDMX	**0.440430**

Esto, como ya mencionamos, podría reflejar la aversión al riesgo que tienen las empresas al incursionar en ciertas Entidades Federativas , ya que el sector informal está asociado a la falta de información de los prospectos, cuyos datos tradicionalmente la banca, a través de sus modelos tratar de reflejar una probabilidad de pago, o el flujo financiero de un acreditado proyectado al futuro, así como una opción de monto a otorgar, a través del uso del historial crediticio, montos manejados con otras instituciones, depósitos documentados, etcétera, sin embargo, esto pocas veces sucede en estos sectores, sin embargo, el Banco Mundial en 2019, hace una serie de sugerencias que pueden ayudar a complementar la documentación de un prospecto, a través de información alterna, tal como el uso de redes sociales, datos de pago, y perfiles psicométricos.

3.3 ¿Qué producir?: El papel de los Credit Scoring Reactivos en Microcrédito.

Cómo lo hemos visto anteriormente, existen dos retos fundamentales para que el acceso al crédito se pueda dar, primeramente, es que los bancos o las instituciones financieras quieran asumir el riesgo que se tiene al perfilar un prospecto con información asimétrica y, en segundo lugar, tratar de minimizar las pérdidas, dado el riesgo que se tiene. Una vez resueltos estos retos, podemos decir que la brecha se rompe, debido a que los bancos o instituciones financieras pueden hacer eficiente su cartera y los demandantes de crédito pueden accesar al crédito, por lo que, consideramos que el puente ideal es a través del scoring de crédito reactivo, el cual se utiliza por primera vez con prospectos completamente desconocidos y es el canal ideal por el cual se puede saber la probabilidad de impago y partiendo de ello, podemos modular el *exposure*, de tal manera que podamos minimizar pérdidas.

3.3.1 Scorings Reactivos

En la literatura existen varios trabajos de investigación enfocados a realizar scorings en el campo del microcrédito, por ejemplo, (Rayo Cantón et al., 2010) realizaron un ejercicio de Credit Scoring reactivo aplicado para una institución financiera en Perú, los resultados que encontraron fue que al realizar una regresión logística con diferentes tipos de variables, llegaron 78.3% de predicción, su aplicación de dicho scoring estribó en calcular la tasa de interés mínima, así como el capital mínimo necesario para la operación en el marco de Basilea II, una de las grandes áreas de oportunidad que nosotros vemos en este trabajo fue que los autores eliminaron las observaciones con *missing* values, lo que sesga totalmente el modelo.

Otro trabajo de investigación de interés es el de Nieto Murillo (2011) realizó un estudio aplicado a créditos revolventes en México utilizando regresiones logísticas de tipo logit, obteniendo resultados que permiten una mejor originación del crédito. Lo más interesante de este documento, es la aplicación de cadenas de Markov para clasificar clientes como; “buenos”, “indeterminados” y “malos”.

Beltrán Pascual (2015) realiza un análisis exhaustivo y aplicación entre las diferentes técnicas, no solamente con los modelos tradicionales sino con técnicas no paramétricas como el algoritmo AID, CART, QUEST, CHAID, C5, redes neuronales, Vector Support Machine, redes bayesianas y regresiones logísticas. Donde hace un contraste, utilizando cada uno de los métodos antes mencionados, resaltando el poder de predicción de las redes bayesianas, no obstante, va más allá utilizando diferentes tipos de multiclasificadores para encontrar el *credit scoring* con mayor poder de predicción.

Ya hemos mencionado técnicas estadísticas asociadas a los credit scoring, sin embargo, un punto de relevancia a llevar a cabo es la inclusión de información de relevancia para la construcción del modelo, Vogelgesang (2003) analiza los efectos macroeconómicos de acreditados en Microfinancieras de Bolivia mediante modelos multivariado e incluye un análisis de supervivencia.

Aunque la mayoría de los credit scoring que se realizan dentro en las instituciones financieras dependen de la información arrojada por las Sociedades de Información Crediticia (SIC's), no es común encontrar metodologías que carezcan de esta información En este sentido Espin-García y Rodríguez-Caballero (2013) ofrecen una metodología para realizar credit scoring sin información de burós crédito basado en técnica CHAID. Por otro lado, Dellien y Schreiner (2005) sintetizan una reflexión entre la forma de operar de un Banco y de las IMF's y de las prácticas que pueden beneficiarse. Schreiner (2003) realiza un análisis donde se predice la probabilidad de impago a 15 días o más a partir de variables como: garantías, giros, asesor que gestiona la originación, género, etc. Sin embargo, este modelo es contemplado sólo para la renovación de un crédito ya que depende significativamente de la experiencia crediticia interna. Por último, Beledo (2007) incorpora en su regresión, variables relacionadas con las características del desembolso, prestatario, historial crediticio y variables propias de la organización.

3.3.2 Capacidad de Pago

En principio diremos que no existen muchos modelos a enfocar o esquematizar las dimensiones que comprende la capacidad de pago, basta con observar que no existen muchas definiciones por órganos reguladores para establecer una definición puntual, por ejemplo la Procuraduría Federal del Consumidor (2006) define la capacidad de pago como:

> *"El ingreso que no se destina a ningún otro concepto del gasto familiar ni deuda".*

Sin embargo, esta definición se encuentra escueta, por ello nos remitimos a la definición del Banco Nacional de Costa rica, donde menciona que:

> "Es la capacidad del deudor para generar dinero en el giro normal de su negocio y/o con el producto del salario que recibe por su trabajo, de forma tal, que le permita atender sus deudas en los términos y condiciones en que fueron pactadas."

De igual manera, la PROFECO establece una fórmula sencilla para determinar la capacidad de pago, donde:

$$\text{Capacidad de pago} = \text{Ingreso Mensual} - \text{Gasto Mensual}. \quad (1)$$

Cuando se habla de valuar la capacidad de pago hay muchas maneras de hacerlo, en general, depende del destino del crédito, del tipo de cliente, de la persona jurídica que lo suscribe, el tipo de proyecto, etcétera. Sin embargo, en la práctica del microcrédito, resulta muy difícil tomar la expresión (1), ya que se presentan los siguientes problemas según (Otero et al., 1998):

1) La información con que el acreedor cuenta es incompleta e imperfecta. Adquirir la información podría incrementar sustancialmente el costo.

2) Al deudor también le resultaría caro suministrar la información al acreedor. Como la información que ambos poseen no es la misma (es asimétrica), el acreedor no confía en todo lo que el deudor afirma.
3) El deudor tiene incentivos para comunicar sólo lo que le conviene y el acreedor no sabe cuándo puede creerle o no. Esto obliga al acreedor a verificar la información independientemente y a crear incentivos para que el deudor pague.
4) El deudor podría cambiar el destino del crédito convenido, podría tomar riesgos adicionales o podría no ser todo lo diligente que se necesita para garantizar el pago.

En este sentido, la literatura existente está más enfocada para modelar y evaluar la capacidad de pago en créditos comerciales, un claro ejemplo es el trabajo de Rodríguez Sandiás *et al*. (1999) con base en flujos proyectados y razones financieras, sin embargo, como lo mencionamos anteriormente, esto no es factible desde el punto de vista del microcrédito ya que los solicitantes del crédito pocas veces llegan a tener este tipo de datos.

La Sociedad Hipotecaria Federal (2008), muestra una relación pago – ingreso en donde resulta ser el equivalente de la capacidad de pago y está en función de la siguiente expresión:

$$Capacidad\ de\ pago = \frac{Mensualidad\ del\ crédito}{Ingreso\ del\ acreditado}$$

Adicionalmente, estipulan umbrales que no deberán ser inferiores al resultado de este cociente, dejando en claro el nivel de capacidad de pago mínimo aceptado por esta institución.

Bedoya (2014) hace una crítica del uso del EBITDA como razón y base para el cálculo de la capacidad de pago ya que considera muy riesgoso utilizarla sin contemplar otro tipo de variables. Por otra parte (Amoroso y Jara, 2009) mencionan que las instituciones bancarias no tienen una metodología definida para evaluar la capacidad de pago de los sujetos de crédito, sino más bien las evaluaciones son subjetivas considerando su situación financiera, sin embargo, desarrollan un modelo discriminante para poder estimar la capacidad de pago con base en algunas variables cualitativas y cuantitativas.

3.3.3 Pricing de Crédito

Una de las decisiones más importantes que enfrentan las instituciones financieras, cuyo principal objetivo es la intermediación financiera, es la definición del precio o la tasa de interés de los créditos que otorga. De ahí que una sobreestimación o subestimación de esta, impacta de manera fundamental los resultados financieros, de posicionamiento y participación en el mercado de estas entidades.

Una tasa de interés alta de colocación en créditos expone a la entidad a la pérdida de clientes, a competir en condiciones fuera de mercado o atender solo los clientes con mayor nivel de riesgo crediticio que no son atendidos por la mayor parte de las entidades, mientras

que la subestimación de la misma, puede comprometer la sostenibilidad o rentabilidad de la institución financiera a largo plazo.

En la estimación de las tasas de interés hay diferentes aproximaciones metodológicas para su cálculo que, de forma general, agregan los componentes de la tasa activa de colocación como referente de tasa a ser asignada al cliente. Una agregación lineal de componentes tales como; el costo de fondos, gastos de originación o niveles de provisión o perdida estimada, entre otros, puede compensar la rentabilidad esperada.

El precio ajustado por riesgo, tal como lo plantean Jung y Strohecker (2009), se define como un estado en el que el deudor paga una prima de riesgo que represente, al menos, la cantidad de su pérdida esperada más un rendimiento adicional del capital regulatorio, que se obliga a los bancos a mantener con el fin de cubrir las pérdidas potenciales. En este sentido Mermelenstein (2008) nos indica una expresión donde podemos llegar mediante el uso del credit scoring.

Consultative Group to Assist the Poor (CGAP) en 2002 nos provee de una guía para la definición de una tasa sostenible para las Microfinancieras que no tengan un sistema avanzado de pricing que contempla componentes básicos y fáciles de manejar. Yang (2013) propone dos fórmulas enfocadas al pricing en Microfinancieras; una para obtener la tasa base en microcréditos y otra para el precio de la tasa de interés en clientes específicos.

3.3.4 El problema de altas tasas de interés: Las Reservas y Capacidad de Pago

Como lo mencionamos, las instituciones buscarían tener primas de riesgo más altas, dado la probabilidad de incumplimiento del acreditado, sin embargo, el tener una tasa de interés alta, o ajustada por el incumplimiento, puede presionar la capacidad de pago de los prospectos, orillándolos a incumplir el pago del crédito, debido a que su capacidad de pago no es respetada, además, para la institución puede representar un gasto extra en reservas, en caso de que se torne como un crédito irrecuperable, esto debido a que al menos en México, en el Anexo 33 de la CUB de Bancos en su apartado B-6 estipula que:

"Se deberá suspender la acumulación de los intereses devengados de las operaciones crediticias, en el momento en que el saldo insoluto del crédito sea considerado como vencido"

Por lo que, en la mayoría de los casos[31], se tendrá que reservar 90 días de intereses devengados, esto implica que, a mayor tasa de interés, mayor será el porcentaje para reservar, ya que considerando que las reservas son la suma de las pérdidas esperadas del í-ésimo crédito, dada por la siguiente fórmula:

$$\text{(2) Pérdida Esperada} = PI * EI * SP$$

[31] Existen más casos por los que un crédito se puede reconocer como vencido. Consultar anexo B-6 Cartera de Crédito.

donde:

Probabilidad de Incumplimiento (PI): a la Probabilidad expresada como porcentaje de que ocurra cualquiera o ambas de las siguientes circunstancias en relación con un deudor específico:

a) El deudor se encuentra en situación de mora durante 90 días naturales o más respecto a cualquier obligación crediticia importante frente a la Institución.
b) Se considere probable que el deudor no abone la totalidad de sus obligaciones crediticias frente a la Institución.

Exposición al Incumplimiento (EI): Al saldo insoluto a la fecha de cierre, el cual representa el monto de crédito efectivamente otorgado al acreditado, ajustado por los intereses devengados, menos los cobros de principal e intereses, así como por las quitas, condonaciones, bonificaciones y descuentos que, en su caso, se hayan otorgado.

Severidad de la Pérdida (SP): a la intensidad de la pérdida en caso de incumplimiento expresada como porcentaje de la Exposición al Incumplimiento, una vez tomados en cuenta el valor de las garantías y los costos asociados a los procesos de realización (judiciales, administrativos de cobranza y de escrituración, entre otros).

Por lo que el factor de Exposición al Incumplimiento (EI) se ve directamente afectado por la tasa de interés. Considerando un ejemplo muy sintético en saldos insolutos y globales, encontramos que:

$$(3)\ \lim_{i\to\infty} EI_{Insolutos}(i) = \lim_{i\to\infty}\left(\left(\frac{A*i}{1-(1+i)^{-n}}\right)*n\right) = \infty$$

$$(4)\ \lim_{i\to\infty} EI_{Globales}(i) = \lim_{i\to\infty} A((1+(i*n)) = \infty$$

donde,
$A = Monto\ del\ \ Importe\ original\ del\ Crédito$
$i = Tasa\ de\ interés$
$n = Número\ de\ Periodos$

$$A, i, n > 0$$

Por lo que,

$$\lim_{i\to\infty} PerdidaEsperada\ = \lim_{i\to\infty} PI\ *EI*SP\ = \infty$$
$$\leftrightarrow \qquad 0 < PI \leq 1 \qquad y \qquad 0 < SP \leq 1$$

De manera que, consideramos que las altas tasas de interés, con probabilidades de impago muy altas, implicarían mayores gastos de reserva, además que no serían competitivas en el mercado de crédito.

3.4 ¿Cómo producirlo?: Revisión de los elementos necesarios de casos de éxito en Credit Scoring

Las investigaciones citadas anteriormente sostienen buenos resultados, ya sea a través de lograr buenas predicciones en tanto a probabilidad de impago que se transforman en políticas duras, además de que en muchas ocasiones los autores se enfrentaron a distintos retos, como información incompleta, relaciones a contracorriente, etcétera, por lo que realizamos una síntesis de los trabajos que hemos revisado con el fin de encontrar variables, metodologías y puntos de referencia que nos ayudarán a realizar nuestro modelo.

Cuadro 36. Modelos reactivos de *credit scoring*

No.	Autores	Tipo de Modelo	Año	Método estadístico usado	Aporte
1	Nieto Murillo y Pérez Salvador	Credit Scoring	2011	Logit	Uso de Cadenas de Markov para definir *buenos* y *malos*
2	Espin García y Rodríguez Caballero	Credit Scoring	2013	CHAID QUEST Cart	Credit scoring sin referencias crediticias.
3	Ospina Cardona	Credit Scoring	2015	Logit	Riesgo por incumplimiento Riesgo por exposición Pérdidas esperadas
4	Rayo Cantón	Credit Scoring	2010	Logit Árboles decisión Redes Neuronales	Inclusión de variables macroeconómicas y del proceso de crédito
5	Figueroa	Credit Scoring	2006	Análisis discriminante Análisis de Fisher Regresión Logística Árboles de clasificación Redes Neuronales	Desempeño de diferentes técnicas estadísticas
6	Salamanca Edwin	Credit Scoring	2014	Logit	Macro-Scoring incluyendo variables macroeconómicas y pérdida esperada
7	Lara Rubio	Credit Scoring	2010	Análisis Discriminante Modelos de Probabilidad Lineal Logit	Desempeño de diferentes técnicas estadísticas

				Probit Redes Neuronales Árboles de decisiones	
8	Serrano Cinca *et al*.	Credit Scoring	2016	Evaluación multicriterio	Solicitud de préstamo denegada
9	Beltrán Pascual	Credit Scoring	2015	Análisis Discriminante Modelos de Probabilidad Lineal Logit Probit Redes Neuronales Árboles de decisiones Support Vector Machines Algoritmos Evolutivos Redes Bayesianas Fuzzy Logit Modelos Híbridos	Implementación de un multiclasificador a partir de la minería de datos incorporando diferentes técnicas estadísticas
10	López S. *et al*.	Capacidad de Pago	1999	Modelo de Cobertura Temporal	Uso de razones financiera para el cálculo de capacidad de pago

Fuente: Elaboración propia

Cuadro 37. Modelos reactivos de credit scoring parte II

No.	Autores	Tipo de Modelo	Año	Método usado	Aporte
11	Sociedad Hipotecaria Federal	Capacidad de Pago	2008	Modelo mediante cocientes.	Cálculo de pago mediante la relación ingreso - valor de la cuota. Aun cuando no existen ingresos formales.
12	Procuraduría Federal del Consumidor	Capacidad de Pago	2006	Modelo de identidad	Cáculo de capacidad de pago simplificada
13	Hernández Corrales *et al*.	Capacidad de Pago	2005	Modelo de Descuento	Cálculo de capacidad de pago mediante EBITDA
14	Nieto Amoroso et al.	Capacidad de Pago	1997	Análisis Discriminante	Inclusión de un modelo econométrico para el cálculo de capacidad de pago
15	CGAP	Pricing de	2002	Modelo mediante	Establece la tasa

		Crédito		costos	sostenible mínima requerida mediante la inclusión de costos fundamentales para intermediación.
16	Rozo *et al*.	Pricing de Crédito	2014	Modelo mediante costos	Establece la tasa sostenible mínima requerida mediante la inclusión de costos fundamentales para intermediación. Compara los niveles de *Pricing* entre bancos.
17	Abanto y Chávez	Pricing de Crédito	2004	Panel de Datos	Realiza un análisis externo a partir de la maximización del beneficio para Microfinanciera dadas sus peculiaridades del sector.
18	Mermelestein	Pricing de Crédito	2008	Ecuación de igualdad	Propone una ecuación relacionada directamente con la Probabilidad de Default.
19	Yang	Pricing de Crédito	2013	Ecuación de igualdad	Propone dos ecuaciones para obtener una tasa base y una tasa específica por tipo de cliente.

Fuente: Elaboración propia

Cuadro 38. Modelos reactivos de scoring. (Continuación)

No.	Muestra entreno	Muestra *Test*	Tipo de Contrato*	Institución	Tipo de Variables	Definición de Default
1	26820	17810	Crédito al consumo revolvente	Banco	Sociodemográficas Tipo Seguridad Social Empleo Ingresos	
2	3252	812	Crédito al consumo revolvente	Banco	Sociodemográficas Académicas Ingresos	

3	2880	720	Sin Información	Cartera de Crédito	Sociodemográficas Ingresos Historial crediticio Ubicación geográfica	
4	4088	1363	Crédito al consumo no revolvente: Microcrédito	Microfinanciera	Sociodemográficas Historial Crediticio Ingresos Macroeconómicas Garantías Condiciones del crédito	Atraso de 15 días o más
5	4332	2165	Sin Información	Institución Financiera	Sociodemográficas Condiciones del crédito Ubicación Geográfica	Sin Información
6	444	189	Crédito Comercial	Banco	Razones Financieras Demográficas Macroeconómicas	Ratio mora superior al promedio
7	12118	4039	Crédito al consumo no revolvente: Microcrédito	Microfinanciera	Sociodemográficas Condiciones del crédito Ubicación Geográfica	Un atraso que represente un costo para la entidad
8	S/I	S/I	Crédito al consumo no revolvente: Microcrédito	Microfinanciera	Sociodemográficas Condiciones del crédito Ubicación Geográfica	
9	1609	177	Crédito no revolvente: Microcrédito	Caja Rural	Sociodemográficas Familiares Condiciones del crédito Ubicación Geográfica	El crédito se regresa o no.

Fuente: Elaboración propia

Cuadro 39. Modelos reactivos de credit scoring parte II. Continuación

No.	Autores	Tipo de Modelo	Clasificación	Año	Método usado	Aporte
10	N/A	N/A	Crédito Hipotecario Crédito al consumo no revolvente Crédito comercial	No fue aplicado	N/A	N/A
11	N/A	N/A	Crédito Hipotecario	Hipotecaria	N/A	N/A
12	N/A	N/A	No se especifica	No fue aplicado	N/A	N/A
13	N/A	N/A	Crédito Comercial	Banco	N/A	N/A
14	63	37	Crédito al consumo no revolvente.	Banco	Sociodemográficas Ingresos Historial crediticio Ubicación geográfica	
15	N/A	N/A	Crédito al consumo no revolvente: Microcrédito	Microfinanciera	N/A	N/A
16	N/A	N/A	Crédito Comercial	Banco	N/A	N/A
17	ND	ND	Crédito al consumo no revolvente: Microcrédito	Microfinanciera	N/A	N/A
18	N/A	N/A	No se especifica	No fue aplicado	N/A	N/A
19	N/A	N/A	Crédito al consumo no revolvente: Microcrédito	Microfinanciera		

Fuente: Elaboración propia

3.5 El uso de Redes Sociales como un factor contraproducente en el uso de Credit Scoring de Microcrédito.

Ya comentábamos que la principal sugerencia el Banco Mundial, a través del *Credit Reporting Knowledge Guide* (2019) que hace para evitar la falta de información a la hora hacer credit scoring, primordialmente, en el mercado informal, se basa en obtener datos alternos del posible acreditado, como lo son:

1.) Perfiles de redes sociales
2.) Uso de transacciones financieras no crediticias
3.) Datos de pagos con proveedores
4.) Psicometría

Existe gran parte de la literatura que documenta el uso de las redes sociales como *inputs* en *credit scoring*, por ejemplo (Guo et al., 2016) realizan un análisis a través de la minería de datos, usando como principal soporte características sociodemográficas, así como variables asociadas a redes sociales, utilizó como base la información de *twitter*, lo que encontró, fue que existe bastante sesgo al tratar de usar dicha información, ya que la información puede no ser fidedigna, y puede no reflejar el verdadero comportamiento crediticio de una persona, además de que las conexiones son difíciles de manipular, ya que los *tweets* o perfiles personales pueden ser socavados por usuarios maliciosos.

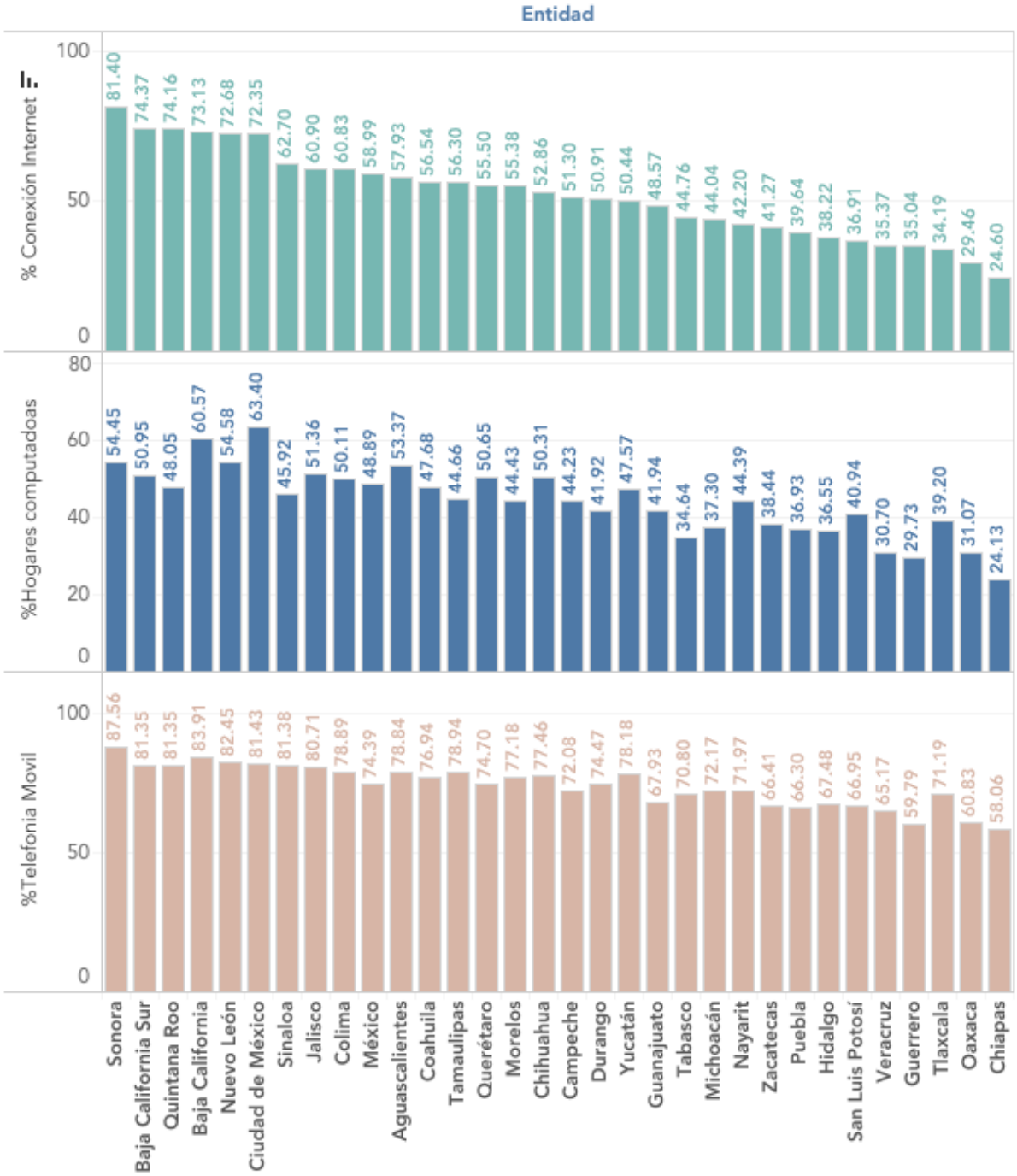
(Niu et al., 2019) utilizaron técnicas de machine learning para predecir el default crediticio de una persona, utilizando variables sociodemográficas, así como variables asociada a *Facebook*, en comparación con el trabajo anteriormente citado, encontraron que el uso de variables a redes sociales, mejora la capacidad de predicción de los credit scoring, por ejemplo, se documentó que la importancia de las variables en modelos como Bosques Aleatorios, *Adaboost* y *LightGBM*, fueron rankeadas en los últimos lugares, sin embargo contribuyeron a un mejor *accuracy*. Este trabajo concluye que las variables de la red social extraídas a través de teléfonos móviles podrían utilizarse para mejorar la precisión de la predicción de préstamos. La problemática de hacer uso de este tipo de técnicas, al menos en México, radica en que, gran porcentaje de la población no tiene acceso a internet y/o a *smartphones*. El INCAE (2019) a través de su trabajo de investigación sostiene que el Índice de Progreso social, cuya metodología radica en evaluar diferentes aspectos de la calidad de vida por Estado, aun es muy bajo en ciertas Entidades, en este mismo trabajo, se estipulan porcentajes de acceso a la población en materia de Internet y Teléfonos Móviles, sin embargo no hay una garantía de que estos teléfonos móviles entren en categoría de Smartphones.

Gráfica 18. Tasa de acceso a la información

Fuente: Elaboración propia con Datos INCAE 2019

Observamos que, a la par se tiene una falta de acceso enorme a la información en las Entidades sureñas del país, si miramos como se desagrega ese indicador, observamos lo siguiente.

Figura 19. Desagregación Indicador de Acceso Comunicaciones



Fuente: Elaboración propia con Datos INCAE 2019

De manera que observamos que 15 Entidades Federativas su población, cuentan con acceso de conexión a internet en una proporción menor al 50%, y en la gran mayoría, menos de un 78% no cuenta con algún dispositivo móvil, que no necesariamente implica que sea un *smartphone* y que tenga la capacidad recolectar datos necesarios que sirvan como inputs a la hora de generar *alternative* data. Esto representa que en México no existe la suficiente infraestructura para recolectar estos datos, que además, puede que el uso de esta información no revele verdaderas conexiones de un prospecto a acreditar y generaríamos relaciones espurias.

3.6 ¿Quién lo produce?: El Papel de los organismos pertenecientes a la LACP

En todo momento a lo largo de este trabajo hemos considerado como eje, la parte de Banca Múltiple, sin embargo, existen Instituciones Financieras que fueron diseñadas con este objetivo. Uno de los grandes avances en materia de Microfinanzas en México fue la aprobación de la ley de Ahorro y Crédito Popular (LACP) cuyo fin era crear un marco legal que estableciera mecanismos que facilitaran la organización e impulsaran credibilidad en las actividades financieras que desarrollan los organismos que captan ahorro popular y otorgan microcrédito, esto, mediante la incorporación de normas de organización, operación y funcionamiento adecuadas para los mismo. Dicha regulación establece tres figuras jurídicas

1.) Sociedades Financieras Comunitarias (SOFINCOS)
2.) Sociedades Financieras Populares (SOFIPOS)
3.) Organismos de Integración Financiera Rural

A las cuales se les conocen como Entidades de Ahorro y Crédito Popular (EACPs), estas a su vez van desde el nivel I hasta el nivel IV y respalda las operaciones que pueden realizar, hemos retomado de la CNBV (2020) lo siguiente:

Tabla 40. Niveles de Operación EACP's

	Nivel de Operaciones (valor en UDIS)	Nivel de Operaciones (valor en UDIS)	Nivel de Operaciones (valor en UDIS)	Nivel de Operaciones (valor en UDIS)
	I	II	III	IV
Activos totales	≤ 15 millones	> 15 millones y hasta 50 millones	> 50 millones y hasta 280 millones	superiores a 280

Fuente: Retomado CNBV

De manera que, margen de maniobra no es muy diferente para cada una de ellas, ya que en la mayoría se puede hacer captación y otorgar préstamos.

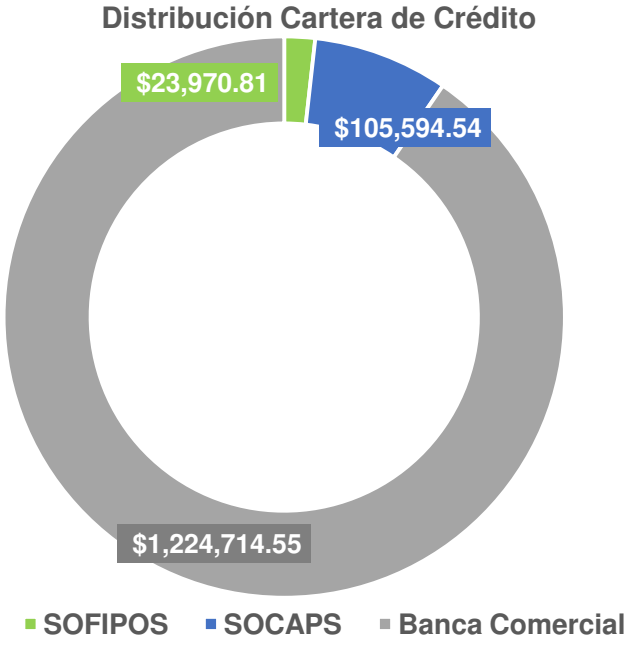
Cuadro 41. Operaciones por nivel EACP's

Operaciones	Nivel de Operaciones			
	I	II	III	IV
Recibir depósitos de dinero a la vista, de ahorro, a plazo, retirables en días preestablecidos y retirables con previo aviso.	●	●	●	●
Recibir préstamos y créditos de instituciones de crédito nacionales o extranjeras, fideicomisos públicos, organismos e instituciones financieras internacionales, así como de instituciones financieras extranjeras.	●	●	●	●
Otorgar préstamos o créditos a sus clientes.	●	●	●	●
Descontar, dar en garantía o negociar títulos de crédito, y afectar los derechos provenientes de los contratos de financiamiento que realicen con sus Clientes.	●	●	●	●
Distribuir seguros que se formalicen a través de contratos de adhesión, por cuenta de alguna institución de seguros o Sociedad mutualista de seguros, debidamente autorizada.	●	●	●	●
Distribuir fianzas, en términos de las disposiciones aplicables a dichas operaciones.	●	●	●	●
Celebrar contratos de arrendamiento financiero.	●	●	●	●
Realizar operaciones de factoraje financiero con sus clientes o por cuenta de éstos.		●	●	●
Ofrecer el servicio de abono y descuento en nómina.		●	●	●
Celebrar contratos de arrendamiento financiero con sus Clientes.			●	●
Prestar servicios de caja y tesorería.			●	●
Expedir tarjetas de crédito.				●
Ofrecer y distribuir, entre sus Socios las acciones de Sociedades de inversión operadas por las Sociedades Operadoras de Sociedades o por aquellas en cuyo capital participen indirectamente, así como promocionar la afiliación de trabajadores a las Administradoras de Fondos para el Retiro en cuyo capital participen directa o indirectamente.				●

Fuente: Retomado CNBV

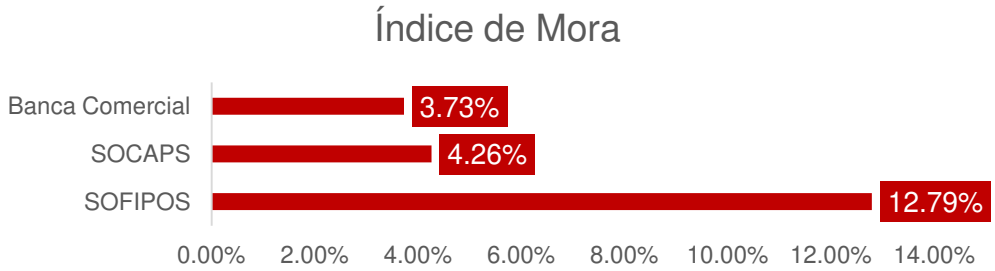
Sin embargo, estos esfuerzos no han sido suficientes para romper la brecha del crédito, esto debido a que existe poco capital invertido a este tipo de instituciones, considerando que existen muchas organizaciones adscritas a algunas de estas figuras jurídicas, pero no han tenido el impacto deseado. Si comparamos la cartera con todo tipo de productos de las EACP's, con respecto solamente de la cartera al consumo de la banca nos vamos a encontrar que aun comparando toda la cartera reportada de SOCAPS y SOFIPOS, representas juntas sólo el 10% de los saldos totales del sistema.

Gráfica20. Cartera Crédito EACP's y Banca



En tanto a riesgo, el índice de mora, al menos en SOCAP's parece estar controlado, mientras que, en las SOFIPOS, tiene cerca del 12.80% en cartera vencida.

Gráfica 21. IMOR Banca y EACP's



Esto parece ser indicativo de que las operaciones al menos en SOFIPOS pueden ser más riesgosas. Aunado a esto, su penetración por cada 10,000 habitantes parece ir en la misma dirección teniendo muy pocos contratos de crédito.

Cuadro 22. Penetración geográfica. Contratos por cada 10K habitantes EACP y Banca

Estado	Contratos EACP 10K	Contratos Banca 10K	Estado	Contratos EACP 10K	Contratos Banca 10K
Aguascalientes	326	6,121	**Morelos**	411	6,379
Baja California	15	5,547	**Nayarit**	710	5,381
Baja California Sur	43	6,113	**Nuevo León**	387	8,676
Campeche	106	5,663	**Oaxaca**	522	4,078
Coahuila	162	6,538	**Puebla**	149	4,803
Colima	1,195	6,042	**Querétaro**	1,217	6,537
Chiapas	118	3,201	**Quintana Roo**	102	6,210
Chihuahua	203	5,700	**San Luis Potosí**	688	5,086
Ciudad de México	131	12,272	**Sinaloa**	116	6,305
Durango	632	4,916	**Sonora**	27	6,055
Guanajuato	1,273	5,096	**Tabasco**	51	5,488
Guerrero	150	4,888	**Tamaulipas**	117	6,275
Hidalgo	204	5,166	**Tlaxcala**	92	5,762
Jalisco	787	5,818	**Veracruz**	219	5,587
México	128	5,668	**Yucatán**	339	6,087
Michoacán	665	5,316	**Zacatecas**	406	4,012

Existen muchos factores por lo que las Instituciones de Crédito enfocadas a atender microcrédito no se expandan pese que existen las condiciones para hacerlo, esto se puede deber a la dificultad en las maneras de operar el microcrédito.

3.6.1 Procesos de Crédito de las EACP's

Hasta el día de hoy, nos encontramos con una estructura generalizada muy laxa de lo que son los procesos de concesión del microcrédito por parte del regulador, cabe mencionar que al igual que las definiciones ambiguas de microcrédito por parte de la autoridad, como lo comentábamos en la sección 1, no existen procesos robustos para la otorgación crediticia, basta con mirar los requisitos mínimos de las disposiciones de carácter general aplicables a entidades de ahorro y crédito popular la cual citamos textualmente:

"I. Promoción y otorgamiento de crédito.

a) Las Entidades podrán establecer en los manuales de crédito, procesos de autorizaciones automáticas de créditos que permitan otorgar el crédito correspondiente a cualquier solicitante, siempre y cuando se reúnan las condiciones que se indican a continuación:

1. Documentación mínima a ser entregada por tipo de crédito;
2. Identificación del solicitante, así como finalidad para la cual se solicita el crédito o, en su caso, características de los depósitos que el solicitante mantenga en la Entidad;
3. Monto máximo a otorgar según el resultado de la información entregada, y
4. Tasas de interés conforme a sus políticas.

b) Adicionalmente, las Entidades podrán establecer metodologías para la aprobación y otorgamiento de créditos cuyo monto sea considerable, según las características de las operaciones que realice la Entidad, para lo cual deberá tomarse en cuenta, por lo menos, lo siguiente:

1. Contar con la documentación mínima indispensable que establezca el propio manual de crédito;
2. La información que valide la experiencia de ahorro o de pago del acreditado;
3. La capacidad del acreditado para cumplir con sus obligaciones, y
4. La determinación de un parámetro o escala de medición que indique el riesgo del potencial acreditado.

c) El Comité de Crédito o su equivalente, será la instancia responsable de la aprobación de los créditos solicitados a la Entidad, aunque podrá delegar sus funciones en subcomités ya sea regionales o por sucursales, siempre y cuando la existencia de dichos subcomités esté prevista en el manual de crédito de la Entidad, los cuales deberán estar integrados por funcionarios de la propia Entidad. Para dicha aprobación deberán seguir los lineamientos que al efecto se establezcan en el manual.
d) La Entidad que cumpla con los requerimientos de capitalización por riesgos de crédito correspondientes y en general con lo establecido en la presente Sección, quedará relevada de la obligación de contar con la aprobación del Comité de Crédito o su equivalente, cuando el importe total de los créditos otorgados por dicha Entidad a la persona solicitante, incluyendo a sus dependientes económicos, no sea mayor a 5,000 (cinco mil) UDIS, y siempre y cuando su manual de crédito prevea procesos de autorizaciones automáticas, de conformidad con lo establecido en el inciso a) anterior

Para nosotros el punto fundamental es el punto 4 del inciso b debido a que la regulación lo deja a libre concepción, cuando en la mayoría de los casos, las entidades enfocadas al otorgamiento del microcrédito no cuentan con la experiencia fundamental para realizar *scorings* de créditos robustos, sus procesos aun son muy artesanales o jamás se sofistican y sus criterios de riesgo regularmente está basado en scores genéricos de Sociedades de Información Crediticia (SIC's). Además que por regulación, no existen parámetros estadísticos mínimos aceptables para la determinación de potencial riesgo en el proceso de crédito, como por ejemplo un mínimo de KS o Gini en el desarrollo de metodologías internas que permitan hacer decisiones más inteligentes con los acreditados a través de scorings, con esto no nos incentivamos la idea de excluir prospectos dada una probabilidad de incumplimiento, sino por el contrario, saber que ciertos prospectos cuentan con probabilidades de impago altas, por lo tanto tener estrategias adecuadas para aceptarlos y mantener procesos internos de seguimiento que permitan mitigar el riesgo. En un estudio

realizado por Cusquer y Maldonado (2011) evidenció que el 82.9% de las entidades financieras carecían de metodologías adecuadas para la otorgación de microcréditos.

3.6.2 Fondeo de EACP

El fondeo en una institución financiera es fundamental para garantizar la continuidad sus operaciones de crédito, es por ello que es un punto relevante de este trabajo debido a que hay una diferencia contra las instituciones de Banca Múltiple que impacta directamente el producto de microcrédito, y es que, al menos los grandes bancos pueden captar grandes cantidades recursos del público a precios muy baratos, es decir, con tasas de rendimiento muy bajas, mientras que para las EACP's es más difícil, puesto que el fondeo se constituye primordialmente de 4 ejes:

1.) Captación del Público
2.) Patrimonio
3.) Fondeo Público
4.) Fondeo Privado

Lo que implica una restricción continua para dichas instituciones, ya que al no ser tan representativas o no ser economías de escala, presentan dificultades para captar del público o conseguir fondeo privado, esto representa un problema cíclico, porque pueden tener infraestructura para la otorgación de créditos, pero no los recursos para dispersarlos, lo que se convierte en un *loop* para dichas instituciones, en un trabajo de Marulanda Consultores en colaboración con DAI (2010) sostuvieron que gran parte de las 6 Instituciones Microfinancieras analizadas tenían gran concentración con fuentes de Fondeo Privado Internacional, el problema con este tipo de fondeo, es que en muchas ocasiones la deuda suscrita es moneda extranjera, lo que obliga a las instituciones a prever tasas de interés más altas para cubrir posibles pérdidas por riesgo de mercado. Sin embargo, uno de los grandes soportes en este rubro, es el apoyo del Gobierno Federal a través de un Programa Nacional de Financiamiento al Microempresario (PRONAFIM) lanzado en 2001, cuyo objetivo es fomentar la actividad de microempresarios a través del crédito y está divido fundamentalmente en dos fondos FORMUR y FINAFIM, el primero enfocado a otorgación de créditos a mujeres del medio rural y el segundo surgió como un fondo que otorgaba créditos para apoyar la creación y crecimiento de Microfinancieras sin definir un tipo de parámetro, sino más bien, asimilándolo a un producto de crédito como lo es el personal, lo que sí requería, era un mínimo tiempo de operación y una plataforma tecnológica que permitiera hacerle un adecuado seguimiento a la cartera. Hasta la fecha, siguen existiendo estos apoyos para el Fondeo, sin embargo, una de las problemáticas que surgen, es que el fondeo público, está sujetos a tasas de interés variables, tal es así que Corporación Financiera Internacional (IFC) *et al*. (2017), hace un análisis en retrospectiva del fondeo de 2006 a 2015 donde se remarca el impacto de la tasa CETE, en el fondeo promedio por año y a medida que crece o disminuye la CETE, se observa el mismo comportamiento en el fondeo, esto también sugiere que se deben prever en el largo plazo, shocks ante cambios en tasas de interés derivados de los cambios en la economía, ya que en la mayoría de los créditos en Microfinancieras, los créditos se pactan a tasas fijas, lo que implicaría

problemas serios de liquidez en caso de que la tasa de los créditos no contemple estos factores.

3.7 Costos del Microcrédito

Como parte final de esta sección, es importante recalcar que la estructura de los costos en microcréditos, va más allá del riesgo que representa la probabilidad de impago y del fondeo al cuál se pueden conseguir los recursos, y es que la forma de operar de las instituciones dedicadas al microcrédito también es un factor fundamental en la estructura de costos, es por ello que citando de nuevo a Corporación Financiera Internacional (IFC) *et al*. (2017), a través de un panel de datos explica que el primer factor que impacta la tasa de interés, es el fondeo, seguido de los costos operativos ajustados por riesgo de crédito, es decir, que a medida que las instituciones mejoren sus procesos de otorgamiento de crédito, la calidad de la cartera progresivamente va a mejorar y, con ello, bajar los costos por cobranza en sus diferentes etapas, ya que en Microcrédito, la forma de cobrar, juega un papel fundamental, ya que se debe mantener un contacto continuo con el cliente para el repago del crédito, esto en efecto dominó, podría disminuir el nivel de reservas y potencialmente contribuir a bajar las tasas de interés, este último punto de las reservas comentábamos, se puede modular a través del *exposure* y la tasa de interés ex ante.

Estos análisis, nos dan pauta para enfocarnos al microcrédito, ya que es un producto financiero que nadie oferta, pero muchos necesitan y que en gran medida, como lo explicamos en la sección 1, se da debido a que no hay información que respalde el scoring de crédito tradicional o sus atributos ex ante de la operación, catalogan a los prospectos con altas probabilidades de impago

3.8 Reflexiones de la sección

1.) Insistimos en que la brecha de financiera se rompe cuando los bancos o instituciones financieras pueden hacer eficiente su cartera (dados ciertos riesgos) y los demandantes de crédito pueden accesar al crédito.
2.) El principal problema del microcrédito, es la informalidad puesto está seriamente relacionada con la información asimétrica, por lo que para las instituciones oferentes de crédito es difícil aceptarlos como parte de su cartera, debido a que en la evaluación crediticia, no pueden revelar el nivel de riesgo al que se enfrentan.
3.) En el caso de México existe una clara relación inversamente proporcional entre los saldos de cartera por Entidad Federativa y la tasa de informalidad (TIL 1), por lo que consideramos que este es uno de los factores céntricos a resolver para poder otorgar crédito.
4.) El puente ideal para resolver la brecha financiera, asumiendo la informalidad e información asimétrica es a través del uso de credit scorings reactivos potentes haciendo el uso de toda la información alterna disponible.
5.) La única data alterna que no consideramos viable es la información proveniente de redes sociales, puesto que en la práctica no adicionan gran valor de predicción, además de que es muy subjetiva su interpretación.

6.) En México será difícil enfrentar la brecha de crédito a través de la tecnología, ya que el nivel de acceso a la información en ciertas Entidades Federativas, particularmente las del sur, no tienen la infraestructura suficiente, lo que es indicativo que estaríamos en un constante *loop* al tratar de incursionar en el microcrédito debido a que no existen las condiciones tecnológicas necesarias para que los demandantes del crédito puedan accesar al sistema financiero.
7.) Las tasas de interés, aun con riesgo de impago alto no deberían ser tan altas, puesto que a nivel Exposición al Incumplimiento es mucho mayor, esto implica, que al menos en el caso de México se debiera reservar más, lo que implica un doble costo en caso de irrecuperabilidad, debido al impacto directo en los intereses devengados no cobrados por 90 días. Además, que, a nivel competencia, sería muy contraproducente si se aspira a ser una economía de escala.
8.) Las instituciones ideales para administrar el microcrédito, debieran ser las asociadas a la LACP, ya que originalmente fueron diseñadas para eso.
9.) Dichas instituciones carecen de procesos robustos en la "originación" de los créditos, ya que por regulación, no se tienen parámetros estadísticos definidos con respecto a las decisiones de crédito, que posteriormente se traducen en problemas de cartera vencida lo que dificulta su sostenibilidad.
10.) Las tasas de interés en microcrédito se ven fundamentalmente impactadas por el fondeo (más, cuando son pactadas a tasas variables) y los costos operativos, por lo que en mayor proporción de estos, implicarían tasas de interés mayores

4. Herramientas estadísticas y de probabilidad

4.1 Cadenas de Markov y clasificación de estados

Una cadena de Markov es un proceso en tiempo discreto en el que una variable aleatoria X_n cambia con el tiempo. Las matrices de Markov tienen la propiedad de que la probabilidad X_n = j, X sólo dependen del estado inmediato anterior X_{n-1} en que se considere, *n*,

$$P\left(X_n = j \mid X_{n-1} = i\right)$$

Se denomina cadena homogénea, es decir, que cuenta con las mismas probabilidades.

Clasificación de las cadenas de Markov

A) Estado Absorbente.

Un estado es absorbente cuando una vez que entra en él no se puede salir del mismo. Es decir, un estado E_i es absorbente sí

$$p_{ii} = 1$$
$$p_{ij} = 0 \text{ para } i \neq j \mid j = 1,...,m$$

en la $i-ésima$ fila de T.

B) Estado periódico.

La probabilidad de que se regrese al estado E_i en el paso n es $p_{ii}^{(n)}$. Sea $t \in \mathbb{Z}$ además $t > 1$. Supongamos que

$$p_{ii}^{(n)} = 0 \text{ para } n \neq t, 2t, 3t, ...$$

$$p_{ii}^{(n)} \neq 0 \text{ para } n = t, 2t, 3t, ...$$

En este caso, se dice que el estado E_i es periódico de periodo t. Si para un estado no existe dicho valor de t, entonces se dice que el estado es *aperiódico.*
Alternativamente, definimos

$$d(i) = mcd\left\{n \mid p_{ii}^{(n)} > 0\right\},$$

es decir, el máximo común divisor (mcm) del conjunto de los números enteros n para los que $p_{ii}^{(n)} > 0$. Entonces, el estado de E_i, es periódico si $d(i) > 0$ y aperiódico si $d(i) = 1$.

C) Estado Recurrente

Definimos como $f_j^{(n)}$ a la probabilidad de que la primera visita al estado E_j ocurra en la etapa n. Esta probabilidad no es la misma que $p_{jj}^{(n)}$, es decir, es la probabilidad de que se produzca un retorno en el $n-ésimo$ paso y esto incluye a los posibles retornos en los pasos $1,2,3,\ldots,n-1$ también. De lo cual se deduce que

$$p_{jj}^{(1)} = f_j^{(1)}$$

$$p_{jj}^{(2)} = f_j^{(2)} + f_j^{(1)} p_{jj}^{(1)}$$

$$p_{jj}^{(3)} = f_j^{(3)} + f_j^{(1)} p_{jj}^{(2)} + f_j^{(2)} p_{jj}^{(1)}$$

es decir, la probabilidad de un retorno en el paso 3 es igual a la probabilidad de un primer retorno en el paso 3 o la probabilidad de un retorno en el segundo paso y un retorno un paso después.

En general,

$$p_{jj}^{(n)} = f_j^{(n)} + \sum_{r=1}^{n-1} f_j^{(r)} p_{jj}^{(n-r)}$$

se puede expresar en términos de $f_j^{(n)}$

$$f_j^{(1)} = p_{jj}^{(1)}$$

$$f_j^{(n)} = p_{jj}^{(n)} - \sum_{r=1}^{n-1} f_j^{(r)} p_{jj}^{(n-r)}$$

La probabilidad de regresar en algún paso al estado E_j es

$$f_j = \sum_{n \in \mathbb{N}} f_j^{(n)},$$

si $f_j = 1$, entonces seguro que se regresa a E_j estado recurrente.

D) Estado transitorio

En un estado recurrente, la probabilidad de que se regrese por primera vez a ese estado en algún paso es 1, pero para otros estados sucede que;

$$f_j = \sum_{n\in\mathbb{N}} f_j^{(n)} < 1$$

lo que significa que no regresa al estado E_j de manera segura. A esté estado lo llamaremos *transitorio.*

E) Estado ergódico

Un estado ergódico, es el estado que cumple con ser recurrente, no nulo y aperiódico, su principal función es la clasificación de cadenas y probar la existencia de distribuciones de probabilidad de límite.

4.1.2 Probabilidades de Transición de *n* pasos

Una probabilidad de transición es una cadena homogénea finita, es decir, con *m* posibles estados $E_1, E_2, ..., E_m$ podemos decir que;

$$p_{ij} = P\left(X_n = j \mid X_{n-1} = i\right),$$

donde $i, j = 1, 2, ..., m.$ Si $p_{ij} > 0$ entonces se dice que el estado E_i puede *comunicar* con E_j. La comunicación puede ser mutua si también $p_{ji} > 0.$

Para cada i fijo, la serie de valores $\left\{p_{ij}\right\}$ es una distribución de probabilidad, debido a que en cualquier paso puede ocurrir alguno de los eventos $E_1, E_2, ..., E_m$ y son mutuamente excluyentes. Los valores p_{ij} se denominan *probabilidades de transición*[32] que satisfacen la siguiente condición;

$$Si\ p_{ij} > 0,\ \text{además}\ \sum_{j=1}^{m} p_{ij} = 1 \text{ para todo i } \in \{1, 2, ..., m\}$$

De tal manera, que estos valores se combinan formando una *matriz de transición*[33]T de tamaño $m \times m$, donde

$$T = \left[\, p_{ij} \,\right] = \begin{bmatrix} p_{11} & p_{21} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}$$

Podemos observar que cada fila es una distribución de probabilidad, es decir;

$$\sum_{j=1}^{m} p_{ij} = 1\,.$$

32

33

Probabilidad $p_j^{(n)}$

Una probabilidad de mayor interés es la probabilidad de llegar a E_j después de n pasos, con una distribución de probabilidad $\left\{p_i^{(0)}\right\}$.

Observamos que $\left\{p_i^{(0)}\right\}$ es la probabilidad de que el sistema ocupe inicialmente el estado E_i, de modo que

$$\sum_i^m p_i^{(0)} = 1.\ \forall i$$

Ahora, supongamos que $p_j^{(1)}$ es la probabilidad de alcanzar E_j en un solo paso, entonces, sabemos por el *Teorema de probabilidad total* [34]que sí;

$$p_j^{(1)} = \sum_{i=1}^m p_i^{(0)} p_{ij}.$$

Entonces se puede expresar de manera vectorial. Sean $p^{(0)}$ y $p^{(1)}$ los vectores fila de probabilidad dados por

$$p^{(0)} = \left(p_1^{(0)},..., p_m^{(0)}\right) \quad \text{y} \quad p^{(1)} = \left(p_1^{(1)},..., p_m^{(1)}\right),$$

donde $p^{(0)}$ es la distribución de probabilidad inicial y que $p^{(1)}$ es la probabilidad de que se alcance cada uno de los estados $E_1, E_2, ..., E_m$ después de un paso. Con la misma notación, vemos que;

$p^{(1)} = \left[\, p_j^{(1)} \right] = \left[\sum_{i=1}^m p_i^{(0)} p_{ij} \right] = p^{(0)}T$ donde T pertenece a las matrices de transición.

Del mismo modo,

$$p^{(2)} = p^{(1)}T = p^{(0)}T^2$$

y en n pasos,

$$p^{(n)} = p^{(n-1)}T = p^{(0)}T^n$$

donde;

$$p^{(n)} = \left(p_1^{(n)},..., p_m^{(n)}\right)$$

y de manera general,

$$p^{(n+r)} = p^{(r)}T^n$$

Observe que $p_j^{(n)}$ es la probabilidad incondicional de estar en el estado E_j en el $n-ésimo$ paso, dado que la probabilidad inicial es $p^{(0)}$, esto es,

$$P\left(X_n = j\right) = p_j^{(n)},$$

tal que,

34

$$\sum_{j=1}^{m} p_j^{(n)} = 1.$$

Se define $p_{ij}^{(n)}$ como la probabilidad de que la cadena esté en el estado E_j, después de n pasos, dado que la cadena empezó en el estado E_i, se tiene que;

$$p_{ij}^{(n)} = P\left(X_n = j \mid X_0 = i\right)$$

por la propiedad markoviana [35] se tiene que

$$p_{ij}^{(n)} = \sum_{k=1}^{m} P\left(X_n = j, X_{n-1} = k \mid X_0 = i\right), \text{ para } n \geq 2$$

ya que la cadena debe haber pasado por uno de los m posibles estados en la etapa $n-1$.

Nota: Si tiene la siguiente igualdad, para tres posibles sucesos A, B y C:

$$P\left(A \cap B \mid C\right) = P\left(A \mid B \cap C\right) \cdot P\left(B \mid C\right)$$

Si sustituimos:

$$A \rightarrow \left(X_n = j\right)$$
$$B \rightarrow \left(X_{n-1} = k\right)$$
$$C \rightarrow \left(X_0 = i\right)$$

entonces, $p_{ij}^{(n)}$

$$= \sum_{k=1}^{m} P\left(X_n = j, X_{n-1} = k \mid X_0 = i\right)$$
$$= \sum_{k=1}^{m} P\left(X_n = j \mid X_{n-1} = k, X_0 = i\right) P\left(X_{n-1} = k \mid X_0\right)$$
$$= \sum_{k=1}^{m} P\left(X_n = j \mid X_{n-1} = k\right) P\left(X_{n-1} = k \mid X_0 = i\right)$$
$$= \sum_{k=1}^{m} p_{kj}^{(1)} p_{ik}^{(n-1)} = \sum_{k=1}^{m} p_{ik}^{(n-1)} p_{kj}^{(1)}$$

en donde se ha utilizado la propiedad markoviana nuevamente. La ecuación anterior se denomina de Chapman-Kolmogorov. Haciendo n igual a 2,3,... se obtiene que las matrices con esos elementos son

$$\left[p_{ij}^{(2)}\right] = \left[p_{ik}^{(1)} p_{kj}^{(1)}\right] = T^2$$
$$\left[p_{ij}^{(3)}\right] = \left[p_{ik}^{(2)} p_{kj}^{(1)}\right] = T^3$$
$$\vdots \qquad\qquad \vdots$$
$$\left[p_{ij}^{(n)}\right] = T^n.$$

35

4.2 Modelo Logit

Existen modelos de elección discreta en los que el conjunto solamente tienen dos eventos mutuamente excluyentes con una variable Bernoulli de parámetro *p*.

$$P(yi) = p_i^{y_i}\left(1-p_i\right)^{(1-y_i)} \text{ donde los valores de } y_i = 0,1.$$

La función de máxima verosimilitud para una muestra aleatoria de n datos (x_i, y_i) se calcula como

$$P(y_i,...,y_n) = \prod_{i=1}^{n} p_i^{y_i}\left(1-p_i\right)^{(1-y_i)},$$

aplicando el algoritmo de ambos lados de la igualdad, obtenemos que:

$$\log P(y) = \sum_{i=1}^{n} y_i \log p_i + \sum^{n}(1-y_i)\log(1-p_i)$$

asimismo, la función log verosimilitud se escribe como

$$\log P(y) = \sum_{i=1}^{n} y_i \log\left(\frac{p_i}{(1-p_i)}\right) + \sum_{i=1}^{n}\log(1-p_i) \tag{0.1}$$

Consideremos $\beta^T = (\beta_0, \beta_1, ..., \beta_p)$ y $x_i^T = (1, x_{i1}, x_{i2}, ..., x_{ip})$ para escribir el modelo de la forma

$$\log\left(\frac{p_i}{(1-p_i)}\right) = x_i^T\beta \tag{0.2}$$

Ahora, la ecuación (2.2) la sustituimos en (2.1). Donde obtenemos la función de verosimilitud en términos de logaritmos, con valores de β dada por

$$L(\beta) = \sum_{i=1}^{n} y_i x_i^T\beta - \sum_{i=1}^{n}\log\left(1+e^{x_i^T\beta}\right). \tag{0.3}$$

Para obtener los estimadores β de máxima verosimilitud derivamos $L(\beta)$ con respecto de cada uno de los parámetros de β_j con $j \in \{1, 2, ..., p\}$ e igualamos a cero.

En términos de matrices

$$\begin{bmatrix} \frac{\partial L(\beta)}{\partial \beta_0} \\ \frac{\partial L(\beta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_j} \\ \vdots \\ \frac{\partial L(\beta)}{\partial \beta_p} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i (1) \\ \sum_{i=1}^{n} y_i x_{i1} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{ij} \\ \vdots \\ \sum_{i=1}^{n} y_i x_{ip} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{n} (1) \left(\frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}} \right) \\ \sum_{i=1}^{n} x_{i1} \left(\frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}} \right) \\ \vdots \\ \sum_{i=1}^{n} x_{ij} \left(\frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}} \right) \\ \vdots \\ \sum_{i=1}^{n} x_{ip} \left(\frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}} \right) \end{bmatrix} \tag{0.4}$$

cada una de estas derivadas se expresan en un vector columna de la forma

$$\frac{\partial L(\beta)}{\partial \beta} = \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i \left(\frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}} \right), \tag{0.5}$$

igualando (2.4) al vector cero, obtenemos:

$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} x_i \left(\frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}} \right) = \sum_{i=1}^{n} x_i p_i. \tag{0.6}$$

Si β es el vector de parámetros que cumplen que el sistema (2.4), calculamos p_i en términos de esos estimadores y obtenemos una estimación y_i tal que $\hat{y}_i = \hat{p}_i$, así

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} x_{ij} \hat{y}_i$$

de aquí

$$\sum_{i=1}^{n} x_{ij} e_i = \sum_{i=1}^{n} x_{ij} \left(y_i - \hat{y}_i \right) = 0$$

donde e_i representa los residuos del modelo son equivalentes a la regresión estándar (mínimos cuadrados).

5. Técnicas de modelación *Machine Learning*

5.1 Máquinas de soporte vectorial

Los fundamentos teóricos de las Máquinas de soporte vectorial (*SVM*) fueron presentados en el año de 1992 en la conferencia COLT (*Computacional Learning Theory*) por Boser *et al.* (1992) . Las máquinas de soporte vectorial pertenecen a la familia de los clasificadores

lineales dado que inducen hiperplano o separadores lineales de dimensiones grandes introducidos por funciones núcleo o kernel, es decir, este modelo busca una dimensión mayor en la cual los puntos puedan separarse linealmente

5.5.1 SVM Lineal

Consideremos un conjunto de datos de vectores característicos que queremos clasificar:

$$X = \{\bar{x}_1, \bar{x}_2, ..., \bar{x}_n\} \text{ donde } \bar{x}_i \in R^m$$

Por simplicidad, suponemos que estamos trabajando con una clasificación bipolar (en todos los demás casos, es posible utilizar automáticamente la estrategia de uno contra todos) y fijemos las etiquetas de clase como -1 y 1:

$$Y = \{y_1, y_2, ..., y_n\} \text{ donde } y_i \in \{-1,1\}$$

El objetivo es encontrar el mejor hiperplano de separación, para lo cual la ecuación es como sigue:

$$\bar{\omega}^T \cdot \bar{x} + b = 0 \text{ donde } \bar{\omega} = (\omega_1, \omega_2, ..., \omega_m)^T$$

Podemos reescribirlo como sigue:

$$\bar{y} = f(\bar{x}) = signo(\bar{\omega}^T \cdot \bar{x} + b)$$

Normalmente dos clases están normalmente separadas por un margen con dos límites donde se encuentran unos pocos elementos. Esos elementos se denominan vectores de apoyo y el nombre del algoritmo deriva de su peculiar función. Para una expresión matemática genérica, es preferible renormalizar nuestro conjunto de datos de manera que los vectores de apoyo se encuentren en dos hiperplanos con las siguientes ecuaciones:

$$\bar{\omega}^T \cdot \bar{x} + b = -1$$

$$\bar{\omega}^T \cdot \bar{x} + b = 1$$

Cuando al maximizar la distancia entre dos hiperplanos fronterizos para reducir la probabilidad de una clasificación errónea (que es mayor cuando la distancia es corta). Suponiendo que los límites son paralelos, la distancia entre ellos está definida por la longitud del segmento perpendicular a ambos y que conecta dos puntos, considerando los puntos como vectores, tenemos lo siguiente:

$$\bar{x}_2 - \bar{x}_1 = t\bar{\omega}$$

Ahora, manipulando las ecuaciones del hiperplano límite, obtenemos esto:

$$\bar{\omega}^T \cdot \bar{x}_2 + b = \bar{\omega}^T \cdot (\bar{x}_1 + t\bar{\omega}) = 1 \Rightarrow (\bar{\omega}^T \cdot \bar{x}_1 + t\bar{\omega}) + t\|\bar{\omega}\|^2 = 1$$

El primer término de la última parte es igual a -1, así que resolveremos para t, y obtenemos esta ecuación:

$$t = \frac{2}{\|\bar{\omega}\|^2}$$

La distancia entre x_1 y x_2 es la longitud 1 del segmento t; sustituyendo la expresión anterior, podemos derivar otra ecuación:

$$d(\bar{x}_1, \bar{x}_2) = t\|\bar{\omega}\| = \frac{2}{\|\bar{\omega}\|}$$

Considerando todos los puntos del conjunto de datos, podemos imponer la siguiente restricción:

$$y_i\left(\bar{\omega}^T \cdot \bar{x}_i + b\right) \geq 1 \quad \forall\left(\bar{x}_i, y_i\right)$$

Esto se garantiza usando -1, 1 como etiquetas de clase y márgenes de límite. La igualdad es verdadera únicamente para los vectores de apoyo, mientras para todos los otros puntos será mayor que 1. Es importante considerar que el modelo no toma en cuenta vectores más allá de este margen.

En muchos casos, esto puede dar lugar a un modelo robusto, pero en muchos conjuntos de datos, esto también puede ser una seria limitación.

Usemos lo siguiente para evitar la rigidez de este modelo, mientras mantenemos la misma técnica de optimización.

En este punto, definamos la función a minimizar, para entrenar un SVM (lo que equivale a maximizar la distancia):

$$\begin{cases} \min \frac{1}{2}\|\bar{\omega}\| \quad \text{s.a} \quad y_i\left(\bar{\omega}^T \cdot \bar{x}_i + b\right) \geq 1 \forall\left(\bar{x}_i, y_i\right) \end{cases}$$

Esto puede simplificarse aún más (eliminando la raíz cuadrada de la norma) en el siguiente problema de programación cuadrática:

$$\begin{cases} \min \frac{1}{2}\bar{\omega}^T \bar{\omega} \text{ s. a } y_i\left(\bar{\omega}^T \cdot \bar{x}_i + b\right) \geq 1 \forall\left(\bar{x}_i, y_i\right) \end{cases}$$

Este problema es equivalente a la minimización de la función de pérdida de la bisagra:

$$L = \max\left(0, 1 - \left(\bar{\omega}^T \cdot (\bar{x}_i + b) y_i\right)\right)$$

De hecho, el objetivo no es sólo encontrar el hiperplano de separación óptimo, sino también maximizar la distancia entre los vectores de apoyo (que son los delimitadores de los extremos) cuando una muestra X_i está correctamente clasificada pero su distancia al hiperplano es inferior a 1, L > 0 y el algoritmo se ve obligado a actualizar el vector del parámetro ω, mientras que permanece pasivo si L=0 (condición que cumple todas las muestras correctamente clasificadas cuya distancia al hiperplano de separación es superior a 1). Hasta cierto punto, los SVM son modelos muy económicos, porque explotan las propiedades geométricas de un conjunto de datos. Como los vectores de apoyo son los puntos diferentes más cercanos (en términos e clase), no hay necesidad de preocuparse por todas las demás muestras. Cuando se ha encontrado el mejor hiperplano (sólo los vectores de apoyo contribuyen a su ajuste), L=0 y no se necesitan otras correcciones.

5.2 Clasificación lineal (Clasificación basada en el núcleo)

Cuando se trabaja con problemas no lineales, es útil transformar los vectores originales proyectándolos en un espacio (a menudo de mayor dimensión) donde puedan ser separados linealmente. Considere una función de mapeo del espacio de muestra de entrada X a otro, V:

$$\phi(\bar{x}) : X \to V \quad \forall \bar{x} \in X$$

Suponiendo que cuando las muestras se han proyectado en V pueden separarse fácilmente. Sin embargo, ahora hay un problema de complejidad que tenemos que superar. La formulación matemática, de hecho, se convierte en lo siguiente:

$$\min \frac{1}{2}\bar{\omega}^T\bar{\omega} + C\sum_i \zeta_i \quad \text{s.a} \quad y_i\left(\bar{\omega}^T \cdot \phi\left(\bar{x}_i\right) + b\right) \geq 1 - \zeta_i \quad \forall\left(\bar{x}_i, y_i\right) \text{ y } \zeta_i \geq 0$$

Podemos ver que parámetro ν está delimitado entre 0 (excluido) y 1, puede utilizarse para controlar al mismo tiempo el número de vectores de apoyo (los valores más altos aumentarán su número) y el error de entrenamiento(los valores más bajos reducen la fracción de errores). La prueba formal de estos resultados requiere que expresemos el problema utilizando el método de Lagrange; sin embargo, es posible comprender la dinámica intuitivamente, considerando los casos límite.

Cuando $\vartheta \to 0$, la variable τ no tiene más efecto sobre la función objetivo. Si η >1, las variables de holgura se penalizan como en C-SVM con C<1. En este caso, el número de vectores de apoyo se convierte en el mínimo y al mismo tiempo, el error aumenta. Por otro lado, cuando ϑ>0, la penalización en la función objetivo se convierte en lo siguiente:

$$\max\left(\sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \phi\left(\bar{x}_i\right)^T \cdot \phi\left(\bar{x}_j\right)\right) \quad \text{s.a } \sum_i \alpha_i y_i = 0$$

Por lo tanto, para cada par de vectores, es necesario el siguiente producto punto:

$$\phi\left(\bar{x}_i\right)^T \cdot \phi\left(\bar{x}_j\right)$$

Ahora, usaremos el truco del núcleo. Hay funciones particulares llamadas kernel que tienen la siguiente propiedad:

$$K\left(\bar{x}_i, \bar{x}_j\right) = \phi\left(\bar{x}_i\right)^T \cdot \phi\left(\bar{x}_j\right)^T$$

En otras palabras, el valor del núcleo de dos vectores de características es el producto de los dos vectores proyectados. La buena noticia es que, según el Teorema de Mercer, la función φ(x) siempre existe cuando el núcleo satisface la condición de Mercer en X. Si consideramos K como una matriz, es necesario que:

$$\sum_i \sum_j \bar{x}^{(i)} K_{ij} \bar{x}^{(j)} \geq 0 \quad \forall \bar{x} \in X$$

Significa que K debe ser semidefinido positivo. También existe una condición análoga para las funciones continuas, pero es este contexto, es útil saber que encontrar núcleo no están complejo como se imagina; por lo tanto, su aplicación se ha ido extendiendo cada vez más. Utilizando el truco del núcleo la complejidad computacional sigue siendo casi la misma, pero es posible beneficiarse de la potencia de las proyecciones no lineales, incluso en un número muy grande de dimensiones.

5.3 Función de base radial

El núcleo de la Función de Base Radial (FBR) es el valor por defecto para la SVM y se basa en la siguiente función:

$$K\left(\bar{x}_i, \bar{x}_j\right) = e^{-\gamma\left|\bar{x}_i - \bar{x}_j\right|^2}$$

El parámetro γ determina la amplitud de la función, que no está influenciada por la dirección sino sólo por la distancia del origen. Este núcleo es particularmente útil cuando

los conjuntos son cóncavos y se intersectan, por ejemplo, cuando un subconjunto perteneciente a una clase está rodeado por otro perteneciente a otra clase.

5.3.1 Núcleo polinómico

El núcleo polinómico se determina con la siguiente función:

$$K\left(\overline{x}_i, \overline{x}_j\right) = \left(\gamma \overline{x}_i^T \cdot \overline{x}_j + r\right)^c$$

El exponente c se especifica a través del parámetro grado, mientras que el término constante r se llamada coef0. Esta función puede expandir fácilmente la dimensión con un gran número de variables de apoyo y superar problemas no lineales.
Los requisitos en términos de recursos son normalmente más altos, pero considerando que una función no lineal puede a menudo aproximarse bastante para un área delimitada.

5.3.2 Núcleo sigmoideo

El núcleo sigmoideo está definido como:

$$K\left(\overline{x}_i, \overline{x}_j\right) = \frac{1 - e^{-2\left(\gamma \overline{x}_i^T \cdot \overline{x}_j + r\right)}}{1 + e^{-2\left(\gamma \overline{x}_i^T \cdot \overline{x}_j + r\right)}} = \tanh\left(\gamma \overline{x}_i^T \cdot \overline{x}_j + r\right)$$

El término constante r se especifica a través del parámetro coef0. Si γ<<1 y r<0, el núcleo sigmoideo se comporta como uno RBF, sin embargo, su rendimiento nunca es dominante con respecto a los núcleos RBF o polinómicos. Por lo tanto, es preferible probar los dos métodos antes de probar este.

5.3.3 Funciones kernel

Una función kernel es útil para calcular productos escalares en el espacio de las características. Es una función $K : X \cdot X \to R$ tal que $K\left(X, Y\right) = <\Phi\left(x\right), \Phi\left(z\right)>$ donde Φ es una transformación de X en un espacio de Hilbert.
Sin embargo, hay una gran cantidad de posibles funciones núcleo que pueden ser utilizadas para crear tal espacio de características de alta dimensional.
Algunas funciones núcleo inicialmente utilizadas y de propósito general son las siguientes:

1. Polinómica:

$$K\left(x_i, y_j\right) = \left(x_i \cdot y_j + 1\right)^p$$

2. Gaussiana:

$$K\left(x_i, y_j\right) = \exp\left(-\frac{\left\|x_i - y_j\right\|^2}{2\sigma^2}\right)$$

3. Sigmoidal o tangente:

$$K\left(x_i, y_j\right) = \tanh\left(a x_i \cdot y_j + b\right) \quad a, b \in R$$

4. Multicuadrática inversa:

$$K\left(x_i, y_j\right) = \frac{1}{\sqrt{\left\|x_i - y_j\right\|^2 + c^2}} \quad c \geq 0$$

Además de las funciones gaussiana y sigmoidea, Ivanciuc (2007) presenta otras funciones núcleo que, como las anteriores, dependen de algunos parámetros que pueden ser calculados mediante diferentes métodos empíricos o estadísticos. Estas funciones kernel se especifican de la siguiente forma:

1. Anova Kernel

$$K\left(x_i, y_j\right) = \left(\sum_i \exp\left(-\gamma\left(x_i - y_j\right)\right)\right)^d$$

2. Fourier series kernel

$$K\left(x_i, y_j\right) = \frac{sen\left(N + 1/2\right)\left(x_i - y_i\right)}{sen\left(1/2\left(x_i - y_j\right)\right)}$$

3. Spline Kernel

$$K\left(x_i, y_j\right) = \left(\sum_{r=0}^{k} x_i^r y_j^r\right) + \left(\sum_{s=1}^{N} x_i - t_s\right)^k + \left(y_j - t_s\right)^k$$

4. Additive Kernel

$$K\left(x_i, y_j\right) = \sum_n K_n\left(x_i, y_j\right)$$

5. Tensor Product kernel

$$K\left(x_i, y_j\right) = \prod_n K_n\left(x_i, y_j\right)$$

5.4 Árboles de decisión

Los árboles de decisión son algoritmos para clasificar utilizando particiones de un conjunto de datos que maximizan las diferencias de la variable dependiente.

Los árboles de decisión representan de forma gráfica sobre un conjunto de reglas denominada *clustering* jerárquico la cual utiliza un proceso de fragmentación en secuencia e iterativo, partiendo de una variable dependiente que se pretende explicar, a través de combinaciones de variables independientes.

Los árboles de decisión se componen a través de nodos, ramas y hojas. Los nodos son los inputs de entrada, las ramas representan los posibles valores de la variable de entrada y las hojas son los posibles valores de la variable de salida.

El primer elemento de un árbol de decisión lo llamamos nodo raíz que va a representar a la variable de mayor relevancia en el proceso de clasificación.

Los elementos y las herramientas de los algoritmos que determinan la construcción de un árbol son varios:

1) El criterio con el que delimita
2) La regla que declara un nodo terminal.
3) La definición de clase a cada nodo.
4) Partición. Selección del punto de división. La variable utilizada para dividir el conjunto de todos los datos se elige por comparación con todas las demás.
5) Poda. Se eliminan las ramas que añaden poco valor de predicción del árbol.
6) La evaluación del clasificador, es decir, la estimación de la validación del árbol y el cálculo del riesgo.

5.4.1 Criterios de selección: Índice de divergencia de Gini

El índice de Gini es una medida de divergencia de las clases en un nodo del árbol que se utiliza. Este índice se emplea en diferentes algoritmos de árboles de decisión:

$$G(A_i) = \sum_{j=1}^{M_i} p(A_{ij}) G(C \mid A_{ij})$$

Siendo $G(C \mid A_{ij})$ igual a:

$$G(C \mid A_{ij}) = -\sum_{j=1}^{M_i} p(C_k \mid A_{ij}) p(1 - p(C_k \mid A_{ij}))$$

A_{ij} es el atributo empleado para ramificar el árbol, J es el número de clases, M_i es el de valores distintos que tiene el atributo A_i y $p(A_{ij})$ constituye la probabilidad A_i tome su $j-ésimo$ valor y $p(C_k \mid A_{ij})$ representa la probabilidad de que un ejemplo sea de la clase C_k cuando su atributo A_i toma su $j-ésimo$ valor.

El índice de divergencia de Gini toma el valor cero cuando un grupo es completamente homogéneo y el mayor valor lo alcanza cuando todas las $p(A_{ij})$ son constantes, entonces el índice es $(J-1)/J$.

5.5 Criterios de selección: Ganancia de información

Las medidas intentan maximizar la ganancia de información que consigue el atributo para ramificar el árbol de clasificación mediante la siguiente función I:

$$I(A_{ij}) = \sum_{j=1}^{M} p(A_{ij}) H(C \mid A_{ij})$$

La entropía es una medida de la incertidumbre que hay en un sistema, es decir, trata de medir ante una situación determinada la probabilidad de que ocurra cada uno de los posibles resultados. La entropía de clasificación se define como:

$$H(C_k \mid A_{ij}) = -\sum_{k=1}^{j} p(C_k \mid A_{ij}) \log_2 (C_k \mid A_{ij})$$

La ganancia de información que se produce al dividir T en los subconjuntos T_j viene dada por:

$$H(T) - \sum p(T_j) H(T) \text{ donde } H(T) \text{ es la entropía } T.$$

5.4.2 Regresión del árbol de decisión

Los árboles de decisión también pueden emplearse para resolver problemas de regresión. Sin embargo, en este caso, es necesario considerar una forma ligeramente diferente de dividir los nodos. En lugar de considerar una medida de impureza, una de las opciones más

comunes es elegir la característica que minimiza el error cuadrado medio, considerando la predicción media de un nodo. Supongamos que un nodo, *i*, contiene m muestras. La predicción media es la siguiente:

$$\bar{y}_i = \frac{1}{m}\sum_j y_j$$

En este punto, el algoritmo tiene que buscar todas las divisiones binarias para encontrar la que minimiza la función objetivo;

$$MSE_i = \frac{1}{m}\sum_j \left(y_j - \bar{y}_i\right)^2$$

Análogamente a los árboles de clasificación, el procedimiento se repite hasta que MSE está por debajo de un umbral fijo, λ.

Incluso si no es correcto, podemos pensar en un nivel de impureza inaceptable cuando la predicción de un nodo tiene una baja precisión. De hecho, en un árbol de clasificación, un nodo impuro contiene más de una clase y la incertidumbre podría ser demasiado alta para tomar una decisión razonable. De la misma manera, un nodo cuya $MSE \succ \lambda$ es todavía demasiado incierto sobre la salida correcta significa que se requieren más divisiones. Un enfoque alternativo se basa en el error absoluto medio (MAE), pero en la mayoría de los casos, el MSE es la elección óptima.

5.5 Bosques aleatorios

Un bosque aleatorio es un método de combinación de árboles de decisión. Si tenemos clasificadores N_c, el conjunto de datos original se divide en subconjuntos N_c (con reemplazo) llamados muestras *bootstrap:*

$$\left(X_i, Y_i\right) = \left\{\left(\bar{x}_j, y_j\right) j = 1,...,k\left(\bar{x}_j, y_j\right) \subset \left(X, Y\right)\right\}$$

Contrario a un árbol de decisión único, en un bosque al azar la política de división se basa en un nivel medio de aleatoriedad. De hecho, en lugar de buscar la mejor opción utiliza un subconjunto aleatorio de características es usado tratando de encontrar el umbral que mejor separa los datos. Como resultado, habrá muchos árboles que se entrenan de una manera más débil y cada uno de ellos producirá una predicción. Al mismo tiempo, cada árbol estará más especializado en una porción del espacio de muestra, mientras que en otras regiones se obtiene predicciones inexactas.

Hay dos caminos para interpretar estos resultados. La más común se basa en el voto mayoritario (la clase más votada, obtenida a través de un $\arg\max(\cdot)$ binario, se considera correcta:

$$\tilde{y} = \arg\max_j \left(c_j\left(\bar{x}_i\right)\right)$$

El concepto de importancia de las características aplicado a los Bosques aleatorios, se obtiene, calculando el promedio de todos los árboles del bosque, como se muestra a continuación:

$$\text{Importancia}\left(\bar{x}^{(i)}\right) = \frac{1}{N_{árboles}} \sum_i \sum_k \frac{N_k}{N} \Delta I_{\bar{x}^{(i)}}$$

Los rasgos importantes de este modelo son: el bloque de rasgos de importancia media y una cola que contiene rasgos que tienen poca influencia en las predicciones Utilizando árboles de decisión o bosques aleatorios, es posible evaluar la importancia real de todas las características y excluir todos los elementos bajo un umbral fijo. De esta manera, un proceso de decisión complejo puede ser simplificado y, al mismo tiempo, parcialmente denotado.

5.6 Boosting

Boosting es un método de *ensamble* que se puede emplear con grupos de métodos de *statistical learning,* como son los árboles de decisión. El objetivo de este modelo es ajustar, de forma de partición. Cada nuevo modelo emplea información del modelo anterior para aprender de sus errores, mejorando iteración a iteración. Esto se consigue utilizando árboles con una o pocas ramificaciones. A diferencia del método de *bagging,* el *boosting* no hace uso de muestreo repetido (*bootstrapping),* por lo que cada árbol construido depende en gran de los árboles previos. Tres de los algoritmos de *boosting* más empleados son *AdaBoost, Gradient Boosting* y *Stochastic Gradient Boosting.*

5.6.1 AdaBoost

La estructura básica detrás de esto puede haber un árbol de decisión, pero el conjunto de datos utilizado para el entrenamiento es continuamente adaptado para forzar al modelo a centrarse en aquellas muestras que están mal clasificadas. Además, los clasificadores se añaden de forma secuencial, de modo que uno nuevo potencia el anterior mejorando el rendimiento en aquellas áreas en las que no era tan preciso como se esperaba. En cada iteración se aplica un factor de peso a cada muestra para aumentar la importancia de las muestras que se predicen erróneamente y disminuir la importancia de las demás. En otras palabras, el modelo es repetidamente, comenzando como un aprendiz muy débil hasta que se alcanza el máximo número de n-estimadores. Las predicciones en este caso, son obtenidas por mayoría de votos.

Este modelo está basado en un conjunto de peso actualizado dinámicamente (considerando n muestras):

$$W^{(t)} = \left\{\omega_1^{(t)}, \omega_2^{(t)}, ..., \omega_n^{(t)}\right\} \text{ donde } \omega_i^{(t)} \geq 0$$

El valor inicial $W^{(0)}$, se establece igual a $\frac{1}{n}$ para todos los pesos, de modo que no se expresa ninguna muestra. Antes de comenzar el proceso de entreamiento, se establece cada muestra del clasificador, considerando los pesos existentes. Después de cada paso del entrenamiento, se calcula una función indicadora

$$\varepsilon^{(t)} = \frac{\sum_{c_t(\bar{x}_i) \neq y_i} \omega_i}{\sum_i \omega_i} \text{ donde } v \sim \varepsilon^{(t)}$$

Si no se han producido clasificaciones erróneas, $\varepsilon^{(t)} = 1$ mientras esto es igual a cero, si todas las muestras hay sido asignadas a la clase equivocada. Para volver a ponderar el conjunto de datos para la iteración subsiguiente, se emplea una función especial:

$$\alpha^{(t)} = \log\left(\frac{1-\varepsilon^{(t)}}{\varepsilon^{(t)}+v}\right) \text{ donde } v \sim \varepsilon^{(t)}$$

La constante *v* se añade normalmente para mejorar la estabilidad numérica; sin embargo, supongamos por simplicidad que v es nula.

Cuando $\varepsilon^{(t)} \to 0, \alpha^{(t)} \to +\infty, \varepsilon^{(t)} \to 1, \alpha^{(t)} \to -\infty$. En este caso de una suposición aleatoria binaria, $\varepsilon^{(t)} = \frac{1}{2}$ y $\alpha^{(t)} = 0$. Incluso si no es intuitivo, el principal objetivo es excluir los clasificadores que son oráculos aleatorios ($\varepsilon^{(t)} = \frac{1}{2}$), que representan la peor situación, y su predicción está completamente excluida. Cuando $\varepsilon^{(t)} \to 1$, el clasificador debería recibir un impulso negativo, pero esto estaría en contraste con el propósito inicial del algoritmo.

Por lo tanto, en este caso, la salida se invierte (también si se trata de una clasificación errónea) y el impulso se convierte en positivo (por ejemplo, si $\varepsilon^{(t)} = 0.75$, se transforma en $\varepsilon^{(t)} = 0.25$), lo que no cambia el valor absoluto de $\alpha^{(t)}$.

Esto puede lograrse utilizando una función de decisión global ponderada:

$$d\left(\overline{x}_i\right) = sign\left(\sum_{j=1}^{N_c} \alpha^{(j)} c_j\left(\overline{x}_i\right)\right)$$

De esta manera, los clasificadores que permanecen atascados con una precisión cercana a $\frac{1}{2}$ son automáticamente descarados, mientras que todos los demás son potenciados. El procedimiento de refuerzo es muy simple y se basa en el resultado de cada clasificación.

Consideremos la muestra x_i, introduciendo una variable auxiliar o_i:

$$o_i \begin{cases} 1 \text{ si } c^{(t)}\left(\overline{x}_i\right) \neq y_i \\ -1 \text{ si } c^{(t)}\left(\overline{x}_i\right) = y_i \end{cases}$$

El peso correspondiente, ω_i, se actualiza considerando la siguiente regla:

$$\omega_i^{(t+1)} = \omega_i^{(t)} e^{\alpha_i^{(t)} o_i}$$

Es sencillo comprender que un peso sólo aumenta cuando se ha producido una clasificación errónea, y disminuye si la muestra se ha clasificado correctamente, Además el parámetro $\alpha(t)$ tiene en cuenta el comportamiento general del estimador. Si $\varepsilon^{(t)} \to \frac{1}{2}$, la reponderación no altera la configuración existente, mientras que se vuelve más agresivo cuando se han obtenido valores más grandes de $\alpha(t)$. El resultado de este proceso es una nueva distribución, en la que las muestras más problemáticas tendrán cada vez más

probabilidades de ser muestreadas, mientras que las más sencillas podrían ser descartadas después de las primeras iteraciones.

5.6.2 Gradient Boost

Gradient Boost es una técnica que permite construir un conjunto de árboles paso a paso (el método también se conoce como modelado aditivo en etapas avanzadas), con el objetivo de minimizar la función de pérdida de objetivos. El resultado genérico del conjunto puede representarse de la siguiente manera:

$$y_E = \sum_j \alpha_j c_j\left(\bar{x}\right) = \sum_j \alpha_j f\left(\bar{x};\bar{\theta}_i\right)$$

En este caso, $c_j\left(x\right)$ es una función que representa a un aprendiz débil (en este caso particular, siempre es un árbol de decisión que puede ser modelado como una única función parametrizada $f\left(\cdot\right)$, donde el vector θ_i agrupa todas las tuplas divisorias del árbol *i*-ésimo. El algoritmo se basa en el concepto de añadir un nuevo árbol de decisión en cada paso para minimizar una función de coste global (basado en una pérdida predefinida $L\left(\cdot\right)$ utilizando el método de descenso de gradiente más profundo. Teniendo en cuenta los clasificadores están parametrizados, utilizando un vector θ, y todos los coeficientes de ponderación se agrupan en un solo conjunto, la función de coste global se define como:

$$C\left(X,Y;\bar{\alpha},\bar{\theta}\right) = \sum_i L\left(y_i,\alpha_i c\left(\bar{x}_i;\bar{\theta}\right)\right)$$

Por lo tanto, el objetivo es encontrar la tupla óptima para que:

$$\left(\bar{\alpha}^*,\bar{\theta}^*\right) = \arg\min_{\bar{\alpha},\bar{\theta}} C\left(X,Y;\bar{\alpha},\bar{\theta}\right)$$

Considerando este objetivo, el procedimiento incremental puede reescribirse:

$$c_i\left(\bar{x}\right) = c_{i-1}\left(\bar{x}\right) + \arg\min_f \sum_j L\left(y_j, c_{i-1}\left(\bar{x}_j\right) + f\left(\bar{x}_j;\bar{\theta}_i\right)\right)$$

En la expresión anterior, la pérdida se calcula considerando las contribuciones anteriores y se optimiza con respecto al nuevo clasificador. Desafortunadamente, aunque formalmente es claro, este problema es extremadamente complejo y requiere un costo computacional inaceptable. Sin embargo, introduciendo un gradiente, podemos reescribir la expresión anterior para transformar el modelo aditivo en un procedimiento de optimización:

$$c_i\left(\bar{x}\right) = c_{i-1}\left(\bar{x}\right) - \eta\alpha_i \sum_j \nabla_c L\left(y_j, c_{i-1}\left(\bar{x}_j\right)\right)$$

Dado que el objetivo es minimizar la función de costo global. Entonces, al moverse en la dirección opuesta al gradiente, el nuevo clasificador se construye con el propósito de reducir la función de costo global con respecto a sus predecesores. Notemos que η es la tasa de aprendizaje, que debe ser elegida mediante una búsqueda en la cuadrícula para evitar una convergencia muy lenta o la inestabilidad con el consiguiente sub-óptimo. Las ponderaciones, α_i se calculan utilizando un algoritmo de búsqueda de líneas, después del cálculo del gradiente:

$$\alpha_i = \arg\min_\alpha \sum_j L\left(y_j, c_i\left(\bar{x}_j,\alpha\right)\right) = \arg\min_\alpha \sum_j L\left(y_j, c_{i-1}\left(\bar{x}\right) - \eta\alpha \sum_j \nabla_c L\left(y_i, c_{i-1}\left(\bar{x}_j\right)\right)\right)$$

En este modelo, se elige un estimador para mejorar el predecesor (en un escenario perfecto, la función de costos debe disminuir hasta el mínimo), pero no

5.7 K-Nearest Neigbour (KNN)

El modelo *K-Nearest Neigbor (*Vecinos más cercanos) es uno de los algoritmo de *machine learning* más simples. Básicamente es un modelo fundamentado en la idea de identificar observaciones en el conjunto de entrenamiento que se asemejen a la observación de *test* (observaciones vecinas) y asignarle como valor predicho la clase predominante entre dichas observaciones. Además este algoritmo se basa completamente en los datos y su estructura subyacente, para ello consideremos un conjunto de datos:

$X = \{\overline{x}_1, \overline{x}_2, ..., \overline{x}_n\}$ donde $\overline{x}_i \in R^m$

Para obtener la similitud, introduzca una función de distancia. La elección más común es la métrica de Minkowski, que se define como:

$$d_p\left(\overline{x}_1, \overline{x}_2\right) = \left(\sum_j \left|\overline{x}_1^{(j)} - \overline{x}_2^{(j)}\right|^p\right)^{\frac{1}{p}}$$

$p = 1, d_1(\cdot)$ es la distancia de Manhattan, mientras que $p = 2, d_{2(\bullet)}$ es la distancia euclidiana. Para valores de p más grandes conducen a medidas más cortas y, para $p \to \infty, d_p(\cdot)$, converge a la mayor diferencia absoluta del componente, $\left|x_1^{(k)} - x_2^{(k)}\right|$ (suponiendo que k es el índice correspondiente a la mayor diferencia). En muchas aplicaciones, la distancia euclidiana es la elección óptima; sin embargo, el valor asignado p puede afectar a la semántica de la propia métrica. De hecho, cuando $p = 1$, todos los componentes se tienen en cuenta en la misma dirección. Por otro lado, el aumento de p reducirá proporcionalmente el impacto de todas las pequeñas diferencias de los componentes, obligando a la medida a representar sólo la más relevante. Encontrar el valor más apropiado para p requiere un análisis previo del conjunto de datos y un conocimiento completo del dominio (por ejemplo, en un dominio concreto, dos muestras que tienen una gran diferencia en un solo componente deben considerarse disímiles, mientras que en otra, la distancia euclidiana es la forma más precisa de medir su similitud).

Una vez elegida la función de distancia, es fácil definir los vecinos. En general, se emplean dos enfoques. El primero se basa en el número de vecinos más cercanos, por lo tanto, dada una muestra x_i , la vecindad se define como:

$$N_k\left(\overline{x}_i\right) = \arg\min_j^k d_p\left(\overline{x}_i, \overline{x}_j\right)$$

En la formula anterior, la función $\arg\min_j^k(\cdot)$ selecciona los k último índices j correspondientes a las menores distancias del centro x_i . Sin embargo, en algunos casos es útil obtener el conjunto de todos los vecinos cuya distancia es menor que un radio prefijado, R:

$$N_R\left(\overline{x}_i\right) = \left\{\overline{x}_j : d_p\left(\overline{x}_i, \overline{x}_j\right) \le R\right\}$$

Ambos enfoques son perfectamente compatibles; sin embargo, el segundo sólo puede emplearse cuando el científico de los datos tiene un conocimiento completo de las distancias. De hecho, para conjuntos de datos de altas dimensiones, a menudo es más simple establecer el número máximo de vecinos que encontrar el radio correcto. Sin embargo, este es un problema de contexto y la elección correcta no puede ser fácilmente generalizada.

5.8 Redes Neuronales

Las redes neuronales cuentan con la habilidad para procesar bases de datos con ruido o incompletas y su tolerancia a fallos permite que las redes operen en tiempo real por su operatividad en paralelo. Las redes neuronales pueden describirse mediante cuatro conceptos: el tipo de modelo de la red neuronal; las unidades de procesamiento que recogen información, la procesan y arrojan un valor; la organización del sistema de nodos para transmitir las señales desde los nodos de entrada a los nodos de salida y la función de aprendizaje a través de la cual el sistema se retroalimenta.
Se considera una red neuronal la ordenación secuencial de tres tipos básicos de nodos o capas: nodos de entrada, nodos de salida y nodos intermedios (capa oculta o escondida). El nodo es la unidad de procesamiento que actúa en paralelo con otros nodos de la red.
La primera tarea del nodo es procesar los datos de entrada creando un valor resumen que es la suma de todas las entradas multiplicadas por sus ponderaciones. Este valor se procesa a continuación mediante una función de activación para generar una salida que se envía al siguiente nodo del sistema.
Las cuatro funciones de activación más utilizadas son:

a) Función escalón.

   Considere que la activación debe llegar a un determinado nivel U (umbral de activación), esta función adopta la forma:

b) $x_j = f_j\left(neta_j\right) = \begin{cases} 1 & \forall neta_j \geq U \\ 0 & \forall neta_j < U \end{cases}$

   Para evitar la discontinuidad se utilizan frecuentemente las siguientes tres funciones.

c) Función identidad.

$$y\ \left(neta_j\right) = neta_j$$

d) Función sigmoidea o $x_j = neta_j$ logística.

$$x_j = \frac{1}{1+e^{-aneta_j}} \text{ si } x_j \in [0,1] \text{ es la función sigmoidea asimétrica.}$$

e) Tangente hiperbólica.

$$x_j = 1 - \frac{1}{e^{2neta_j+1}} \text{ tal que } x_j \in [-1,1]$$

Los nodos de salida reciben entradas y calculan el valor de salida (no van a otro nodo). En casi todas las redes existe una tercera capa denominada oculta. Este conjunto de nodos utilizados por la red neuronal, junto con la función de activación posibilita a las redes

neuronales representar fácilmente las relaciones no lineales, que son muy problemáticas para las técnicas multivariantes.

La red perceptrón multicapa fue desarrollada en 1957 por Frank Rosenblatt, popularizada en 1986 por Rumelhart. Es la red más utilizada en las aplicaciones prácticas debido a que es el aproximado más preciso de funciones. La verdadera razón de su tremenda utilidad radica en su capacidad de organizar una representación interna del conocimiento en las capas ocultas de neuronas a fin de aprender la relación entre un conjunto de datos entradas y salidas.

I. Etapa de *test*.

El desarrollo de la red consta de una fase de entrenamiento y otra de *test*. En la primera etapa cuando se presenta un patrón de entrada $X_p : x_{p1},...,x_{pi},...x_{pn}$ se transmite a la red a través de los pesos w_{ji} desde la capa de entrada a la capa oculta. Las neuronas de esta capa transforman las señales a través de la función de activación proporcionando un valor de salida. Este valor se transmite a su vez a través de los pesos v_{kj} a la capa de salida donde aplicando la función de activación obtenemos un valor de salida.
Suponga que la entrada total o neta de una neurona oculta j la expresamos cómo $neta_{pi}$, entonces matemáticamente la podemos expresar de la siguiente manera:

$$net_{pj} = \sum_{j=1}^{N} w_{ji} x_{pi} + \theta_j$$

θ es el umbral de la neurona que se considera como un peso asociado a una neurona ficticia con valor de salida igual a 1. El valor de salida de la neurona oculta j, b_{pj} lo tenemos aplicando la función de activación $f(\cdot)$ sobre la entrada net.

$$b_{pj} = f\left(net_{pj}\right)$$

La entrada net que recibe una neurona de salida k se puede expresar cómo:

$$net_{pk} = \sum_{j=1}^{L} v_{kj} b_{pj} + \theta_k$$

El valor de salida de la neurona k, y_{pk} es el siguiente:

$$y_{pk} = f\left(net_{pk}\right)$$

II. Etapa de entreno

El objetivo en esta está es minimizar la discrepancia o error entre la salida de la red y el valor real que presenta el usuario. La función a optimizar es:

$$E_p = \frac{1}{2}\sum_{k=1}^{M}\left(d_{pk} - y_{pk}\right)^2$$

donde d_{pk} es la salida presentada por la red de la neurona k ante la presentación del patrón p. La medida general del error es la suma de todos los errores para todos los patrones

$$E = \sum_{p=1}^{p} E_p$$

Es decir, en este análisis el objetivo es minimizar el cuadrado de los errores entre el valor real y el de la variable salida. Esta función depende de dos parámetros; el conjunto de variables de entrada, las variables explicativas del modelo, y el conjunto de pesos sinápticos.

Disponemos de un buen número de métodos o algoritmos que permiten estimar los parámetros del modelo de RNA minimizando E_p . En los métodos no lineales la función E_p es una función continua y diferenciable así que podemos aplicar los métodos del gradiente. En estos métodos el proceso de búsqueda de la solución óptima puede ser descrito como:

$$\vartheta_{t+1} = \vartheta_t + \lambda_t \Delta_t$$

Donde ϑ_t es la solución que es viable, λ_t es el tamaño del peso y $\Delta_t = M_t g_t$ es el vector de dirección. Para el vector de dirección se tiene que M_t es una matriz definida positiva y $g_t = g(\vartheta_t)$ es el vector gradiente.

En RNA la técnica más utilizada es la del gradiente decreciente, Rumelhart, (1986), denominada algoritmo *backpropagation* debido a la forma de modificación de los pesos. Este gradiente toma la dirección que determina el incremento más rápido en el error. Así que el error puede reducirse ajustando cada peso en la siguiente dirección:

$$-\sum_{p=1}^{p} \frac{\partial E_p}{\partial w_{ji}}$$

Entre los métodos gradientes se encuentra el método de Newton-Raphson. Este algoritmo utiliza la inversa de la matriz hessiana como matriz M. En este procedimiento puede ocurrir que el valor del punto inicial ϑ_0 no esté próximo al punto óptimo lo que dificultaría el cálculo de la matriz hessiana, además de que puede suceder que en algunas aplicaciones esta matriz sea difícil de calcular, por lo que diferentes investigadores han desarrollado procedimientos de estimación que se conocen como métodos Quasi-Newton. Estos métodos usan una aproximación iterativa de la matriz hessiana:

$$M_{t+1} = M_t + N_t$$

Donde N_t es una matriz definida positiva, lo que garantiza que en cada paso del proceso iterativo la aproximación sea también definida positiva, al ser de dos matrices positivas. En esta implementación tenemos que seleccionar el punto inicial ϑ_0 y una matriz M_0 que deberá ser definida positiva.

Otro parámetro que es necesario controlar adecuadamente es la tasa de disminución de aprendizaje, puesto que comúnmente la tasa de aprendizaje va decayendo a medida que se realiza el entrenamiento. Esto tiene el efecto importante de aplanar la superficie de búsqueda y de esta manera permitir encontrar los mínimos globales de una forma más fácil.

En la práctica la forma de modificar los pesos de forma iterativa se realiza aplicando la regla de la cadena a la expresión del gradiente y añadir una tasa de aprendizaje que se denomina η . Para una neurona de salida obtenemos la siguiente expresión:

$$\Delta v_{kj}\left(n+1\right)=\eta\sum_{p=1}^{p}\delta_{pk}b_{pj}$$

Para el peso de una neurona oculta:

$$\Delta w_{ji}\left(n+1\right)=\eta\sum_{p=1}^{p}\delta_{pj}x_{pi}$$

donde $\delta_{p,j}$ es igual a:

$$\delta_{pj}=f^{'}\left(net_{pj}\right)\sum_{k=1}^{M}\delta_{pk}v_{kj}$$

Para acelerar el proceso de convergencia de los pesos se sugiere añadir a la expresión un factor momento, α el cual tiene en cuenta la dirección del incremento tomado en la iteración anterior. Por ejemplo, para el peso de una neurona de salida, la expresión es la siguiente:

$$\Delta w_{ji}\left(n+1\right)=\alpha\left(n\right)\left(x_{pi}-w_{ji}\left(n\right)\right)$$

Y para el peso de una neurona oculta:

$$K\left(x_i,y_j\right)=\left(\sum_{r=0}^{k}x_i^r y_j^r\right)+\left(\sum_{s=1}^{N}x_i-t_s\right)^k \qquad \Delta v_{kj}\left(n+1\right)=\eta\left(\sum_{p=1}^{p}\delta_{pk}b_{pj}\right)+\alpha\Delta v_{kj}\left(n\right)$$

5.9 Selección del modelo, *Cross-Validation*

El término *Cross-Validation* muestra distintas estrategias para estimar el *test error rate*. En este proceso se excluye una serie de observaciones del *training data set* disponible, se ajusta el modelo y finalmente se evalúa con los datos excluidos.
Validación simple.
El método más sencillo consiste en dividir aleatoriamente las observaciones disponibles en dos grupos, uno se emplea para entrenar al modelo y otro para evaluarlo. Sin embargo, se enfrenta a dos problemas importantes:

a. La estimación del *test error rate* es altamente variable dependiendo de qué observaciones se incluyan como set de entrenamiento y cuáles como set de validación (problema de varianza).

b. Al excluir parte de las observaciones disponibles como datos de entrenamiento (generalmente la mitad), se dispone de menos información con la que crear el modelo y por lo tanto se reduce su capacidad. Esto suele tener como consecuencia una sobrestimación del *test error* comparado al que se obtendría si se emplearan todas las observaciones para el entrenamiento (problema de *bias*).

5.9.1 Leave One Out Cross-Validation

El método LOOCV en un método iterativo que se inicia empleando como *training* data set todas las observaciones disponibles excepto una, que se excluye para emplearla como *test*. Si se emplea una única observación para calcular el *test* error, este varía mucho

dependiendo de qué observación se haya seleccionado. Para evitarlo, el proceso se repite tantas veces como observaciones disponibles, excluyendo en cada iteración una observación distinta, ajustando el modelo con el resto y calculando el error con dicha observación. Finalmente, el *test* error rate estimado por el LOOCV es el promedio de todos los i errores calculados.
En el caso de variables continuas en las que el error se mide como MSE:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(MSE_i\right)$$

En el caso de variables cualitativas en las que el error se mide por número de errores de clasificación:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(Err_i\right) \text{ tal que } Err_1 = I\left(y_1 \neq \hat{y}_i\right).$$

El método LOOCV permite reducir la variabilidad que se origina si se divide aleatoriamente las observaciones únicamente en dos grupos, *training* y *test*. Esto es así porque al final del proceso de LOOCV se acaban empleando todos los datos disponibles tanto como entrenamiento como validación. Al no haber una separación aleatoria de los datos, los resultados de LOOCV son totalmente reproducibles.
La principal desventaja de este método es su coste computacional. El proceso requiere que el modelo sea reajustado y validado tantas veces como observaciones disponibles (n) lo que en algunos casos puede ser muy complicado. Excepcionalmente, en la regresión por mínimos cuadrados y regresión polinómica, por sus características matemáticas, solo es necesario un ajuste, lo que agiliza mucho el proceso.

5.9.2 K-Fold Cross Validation

El método *K-Fold Cross-Validation* es también un proceso iterativo. Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño $k-1$ grupos se emplean para entrenar el modelo y uno de los grupos se emplea como test, este proceso se repite k veces utilizando un grupo distinto como *test* en cada iteración. El proceso genera k estimaciones del *test error* cuyo promedio se emplea como estimación final.

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\left(MSE_i\right)$$

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\left(Err_i\right)$$

6. Desarrollo de los Modelos

Para la creación de nuestro Scoring Reactivo utilizaremos fundamentalmente bases de datos de créditos simples que por confidencialidad cambiaremos algunas variables, fechas, etcétera. Llamaremos a la fuente "Institución de Crédito" para referirnos a la institución donde se generaron los datos.

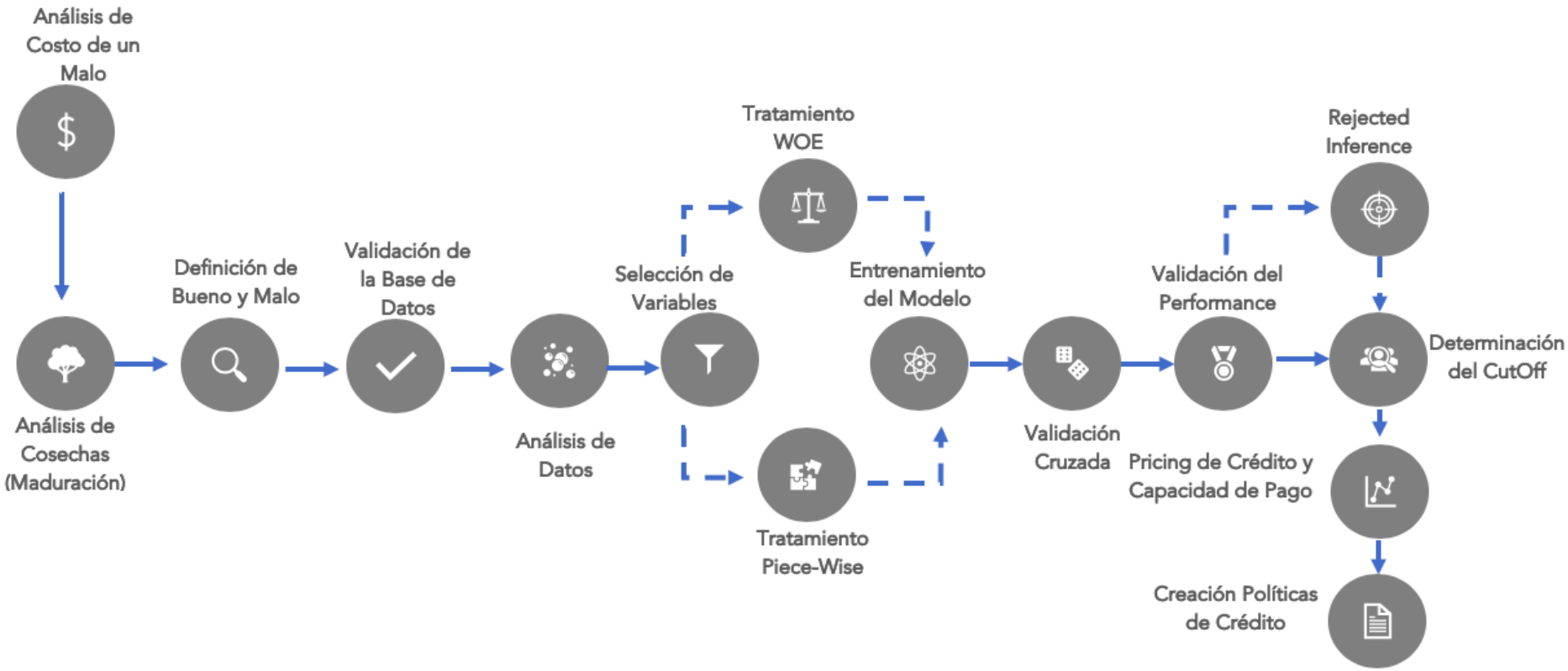
Esta información contiene población total de créditos otorgados desde Enero 2014 hasta Enero de 2019 cuya población es específicamente el segmento de Microcrédito Individual, dentro de esta población, se pueden distinguir en dos segmentos muy peculiares: Clientes Hit y Clientes NO HIT.

El primer grupo concentra una población de clientes que han tenido experiencia crediticia previa en alguna otra institución de crédito y que actualmente mantiene el historial de crédito activo con cuentas que son válidas para generar score genérico[36]. El segundo grupo tiene como principal característica que son clientes que no han tenido ninguna experiencia de crédito previa (en su mayoría), sin embargo también contiene aquella población que no cuentan con el historial crediticio robusto para generar score genérico.

Esta información se conformó a través de dos bases de datos, una, que contiene los días de atraso a cierre de mes por crédito, y la segunda, contiene las características, atributos de los acreditados, algunas variables asociadas con el historial de crédito y condiciones generales de los contratos con los cuáles fueron concedidos los créditos. Todo el análisis se realizó a través del software R que es un *open source*, el cuál es muy flexible para hacer cada uno de los cálculos y procesos necesarios para el desarrollo de modelos.

[36] Por score genérico nos referimos al BC Score de Buró de Crédito.

Grosso modo explicaremos mediante el siguiente esquema la metodología del proceso de scoring reactivo de crédito, el cual nos ayudará a esclarecer cada uno de los pasos a seguir para el desarrollo del modelo.



6.1 Costo de un Malo

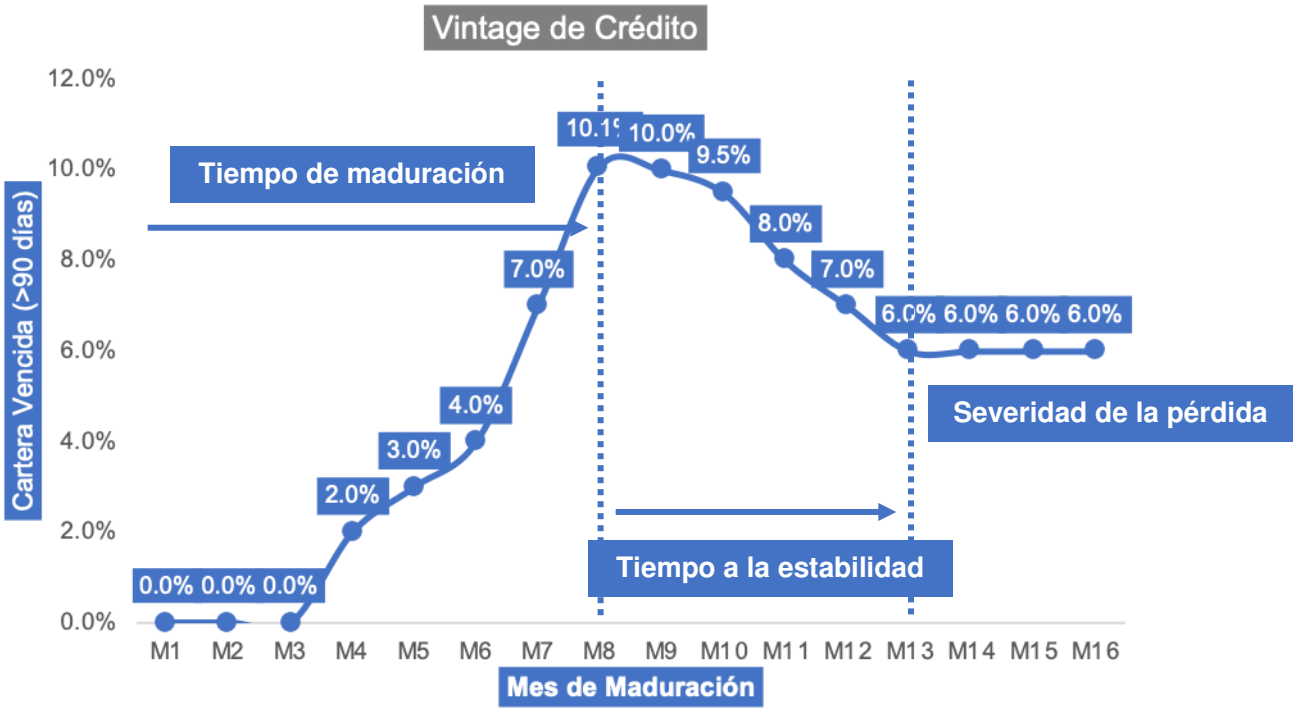
En el análisis de riesgo, es fundamental saber qué costo tiene un crédito "Malo", ya que de esto depende la sustentabilidad de cualquier entidad financiera, en primera instancia debemos definir como "Malo" *a aquél crédito que este en un estado no deseado para la institución que pueda provocar problemas en el futuro* y esto es muy variable, puesto que cada institución tiene apetitos diferentes de riesgo, sin embargo siempre es bueno saber hasta qué punto es costeable tener un crédito "Malo". Consideremos un ejemplo donde existen cinco créditos, de los cuáles, se darán en las mismas condiciones y plazos y de los cuáles uno será malo en su totalidad, es decir, que no paga ni una parte exigible de todo el crédito:

Crédito	Monto Desembolsado (A)	Interés Exigible (B)	Tipo de cliente	Capital Recuperado (C)	Utilidad por crédito (D)	Utilidad Acumulada Acum(C+D-A)
Crédito 1	$10,000.00	$2,000.00	Bueno	$10,000.00	$2,000.00	$2,000
Crédito 2	$10,000.00	$2,000.00	Bueno	$10,000.00	$2,000.00	$4,000
Crédito 3	$10,000.00	$2,000.00	Bueno	$10,000.00	$2,000.00	$6,000
Crédito 4	$10,000.00	$2,000.00	Bueno	$10,000.00	$2,000.00	$8,000
Crédito 5	$10,000.00	$2,000.00	Malo			-$2,000

Por lo que, pese a que tuvimos un buen acierto en otorgar cuatro créditos "Buenos", el quinto visto como "Malo", al no poder recuperar ni una parte del capital, nos pone en una situación de pérdida para la Institución Financiera, es por ello, que debemos saber elegir a los acreditados, o diseñar una estrategia muy eficiente para hacer sostenible la Institución Financiera pese a que tenga un nivel muy alto de "malos", que es el objetivo fundamental de este trabajo. En esta sección no profundizaremos con análisis minuciosos en este aspecto en particular, ya que esto formará parte del *Pricing* y Capacidad de Pago de un crédito.

6.2 Análisis de Cosechas

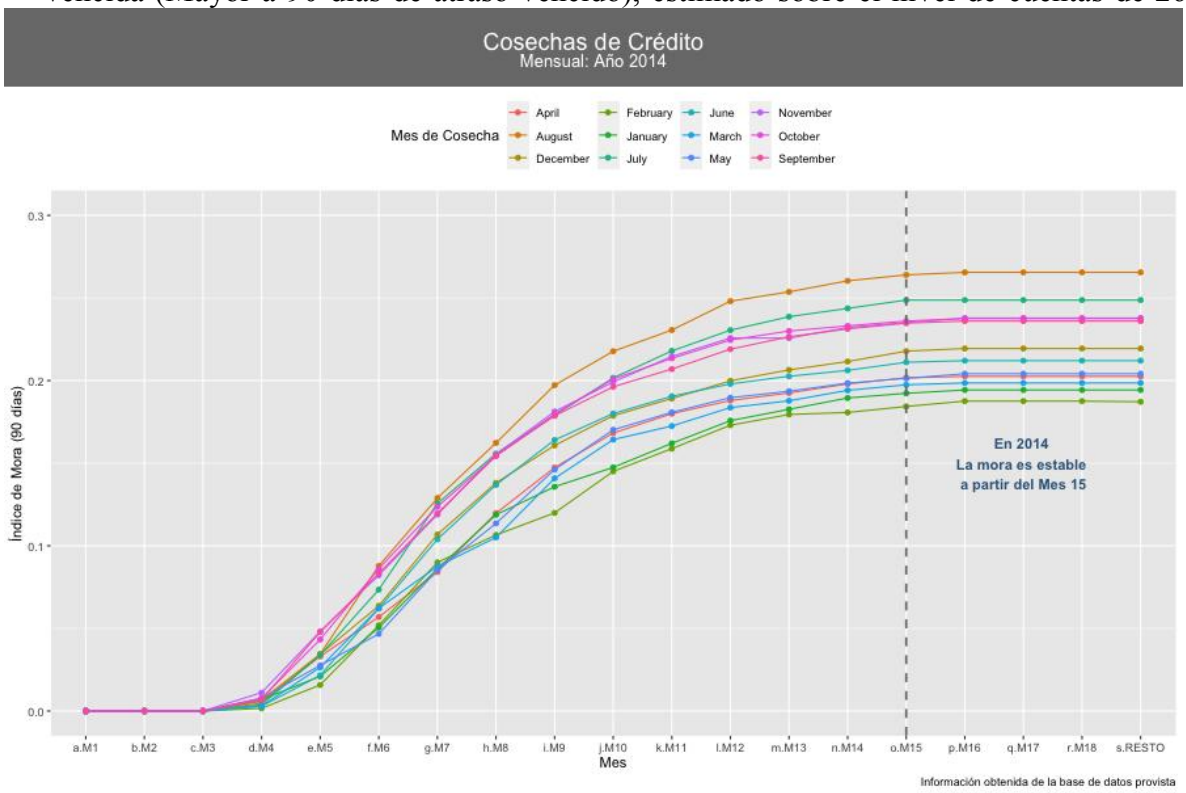
En el Riesgo de Crédito, así como para el desarrollo de modelos de *scoring*, es fundamental analizar el momento en el cuál una serie de créditos desembolsados en un período en específico (cohortes) se tornan como "maduros", es decir que llegan a un punto específico de "cartera vencida". Estos análisis funcionan primordialmente para determinar que ciertas características o atributos de un prospecto tienen cierto efecto a lo largo del tiempo, sin embargo se van desarrollando en diferentes etapas, cuando este se torna como un máximo (*bad rate*, cartera vencida o la métrica que estemos utilizando), se dice la cosecha está "madura", es decir, se ha llegado a su máximo y no puede empeorar un estado de mora, por lo que se puede hacer uso de los atributos específicos de las observaciones y determinar que un crédito ha sido "Bueno" o "Malo". Es por ello que previo a hacer los análisis de maduración, mostraremos la anatomía de la *Vintage*.



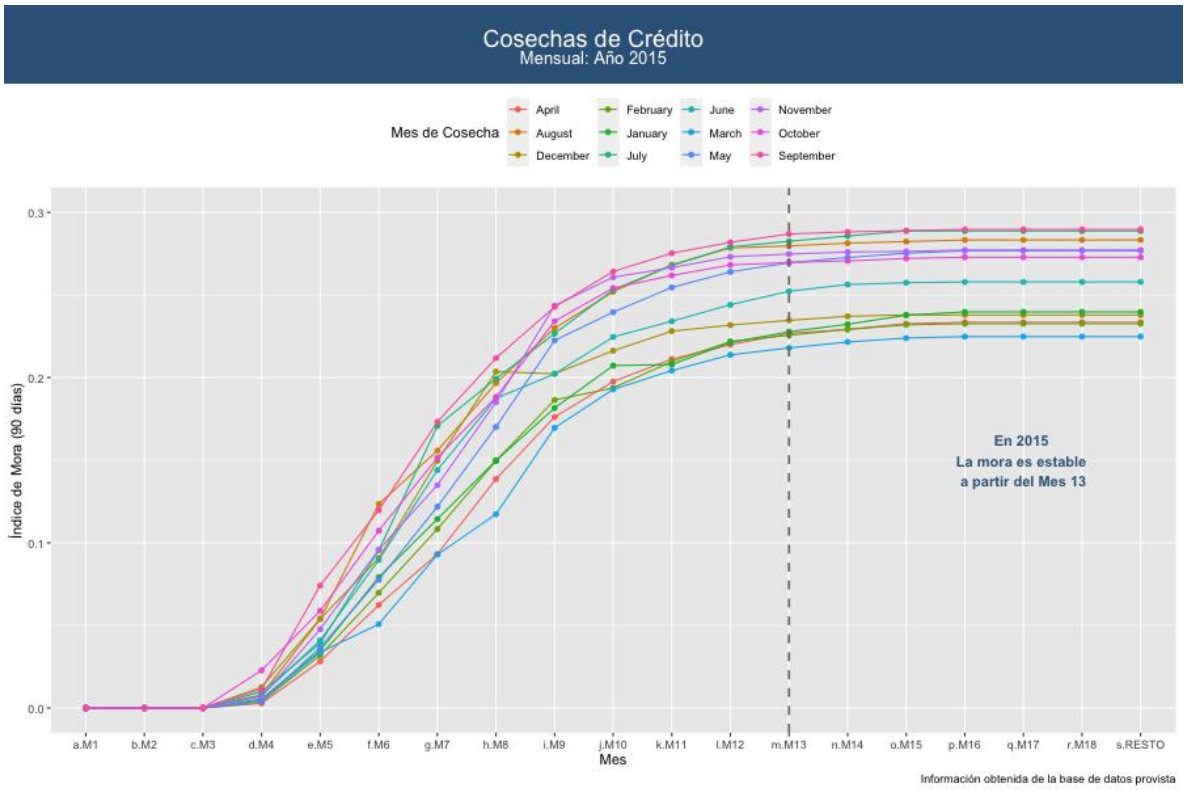
De primera instancia podemos mencionar que en este ejemplo. cosecha está midiendo la mora de cierto cohorte a más de 90 días de vencimiento, por lo que los 3 primeros meses, es imposible que un crédito haya caído dentro de ese rubro, es la razón, por lo que la cosecha está en 0%, sin embargo, a partir del mes 4 se dispara al 2%, continua el crecimiento hasta el mes 8 cuando se torna como el máximo de cartera vencida, bajo este supuesto podemos mencionar que el cohorte o la cosecha está madura al mes 8, porque más allá de ese mes la mora no crece, por lo que no hay riesgo en utilizar atributos de esa población para hacer análisis, por otra parte, el tiempo que necesita la cosecha una vez madura, para llegar a su punto final, es de 5 meses, puesto que comienza a bajar su nivel de mora hasta que se hace estable, por último, ya cuando no hay posibilidades de bajar el nivel de mora, podemos considerarlo como la severidad de la pérdida, puesto que esa será la intensidad final, respecto al mes 8 de mora por la pérdida irrecuperable del crédito.

En el caso de nuestros datos, la Institución Financiera nos otorgó registros de todos los créditos desde Enero 2014 hasta Enero 2019, por lo que estudiaremos las cosechas por año, sin embargo no todas cuentan con más de 18 meses de historia, es por eso que sólo llegamos a analizar hasta mayo 2017. Estas cosechas toman como referencia el número de cuentas en atraso, más no el saldo insoluto.

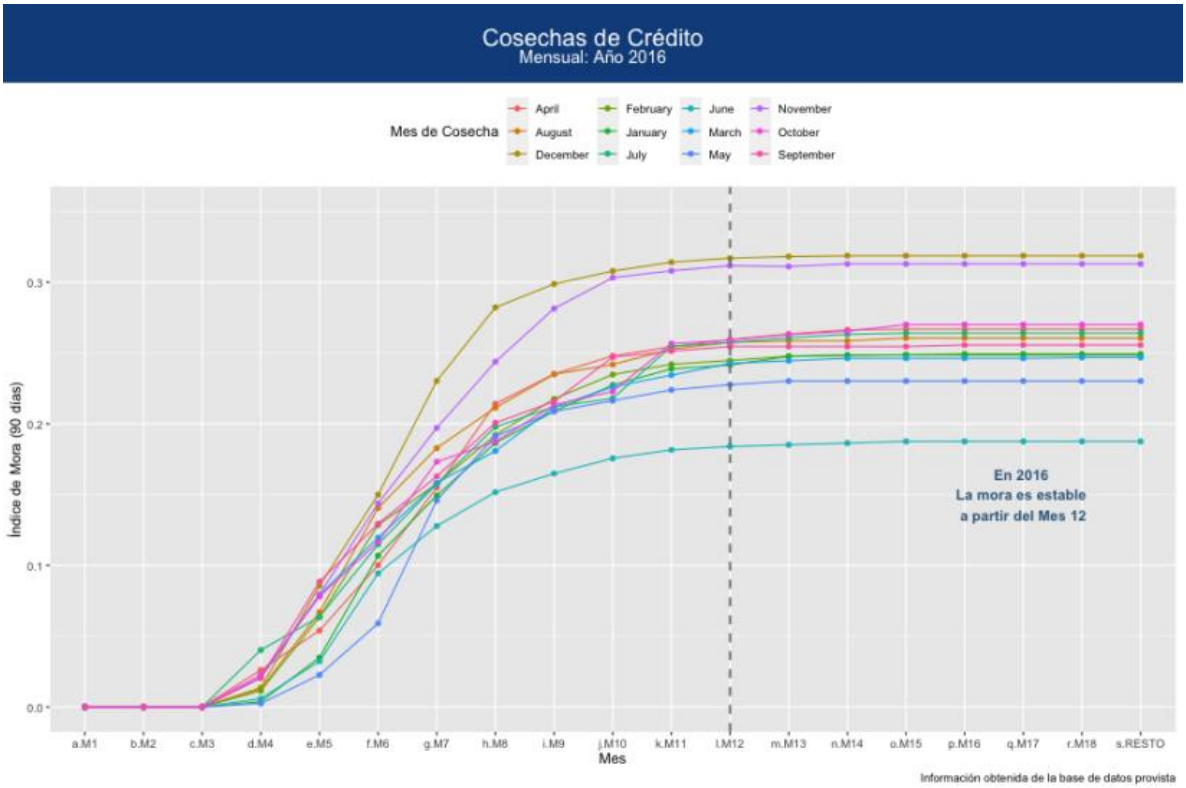
En la *vintage* correspondiente al año 2014 podemos observar que las cosechas se tornan como maduras a partir del Mes 15, llegando como máximo a tener un porcentaje de cartera vencida (Mayor a 90 días de atraso vencido), estimado sobre el nivel de cuentas de 26%,



cuya cosecha o cohorte de ese máximo corresponde al mes de Abril.

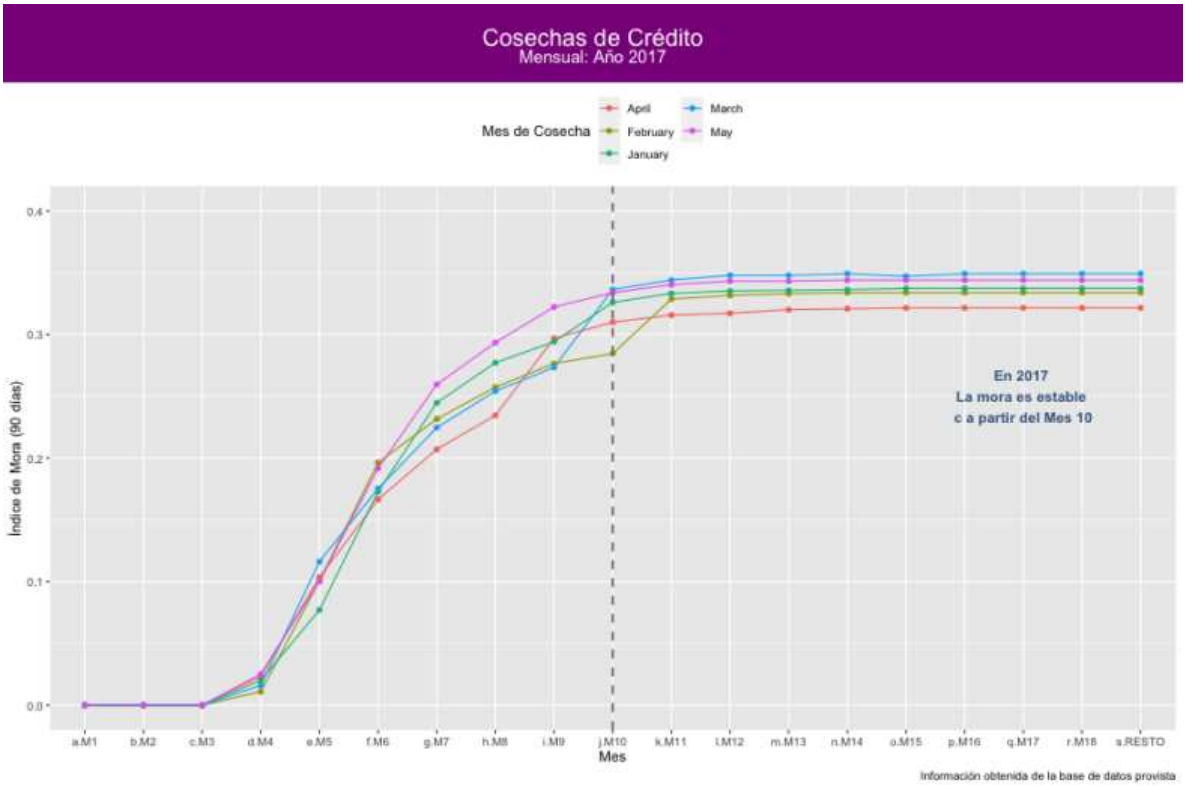


Para el año 2015, observamos que la cosecha se vuelve madura a partir del mes 13. A la par observamos, que incrementa sustancialmente la cartera vencida en alrededor de 3 puntos porcentuales (cerca de un 28%, tomando como referencia al mes de Julio, como el máximo de índice de mora por cuentas dentro de las cosechas colocadas en 2015) esto puede ser explicado básicamente por un mayor apetito de riesgo por parte de la Institución Financiera o por un deterioro en las condiciones del mercado.



En 2016, tenemos un panorama con dos meses con un rango de desviación más grande que en las dos cosechas vistas anteriormente, podemos observar que la maduración se da a partir del Mes 12 y que la mora máxima supera el 30% de cuentas en cartera vencida.

Por último, en 2017 podemos observar, que hubo un incremento de mora por cuenta de casi 3 puntos porcentuales si tomamos como referencia el máximo cohorte (cosecha) que es el mes de Marzo, cuyo índice de mora por cuenta llega a ser del 35%, sin embargo el nivel de maduración de los créditos, se va reduciendo al mes 10, una vez colocado el crédito, esto representa que existió un cambio muy drástico en el apetito de riesgo de la institución financiera, volviéndose menos conservadora. Cabe aclara que el que aumente el índice de mora o el *bad rate* por cuenta, no necesariamente representa que haya un aumento de la misma magnitud en el *exposure* de la institución, ya que depende de la situación de cada crédito y su evolución a través del tiempo.



A manera de conclusión podemos decir que, para que nuestro modelo surta efecto y podamos considerar un crédito como "maduro" deberá este cumplir con una maduración mínima de 15 meses. Pese a que este indicador pertenece al *vintage* más antiguo (2015), podemos decir que es el mínimo tiempo que una cuenta necesita para poder revelar su verdadero estatus dentro de la cartera.

6.3 Definición de Bueno o Malo

Como lo hemos dicho anteriormente, debemos definir el estado donde un crédito se torna como "Malo" y con ello nos referimos a un estado máximo que estamos dispuestos a permitir antes de que convierta en un "problema" que amenace con la sustentabilidad de la Institución Financiera, para ello haremos uso de las probabilidades de transición descritas antes a través de Cadenas de Markov, donde para una muestra de *N* créditos (incluyendo "originaciones" y renovaciones) cuyas transiciones entre los diferentes estados transcurren en tiempos discretos $t = 0, \ldots . T$ donde t es medido en periodos homogéneos de 90 días cuyos estados los describimos a continuación:

Estado	Días de Mora
A	0 días
B	1 a 7 días
C	8 a 15 días
D	16 a 30 días
E	31 a 45 días
F	46 a 60 días

G	61 a 75 días
H	76 a 90 días
I	91 a 105 días
J	106 a 121 días
K	Más de 121 días

donde
X_t Es la variable aleatoria discreta del estado del crédito en el período t
$b_i(t)$ Es el número de créditos en el estado i en el periodo t
$b_{ij}(t)$ Es el número de créditos que tuvieron una transición de i en el periodo $t-1$ a j en el periodo t
$B_i(T) = \sum_{t=0}^{T-1} b_i(t)$ es el número total de transiciones que se encontraban en el estado i al principio de los estados de transición
$B_{ij}(T) = \sum_{t=1}^{T} b_{ij}(t)$ es el número total de las transiciones observadas de i a j a lo largo de todo el periodo
$U_x = \{A, B, C, D, E, F, G, H, I, J, K\}$ Es el conjunto de la variable discreta X_n, es decir, todos los posibles estados que puede presentar un crédito.

Por lo que hemos diseñado diferentes cadenas de Markov en períodos trimestrales, durante el período de maduración que mencionamos anteriormente (15 meses) iniciando en el mes 3 de maduración una vez colocado el cohorte, y visto en diferentes meses, con esas misma diferencia de tiempo, es decir, comparamos el mes 3 de maduración, con el mes 6 y en una segunda cadena de Markov, vemos el mes 4 de maduración contra el mes 7, y así sucesivamente hasta llegar al décimo mes de maduración, esto debido a que ya no aparecían todas las categorías de los estados, por lo que los estados de transición desaparecían, provocando sesgo en nuestro análisis.

		Mes 6											
	Estados	A	B	C	D	E	F	G	H	J	K	Total	Prob Mig
Mes 3	A	86%	6%	3%	2%	1%	1%	1%	0%	0%	0%	100%	14%
	B	57%	9%	9%	6%	6%	6%	3%	4%	0%	0%	100%	34%
	C	33%	6%	5%	9%	11%	12%	13%	8%	1%	0%	100%	55%
	D	18%	1%	1%	5%	6%	7%	14%	21%	26%	1%	100%	75%
	E	8%	1%	2%	4%	3%	3%	3%	4%	21%	50%	100%	82%
	F	3%	3%	4%	8%	7%	14%	9%	3%	3%	44%	100%	59%
	G	3%	1%	2%	4%	4%	5%	4%	3%	2%	72%	100%	77%
	H	7%	0%	0%	0%	0%	0%	0%	0%	0%	93%	100%	93%

Cobranza de la Cartera
Deterioro Cartera (Pmig)
Ever

Observamos que esta Cadena de Markov del mes 3 de maduración contra el mes 6, hay una importante probabilidad de migración a estados superiores en el estado E (30-45 días), donde todo lo que toca ese estado, sube de fase en un 82%, por lo que no es conveniente que ningún crédito llegue a ese nivel, puesto, que a nuestro criterio, lo consideraríamos como "Malo", debido a que el nivel de migración es insostenible a partir de ese estado, sin embargo para tener un panorama más amplio, veremos la migración intermensual de los cohortes.

Estado Inicial	Primg_ M3-M6	Primg_ M4-M7	Primg_ M5-M8	Primg_ M6-M9	Primg_ M7-M10	Primg_ M8-M11	Primg _9-M12	Primg_ M10-M13	Promedio	
A	13.6	11.6	9.4	6.0	4.8	1.3	1.1	0.3	6.0	Bueno
B	34.0	30.4	30.5	26.9	28.6	25.8	24.7	17.3	27.3	
C	54.8	47.6	44.7	39.2	40.3	37.9	38.1	34.0	42.1	Indeterminado
D	75.4	68.0	60.1	54.0	52.9	48.5	47.3	39.5	55.7	
E	81.6	80.6	75.8	71.0	69.8	63.0	68.9	66.1	72.1	
F	59.5	89.0	85.0	80.5	80.0	76.0	79.0	75.4	78.0	Malo
G	77.5	83.9	86.9	84.5	84.7	84.5	83.6	83.2	83.6	
H	92.7	93.4	93.7	92.2	91.8	92.2	91.7	91.4	92.4	
I		86.4	90.2	91.8	91.4	93.8	99.1	96.8	92.8	
J		92.2	96.4	96.2	97.0	96.6	99.6	98.9	96.7	
K			100.0	100.0	100.0	100.0	100.0	100.0	100.0	

Consideramos entonces, bajo el enfoque de cadenas de Markov que un cliente "Malo" será aquel que tenga la probabilidad de migrar a un estado superior en más del 75%, por lo que esto se cumple a partir del estado F (45 a 60 días) y un cliente "Bueno", será aquel que esté por debajo de una probabilidad de migración de 30%, los demás los consideraremos "Indeterminados" puesto que no tenemos una dirección clara de su probabilidad de migración, por ejemplo, observamos en el estado C, que el 5% de los clientes se mantienen en ese estado, 39% se cobran y un 54% migra a un estado superior, por lo que a nivel *atributos* para el diseño de un *scoring* reactivo, necesitamos tener un perfil bien definido de "Bueno" y "Malo", ya que si consideramos los atributos de los "Indeterminados" pueden sesgar de manera importante la predicción del modelo. De manera que, la definición formal de "Bueno", "Malo" e "Indeterminado" según nuestra muestra se dará en la siguiente clasificación.

Máximo día de Atraso en toda la vida del Crédito	Clasificación
0 a 15 días	Bueno
16 a 45 días	Indeterminado
46 en adelante	Malo

Cabe mencionar que esta definición no es siempre válida para la realización de un *scoring*, puesto que depende principalmente del segmento que se atiende, así como, las condiciones en las que habitualmente la Institución Financiera otorga los créditos, por ejemplo, para este caso en particular, la Institución Financiera promueve créditos en frecuencias semanales y catorcenales, por lo que, en empresas donde el segmento es en frecuencias mensuales, esta definición no sería lo más adecuada, para estipular la clasificación de tipo de cliente.

7. Modelo HIT

Primordialmente iniciaremos con el segmento HIT, el cuál como definíamos en un inicio de esta sección se hace referencia al grupo de personas que cuentan previamente con historial de crédito, cuyas cuentas activan el score genérico, es decir, generan algún tipo de probabilidad de cumplimiento ya que tienen un historial de crédito "robusto" que permite una "mejor" selección del riesgo de crédito. Una vez aplicado los criterios de "Bueno" y "Malo" a nuestra base de datos, obtenemos lo siguiente:

Buenos (A)	Malos (B)	Indeterminados (C)	Total (A+B+C)	Califican Modelo D= (A+B)	Bad Rate B/D
11,442	4,252	3,572	19,266	15,694	27.09%

De los cuáles podemos determinar a primera vista que se trata de un problema desbalanceado, esto representa a nivel Minería de Datos, que es un problema donde tenemos observaciones en forma minoritaria, lo que provoca un desbalanceo en los datos que utilizaremos para la creación de nuestro modelo.

7.1 Validación de Base de Datos

En esta sección veremos sintéticamente los datos descriptivos de la base de datos, en donde el objetivo es reconocer:

1.) Variables con alto grado de *missings*
2.) Los *outliers* de las variables continuas y discretas
3.) Sesgo de las variables discretas y continuas (Sí la mediana está muy alejada del promedio)

A continuación se mostrará la descripción de las variables que se revisarán, el proceso se realiza de manera automatizada a través de la plataforma R, cuyo script es detallado en el anexo II.

Para efectos de este trabajo, dejaremos en claro el tipo de clases de variables que tenemos en nuestra base de datos:

1.) **Numérica o cuantitativa:** Son números que representan una medición o magnitud sobre la observación. Este tipo de clases de variables se pueden someter a un análisis bivariante a través de funciones de densidad.
2.) **Categórica**: Representa las categorías, grupos o atributos cualitativos que puede tener una observación, estas pueden tener un orden lógico o no. Es por ello que a este tipo de variables no las someteremos a un análisis riguroso.

Variable	Tipo de variable	Clase	Definición
Score Genérico	Sin transformar	Numérica	Puntaje del score genérico en fecha de solicitud del prospecto
Edad	Sin transformar	Numérica	Edad con la que se registró el prospecto en fecha de

			solicitud
Ingreso_Mensual	Sin transformar	Numérica	Ingresos estimados mensuales fijos comprobables del mes del solicitante
Total_Egresos	Sin transformar	Numérica	Egresos estimados totales del prospecto en la fecha de solicitud
Deuda_Buro_credito	Sin transformar	Numérica	Deuda reportada por instituciones financieras a través del reporte de Crédito en fecha de solicitud
Total_Ingresos	Sin transformar	Numérica	Ingreso mensual total estimado del acreditado
Hipoteca	Sin transformar	Numérica	Cantidad de dinero mensual destinada al pago de la hipoteca y/o renta del prospecto
Remanente	Artificial	Numérica	Diferencia entre Total_Ingresos y Total_Egresos
Inventario	Sin transformar	Numérica	Cantidad aproximada en dinero del stock de ventas del micro negocio
SaldoOtorganteBancos	Sin transformar	Numérica	Suma de saldo total agregado del prospecto a la fecha de corte de productos de crédito otorgados por Banco
SaldoOtorganteComunicaciones	Sin transformar	Numérica	Suma de saldo total agregado del prospecto a la fecha de corte de productos de crédito otorgados por establecimientos de Comunicaciones
SaldoOtorganteFinancieras	Sin transformar	Numérica	Suma de saldo total agregado del prospecto a la fecha de corte de productos de crédito otorgados por Financieras
SaldoOtorganteTiendaComerc	Sin transformar	Numérica	Suma de saldo total agregado del prospecto a la fecha de corte de productos de crédito otorgados por establecimientos de Servicios
SaldoOtorganteServicios	Sin transformar	Numérica	Suma de saldo total agregado del prospecto a la fecha de corte de productos de crédito otorgados por Tiendas Comerciales
SaldoOtorganteAutos	Sin transformar	Numérica	Suma de saldo total agregado del prospecto a la fecha de corte de productos de crédito otorgados por establecimientos

			Automotrices
SaldoOtorganteHipoteca	Sin transformar	Numérica	Suma de saldo total agregado del prospecto a la fecha de corte de productos de crédito otorgados por Hipotecarias
VencimientoOtorganteBancos	Sin transformar	Numérica	Suma de saldo vencido (>1 día atraso) agregado del prospecto a la fecha de corte de productos de crédito otorgados por Banco
VencimientoOtorganteComunicaciones	Sin transformar	Numérica	Suma de saldo vencido (>1 día atraso) agregado del prospecto a la fecha de corte de productos de crédito otorgados por establecimientos de Comunicaciones
VencimientoOtorganteFinancieras	Sin transformar	Numérica	Suma de saldo vencido (>1 día atraso) agregado del prospecto a la fecha de corte de productos de crédito otorgados por Financieras
VencimientoOtorganteTiendaComercial	Sin transformar	Numérica	Suma de saldo vencido (>1 día atraso) agregado del prospecto a la fecha de corte de productos de crédito otorgados por establecimientos de Servicios
VencimientoOtorganteServicios	Sin transformar	Numérica	Suma de saldo vencido (>1 día atraso) agregado del prospecto a la fecha de corte de productos de crédito otorgados por Tiendas Comerciales
VencimientoOtorganteAutos	Sin transformar	Numérica	Suma de saldo vencido (>1 día atraso) agregado del prospecto a la fecha de corte de productos de crédito otorgados por establecimientos Automotrices
RatioExcedente	Artificial	Numérica	Remantente/Ingreso_Mensual
RatioFinancieras	Artificial	Numérica	SaldoFinanciera/Ingreso_Mensual
RatioAutos	Artificial	Numérica	SaldoAutos/Ingreso_Mensual
RatioDeuda	Artificial	Numérica	DeudaBuroCredito/Ingreso_Mensual
Dependientes	Sin transformar	Numérica	Número de dependientes directos que registró prospecto en el momento de la solicitud

Tipo_Negocio	Sin transformar	Categórica	Tipo de local en el que desempeña su actividad prospecto
Tipo_Vivienda	Sin transformar	Categórica	Tipo de contrato de vivienda que prospecto acredita al momento solicitud
Giro	Sin transformar	Categórica	Actividad económica que desempeña acreditado
CP	Sin transformar	Categórica	Código postal de residencia del acreditado
Colonia	Sin transformar	Categórica	Colonia de residencia del acreditado
Municipio	Sin transformar	Categórica	Municipio de residencia del acreditado
Estado	Sin transformar	Categórica	Estado de residencia del acreditado
mop_actBancos	Sin transformar	Categórica/ Numérica	Nivel de atraso de clientes al momento de solicitud incurrido con otorgante Banco
mop_actComunicaciones	Sin transformar	Categórica/ Numérica	Nivel de atraso de clientes al momento de solicitud incurrido con otorgante Comunicaciones
mop_actFinancieras	Sin transformar	Categórica/ Numérica	Nivel de atraso de clientes al momento de solicitud incurrido con otorgante Financieras
mop_actTiendaComerc	Sin transformar	Categórica /Numérica	Nivel de atraso de clientes al momento de solicitud incurrido con otorgante Tienda Comercial
mop_actServicios	Sin transformar	Categórica/ Numérica	Nivel de atraso de clientes al momento de solicitud incurrido con otorgante Servicios
mop_actAutos	Sin transformar	Categórica/ Numérica	Nivel de atraso de clientes al momento de solicitud incurrido con otorgante Automotriz
mop_histBancos	Sin transformar	Categórica/ Numérica	Nivel de atraso de clientes al momento de solicitud incurrido con otorgante Banco
mop_histComunicaciones	Sin transformar	Categórica/ Numérica	Nivel máximo de atraso registrado en toda vida del crédito de clientes al momento de solicitud incurrido con otorgante Comunicaciones
mop_histFinancieras	Sin transformar	Categórica/ Numérica	Nivel máximo de atraso registrado en toda vida del crédito de clientes al

			momento de solicitud incurrido con otorgante Financieras
mop_actHipoteca	Sin transformar	Categórica/ Numérica	Nivel máximo de atraso registrado en toda vida del crédito de clientes al momento de solicitud incurrido con otorgante Tienda Comercial
mop_histServicios	Sin transformar	Categórica/ Numérica	Nivel máximo de atraso registrado en toda vida del crédito de clientes al momento de solicitud incurrido con otorgante Servicios
mop_histAutos	Sin transformar	Categórica/ Numérica	Nivel máximo de atraso registrado en toda vida del crédito de clientes al momento de solicitud incurrido con otorgante Automotriz
mop_histHipoteca	Sin transformar	Categórica/ Numérica	Nivel máximo de atraso registrado en toda vida del crédito de clientes al momento de solicitud incurrido con otorgante Hipotecario
mop_histTiendaComerc	Sin transformar	Categórica/ Numérica	Nivel máximo de atraso registrado en toda vida del crédito de clientes al momento de solicitud incurrido con otorgante Tienda Comercial
Estado_Civil	Sin transformar	Categórica	Estado de civil del acreditado
Sexo	Sin transformar	Categórica	Sexo del acreditado
Avales	Sin transformar	Categórica/ Numérica	Número de avales solicitados operación
Consultas	Sin transformar	Categórica/ Numérica	Número de consultas reportadas del historial crediticio al momento de solicitud
Est_Giros	Artificial	Categórica	Estratificación de giros según su nivel de mora
Est_Mun	Artificial	Categórica	Estratificación de municipios según su nivel de mora

Validación de BD HIT Cuantitativas

Variable	Missing	Completos	Conteo	Media	SD	PC0	P25	PC50	PC75	PC100	Tasa Missing	CoefVar
ScoreGenerico	2	19264	19266	687	42	500	661	696	717	815	0%	6%
Edad	0	19266	19266	40	11	2	31	39	48	110	0%	27%
Ingreso_Mensual	0	19266	19266	27565	28142	0	16000	22400	32000	1800000	0%	102%
Total_Egresos	0	19266	19266	5675	9294	-23224	2996	4432	6659	1043326	0%	164%
Deuda_Buro_credito	0	19266	19266	15534	38682	0	1277	6811	16867	2149203	0%	249%
Total_Ingresos	0	19266	19266	10779	9116	-40000	6738	9381	12538	414320	0%	85%
Hipoteca	0	19266	19266	6504	43543	0	0	0	9500	5000000	0%	670%
Remanente	0	19266	19266	7939	7171	-17722	4512	7012	9439	323016	0%	90%
Inventario	0	19266	19266	10162	61259	0	426	2720	8458	4754977	0%	603%
SaldoOtorganteBancos	3251	16015	19266	6232	11484	0	0	3598	7594	432209	17%	184%
SaldoOtorganteComunicaciones	15646	3620	19266	607	1517	0	0	0	560	29818	81%	250%
SaldoOtorganteFinancieras	9146	10120	19266	5379	10804	0	0	2765	6965	376677	47%	201%
SaldoOtorganteTiendaComerc	17646	1620	19266	1699	4206	0	0	0	1330	59419	92%	247%
SaldoOtorganteServicios	15367	3899	19266	127	771	0	0	0	0	17400	80%	609%
SaldoOtorganteAutos	19055	211	19266	27394	61161	0	0	0	26442	566747	99%	223%
SaldoOtorganteHipoteca	19214	52	19266	3555	9775	0	0	0	3742	62998	100%	275%
VencimientoOtorganteBancos	3251	16015	19266	685	3616	0	0	0	0	150606	17%	528%
VencimientoOtorganteComunicaciones	15646	3620	19266	393	1304	0	0	0	0	29818	81%	332%
VencimientoOtorganteFinancieras	9120	10146	19266	449	2915	0	0	0	0	166005	47%	650%
VencimientoOtorganteTiendaComercial	17646	1620	19266	254	1305	0	0	0	0	19519	92%	514%
VencimientoOtorganteServicios	15367	3899	19266	76	676	0	0	0	0	16668	80%	890%
VencimientoOtorganteAutos	19056	210	19266	1639	7899	0	0	0	0	70587	99%	482%
RatioExcedente	0	19266	19266	Inf	Inf	-5	0	0	0	Inf	0%	NA
RatioFinancieras	9146	10120	19266	0	0	0	0	0	0	21	47%	199%
RatioAutos	19055	211	19266	1	3	0	0	0	1	31	99%	309%
RatioDeuda	0	19266	19266	Inf	Inf	0	0	0	1	Inf	0%	NA

Datos con error en la información (outliers sin sentido)
Insuficiencia de datos en la variable
Alto grado de missings en la variable
Desviación estándar dos o veces más grande promedio

Validación de BD HIT Cuantitativas

Variable	Missing	Completos	Conteo	% Missing	Min_ longitud	Máximo Longitud	Únicos
Dependientes	0	19266	19266	0%	1	2	15
Tipo_Negocio	0	19266	19266	0%	4	14	3
Tipo_Vivienda	0	19266	19266	0%	9	19	4
Giro	0	19266	19266	0%	5	100	475
CP	0	19266	19266	0%	5	5	2250
Colonia	0	19266	19266	0%	4	49	1971
Municipio	0	19266	19266	0%	4	36	318
Estado	0	19266	19266	0%	6	31	11
mop_actBancos	3250	16016	19266	17%	1	2	11
mop_actComunicaciones	15646	3620	19266	81%	1	2	9
mop_actFinancieras	9119	10147	19266	47%	1	2	11
mop_actTiendaComerc	17646	1620	19266	92%	1	2	11
mop_actServicios	15367	3899	19266	80%	1	2	10
mop_actAutos	19054	212	19266	99%	1	2	9
mop_histBancos	3250	16016	19266	17%	1	2	13
mop_histComunicaciones	15646	3620	19266	81%	1	2	13
mop_histFinancieras	9119	10147	19266	47%	1	2	13
mop_actHipoteca	19214	52	19266	100%	1	2	13
mop_histServicios	15367	3899	19266	80%	1	2	13
mop_histAutos	19054	212	19266	99%	1	2	13
mop_histHipoteca	19214	52	19266	100%	1	2	13
mop_histTiendaComerc	17646	1620	19266	92%	1	2	12
Estado_Civil	19	19247	19266	0%	1	1	5
Sexo	2167	17099	19266	11%	1	1	2
Avales	19	19247	19266	0%	1	1	5
Consultas	919	18347	19266	5%	1	2	51

Alto grado de missings en la variable

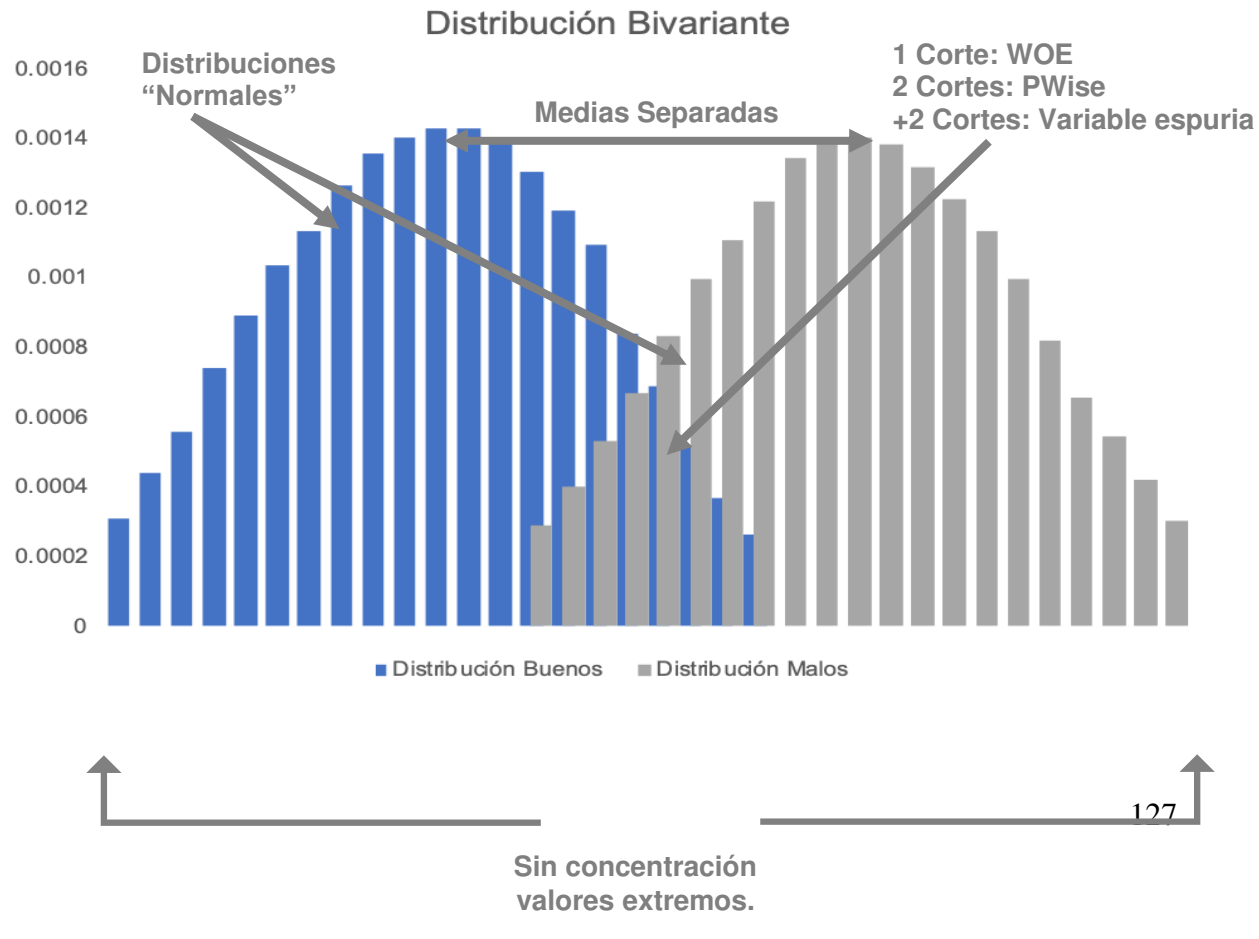
El detalle del análisis de “Datos con errores en la información” se sostiene en lo siguiente:

Variable	Inconsistencia
ScoreGenerico	Se encontró en PC100 un outlier, con un Score Genérico >760. El score no es causante de puntajes mayores a 760 puntos.
Edad	La edad por política de crédito no es menor a 18 años y no mayor a 70, por lo que un dato en PC0 =2 y PC100 = 110, es un error de base de datos.
Total_Egresos	Por política de crédito no se registran valores negativos
Total_Ingresos	Por política de crédito no se registran valores negativos
Remanente	Por política de crédito no se registran valores negativos

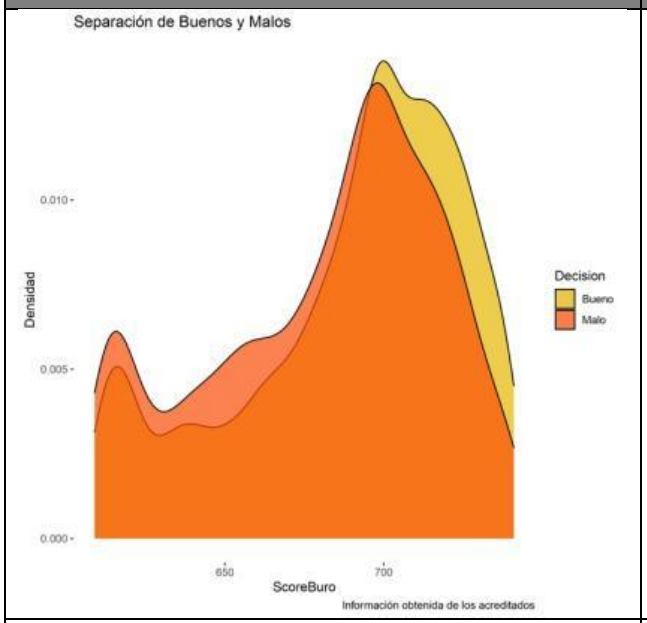
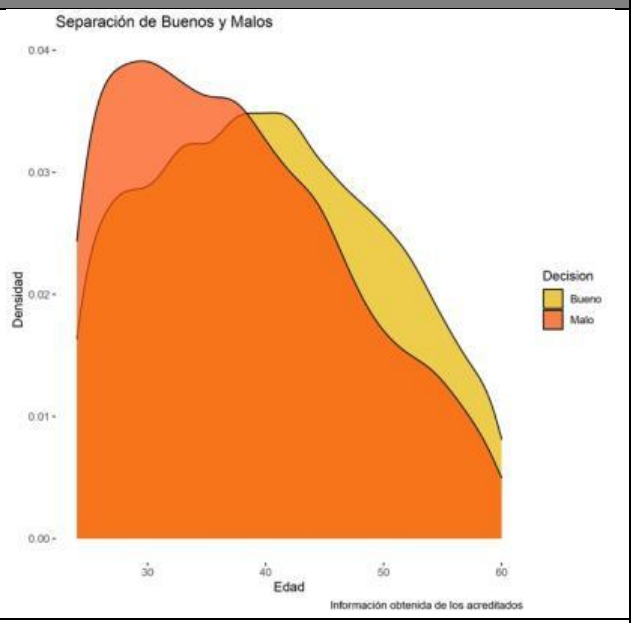
Vemos que la mayoría de las variables discretas y continuas presentan algún tipo de inconsistencia en la base de datos, mientras que el mayor problema en las variables cualitativas son las variables con alto grado de "*missings*", de manera que, al tener datos con sesgo, el universo de variables candidatas a utilizar en un modelo de *scoring* reactivo se van reduciendo.

7.2 Análisis de bases de datos

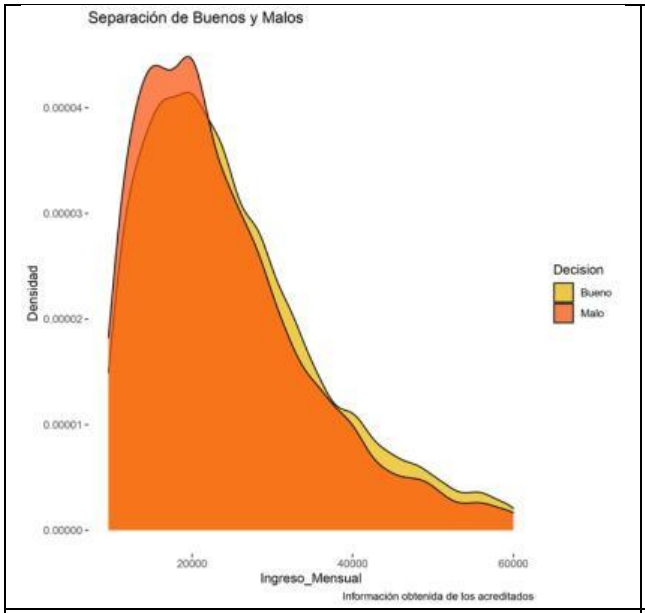
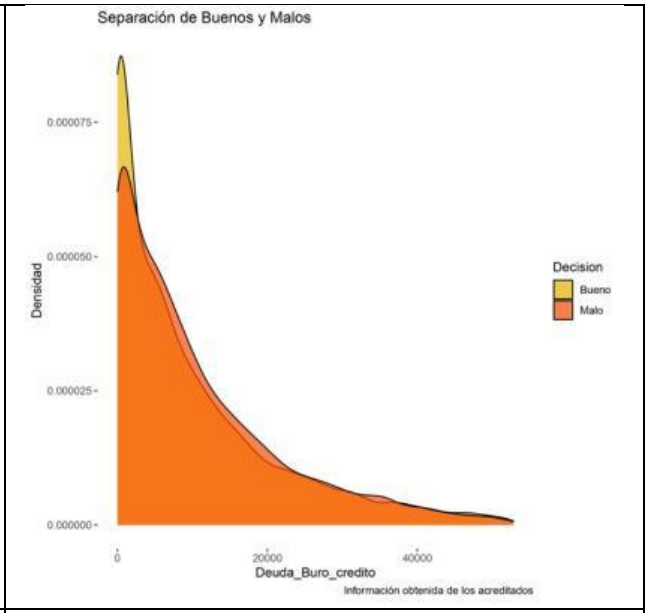
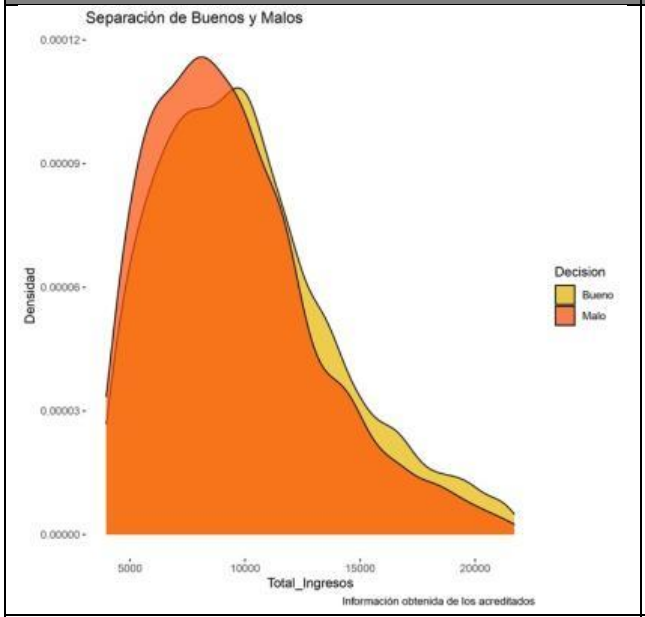
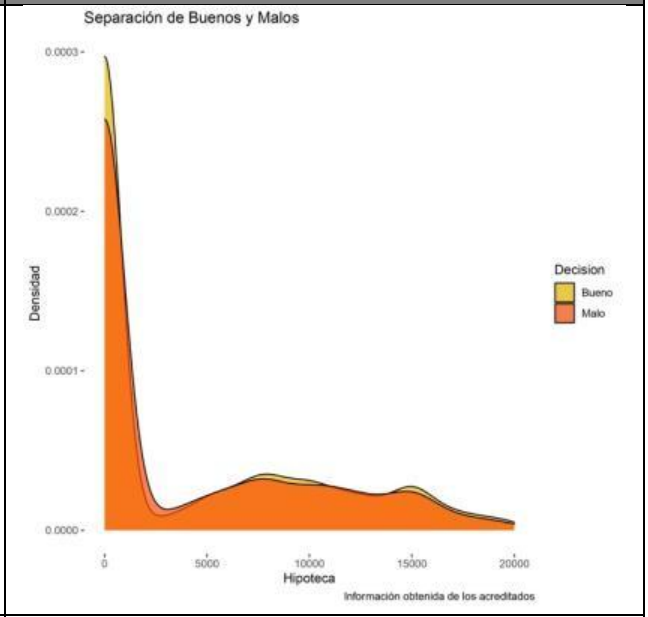
Una vez que hemos detectado en el universo de variables, procederemos a determinar de manera visual a través de un ciclo en R diseñado para realizar n gráficos bivariantes de densidad para variables continuas, en el caso de las variables categóricas directamente pasaremos al análisis WOE. En el siguiente ejemplo, veremos la anatomía de un bivariante continuo, en el cuál determinaremos visualmente elementos para dictaminar una variable como candidata a incluir en un *scoring*.

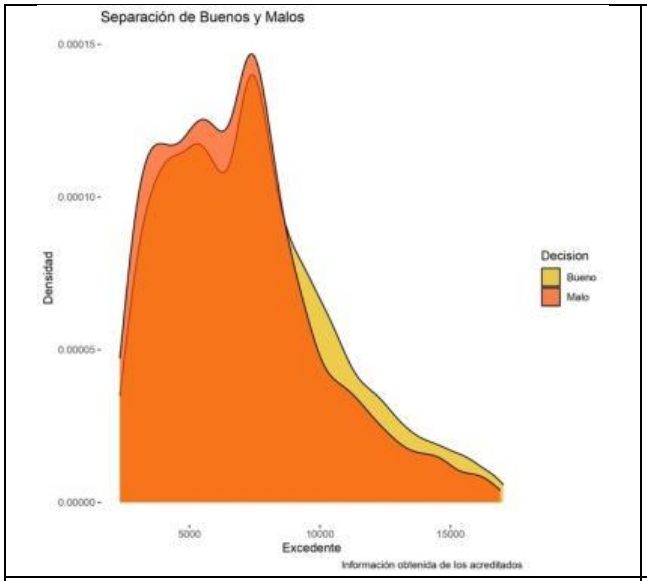
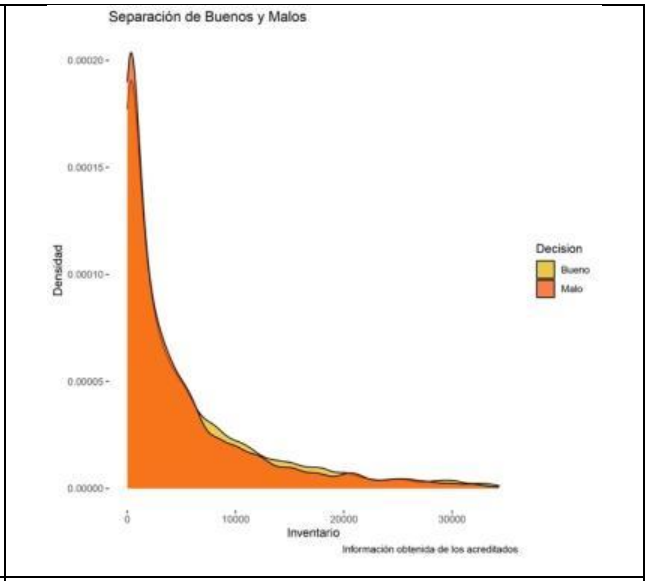
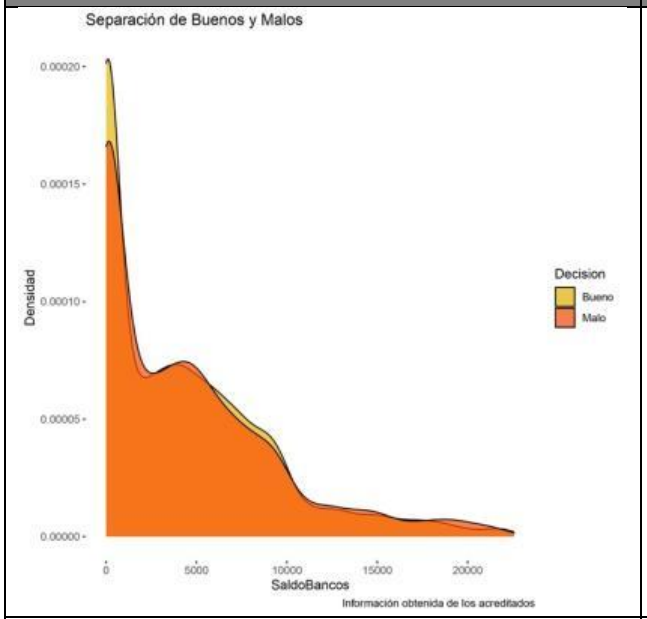
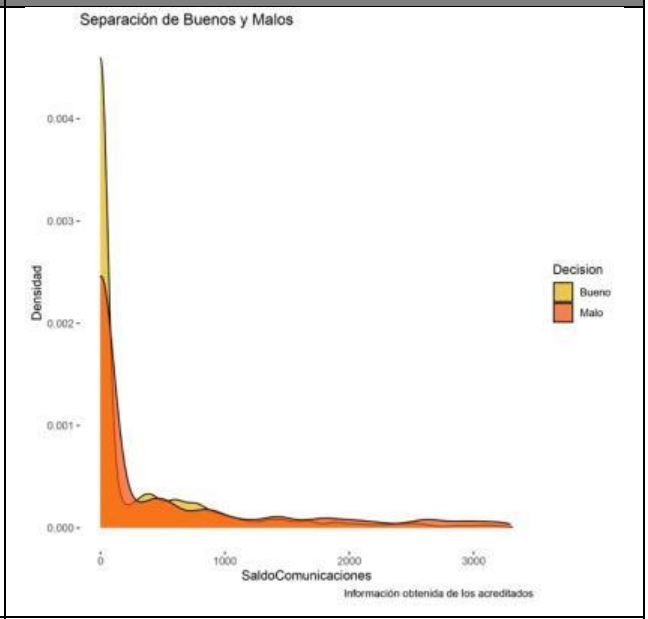


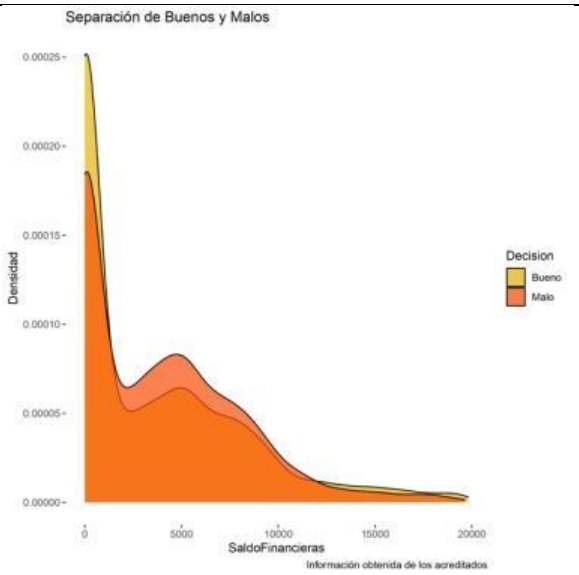
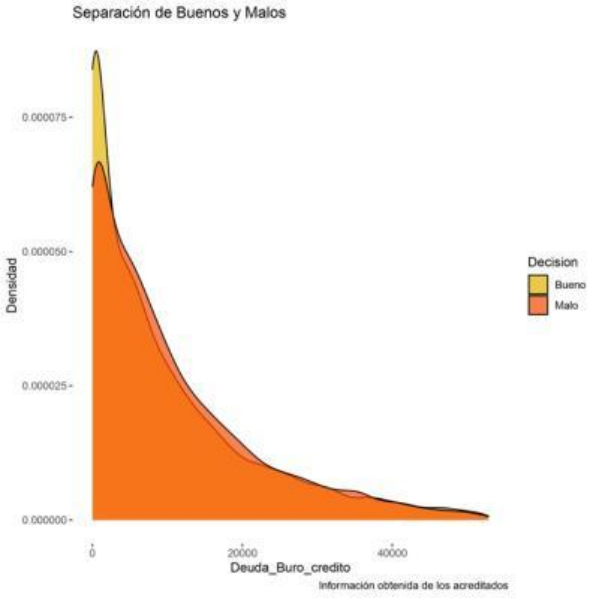
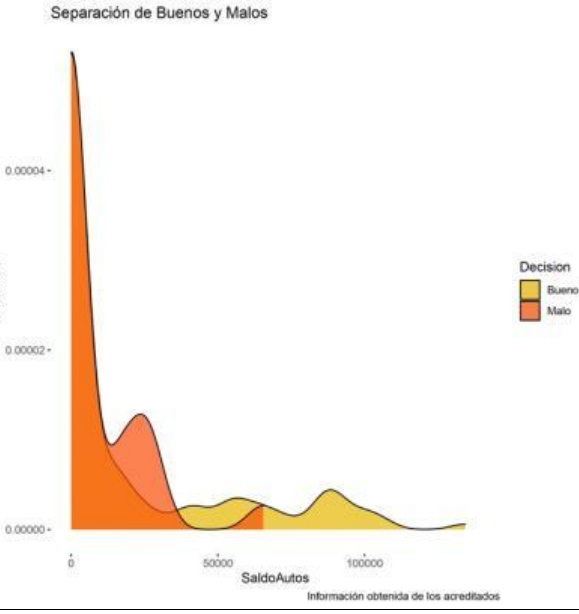
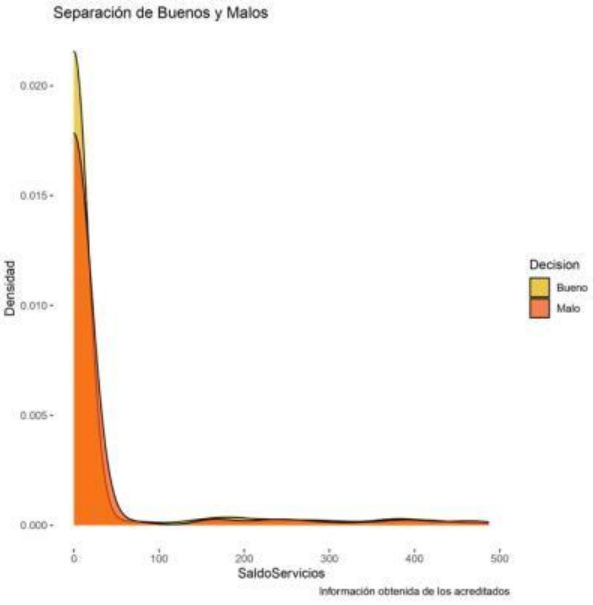
Una de las partes más importantes de este análisis es reconocer el número de veces que corta la función de densidad de buenos, de la de malos, esto debido a que sobre de esos se sustenta la lógica de la variable, es decir, que sigue alguna tendencia de mora.
Adicionalmente, tenemos que encontrar visualmente buenas candidatas que cumplan con el ejemplo antes mencionado, para efectos de eficiencia en este trabajo, mostraremos únicamente las variables que, a nuestro juicio, fueron visiblemente aceptables[37], adicionalmente, en cada gráfico añadiremos el estadístico *SommersD*[38] que nos ayudará a corroborar que la relación de separación de poblaciones se cumple, es decir que entre este estadístico este más cercano a -1 o 1, mejor separación va a tener la variable, es decir, que puede ser candidata.

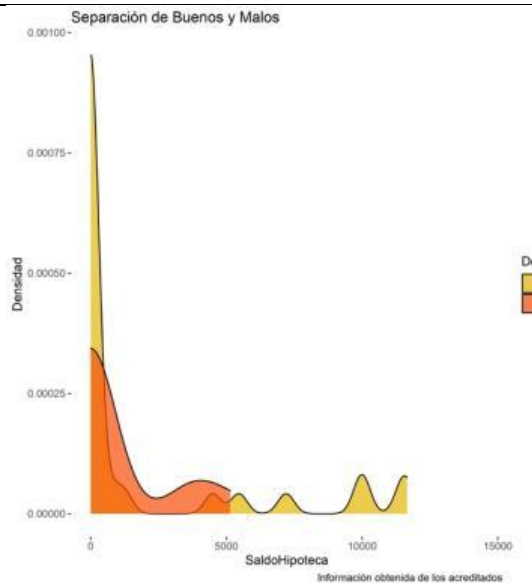
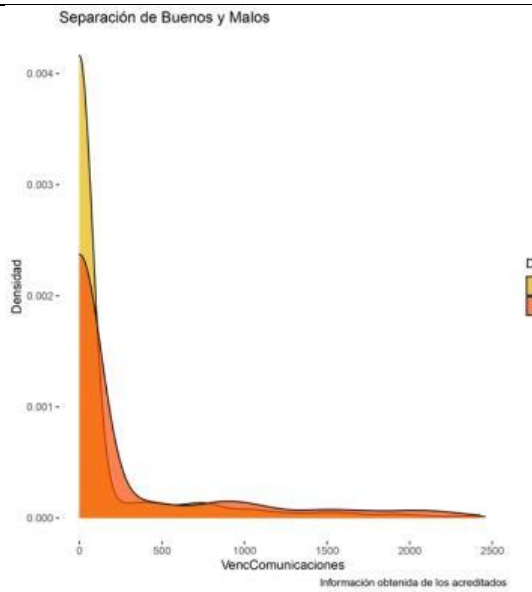
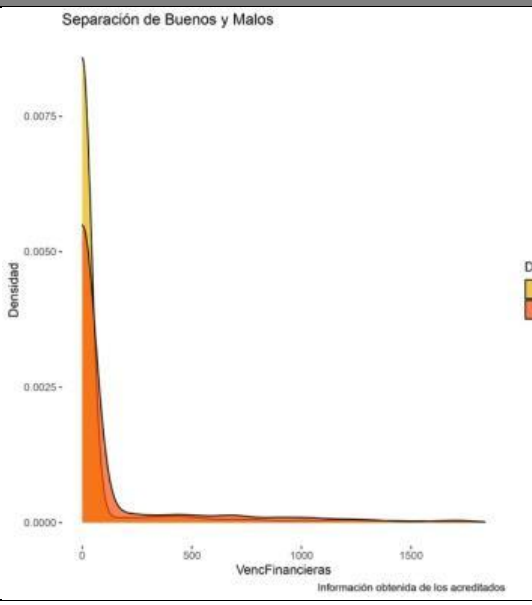
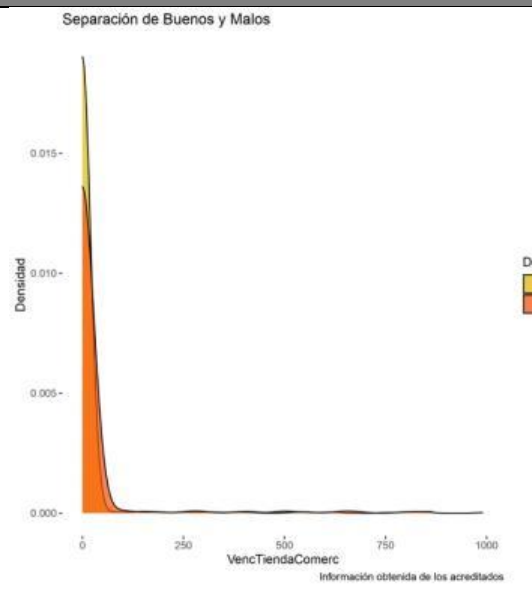
<table>
<tr><th colspan="2">Variable: Score Genérico</th><th colspan="2">Variable: Edad</th></tr>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 1</td><td colspan="2">Número de Cortes: 1</td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.1745</td><td>Missings: 0%</td><td>SommersD: -0.1943</td></tr>
<tr><td>PC50: 696</td><td>Promedio: 686</td><td>PC50: 39</td><td>Promedio: 39</td></tr>
<tr><th colspan="2">Variable: Ingreso Mensual</th><th colspan="2">Variable: Deuda Buró</th></tr>
</table>

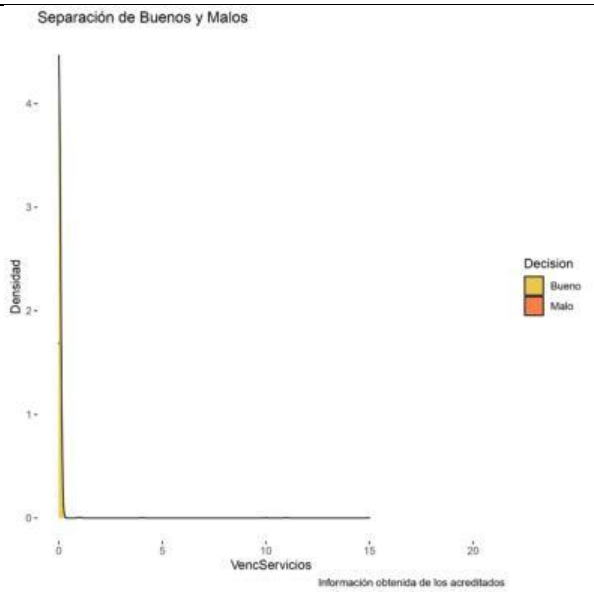
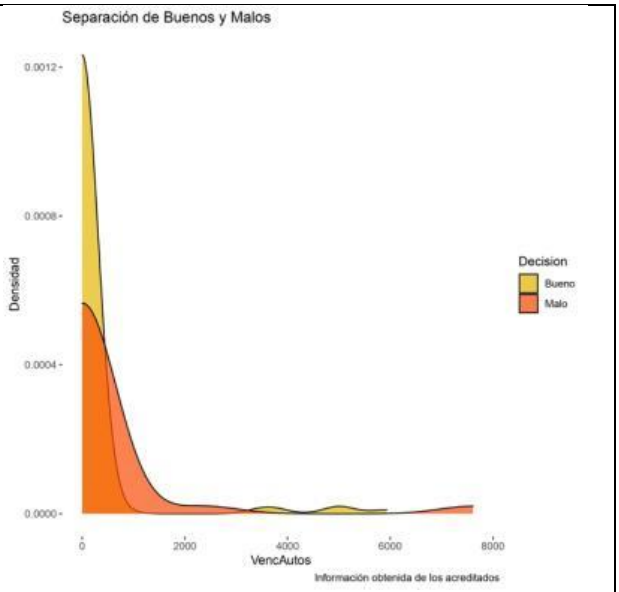
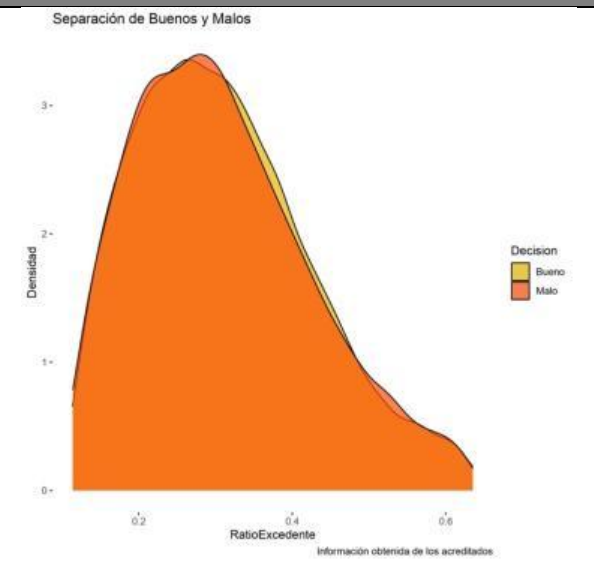
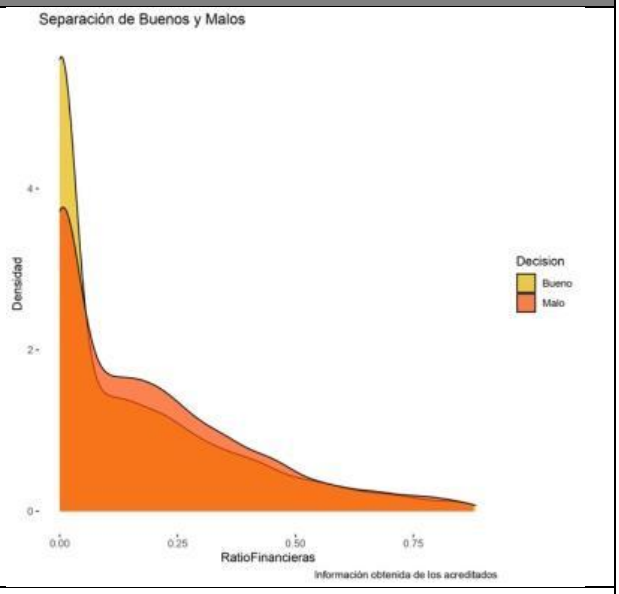
[37] Si se requiere de mayor información, consultar directamente con el autor.
[38] Es un caso particular del estadístico GINI que nos ayuda a establecer la separación de dos variables.

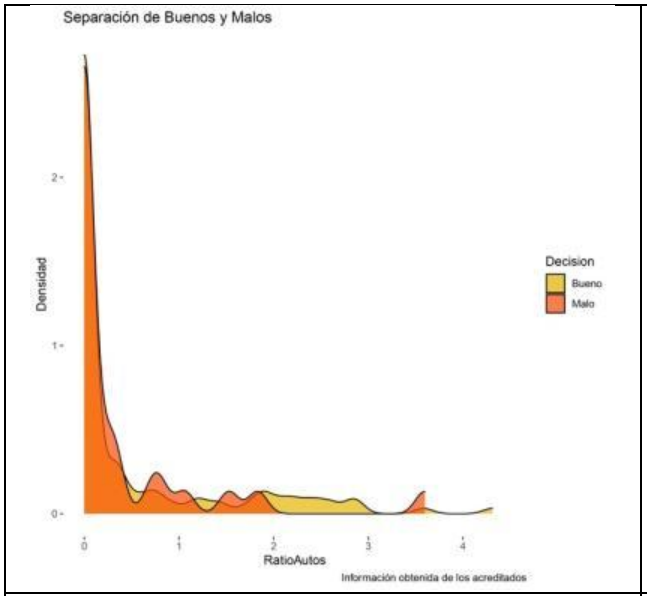
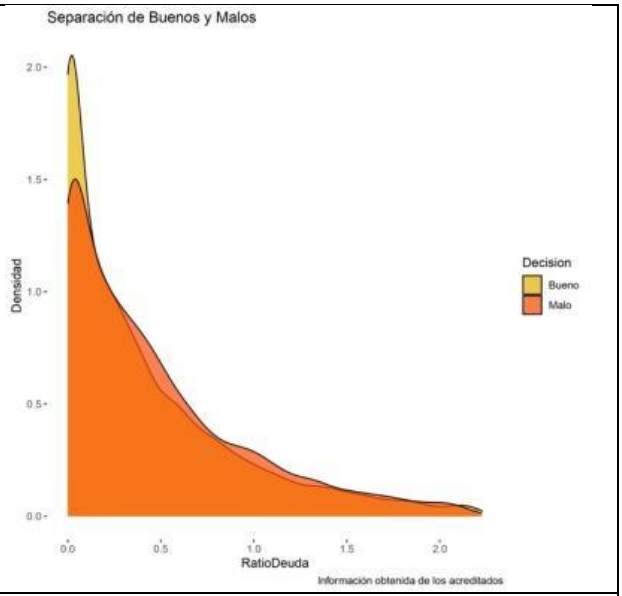
<table>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 1</td><td colspan="2">Número de Cortes: 1</td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.0845</td><td>Missings: 0%</td><td>SommersD: -0.005</td></tr>
<tr><td>PC50: $22,400</td><td>Promedio: $27,565</td><td>PC50: $6,811</td><td>Promedio: $15,334</td></tr>
<tr><td colspan="2">Variable: Total Ingresos</td><td colspan="2">Variable: Hipoteca</td></tr>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 1</td><td colspan="2">Número de cortes: +3</td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.0964</td><td>Missings: 0%</td><td>SommersD: -.03661</td></tr>
<tr><td>PC50: $10,779.44</td><td>Promedio: $9,380</td><td>PC50: $0</td><td>Promedio: $6503</td></tr>
<tr><td colspan="2">Variable: Remanente</td><td colspan="2">Variable: Inventario</td></tr>
</table>

<table>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 1</td><td colspan="2">Número de Cortes: +2</td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.0964</td><td>Missings: 0%</td><td>SommersD: -0.0938</td></tr>
<tr><td>PC50: $7,012</td><td>Promedio: $7,939</td><td>PC50: $2,720</td><td>Promedio: $10,162</td></tr>
<tr><td colspan="2">Variable: SaldoOtorganteBancos</td><td colspan="2">Variable: SaldoOtorganteComunicaciones</td></tr>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: +3</td><td colspan="2">Número de cortes:+3</td></tr>
<tr><td>Missings: 17%</td><td>SommersD: -0.0863</td><td>Missings: 81%</td><td>SommersD: -.2247</td></tr>
<tr><td>PC50: $3,598</td><td>Promedio: $6,232</td><td>PC50: $0</td><td>Promedio: $606</td></tr>
<tr><td colspan="2">Variable: SaldoOtorganteFinancieras</td><td colspan="2">Variable: SaldoOtorganteTiendaCom</td></tr>
</table>

<table>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 2</td><td colspan="2">Número de Cortes: 2</td></tr>
<tr><td>Missings: 47%</td><td>SommersD: -.0607</td><td>Missings: 92%</td><td>SommersD: -.4446</td></tr>
<tr><td>PC50: $2,764</td><td>Promedio:$5,379</td><td>PC50: 0</td><td>Promedio: $1,699</td></tr>
<tr><td colspan="2">Variable: SaldoOtorganteAutos</td><td colspan="2">Variable: SaldoOtorganteServicios</td></tr>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 2</td><td colspan="2">Número de cortes: +3</td></tr>
<tr><td>Missings: 99%</td><td>SommersD: -0.4299</td><td>Missings: 80%</td><td>SommersD: -0.7394</td></tr>
<tr><td>PC50: $0</td><td>Promedio: $27,394</td><td>PC50: $0</td><td>Promedio: $126</td></tr>
<tr><td colspan="2">Variable: SaldoOtorganteHipoteca</td><td colspan="2">Variable: VencComunicaciones</td></tr>
</table>

<table>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 2</td><td colspan="2">Número de Cortes: +2</td></tr>
<tr><td>Missings: 99.9%</td><td>SommersD: -0.5992</td><td>Missings: 81%</td><td>SommersD: -.6708</td></tr>
<tr><td>PC50: $0</td><td>Promedio: $3,554</td><td>PC50: 0</td><td>Promedio: $392</td></tr>
<tr><td colspan="2">Variable: VencFinancieras</td><td colspan="2">Variable: VencTiendaComercial</td></tr>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 2</td><td colspan="2">Número de cortes: +3</td></tr>
<tr><td>Missings: 47%</td><td>SommersD: -0.5641</td><td>Missings: 92%</td><td>SommersD: -0.7724</td></tr>
<tr><td>PC50: $0</td><td>Promedio: $448</td><td>PC50: $0</td><td>Promedio: $253</td></tr>
<tr><td colspan="2">Variable: VencServicios</td><td colspan="2">Variable: VencAuto</td></tr>
</table>

<table>
<tr><td> 
</td><td> 
</td></tr>
<tr><td colspan="2">Número de cortes: NA</td><td colspan="2">Número de Cortes: +3</td></tr>
<tr><td>Missings: 80%</td><td>SommersD: -0.9046</td><td>Missings: 99%</td><td>SommersD: -.8064</td></tr>
<tr><td>PC50: $0</td><td>Promedio: $675</td><td>PC50: $0</td><td>Promedio: $1,638</td></tr>
<tr><td colspan="2">Variable: RatioExcedente</td><td colspan="2">Variable: RatioFinanciera</td></tr>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: +3</td><td colspan="2">Número de cortes: 1</td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.0207</td><td>Missings: 47%</td><td>SommersD: -0.0430</td></tr>
<tr><td>PC50: 0.2959</td><td>Promedio: Inf</td><td>PC50: 0.1067</td><td>Promedio: 0.2368</td></tr>
<tr><td colspan="2">Variable: RatioAutos</td><td colspan="2">Variable: RatioDeuda</td></tr>
</table>

<table>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: +3</td><td colspan="2">Número de Cortes: 1</td></tr>
<tr><td>Missings: 99%</td><td>SommersD: -0.4502</td><td>Missings: 0%</td><td>SommersD: .0293</td></tr>
<tr><td>PC50: 0</td><td>Promedio: 0.9999</td><td>PC50: 0.3000</td><td>Promedio: Inf</td></tr>
</table>

Una vez observados los gráficos bivariantes, podemos observar que muy pocas variables cuentan con las características que buscamos para que sean candidatas, entre ellas las que podemos mencionar son las siguientes:

1.) *Score* Genérico
2.) Edad
3.) Remanente
4.) Ratio Financiera
5.) Ratio Deuda

Esto derivado de que cumplen con las cualidades de tener un solo corte, no tener alto grado de *missings*, las medias no están separadas de la mediana (PC50) y la *SommersD'* lanza un valor aceptable de separación. Para la sección categórica, solo mostraremos

7.3 Selección y transformación de datos

En esta sección haremos uso de la metodología *Weight of Evidence* (WOE), así como de su indicador de separación de variables *Information Value* (IV) con el motivo de seleccionar las variables finales que acompañaran a nuestro modelo. Se realizó una depuración de la base de datos original, de tal manera que sólo mostraremos las variables que tengan un valor superior a un IV de 0.02.

En el caso de las variables continuas, se categorizaron sobre tramos equiprobables a priori, esto significa que todas las posibles categorías tienen las mismas oportunidades de salir, esto es sumamente importante ya que, a la hora de realizar modelos, para mantener el nivel de predicción debemos de mantener una estabilidad poblacional, que se ve seriamente

afectada sí los tramos no son equiprobables, ya que pueden desaparecer categorías que originalmente concebimos en el modelo.

Rango	Total_Buenos	Total_Malos	Total	Bad_Rate	WOE	IV	MAXIV
Score_Genérico							
a)500-619	1187	644	1831	35%	-0.3784	0.018	0.088
b)620-654	1154	556	1710	33%	-0.2597	0.008	0.088
c)655-677	1170	551	1721	32%	-0.2369	0.006	0.088
d)678-691	1240	512	1752	29%	-0.1054	0.001	0.088
e)692-700	1286	513	1799	29%	-0.0709	0.001	0.088
f)701-710	1284	426	1710	25%	0.1134	0.001	0.088
g)711-719	1292	402	1694	24%	0.1776	0.003	0.088
h)719-731	1385	358	1743	21%	0.3630	0.013	0.088
i)Más de 731	1444	289	1733	17%	0.6188	0.036	0.088
Edad							
a)18-29	2109	1152	3261	35%	-0.3852	0.033	0.081
b)30-36	2280	1018	3298	31%	-0.1836	0.007	0.081
c)37-42	2232	779	3011	26%	0.0627	0.001	0.081
d)43-50	2440	721	3161	23%	0.2292	0.010	0.081
e)Más de 50	2381	582	2963	20%	0.4189	0.030	0.081
Remanente							
a)Hasta_4600	2738	1193	3931	30%	-0.1591	0.007	0.031
b)Hasta_7100	2778	1134	3912	29%	-0.0939	0.002	0.031
c)Hasta_9500	2840	1081	3921	28%	-0.0240	0.000	0.031
d)Más de 9420	3086	844	3930	21%	0.3066	0.022	0.031
Est_Giro							
h)Muybajoriesgo	1672	284	1956	15%	0.7829	0.062	0.135
g)BajoRiesgo	1998	568	2566	22%	0.2679	0.011	0.135
f)Bajoaltoriesgo	1144	371	1515	24%	0.1362	0.002	0.135
e)MedioBajoRiesgo	1225	439	1664	26%	0.0363	0.000	0.135
d)MedioRiesgo	1515	589	2104	28%	-0.0452	0.000	0.135
c)MedioaltoRiesgo	1927	842	2769	30%	-0.1620	0.005	0.135
b)AltoRiesgo	942	445	1387	32%	-0.2400	0.005	0.135
a)MuyAltoRiesgo	1019	714	1733	41%	-0.6342	0.050	0.135
Est_ Mun							
h)Muybajoriesgo	1909	175	2084	8%	1.3996	0.176	0.397
g)BajoRiesgo	1604	304	1908	16%	0.6733	0.046	0.397
f)Bajoaltoriesgo	1532	411	1943	21%	0.3258	0.012	0.397
e)MedioBajoRiesgo	1325	447	1772	25%	0.0967	0.001	0.397
d)MedioRiesgo	1526	611	2137	29%	-0.0746	0.001	0.397
c)MedioaltoRiesgo	1299	632	1931	33%	-0.2694	0.009	0.397

b)AltoRiesgo	1246	743	1989	37%	-0.4729	0.031	0.397
a)MuyAltoRiesgo	1001	929	1930	48%	-0.9153	0.120	0.397
Ratio Financieras							
a)Hasta_4%Deuda	2485	834	3319	25%	0.1019	0.002	0.036
b)4%_20%Deuda	1116	542	1658	33%	-0.2677	0.008	0.036
c)21%-39%Deuda	1070	576	1646	35%	-0.3706	0.016	0.036
d)Más de 40%Deuda	1161	486	1647	30%	-0.1191	0.002	0.036
missing	5610	1814	7424	24%	0.1391	0.009	0.036
Ratio Deudas							
a)Hasta_2%deuda	2260	701	2961	24%	0.1807	0.006	0.02
b)2%_20%Deuda	2490	835	3325	25%	0.1027	0.003	0.02
c)21%-43%Deuda	2275	876	3151	28%	-0.0355	0.006	0.02
d)Más de 43%Deuda	4417	1840	6257	29%	-0.1142	0.005	0.02
Tipo_Vivienda							
Propia	3816	286	4102	7%	1.6011	0.426	0.553
Missing	2969	1137	4106	28%	-0.0301	0.000	0.553
Arrendada	1173	705	1878	38%	-0.4808	0.030	0.553
Familiar	3484	2124	5608	38%	-0.4950	0.097	0.553
Estado							
Estado1	1	0	1	0%	0.0000	0.000	0.043
Estado2	2118	525	2643	20%	0.4049	0.025	0.043
Estado3	61	19	80	24%	0.1765	0.000	0.043
Estado4	586	184	770	24%	0.1685	0.001	0.043
Estado5	2048	790	2838	28%	-0.0373	0.000	0.043
Estado6	5066	2014	7080	28%	-0.0675	0.002	0.043
Estado7	1287	526	1813	29%	-0.0951	0.001	0.043
Estado8	2	1	3	33%	-0.2968	0.000	0.043
Estado9	273	191	464	41%	-0.6327	0.013	0.043
Estado_Civil							
Casado	7374	2452	9826	25%	0.1112	0.008	0.030
Soltero	3113	1259	4372	29%	-0.0846	0.002	0.030
Viudo	143	60	203	30%	-0.1214	0.000	0.030
Otros	722	423	1145	37%	-0.4552	0.017	0.030
Divorciado	90	58	148	39%	-0.5505	0.003	0.030
Mop_Actual_Bancos							
a)Sin Experiencia	30	4	34	12%	1.0250	0.002	0.039
b)Mop_1	6523	2108	8631	24%	0.1397	0.010	0.039
c)Más de MOP1	2915	1456	4371	33%	-0.2957	0.026	0.039
Missing	1974	684	2658	26%	0.0700	0.001	0.039
EstConsultas							
a)Menor_3	3646	1088	4734	23%	0.2194	0.014	0.033
b)3_Consultas	1389	528	1917	28%	-0.0227	0.000	0.033

c)Más_3Consultas	5789	2484	8273	30%	-0.1438	0.011	0.033
Missing	618	152	770	20%	0.4127	0.008	0.033

Podemos observar que todas las variables candidatas cumplen con el criterio de *Information Value*, a la par, en las variables elegidas y que contienen *missings* el IV por *missing* nos corrobora que no les asigna mucho valor, lo que representa que podemos hacer uso de las categorías, sin temer porque implique un sesgo en el modelo.

7.4 Validación Cruzada

En esta sección el propósito es poner a competir diferentes técnicas de modelación, así como diferentes tratamientos de datos, previamente a este punto hemos realizado el tratamiento a través de WOE, es decir, sustituiremos el valor resultante por tramo de acuerdo con el que corresponda la variable. Por otra parte, utilizaremos el escalado, normalizado y sin tratar de las mismas variables que fueron relevantes por *Information Value.* Esto con motivo de revisar que tipo de tratamiento es más adecuado para poner en ejecución el score.

El *K-Fold Cross Validation*, lo realizamos a través de generar 5 iteraciones probando cada una de las técnicas estadísticas[39], con 10 diferentes grupos de manera aleatoria de los cuáles el 90% cada grupo dedicado a entrenar y 10% para realizar la prueba. La especificación de las variables dependientes, así como la variable objetivo fueron dadas por las siguientes ecuaciones.

$$Buenos_{WOE} = WOEConsultas + WOEScore + WOEEdad + WOERemanente + WOEEstGiro + WOEEstMunicipio + WOERatioFinancieras + WOETipoVivienda + WOEEstado + WOEEstadoCivil + WOEEMOPBANCOS$$

$$Buenos_{Esc} = EscaladoConsultas + EscaladoScore + EscaladoEdad + EscaladoRemamente + D^{40}EstGiro + DEstMunicipio + EscaladoRatioFinancieras + DTipoVivienda + DEstado + DEstadoCivil + DMOPBANCOS$$

$$Buenos_{Norm} = NormalizadoConsultas + NormalizadoScore + NormalizadoEdad + NormalizadoRemanente + DEstGiro + DEstMunicipio + NormalizadoRatioFinancieras + DTipoVivienda + DEstado + DEstadoCivil + DMOPBANCOS$$

[39] La modelación y los parámetros definidos para cada una de las técnicas estadísticas son detalladas en el anexo 3, a través del script realizado en R.

[40] D Representa una variable dummy, es decir una variable binaria toma valor de (0 ó 1) y revela si un evento sucede o no.

$$Buenos_{sintrat} = Consultas + Score + Edad + Remanente + DEstGiro + DEstMunicipio + RemanenteRatioFinancieras + DTipoVivienda + DEstado + DEstadoCivil + DMOPBANCOS$$

Para cada una de las iteraciones se calcula el error de clasificación del modelo que viene dado por la matriz de confusión:

		Pronostico	
		Bueno	Malo
Real	Bueno	VP (a)	FN (c)
	Malo	FP (d)	VN (b)

donde el error es categorizado como

$$Error = \frac{c + d}{a + b + c + d}$$

Metodología WOE									
Iteración	Logit	Árbol Decisión	Método Bayes	Knn	XG Boosting	Bosques Aleatorios	Adaboost	Red Neuronal	SVM
1	24.07	25.63	26.62	28.36	25.58	25.21	24.12	24.16	24.32
2	24.05	25.54	26.51	28.46	25.70	25.15	24.28	24.34	24.16
3	24.04	25.35	26.54	28.16	25.74	25.20	24.23	24.19	24.25
4	24.02	25.45	26.51	28.34	25.68	25.00	24.23	24.21	24.21
5	23.95	25.66	26.55	28.24	25.70	25.16	24.26	24.24	24.40
Promedio del Error	24.03	25.53	26.55	28.31	25.68	25.14	24.23	24.23	24.27
Coef. Var	0.002	0.005	0.002	0.004	0.002	0.003	0.003	0.003	0.004
Promedio Error/ (1- Coef. Var)	24.07	25.65	26.59	28.42	25.74	25.23	24.29	24.30	24.36
Promedio general				**25.329**					
Coef Var				**0.053**					
Promedio Error/ (1- CoefVar)				**26.747**					
Técnica con menor error				**Logit**					

Los resultados que se obtuvieron bajo el tratamiento de datos por vía WOE, fue en promedio un 25.32% de error bajo diferentes técnicas de modelación, mientras que obtuvimos un coeficiente de variación del 0.053 lo que implica que la muestra estable y no presenta gran volatilidad de error bajo este tratamiento, por lo que no se está ni subestimando o sobreestimando el error ya que independientemente de los datos que quedaron en la muestra de entreno en las diferentes iteraciones, no cambian de manera drástica el resultado final. Como un último dato, vemos que la técnica que genera menor error es la técnica logit, con un error promedio de 24.03% muy por debajo de la media y con el menor coeficiente de variación respecto a otras técnicas de modelación.

Metodología Escalamiento									
Iteración	Logit	Árbol Decisión	Método Bayes	Knn	XG Boosting	Bosques Aleatorios	Adaboost	Red Neuronal	SVM
1	24.35	25.74	28.18	29.94	25.18	24.40	24.11	25.08	25.46
2	24.39	25.87	28.55	29.82	25.15	24.54	24.18	25.01	25.32
3	24.45	25.83	27.69	30.05	25.49	24.62	24.50	25.02	25.51
4	24.37	25.88	27.85	30.17	25.21	24.54	24.57	24.73	25.52
5	24.36	25.99	27.92	29.89	25.35	24.45	24.59	25.09	25.42
Promedio del Error	24.38	25.86	28.04	29.97	25.28	24.51	24.39	24.99	25.45
Coef Var	0.002	0.003	0.012	0.005	0.006	0.004	0.009	0.006	0.003
Promedio Error/ (1-CoefVar)	24.42	25.95	28.38	30.11	25.42	24.60	24.62	25.14	25.53
Promedio general				**25.874**					
Coef Var				**0.071**					
Promedio Error/ (1- CoefVar)				**27.837**					
Técnica con menor error				**Logit**					

Al transformar las variables numéricas a través de escalamientos y utilizando *dummies* para variables categóricas, y someterlo a modelación de diferentes algoritmos, vemos que el error promedio bajo esta transformación de datos es de 25.87% con un coeficiente de variación de 0.071, lo que representa un aumento en la volatilidad del error respecto de la metodología WOE. Otro punto que llama la atención es que la técnica que, en promedio, menor error tiene, es la logit, además de que conserva un coeficiente de variación bajo, seguido del *adaboost*, sin embargo, este presenta una volatilidad más alta, indicando que depende en gran medida de la muestra con la que se entrene para poder predecir mejor o peor, lo que implicaría una sobreestimación o subestimación del error si es que usáramos este algoritmo.

Metodología Normalizado									
Iteración	Logit	Arbol Decisión	Método Bayes	Knn	XG Boosting	Bosques Aleatorios	Adaboost	Red Neuronal	SVM
1	24.51	25.78	28.17	30.23	25.34	24.79	24.30	25.60	25.06
2	24.50	25.92	27.53	30.24	25.50	24.73	24.35	25.07	25.21
3	24.46	25.97	27.64	30.35	25.38	24.81	24.93	24.88	25.05
4	24.48	25.92	28.19	30.17	25.60	24.89	24.62	25.25	25.48
5	24.43	26.03	28.53	29.83	25.08	24.78	24.39	24.89	25.04
Promedio del Error	24.48	25.92	28.01	30.16	25.38	24.80	24.52	25.14	25.17
Coef Var	0.001	0.004	0.015	0.007	0.008	0.002	0.011	0.012	0.007
Promedio Error/ (1-CoefVar)	24.51	26.02	28.43	30.36	25.58	24.86	24.78	25.44	25.35
Promedio general				**25.954**					
Coef Var				**0.071**					
Promedio Error/ (1- CoefVar)				**27.924**					
Técnica con menor error				**Logit**					

Para la transformación de variables, utilizando la normalización de datos y sometiéndola a diferentes técnicas de modelación y algoritmos, obtenemos que, el error promedio bajo este tratamiento es de 25.95%, con un coeficiente de variación cercano a 0.07, lo que lo posiciona aun por detrás de la metodología *WOE* y el escalado de datos, algo que volvemos a notar es que el algoritmo predominante para esta base de datos, es la técnica logit, seguido de igual manera por el *adaboost*, sin embargo podemos notar el mismo patrón en donde es un error por debajo de la media pero con un coeficiente de variación alto, lo que reafirma que el *adaboost*, al menos con esta base de datos, depende en gran medida de la muestra con que se entrene.

Metodología Sin Tratar									
Iteración	Logit	Arbol Decisión	Método Bayes	Knn	XG Boosting	Bosques Aleatorios	Adaboost	Red Neuronal	SVM
1	24.24	25.99	27.71	29.73	25.21	24.43	24.26	35.23	25.51
2	24.43	25.98	27.62	29.80	25.15	24.56	24.70	44.85	25.58
3	24.24	25.92	28.39	29.95	25.21	24.84	24.61	45.78	25.87
4	24.34	25.79	27.53	29.78	25.39	24.65	24.43	44.86	25.19
5	24.35	25.99	28.64	29.60	25.18	24.50	24.44	49.54	25.65
Promedio del Error	24.32	25.94	27.98	29.77	25.23	24.60	24.49	44.05	25.56
Coef Var	0.003	0.003	0.018	0.004	0.004	0.006	0.007	0.120	0.010
Promedio Error/ (1-CoefVar)	24.40	26.02	28.49	29.90	25.32	24.76	24.66	50.07	25.81
Promedio general				**27.992**					
Coef Var				**0.222**					
Promedio Error/ (1- CoefVar)				**35.966**					
Técnica con menor error				**Logit**					

Por último empleamos las diferentes metodologías con una base sin ningún tipo de tratamiento para variables numéricas, y tratando a las variables categóricas como *dummies*, lo que observamos, es un error mucho más pronunciado que en datos transformados con casi un 28% de error en promedio, con un coeficiente de variación de 0.222 lo que representa una volatilidad exagerada en comparación con los resultados obtenidos anteriormente, algo que llama la atención, es que la red neuronal fue la que más retroceso tuvo en el performance, teniendo grados más altos de error, respecto a otras metodologías. Aunque de igual manera observamos que la técnica que menor error posee es la técnica Logit.

7.5 Entrenamiento del Modelo

Una vez que, bajo diferentes metodologías y tratamientos de datos, hemos calculado el error promedio que se tiene por técnica estadística o algoritmo utilizado, y la volatilidad de dicho error a través del coeficiente de variación, en los 180 modelos que se pusieron en marcha, el que cumple con menos volatilidad y error promedio es la técnica logit, con tratamiento de datos tipo WOE, por lo que procederemos a realizar la modelación de manera robusta, es decir que no utilizaremos todas las variables tipo WOE, sino que

rechazaremos aquellas que estadísticamente no tengan sentido, sin temor a que los resultados dependan de la aleatoriedad de la muestra. Tomamos como muestra entreno el 80% de los datos y como prueba el 20% restante. Nuestra estimación econométrica, inicialmente estará dada por la siguiente ecuación:

$$Buenos_{WOE} = WOEConsultas + WOEScore + WOEEdad + WOERemanente + WOEGiro + WOEEstMunicipio + WOERatioFinancieras + WOETipoVivienda + WOEEstEstado + WOEEstadoCivil + WOEEMOPBANCOS$$

Obteniendo los siguientes resultados:

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significancia
(Intercept)	1.00811	0.02335	43.178	2E-16	***
WOEConsultas	0.43348	0.13123	3.303	0.000956	***
WOEScore	0.55927	0.09352	5.98	2.23E-09	***
WOEEdad	0.82714	0.07947	10.408	2E-16	***
WOERemanente	0.49129	0.13028	3.771	0.000162	***
WOEEstGiro	0.93376	0.06253	14.932	2E-16	***
WOEEstMunicipio	0.88665	0.04011	22.103	2E-16	***
WOERatioFinancieras	0.64165	0.12265	5.231	1.685E-07	***
WOETipoVivienda	0.94924	0.03551	26.731	2E-16	***
WOEEstado	0.1739	0.11528	1.508	0.131438	
WOEEstadoCivil	0.57031	0.13073	4.363	1.282E-05	***
WOEEMOPBANCOS	0.43413	0.13394	3.241	0.00119	**

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

AIC 12037

En primera instancia, observamos que los coeficientes de las variables dependientes tienen el mismo signo, esto revela coherencia en la regresión, puesto que bajo el tratamiento de datos WOE, la WOE intrínsecamente carga la relación con respecto de la variable independiente, por lo que ya no es necesario tener hipótesis de signos esperados en las betas. Un segundo punto para observar es que la única variable que no salió significativa es la variable WOEEstado, con un p-value mayor a 0.05, esto principalmente se debe a que está altamente correlacionada con la WOEEstMunicipio, representando el atributo geográfico. En última instancia vemos que el criterio AIC es de 12037, el cual representa que no hay gran sobreajuste en el modelo. Por lo que realizamos una segunda especificación, dada la invalidez de la variable WOEEstado:

$$Buenos_{WOE} = WOEConsultas + WOEScore + WOEEdad + WOERemanente + WOEGiro + WOEEstMunicipio + WOERatioFinancieras + WOETipoVivienda + WOEEstadoCivil + WOEEMOPBANCOS$$

Obteniendo los siguientes resultados

Variable	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.00799	0.02334	43.18	2E-16	***
WOEConsultas	0.4362	0.13119	3.325	0.000884	***
WOEScore	0.5619	0.0935	6.01	1.85E-09	***

WOEEdad	0.8259	0.07945	10.395	2E-16	***
WOERemanente	0.5065	0.1299	3.899	9.622E-05	***
WOEGiro	0.9357	0.06256	14.957	2E-16	***
WOEMunicipio	0.9022	0.03877	23.273	2E-16	***
WOERatioFinancieras	0.6503	0.12252	5.308	1.108E-07	***
WOETipoVivienda	0.9480	0.0355	26.707	2E-16	***
WOEEstadoCivil	0.5789	0.13053	4.435	9.201E-06	***
WOEEMOPBANCOS	0.4264	0.13381	3.187	0.001436	**

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

AIC 12037

Constatamos que todas las variables tienen significancia y que el signo de las betas es el mismo en todas, por lo que consideraremos los resultados de esta regresión como la mejor para ejecutar

7.6 Validación del Desempeño (Performance)

En esta sección mediremos bajo diferentes estadísticos la capacidad predictiva de nuestro modelo, el primer performance a utilizar es a través de la matriz de confusión, en donde se considera como un cliente "Bueno", a aquél que en score de que sea bueno, sea mayor a 500, esta probabilidad deriva del modelo anterior expresado en la siguiente ecuación:

$$Z = 1.0799 + 0.4362 * WOEConsultas + 0.56193 * WOEScore + 0.8259 * WOEEdad + 0.5065 * WOERemanente + 0.9357 * WOEGiro + 0.9022 * WOEEstMunicipio + 0.6503 * WOERatioFinancieras + 0.9480 * WOETipoVivienda + 0.5789 * WOEEstadoCivil + 0.4264 * WOEEMOPBANCOS$$

$$Score = \left(\frac{e^z}{1 + e^z}\right) * 1000$$

Teniendo como resultados por la aplicación de la ecuación en la base de datos general lo siguiente:

	Predicción	
	Buenos	Malos
Buenos	10510	932
Malos	2831	1421

Precisión Global	$\frac{VN+VP}{VN+VP+FP+FN}$	**76.02%**
Error	$\frac{FN+FP}{VN+VP+FP+FN}$	**23.98%**
Precisión Positiva (Sensibilidad)	$\frac{VP}{VP+FN}$	**33.42%**
Precisión Negativa (Especificidad)	$\frac{VN}{VN+FP}$	**91.85%**
Asertividad Positiva	$\frac{VP}{}$	**60.39%**

	VP+FP	
Asertividad Negativa	$\frac{VN}{VN+FN}$	**78.78%**
Falsos Negativos	2831	**66.58%**
Falsos Positivos	932	**8%**

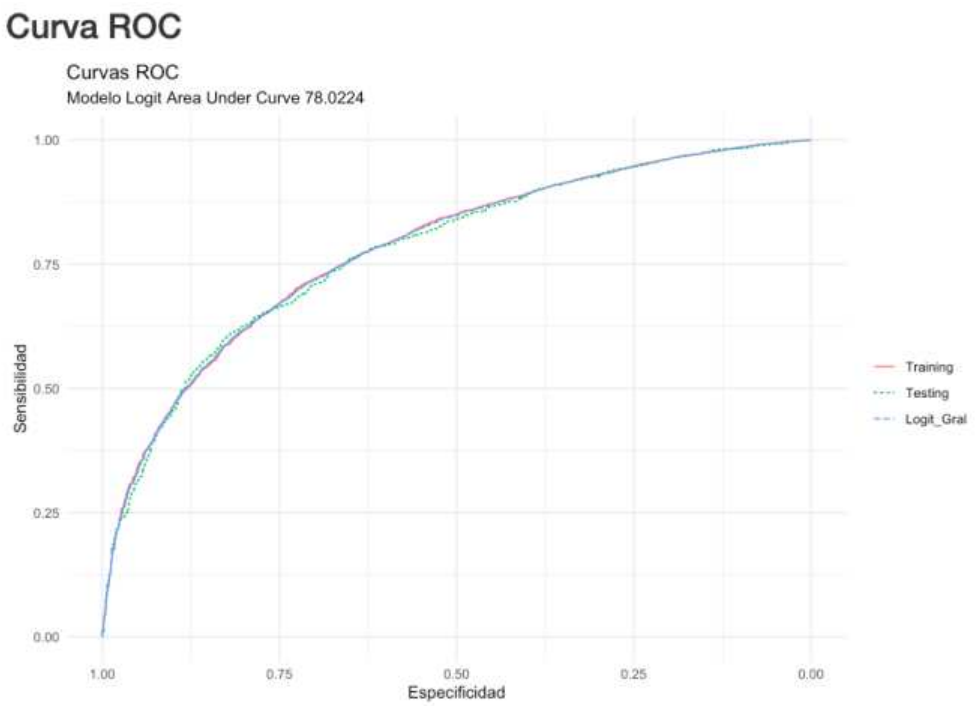
Vemos que el error no dista de lo que a través de las *KFold Cross Validation* nos arrojó inicialmente, lo que nos valida la coherencia de las iteraciones, a la par, también mostramos lo obtenido en la muestra de prueba.

	Predicción	
	Buenos	Malos
Buenos	2046	194
Malos	582	316

Precisión Global	$\frac{VN+VP}{VN+VP+FP+FN}$	**75.27%**
Error	$\frac{FN+FP}{VN+VP+FP+FN}$	**24.73%**
Precisión Positiva (Sensibilidad)	$\frac{VP}{VP+FN}$	**35.19%**
Precisión Negativa (Especificidad)	$\frac{VN}{VN+FP}$	**91.34%**
Asertividad Positiva	$\frac{VP}{VP+FP}$	**61.96%**
Asertividad Negativa	$\frac{VN}{VN+FN}$	**77.85%**
Falsos Negativos	582	**64.81%**
Falsos Positivos	194	**9%**

Dichos resultados son similares, esto nos da pauta a que hay estabilidad en el modelo y que este no depende en gran medida de la muestra con la que se entrene. Algo que remarcamos es que existe un alta especificidad en el modelo, sin embargo no existe una remarcada sensibilidad, que grosso modo representa que, el modelo, dado el punto de corte en 500 la regresión alcanza a captar muy bien a los clientes "Buenos", sin embargo, hay un alto margen de error al captar a los clientes "Malos" esto principalmente se da, porque como lo mencionamos al inicio del proceso de modelación, consideramos que era una muestra desbalanceada, es decir, que no existe la misma proporción de "clases" para el entreno de un modelo.

Esto de gran relevancia, ya que por mucho que se tenga un alto performance, a nivel de toma de decisiones esto representa millones de pesos en *exposure* con un alto grado de riesgo de incobrabilidad, puesto que no detectamos de manera clara a aquellos clientes que no nos pagarán, esto se debe corregir a través de un *cutoff* que nos permita aumentar la sensibilidad y la especificidad, aunque perdamos precisión global.



Otra manera de calcular la predicción de un modelo es a través de la curva ROC (*Receiver Operating Characteristic*) y la AUROC (*Area under ROC curve*) el cuál consideramos los siguientes valores para el nivel de discriminación:

Puntaje ROC	Accuracy
.90-1	Excelente (A)
.80-.90	Bueno (B)
.70-.80	Justo (C)
.60-.70	Pobre (D)
.50-.60	Modelo sin predicción (F)

Por lo que nuestro modelo está en la banda de "justo", sin embargo, está muy cercano de convertirse en "bueno", esto indica que tiene buen poder de predicción. De manera que, siguiendo la siguiente identidad, (Lyn Thomas, Jonathan Crook) obtenemos el GINI:

$$GINI = 2AUROC - 1$$

Por lo tanto

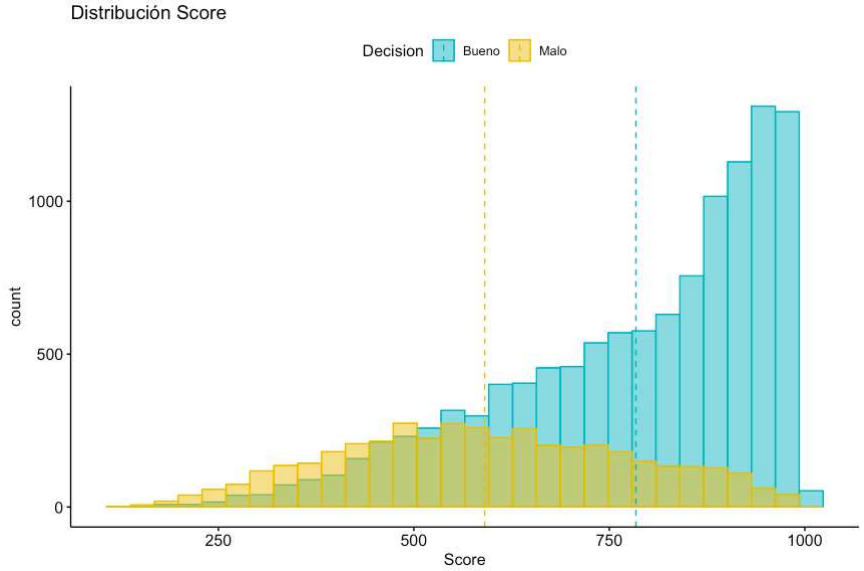
$$GINI = 0.5604$$

Lo que también indica un alto poder de predicción, por último, procederemos a revisar a través de dividir por deciles los resultados del score.

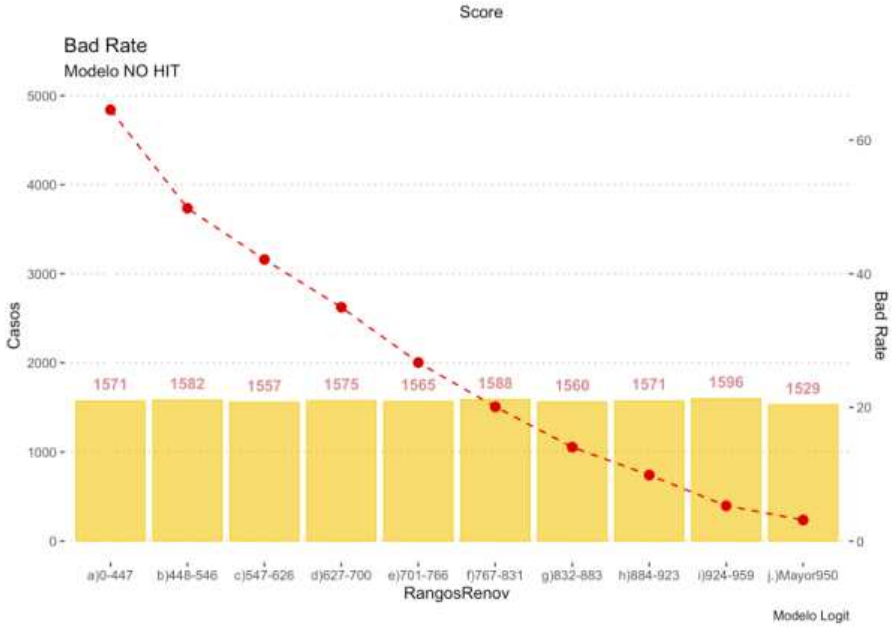
Rangos Score	Total Buenos	Total Malos	Total	Bad Rate	AcumB	DistB	DistM	odds	IV	IVAcum	KS	KS2	MAXIV
a)0-447	557	1014	1571	65%	11442	0.05	0.24	0.20	0.30	1.16	0.00	0.42	1.16
b)448-546	794	788	1582	50%	10885	0.07	0.19	0.37	0.11	0.86	0.19	0.42	1.16
c)547-626	901	656	1557	42%	10091	0.08	0.15	0.51	0.05	0.74	0.31	0.42	1.16
d)627-700	1024	551	1575	35%	9190	0.09	0.13	0.69	0.01	0.69	0.38	0.42	1.16
e)701-766	1147	418	1565	27%	8166	0.10	0.10	1.02	0.00	0.68	0.42	0.42	1.16
f)767-831	1269	319	1588	20%	7019	0.11	0.08	1.48	0.01	0.68	0.42	0.42	1.16
g)832-883	1341	219	1560	14%	5750	0.12	0.05	2.28	0.05	0.66	0.38	0.42	1.16

h)884-923	1416	155	1571	10%	4409	0.12	0.04	3.39	0.11	0.61	0.32	0.42	1.16
i)924-959	1512	84	1596	5%	2993	0.13	0.02	6.69	0.21	0.50	0.23	0.42	1.16
j.)Mayor950	1481	48	1529	3%	1481	0.13	0.01	11.47	0.29	0.29	0.12	0.42	1.16

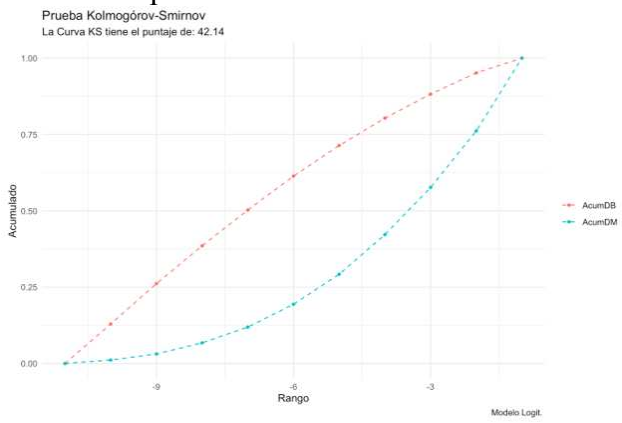
Podemos observar que existe armonía en la distribución del score, esto significa que conforme va aumentando el score, el Bad Rate va disminuyendo y no se pierde lógica en ninguno de los deciles, derivado del cálculo de la prueba Kolmogrov Smirnov (KS), podemos detectar que es de 42 puntos, lo cual indica un grado alto de separación.



Por otra parte observamos que la media del score clasificado como "Malo" es cercano a los 550 puntos, mientras que la media del score clasificado como "Bueno" es cercano a los 760 puntos, esto representa que las distribuciones de buenos y malos están muy bien separadas y lo vemos a través del histograma, en donde observamos que del lado izquierdo, es decir los scores más bajos, concentran las observaciones clasificadas como "Malas", mientras que los "Buenos" se concentran en scores altos.



Por otra parte en el gráfico de *Bad Rate,* observamos gráficamente que existe armonía en el score, lo que nos permite tener un margen de identificación adecuado por decil el nivel de riesgo de las observaciones.
En última instancia vemos de forma gráfica el estadístico de separación KS, en donde observamos la intensidad de la separación de la variable.



Cuyo puntaje según el rango de decil es de 42.14 puntos, lo cual es indicativo de un buen performance para un *scoring* reactivo. De manera que utilizando los indicadores de performance más comunes dentro de la minería de datos, podemos decir que el modelo cuenta con el performance suficiente para ponerse en producción.

8. Modelo NO HIT

8.1 Validación de Base de Datos

Como habíamos mencionado, existe un segundo grupo a evaluar el cuál, es el que no cumple con las características para ser calificado bajo un *score* genérico, es decir que no tienen ningún tipo de experiencia crediticia previa, que no ha tenido ningún crédito activo en los últimos 6 meses o que poseen cuentas que califiquen para clasificarlos como HIT y generen score. Retomaremos la clasificación y el detalle de las variables del modelo HIT, ya que no tienen distinción y provienen de la misma base de datos. Presentamos a continuación la tabla que contiene de acuerdo con nuestros criterios de definición de Bueno, Malo e Indeterminado que son sujetos a modelación

Buenos (A)	Malos (B)	Indeterminados (C)	Total (A+B+C)	Califican Modelo D= (A+B)	Bad Rate B/D
7,820	2,026	344	10,190	9,846	19.88%

Tabla 42. Análisis de BD Cuantitativos NO HIT

Variable	Missing	Completos	Conteo	Media	SD	PC0	P25	PC50	PC75	PC100	Tasa Missing	CoefVar
Edad	0	10190	10190	37.6983317	11.82942476	19	27	36	46	103	0%	31%
Ingreso_Mensual	0	10190	10190	25422.03729	22861.47053	2320	14560	20400	29600	980000	0%	90%
Total_Egresos	0	10190	10190	4428.556527	6075.269095	-1368	2380	3500	5100	469000	0%	137%
Total_Ingresos	0	10190	10190	10767.78891	9260.747052	-70100	6872	9392	12444	524000	0%	86%
Hipoteca	0	10190	10190	6454.486555	39003.37372	0	0	0	9000	2970500	0%	**604%**
Remanente	0	10190	10190	7876.614325	6090.774009	**-6594**	4552	6823.5	9201.5	128855	0%	77%
Inventario	0	10190	10190	8933.144161	30768.6033	0	440	2707	8543.75	1518800	0%	**344%**

Datos con error en la información (outliers sin sentido)

Insuficiencia de datos en la variable

Alto grado de missings en la variable

Desviación estándar dos o veces más grande promedio

Observamos que existen menos variables, esto derivado que las anteriores provenían del reporte de crédito y podrían tener relevancia a la hora de realizar el modelaje.

Al igual que en la tabla de análisis de Base de Datos cuantitativos, vemos una disminución

Variable	Missing	Completos	Conteo	% Missing	Min_ longitud	Máximo Longitud	Únicos
Dependientes	0	10190	10190	0%	1	15	13
Tipo_Negocio	0	10190	10190	0%	4	14	3
Tipo_Vivienda	0	10190	10190	0%	9	19	4
Giro	0	10190	10190	0%	6	100	449
Estado_civil	17	10173	10190	0%	1	1	5
Sexo	0	10190	10190	0%	1	1	2
Avales	17	10173	10190	0%	1	1	4
Consultas	4417	5773	10190	43%	1	2	19
Colonia	0	10190	10190	0%	4	49	1610
Municipio	0	10190	10190	0%	4	36	299
Estado	0	10190	10190	0%	6	31	9


Alto grado de missings en la variable

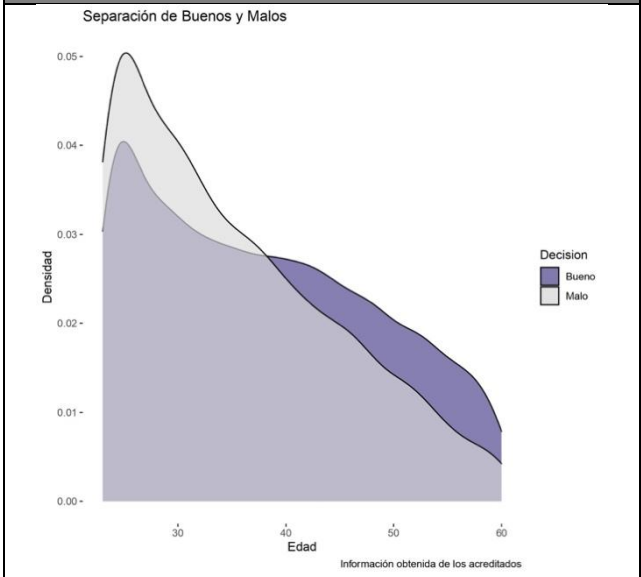
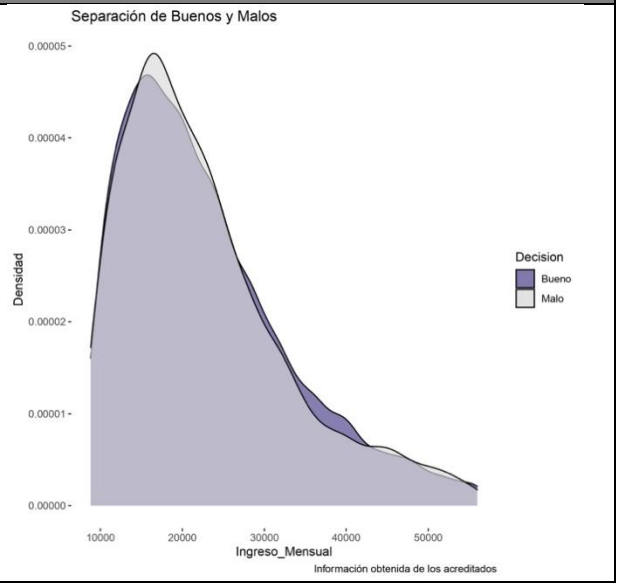
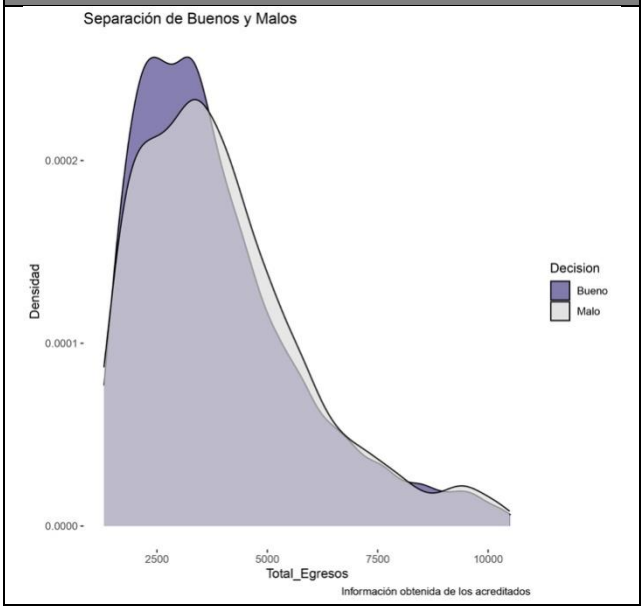
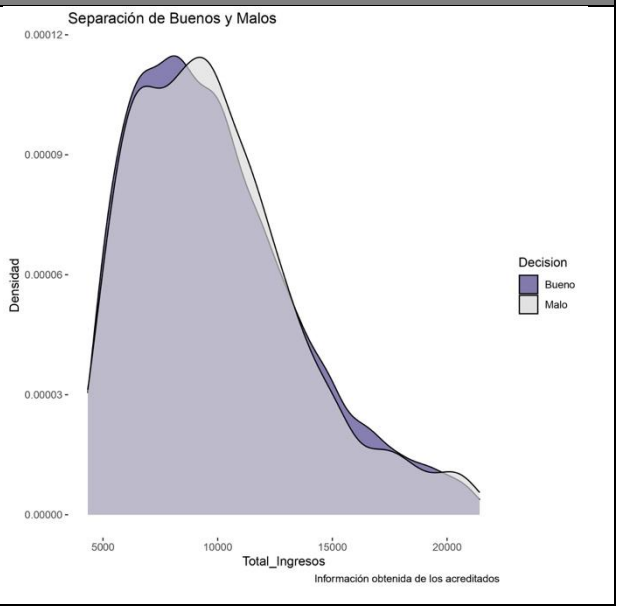
considerable en las variables categóricas, esto derivado de que no se tiene información de buró que nos pueda ser de ayuda. También se encuentran las siguientes inconsistencias en la base de datos:

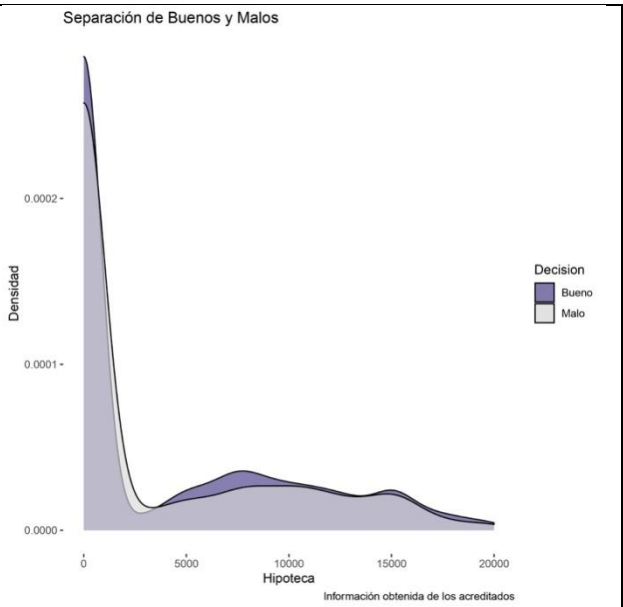
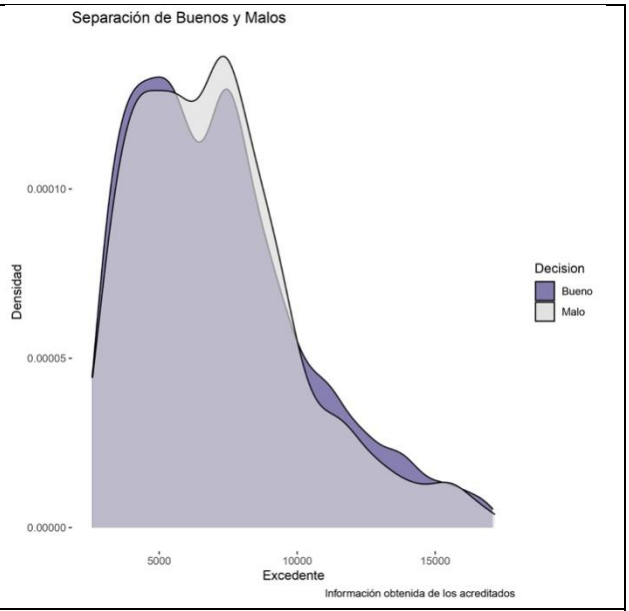
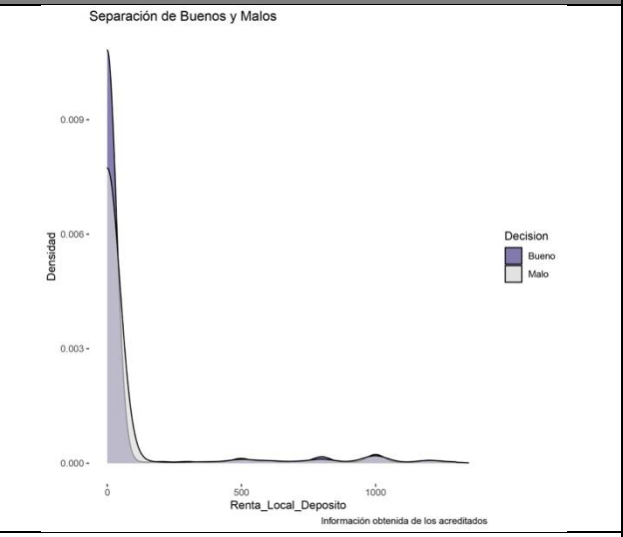
Variable	Inconsistencia
Edad	La edad por política de crédito no es menor a 18 años y no mayor a 70, por lo que un dato en PC100 = 103, es un error de base de datos.
Total_Egresos	Por política de crédito no se registran valores negativos
Total_Ingresos	Por política de crédito no se registran valores negativos
Remanente	Por política de crédito no se registran valores negativos

Vemos que en esta base de datos no se encuentran variables con un alto grado de *missings* a excepción de la variable "Consultas" la cual sólo contiene el 57% de las observaciones completas.

8.2 Análisis de la Base de Datos

Retomamos el análisis y la explicación del análisis gráfico para el modelo HIT, incluyendo de igual manera los estadísticos usados como la *SommersD'*.

<table>
<tr><th colspan="2">Variable: Edad</th><th colspan="2">Variable: Ingreso Mensual</th></tr>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 1</td><td colspan="2">Número de Cortes: +3</td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.1977</td><td>Missings: 0%</td><td>SommersD: -0.0023</td></tr>
<tr><td>PC50: 36</td><td>Promedio: 37</td><td>PC50: $20,400</td><td>Promedio: $25,422</td></tr>
<tr><th colspan="2">Variable: Total Egresos</th><th colspan="2">Variable: Total Ingresos</th></tr>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 2</td><td colspan="2">Número de Cortes: 2</td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.0490</td><td>Missings: 0%</td><td>SommersD: -0.005</td></tr>
<tr><td>PC50: $3,600</td><td>Promedio: $4,428</td><td>PC50: $9,392</td><td>Promedio: $10,767</td></tr>
<tr><th colspan="2">Variable: Hipoteca</th><th colspan="2">Variable: Remanente</th></tr>
</table>

<table>
<tr><td colspan="2"> 
</td><td colspan="2"> 
</td></tr>
<tr><td colspan="2">Número de cortes: 2</td><td colspan="2">Número de cortes: 1</td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.4268</td><td>Missings: 0%</td><td>SommersD: -.0016</td></tr>
<tr><td>PC50: $0</td><td>Promedio: $6,454</td><td>PC50: $6,823</td><td>Promedio: $7,876</td></tr>
<tr><td colspan="2">Variable: Inventario</td><td colspan="2"></td></tr>
<tr><td colspan="2"> 
</td><td colspan="2"></td></tr>
<tr><td colspan="2">Número de cortes: +3</td><td colspan="2"></td></tr>
<tr><td>Missings: 0%</td><td>SommersD: -0.0964</td><td></td><td></td></tr>
<tr><td>PC50: $2,707</td><td>Promedio: $8,933</td><td></td><td></td></tr>
</table>

Bajo la metodología gráfica propuesta, acompañada del *SommersD'* consideramos que existen muy pocas variables para que sean consideradas dentro del modelo, en nuestro caso para el modelo NO hit, sólo la variable edad cuenta con las condiciones estadísticas gráficas, debido a que tiene una *SommersD'* medianamente fuerte, no existe mucha diferencia entre la PC50 y el Promedio, lo que indica que la variable no está sesgada.

8.3 Selección y Transformación de Datos

En esta sección retomaremos el análisis WOE, con el tramo equiprobable correspondiente, acompañado de su indicador de separación de variables (*Information Value*), de las cuáles depuraremos las que tengan un IV menor a 0.02.

Rango	Total Buenos	Total Malos	Total	Bad Rate	WOE	IV	MAXIV
Causa de No Generar Score							
a)Expediente SinCuentas	5759	1487	7246	21%	0.003	0.0000	0.0004
b)Expediente Sin6Meses sinactualizar	2053	538	2591	21%	-0.011	0.0000	0.0004
missing	8	1	9	11%	0.729	0.0004	0.0004
Edad							
a)18-25	1375	487	1862	26%	-0.313	0.0202	0.0919
b)26-30	1213	400	1613	25%	-0.241	0.0102	0.0919
c)31-36	1246	365	1611	23%	-0.123	0.0026	0.0919
d)37-43	1348	332	1680	20%	0.051	0.0004	0.0919
e)44-51	1278	248	1526	16%	0.289	0.0119	0.0919
f)Más de 51	1360	194	1554	12%	0.597	0.0466	0.0919
Est Giro							
h)Muybajoriesgo	1209	88	1297	7%	1.270	0.1411	0.2764
g)BajoRiesgo	1038	171	1209	14%	0.453	0.0219	0.2764
f)Bajoaltoriesgo	984	205	1189	17%	0.218	0.0054	0.2764
e)MedioBajoRiesgo	906	225	1131	20%	0.042	0.0002	0.2764
d)MedioRiesgo	1162	317	1479	21%	-0.052	0.0004	0.2764
c)MedioaltoRiesgo	959	310	1269	24%	-0.221	0.0067	0.2764
b)AltoRiesgo	778	269	1047	26%	-0.289	0.0096	0.2764
a)MuyAltoRiesgo	784	441	1225	36%	-0.775	0.0910	0.2764
Est_Mun							
h)Muybajoriesgo	1199	35	1234	3%	2.183	0.2970	0.5579
g)BajoRiesgo	1151	131	1282	10%	0.823	0.0679	0.5579
f)Bajoaltoriesgo	990	181	1171	15%	0.349	0.0130	0.5579
e)MedioBajoRiesgo	1006	239	1245	19%	0.087	0.0009	0.5579
d)MedioRiesgo	1385	423	1808	23%	-0.165	0.0052	0.5579
c)MedioaltoRiesgo	557	199	756	26%	-0.321	0.0087	0.5579
b)AltoRiesgo	822	348	1170	30%	-0.491	0.0327	0.5579
a)MuyAltoRiesgo	710	470	1180	40%	-0.938	0.1325	0.5579
Tipo de Vivienda							
Propia	2422	114	2536	4%	1.706	0.4323	0.5446
Missing	1794	462	2256	20%	0.006	0.0000	0.5446
Arrendada	996	389	1385	28%	-0.410	0.0265	0.5446
Familiar	2608	1061	3669	29%	-0.451	0.0858	0.5446
Estado							
Estado1	1	0	1	0%	0.000	0.0000	0.0471
Estado2	1	0	1	0%	0.000	0.0000	0.0471

Estado3	1312	212	1524	14%	0.472	0.0298	0.0471
Estado4	43	7	50	14%	0.465	0.0009	0.0471
Estado5	933	209	1142	18%	0.145	0.0023	0.0471
Estado6	237	56	293	19%	0.092	0.0002	0.0471
Estado7	3852	1093	4945	22%	-0.091	0.0043	0.0471
Estado8	1292	386	1678	23%	-0.143	0.0036	0.0471
Estado9	149	63	212	30%	-0.490	0.0059	0.0471
Estado Civil							
missing	17	0	17	0%	0.00	0.00	0.03
V	78	13	91	14%	0.44	0.00	0.03
C	4384	998	5382	19%	0.13	0.01	0.03
O	711	181	892	20%	0.02	0.00	0.03
S	2573	810	3383	24%	-0.19	0.01	0.03
D	57	24	81	30%	-0.49	0.00	0.03
Consultas							
a)1_consulta	2258	564	2822	20%	0.037	0.000	0.067
b)2_Consultas	955	330	1285	26%	-0.288	0.012	0.067
c)3_Consultas	469	174	643	27%	-0.359	0.009	0.067
d)Más_4Consultas	601	246	847	29%	-0.457	0.020	0.067
missing	3537	712	4249	17%	0.252	0.025	0.067
Sexo							
F	4857	1060	5917	18%	0.172	0.017	0.039
M	2963	966	3929	25%	-0.230	0.023	0.039

Vemos que todas las variables cuentan con un *Information Value* mayor a 0.02, sin embargo, dejamos la causa de no generación de *Score* para ver sus implicaciones dentro del modelo, pero dictaminamos sumamente débil para incluirla en el modelo, ya que no alcanza a separar una variable de forma adecuada, además observamos que la mayoría de las variables que tienen algún tipo de *missing*, esa categoría no les agrega mucho valor al *Information Value*, de manera que podemos trabajar con ellas.

8.4 Validación Cruzada NO HIT

De igual manera, a través de *K-Fold Cross Validation*, lo realizamos a través de generar 5 iteraciones probando cada una de las técnicas estadísticas[41], con 10 diferentes grupos de manera aleatoria de los cuáles el 90% cada grupo dedicado a entrenar y 10% para realizar la prueba. Pondremos a competir diferentes técnicas estadísticas con el fin de determinar cuál es la más apropiada para trabajar la base de datos, la cual especificamos a continuación.

$$Buenos_{WOE} = WOEEdad + WOEEstGiro + WOEEstMunicipio + WOETipoVivienda + WOEEstado + WOEEstadoCivil + WOEConsultas + WOESexo$$

$$Buenos_{WOE} = EscaladoEdad + DEstGiro + DEstMunicipio + DTipoVivienda + DEstado + DEstadoCivil + EscaladoConsultas + DSexo$$

[41] La modelación y los parámetros definidos para cada una de las técnicas estadísticas son detalladas en el anexo II, a través del script realizado en R.

$$Buenos_{Norm} = NormalizadoEdad + DEstGiro + DEstMunicipio + DTipoVivienda + DEstado + DEstadoCivil + NormalizadoConsultas + DSexo$$

$$Buenos_{sintrat} = Edad + DEstGiro + DEstMunicipio + DTipoVivienda + DEstado + DEstadoCivil + Consultas + DSexo$$

Y obtuvimos los siguientes resultados

Metodología WOE									
Iteración	Logit	Arbol Decisión	Método Bayes	Knn	XG Boosting	Bosques Aleatorios	Adaboost	Red Neuronal	SVM
1	19.51	20.58	25.77	22.64	20.57	19.76	19.44	19.66	19.78
2	19.54	20.58	25.87	22.37	20.54	19.73	19.77	19.96	19.70
3	19.57	20.58	25.93	22.97	20.81	19.95	19.89	19.69	19.68
4	19.56	20.58	26.00	22.68	20.53	19.88	19.41	19.75	19.82
5	19.50	20.58	25.91	22.71	20.69	19.50	19.54	19.44	19.86
Promedio del Error	19.54	20.58	25.90	22.68	20.63	19.76	19.61	19.70	19.77
Coef Var	0.002	0.000	0.003	0.009	0.006	0.009	0.011	0.009	0.004
Promedio Error/ (1- CoefVar)	19.57	20.58	25.98	22.89	20.75	19.94	19.82	19.89	19.84
Promedio general				**20.906**					
Coef Var				**0.097**					
Promedio Error/ (1- CoefVar)				**23.144**					
Técnica con menor error				**Logit**					

Vemos de primera instancia que con respecto al modelo Hit, este modelo contiene un error menor de clasificación, vemos también que continua la técnica logit como el mejor desempeño, puesto que en promedio tiene 19.54% de error en sus 5 iteraciones, mientras que el *Adaboost* está por encima del *logit* sólo por 7 puntos base, sin embargo la desventaja de *Adaboost* es que tiene un alto coeficiente de variación, esto representa que la desviación es más pronunciada que el promedio, esto se debe a que en esa técnica está dependiendo mucho la muestra de entreno, por lo que nosotros no buscamos eso, sino que sea estable en cualquier momento.

Metodología: Escalamiento									
Iteración	Logit	Arbol Decisión	Método Bayes	Knn	XG Boosting	Bosques Aleatorios	Adaboost	Red Neuronal	SVM
1	19.65	20.58	59.69	23.18	21.04	19.68	19.84	19.84	20.18
2	19.57	20.58	60.32	23.20	20.92	19.70	20.01	26.24	20.48
3	19.59	20.58	59.82	23.15	20.67	19.68	20.00	20.05	20.18
4	19.73	20.58	60.11	23.63	20.68	19.71	19.81	20.13	20.41
5	19.71	20.58	60.19	22.93	20.81	19.57	20.04	19.99	20.20
Promedio del Error	19.65	20.58	60.02	23.22	20.82	19.67	19.94	21.25	20.29
Coef Var	0.004	0.000	0.004	0.011	0.008	0.003	0.005	0.131	0.007

Promedio Error/ (1- CoefVar)	19.72	20.58	60.29	23.48	20.99	19.73	20.05	24.46	20.43
Promedio general				**25.049**					
Coef Var				**0.502**					
Promedio Error/ (1- CoefVar)				**50.311**					
Técnica con menor error				**Logit**					

Bajo el tratamiento de datos, utilizando el escalamiento en variables continuas podemos observar que se dispara el error promedio de todas las técnicas en manera agregada, esto se debe a que, bajo probabilidad Bayesiana, lo que encontramos un grave error de clasificación, algo que no sucedía bajo la metodología WOE, esto lo contrastamos a través del coeficiente de variación, el cuál subió de manera drástica, debido al mismo fenómeno. Aunado a esto, podemos observar que la técnica con menor error es la técnica logit.

Metodología: Normalizado									
Iteración	Logit	Árbol Decision	Método Bayes	Knn	XG Boosting	Bosques Aleatorios	Adaboost	Red Neuronal	SVM
1	19.53	20.58	60.34	23.44	20.58	19.63	19.83	26.28	20.16
2	19.57	20.58	60.32	23.50	21.01	19.68	20.28	20.32	20.33
3	19.45	20.58	60.03	23.48	20.98	19.87	20.00	20.77	20.41
4	19.47	20.58	59.90	23.33	20.92	19.76	20.07	20.20	20.42
5	19.57	20.58	59.82	23.37	20.73	19.79	20.06	26.45	20.22
Promedio del Error	19.52	20.58	60.08	23.42	20.85	19.75	20.05	22.81	20.31
Coef Var	0.003	0.000	0.004	0.003	0.009	0.005	0.008	0.143	0.006
Promedio Error/ (1- CoefVar)	19.58	20.58	60.32	23.50	21.03	19.84	20.21	26.61	20.43
Promedio general				**25.262**					
Coef Var				**0.497**					
Promedio Error/ (1- CoefVar)				**50.219**					
Técnica con menor error				**Logit**					

Cuando analizamos los diferentes métodos estadísticos empleados en la búsqueda del mejor modelo a través de normalizar las variables continuas de los modelos, encontramos que el error incrementó ligeramente con respecto al tratamiento de escalamiento y dista mucho al compararse con la metodología WOE. Por otra parte, algo que es conveniente resaltar, es que el coeficiente de variación disminuyó con respecto al escalado de datos, esto representa que la variación del error en las diferentes iteraciones con respecto de la media disminuyó. Por último, mencionaremos que la técnica con menor error registrado vuelve a ser, bajo la modelación tipo "Logit", la cual registra un error promedio de 19.52%

Metodología: Sin tratar									
Iteración	Logit	Arbol Decisión	Método Bayes	Knn	XG Boosting	Bosques Aleatorios	Adaboost	Red Neuronal	SVM
1	19.88	20.58	60.14	23.22	21.14	20.02	20.13	26.85	20.11
2	19.88	20.58	60.68	23.36	21.26	20.05	20.30	20.33	20.38

3	19.95	20.58	60.78	23.44	21.08	20.02	20.33	25.89	20.40
4	19.89	20.58	60.76	23.75	21.06	20.20	20.27	31.42	20.23
5	19.95	20.58	60.50	23.31	21.19	19.90	19.96	32.09	20.39

Promedio del Error	19.91	20.58	60.57	23.41	21.15	20.04	20.20	27.32	20.30
Coef Var	0.002	0.000	0.004	0.009	0.004	0.005	0.008	0.174	0.006
Promedio Error/ (1- CoefVar)	19.94	20.58	60.84	23.62	21.22	20.15	20.36	33.08	20.44
Promedio general				**25.941**					
Coef Var				**0.489**					
Promedio Error/ (1- CoefVar)				**50.724**					
Técnica con menor error				**Logit**					

Por último, sometemos a prueba, los datos sin ningún tipo de tratamiento y bajo diferentes técnicas estadísticas y encontramos que, al igual que en el modelo HIT, los modelos vuelven a mostrar el peor desempeño, cuando no se realiza ningún tipo de tratamiento de datos, lo cual, es indicativo de que se deberían realizar siempre este tipo de pruebas para asegurar que la técnica y el tratamiento de datos son los más adecuados para realizar un modelo.

8.5 Entrenamiento del Modelo No HIT

Cómo ya hemos verificado a priori de realizar la modelación final de nuestros datos "NO Hit", la técnica que más se acopla, tomando en cuenta, el menor error y volatilidad en las iteraciones es la técnica Logit con tratamiento de datos WOE, por lo que procederemos a realizar la modelación robusta a través de ajustar los parámetros estadísticos para que pueda entrar en producción con la validez estadística suficiente.
La especificación econométrica, será la siguiente:

$$Buenos_{WOE} = WOEEdad + WOEEstGiro + WOEEstMunicipio + WOETipoVivienda + WOEEstEstado + WOEEstadoCivil + WOEConsultas + WOESexo$$

La regresión nos arrojó lo siguiente:

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significancia
(Intercept)	1.367005	0.033405	40.922	0.00	***
WOEEdad	0.795058	0.106477	7.467	0.00	***
WOEEstGiro	0.96845	0.062696	15.447	0.00	***
WOEEstMunicipio	0.936887	0.052648	17.795	0.00	***
WOETipoVivienda	0.880645	0.050795	17.337	0.00	***
WOEEstado	0.10608	0.156861	0.676	0.50	

WOEEstadoCivil	-0.009996	0.195575	-0.051	0.96	
WOEConsultas	0.693242	0.117735	5.888	0.00	***
WOESexo	1.172664	0.15551	7.541	0.00	***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

AIC 6481.5

De primera instancia, podemos encontrar que, en contra de lo que habíamos estipulado en el modelo HIT con respecto a las WOE y el signo esperado, la WOE al carga la relación de la variable dependiente contra la independiente, por lo que no es correcto que tengamos diferentes signos en los estimadores de las betas, aquí se encuentra una variable con un signo diferente en toda la regresión, por lo que de principio, será candidata a retirarla del análisis. Ya revisando la significancia en *p-value*, las variables con menos significancia estadística fueron WOEEstado y WOEEstadoCivil, esto se debe a que al menor WOEEstado está muy correlacionada con WOEEstMunicipio, y al realizar la regresión alguna resulta no estadísticamente significativa, por otra parte WOEEstadoCivil puede deberse a su baja significancia para separar clientes buenos y malos, ya que como lo detallamos en el *PreScoreCard* el *Information Value* de dicha variable, apenas superaba el 0.02. Por lo que procederemos a quitar estas variables no significativas.

$$Buenos_{WOE} = WOEEdad + WOEEstGiro + WOEEstMunicipio + WOETipoVivienda + WOEConsultas + WOESexo$$

Teniendo como resultado lo siguiente:

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Significancia
(Intercept)	1.36617	0.03337	40.946	0.00	***
WOEEdad	0.79362	0.10474	7.577	0.00	***
WOEEstGiro	0.96993	0.06261	15.491	0.00	***
WOEEstMunicipio	0.94345	0.05153	18.307	0.00	***
WOETipoVivienda	0.8801	0.05075	17.341	0.00	***
WOEConsultas	0.69661	0.11762	5.923	0.00	***
WOESexo	1.1712	0.15547	7.533	0.00	***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

AIC 6477.9

Por lo que determinamos que son significativas y estadísticamente el modelo estadísticamente correcto lo hacemos a través de esta regresión.

8.6 Validación del Performance

Una vez que obtuvimos la regresión, es necesario realizar las mismas pruebas de performance que hicimos para el modelo HIT, la cual consiste en transformar la regresión en un score e imputarlo a diferentes pruebas para determinar su solidez estadística para

ponerlo en producción. Comenzaremos diciendo que debemos calcular el score, el cual nos deriva de la regresión anterior.

$$Z = 1.3661 + 0.7936 * WOEEdad + 0.9699 * WOEEstGiro + 0.9434 * WOEEstMunicipio + 0.8801 * WOETIpoVivienda + 0.6966 * WOEConsultas + 1.1712 * WOESexo$$

$$Score = \left(\frac{e^z}{1 + e^z}\right) * 1000$$

Teniendo como resultado por la aplicación en toda la base de datos, lo siguiente:

	Predicción	
	Buenos	Malos
Buenos	7464	356
Malos	1567	459

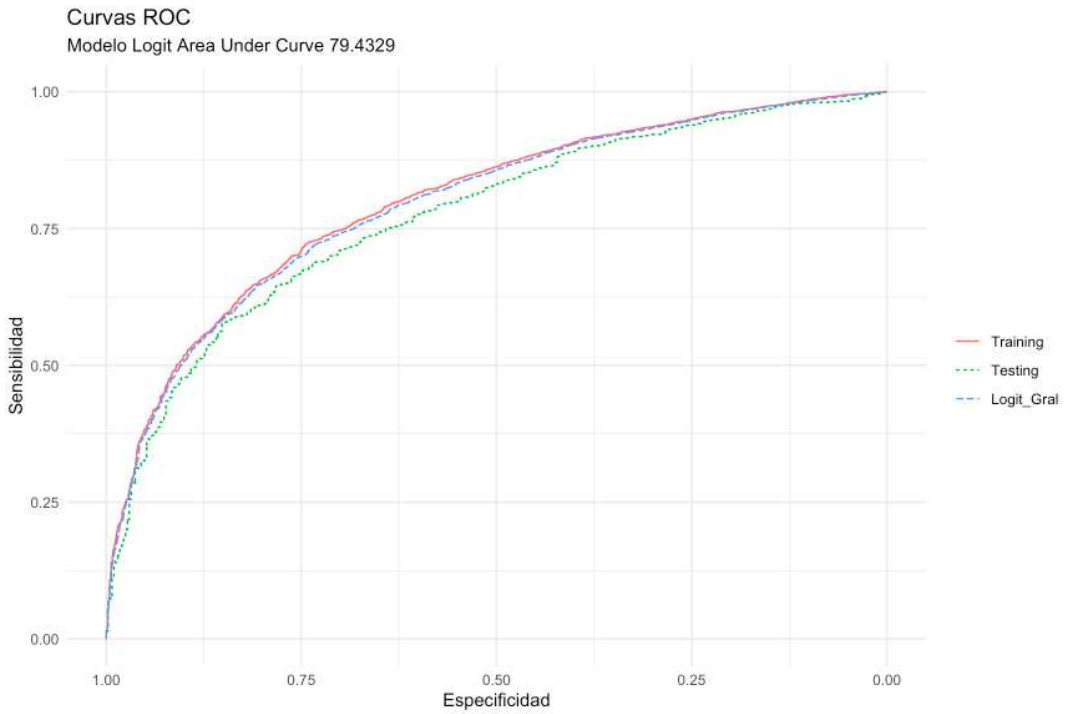
Precisión Global	$\frac{VN+VP}{VN+VP+FP+FN}$	**80.47%**
Error	$\frac{FN+FP}{VN+VP+FP+FN}$	**19.53%**
Precisión Positiva (Sensibilidad)	$\frac{VP}{VP+FN}$	**22.66%**
Precisión Negativa (Especificidad)	$\frac{VN}{VN+FP}$	**95.45%**
Asertividad Positiva	$\frac{VP}{VP+FP}$	**56.32%**
Asertividad Negativa	$\frac{VN}{VN+FN}$	**82.65%**
Falsos Negativos	1567	**77.34%**
Falsos Positivos	356	**5%**

Podemos observar que el modelo tiene un menor error y es muy parecido al que nos arrojaba el KFold Cross Validation, pero al igual que el modelo HIT, cuenta con la particularidad de que es bueno para detectar clientes "Buenos", sin embargo es pésimo para detectar clientes malos, ya que la especificidad, se encuentra en 22.66%.

	Predicción	
	Buenos	Malos
Buenos	1489	75
Malos	323	82

Precisión Global	$\frac{VN+VP}{VN+VP+FP+FN}$	**79.79%**
Error	$\frac{FN+FP}{VN+VP+FP+FN}$	**20.21%**
Precisión Positiva (Sensibilidad)	$\frac{VP}{VP+FN}$	**20.25%**
Precisión Negativa (Especificidad)	$\frac{VN}{VN+FP}$	**95.20%**
Asertividad Positiva	$\frac{VP}{VP+FP}$	**52.23%**
Asertividad Negativa	$\frac{VN}{VN+FN}$	**82.17%**
Falsos Negativos	323	**79.75%**
Falsos Positivos	75	**5%**

Cuando verificamos la muestra de prueba, observamos que existe una proporción similar a la obtenida para la toda base de datos, lo que nos asegura que el modelo es estable para muestras desconocidas. A la par, observamos el mismo problema de una baja especificidad del modelo.



Al analizar la curva ROC, vemos que obtiene un mejor performance en comparación con el modelo HIT, ya que la AUROC es de 79.43, lo que lo posiciona en un modelo de desempeño "Bueno". Siguiendo la identidad para obtener el GINI provista en el modelo HIT, obtenemos lo siguiente:

$$GINI = 2AUROC - 1$$

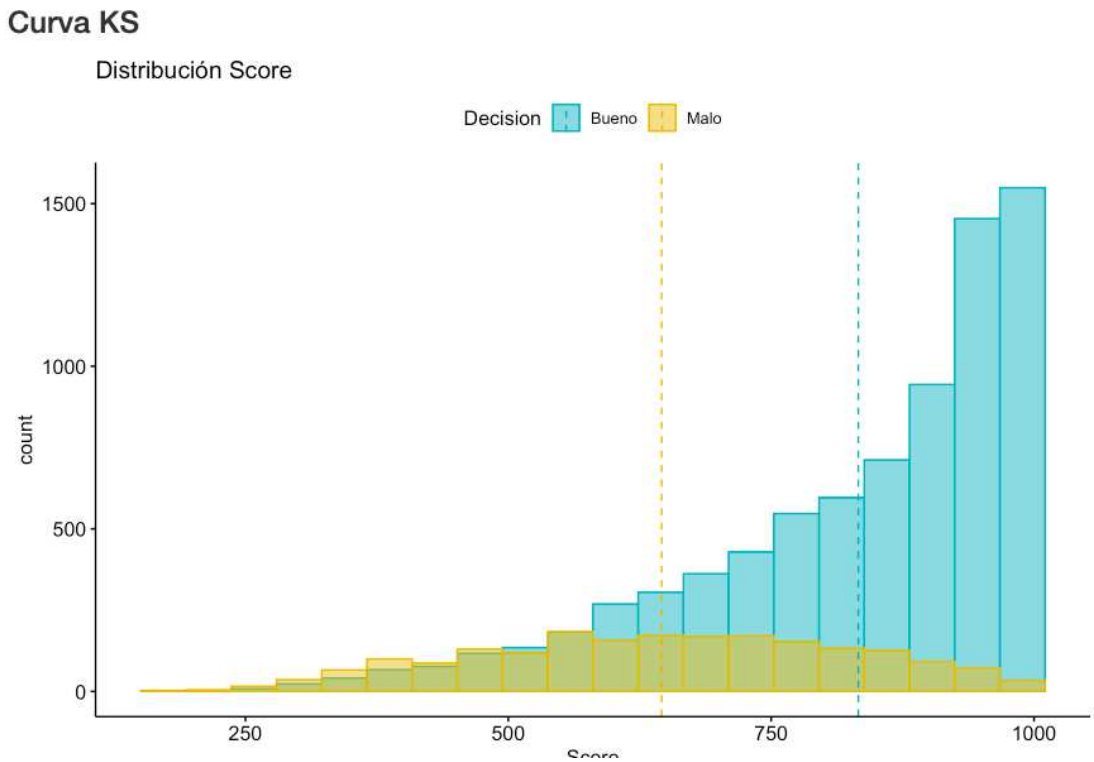
Por lo tanto

$$GINI = 0.5886$$

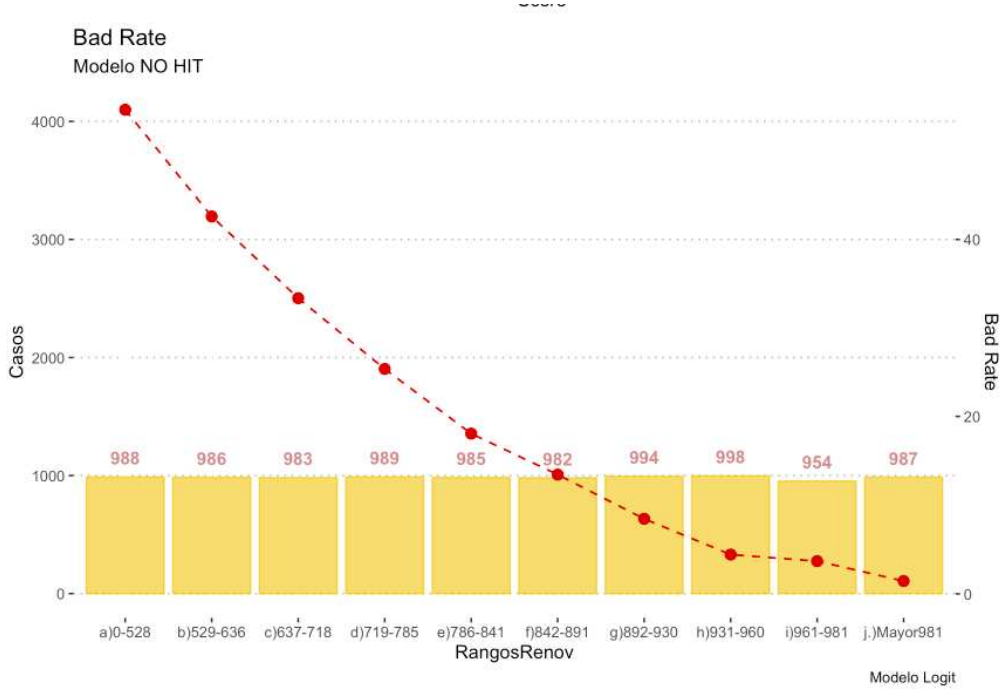
Lo cual indica un alto poder de predicción, por lo que procederemos a realizar el análisis en deciles.

Rangos Score	Total Buenos	Total Malos	Total	Bad Rate	AcumB	DistB	DistM	odds	IV	IVAcum	KS	KS2	MAXIV
a)0-528	448	540	988	55%	7820	0.06	0.27	0.21	0.32	1.33	0.00	0.45	1.33
b)529-636	566	420	986	43%	7372	0.07	0.21	0.35	0.14	1.01	0.21	0.45	1.33
c)637-718	655	328	983	33%	6806	0.08	0.16	0.52	0.05	0.87	0.34	0.45	1.33
d)719-785	738	251	989	25%	6151	0.09	0.12	0.76	0.01	0.82	0.42	0.45	1.33
e)786-841	807	178	985	18%	5413	0.10	0.09	1.17	0.00	0.81	0.45	0.45	1.33
f)842-891	850	132	982	13%	4606	0.11	0.07	1.67	0.02	0.80	0.44	0.45	1.33
g)892-930	910	84	994	8%	3756	0.12	0.04	2.81	0.08	0.78	0.39	0.45	1.33
h)931-960	954	44	998	4%	2846	0.12	0.02	5.62	0.17	0.70	0.32	0.45	1.33
i)961-981	919	35	954	4%	1892	0.12	0.02	6.80	0.19	0.53	0.22	0.45	1.33
j.)Mayor981	973	14	987	1%	973	0.12	0.01	18.01	0.34	0.34	0.12	0.45	1.33

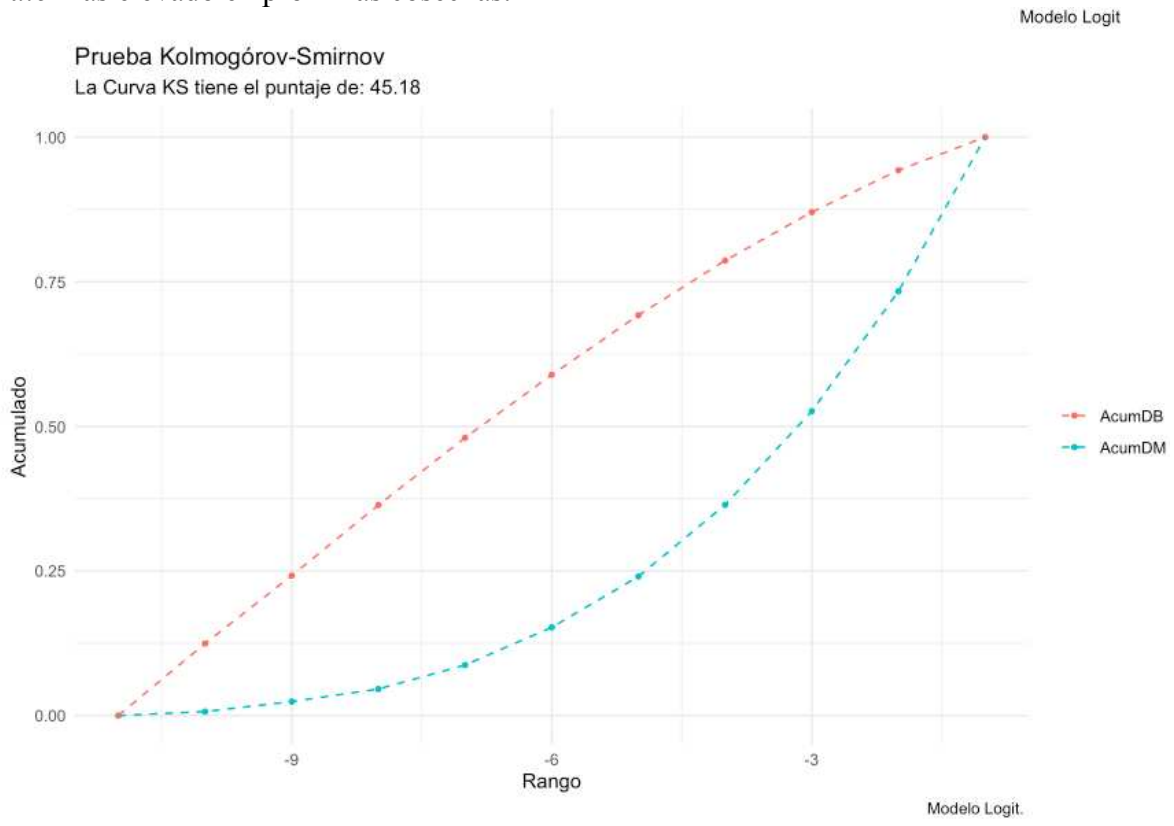
Podemos decir que existe armonía en el score, es decir que conforme el score es mayor, el *Bad Rate* es menor, este se ve frágil en la transición del decil 8 al 9, puesto que tienen un *Bad Rate* semejante, sin embargo es menor la cantidad de malos, lo que la de buenos, aunque no se nota mucho la diferencia. Por otra parte vemos que el estadístico, KS es de 45 puntos, 3 puntos encima del modelo HIT, lo que indica, que también puede separar mejor las poblaciones de Buenos y Malos.



Por otra parte al analizar el histograma, vemos que la mayoría de los clientes malos, se concentran en los scores bajos, mientras que los scores altos contienen a la gran mayoría de los clientes buenos, además de que las medias por clasificación de buenos y malos en los scores se encuentran muy separadas.



Al revisar el gráfico de *Bad Rate,* observamos que es armonioso en todos los deciles, aunque en el decil 8 al 9 observamos una pendiente menos pronunciada del *Bad Rate*, esto no necesariamente significa que es estadísticamente malo, sino, que habrá que poner atención a los segmentos que entren en estos segmentos, ya que pueden presentar un *Bad Rate* más elevado en próximas cosechas.



Por último, observamos la curva KS, en donde verificamos que en todos los puntos, están separados, lo que demuestra gráficamente que el modelo es un buen clasificador.

9. Determinación del Cutoff: Optimización de la Matriz de Confusión Financiera

Ya comentábamos en una sección anterior que el modelo HIT tenía un muy buen performance, pero carecía de un factor fundamental, que era un bajo indicador de sensibilidad, es decir, que el modelo identifica muy bien a los clientes clasificados como "Buenos", sin embargo a los clasificados como "Malos" no los alcanza a identificar, esto sucede porque en los modelos para realizar la clasificación, asumimos que la probabilidad deberá estar encima de un puntaje de 500, sin embargo cuando cambiamos dicho parámetro de probabilidad a rangos superiores o inferiores, cambian todos los indicadores de performance relacionados a la matriz de confusión.

Por lo que un factor clave, es elegir el cutoff que minimice las pérdidas de la Institución Financiera, esto se da sí y sólo sí se puede optimizar la matriz de confusión financiera, la cual consideramos como la matriz de confusión original, pero dándole el sentido económico que representa, y esbozamos a continuación:

		Pronostico	
		Bueno	Malo
Real	Bueno	**VP (a)**	**FN (c)**
	Malo	FP (d)	**VN (b)**

Precisión del Modelo
Costo de oportunidad (Pasar créditos como malos que son buenos)
Missclasification: Costo-Beneficio (Pasar créditos como buenos que son malos)

Por lo que inicialmente deberíamos evaluar los "j" puntos del score y ver las implicaciones que tienen en sensibilidad y en la especificidad, por lo que nosotros proponemos como el punto en donde se maximiza la especificidad y la sensibilidad a la siguiente función:

$$Cutoff_{Max} = Max(Especificidad_i * Sensibilidad_j)$$

Es decir que, dada la evaluación de todos los puntos del score, el cutoff que maximiza la especificidad y sensibilidad se dará cuando el producto de la especificidad por la sensibilidad sea el máximo, en otras palabras, en este punto la identificación de "Buenos" y "Malos" llegan conjuntamente a su máximo posible. Sin embargo, esto no es suficiente para realizar la elección de *cutoff*, debemos considerar, que el costo que de *missclasification*, es el más perjudicial, puesto que es el que indica todos los créditos que son "Malos" y que dado un cutoff pasa como "Buenos", por ejemplo, tomemos los resultados de la matriz de confusión de nuestro modelo HIT:

		Pronostico	
		Bueno	Malo
Real	Bueno	10510	932
	Malo	2831	1421

E imaginemos que cada cuenta, representa $1 desembolsado, lo que tenemos como resultado, es que perderíamos $2,831.00, cerca de un 18.03% de pérdida, sobre el total de cuentas, por un tener un exceso de cuentas en *missclasification*, por lo que el *cutoff* debe estar sujeto a una segunda condición, y es que, los FP sean los menores posibles dados un *cutoff*, esta se da, bajo una segunda función:

$$Cutoff_{Max2} = Max(Especificidad_j * Sensibilidad_j * \left(1 - \left(\frac{FP_j}{FP_j + FN_j}\right)\right))$$

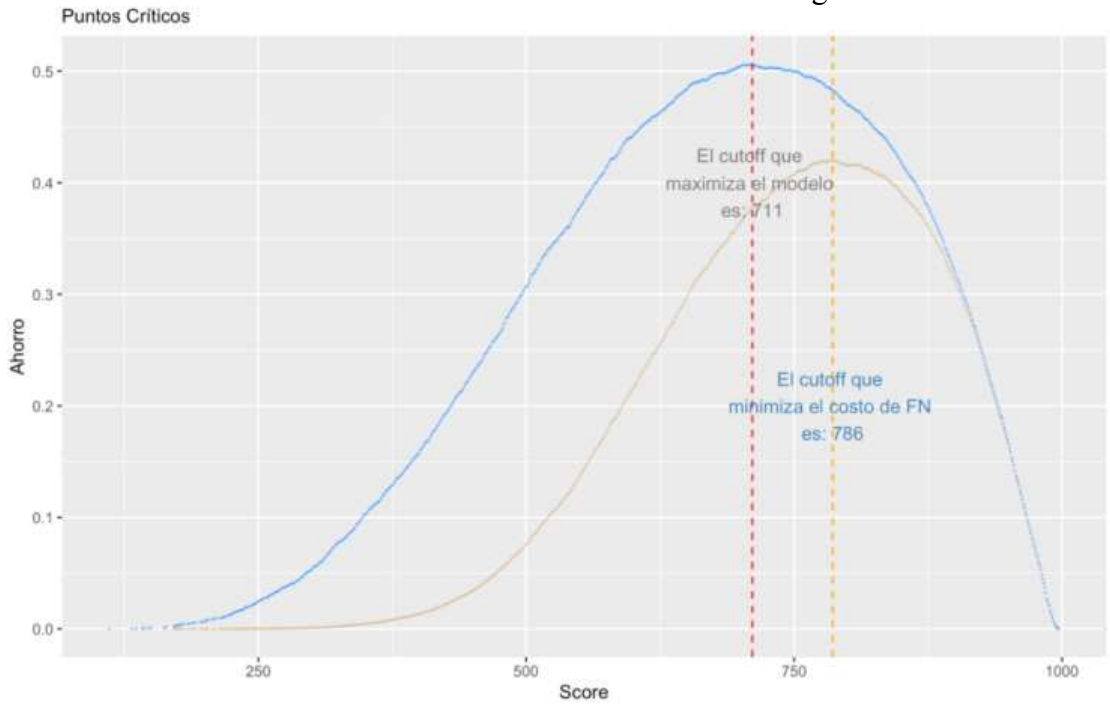
El cuál nos asegura que se ha encontrando el punto mínimo de Falsos Positivos (FP) y por lo tanto, ese el punto de corte que hace rentable a la institución, ya que minimiza los créditos "Malos" que pasan como "Buenos". Por ello hemos evaluado en todos los puntos posibles del score HIT, sin embargo por simplicidad mostraremos en tabla sólo del rango de 800 a 700, encontrando lo siguiente:

Score	VP	FP	VN	FN	Sens	Esp	CutoffMax	CutoffMax2
800	6370	667	3585	5072	55.67%	84.31%	46.94%	41.48%
799	6394	670	3582	5048	55.88%	84.24%	47.08%	41.56%
798	6418	673	3579	5024	56.09%	84.17%	47.21%	41.64%
797	6435	676	3576	5007	56.24%	84.10%	47.30%	41.67%
796	6453	682	3570	4989	56.40%	83.96%	47.35%	41.66%
795	6473	685	3567	4969	56.57%	83.89%	47.46%	41.71%
794	6498	691	3561	4944	56.79%	83.75%	47.56%	41.73%
793	6523	696	3556	4919	57.01%	83.63%	47.68%	41.77%
792	6539	701	3551	4903	57.15%	83.51%	47.73%	41.76%
791	6561	702	3550	4881	57.34%	83.49%	47.87%	41.85%
790	6580	703	3549	4862	57.51%	83.47%	48.00%	41.94%
789	6601	708	3544	4841	57.69%	83.35%	48.08%	41.95%
788	6613	711	3541	4829	57.80%	83.28%	48.13%	41.95%
787	6635	716	3536	4807	57.99%	83.16%	48.22%	41.97%
786	**6643**	**717**	**3535**	**4799**	**58.14%**	**83.11%**	**48.32%**	**42.02%**
785	6667	723	3529	4775	58.27%	83.00%	48.36%	42.00%
784	6683	728	3524	4759	58.41%	82.88%	48.41%	41.98%
783	6702	732	3520	4740	58.57%	82.78%	48.49%	42.00%
782	6724	741	3511	4718	58.77%	82.57%	48.52%	41.94%
781	6739	747	3505	4703	58.90%	82.43%	48.55%	41.90%
780	6764	756	3496	4678	59.12%	82.22%	48.60%	41.84%
779	6788	763	3489	4654	59.33%	82.06%	48.68%	41.82%
778	6812	764	3488	4630	59.54%	82.03%	48.84%	41.92%
777	6826	768	3484	4616	59.66%	81.94%	48.88%	41.91%
776	6849	773	3479	4593	59.86%	81.82%	48.98%	41.92%
775	6866	777	3475	4576	60.01%	81.73%	49.04%	41.92%
774	6882	782	3470	4560	60.15%	81.61%	49.09%	41.90%

773	6903	789	3463	4539	60.33%	81.44%	49.14%	41.86%
772	6920	801	3451	4522	60.48%	81.16%	49.09%	41.70%
771	6943	805	3447	4499	60.68%	81.07%	49.19%	41.73%
770	6968	807	3445	4474	60.90%	81.02%	49.34%	41.80%
769	6981	814	3438	4461	61.01%	80.86%	49.33%	41.72%
768	6994	820	3432	4448	61.13%	80.71%	49.34%	41.66%
767	7019	825	3427	4423	61.34%	80.60%	49.44%	41.67%
766	7036	832	3420	4406	61.49%	80.43%	49.46%	41.60%
765	7056	840	3412	4386	61.67%	80.24%	49.48%	41.53%
764	7078	848	3404	4364	61.86%	80.06%	49.52%	41.47%
763	7093	858	3394	4349	61.99%	79.82%	49.48%	41.33%
762	7109	865	3387	4333	62.13%	79.66%	49.49%	41.26%
761	7128	874	3378	4314	62.30%	79.44%	49.49%	41.15%
760	7147	880	3372	4295	62.46%	79.30%	49.54%	41.11%
759	7160	882	3370	4282	62.58%	79.26%	49.60%	41.13%
758	7177	886	3366	4265	62.73%	79.16%	49.65%	41.11%
757	7200	892	3360	4242	62.93%	79.02%	49.73%	41.09%
756	7219	897	3355	4223	63.09%	78.90%	49.78%	41.06%
755	7241	902	3350	4201	63.28%	78.79%	49.86%	41.05%
754	7258	905	3347	4184	63.43%	78.72%	49.93%	41.05%
753	7279	910	3342	4163	63.62%	78.60%	50.00%	41.03%
752	7291	918	3334	4151	63.72%	78.41%	49.96%	40.92%
751	7303	926	3326	4139	63.83%	78.22%	49.93%	40.80%
750	7317	932	3320	4125	63.95%	78.08%	49.93%	40.73%
749	7334	937	3315	4108	64.10%	77.96%	49.97%	40.69%
748	7355	943	3309	4087	64.28%	77.82%	50.02%	40.65%
747	7380	949	3303	4062	64.50%	77.68%	50.10%	40.61%
746	7398	957	3295	4044	64.66%	77.49%	50.10%	40.52%
745	7408	961	3291	4034	64.74%	77.40%	50.11%	40.47%
744	7429	974	3278	4013	64.93%	77.09%	50.05%	40.28%
743	7441	979	3273	4001	65.03%	76.98%	50.06%	40.22%
742	7461	987	3265	3981	65.21%	76.79%	50.07%	40.12%
741	7478	993	3259	3964	65.36%	76.65%	50.09%	40.06%
740	7493	999	3253	3949	65.49%	76.51%	50.10%	39.99%
739	7514	1004	3248	3928	65.67%	76.39%	50.16%	39.95%
738	7523	1006	3246	3919	65.75%	76.34%	50.19%	39.94%
737	7539	1015	3237	3903	65.89%	76.13%	50.16%	39.81%
736	7561	1020	3232	3881	66.08%	76.01%	50.23%	39.78%
735	7578	1024	3228	3864	66.23%	75.92%	50.28%	39.75%
734	7599	1033	3219	3843	66.41%	75.71%	50.28%	39.63%
733	7614	1039	3213	3828	66.54%	75.56%	50.28%	39.55%
732	7632	1050	3202	3810	66.70%	75.31%	50.23%	39.38%
731	7655	1060	3192	3787	66.90%	75.07%	50.22%	39.24%
730	7675	1063	3189	3767	67.08%	75.00%	50.31%	39.24%
729	7689	1069	3183	3753	67.20%	74.86%	50.31%	39.15%
728	7708	1074	3178	3734	67.37%	74.74%	50.35%	39.10%
727	7731	1084	3168	3711	67.57%	74.51%	50.34%	38.96%
726	7746	1095	3157	3696	67.70%	74.25%	50.26%	38.78%
725	7768	1105	3147	3674	67.89%	74.01%	50.25%	38.63%
724	7783	1111	3141	3659	68.02%	73.87%	50.25%	38.54%
723	7793	1115	3137	3649	68.11%	73.78%	50.25%	38.49%
722	7807	1121	3131	3635	68.23%	73.64%	50.24%	38.40%
721	7821	1124	3128	3621	68.35%	73.57%	50.28%	38.37%
720	7839	1126	3126	3603	68.51%	73.52%	50.37%	38.38%
719	7857	1136	3116	3585	68.67%	73.28%	50.32%	38.21%
718	7871	1140	3112	3571	68.79%	73.19%	50.35%	38.16%

717	7891	1146	3106	3551	68.97%	73.05%	50.38%	38.09%
716	7909	1150	3102	3533	69.12%	72.95%	50.43%	38.04%
715	7926	1153	3099	3516	69.27%	72.88%	50.49%	38.02%
714	7942	1159	3093	3500	69.41%	72.74%	50.49%	37.93%
713	7956	1165	3087	3486	69.53%	72.60%	50.48%	37.84%
712	7973	1165	3087	3469	69.68%	72.60%	50.59%	37.87%
711	**7980**	**1168**	**3084**	**3462**	**69.85%**	**72.51%**	**50.64%**	**37.83%**
710	8010	1179	3073	3432	70.01%	72.27%	50.59%	37.66%
709	8023	1187	3065	3419	70.12%	72.08%	50.54%	37.52%
708	8036	1193	3059	3406	70.23%	71.94%	50.53%	37.42%
707	8054	1200	3052	3388	70.39%	71.78%	50.52%	37.31%
706	8077	1207	3045	3365	70.59%	71.61%	50.55%	37.21%
705	8097	1214	3038	3345	70.77%	71.45%	50.56%	37.10%
704	8117	1221	3031	3325	70.94%	71.28%	50.57%	36.99%
703	8132	1228	3024	3310	71.07%	71.12%	50.55%	36.87%
702	8148	1235	3017	3294	71.21%	70.95%	50.53%	36.75%
701	8166	1243	3009	3276	71.37%	70.77%	50.51%	36.61%
700	8174	1249	3003	3268	71.44%	70.63%	50.45%	36.50%

Esto se ve de manera más clara cuando los analizamos de forma gráfica:



Al evaluar en los n puntos del score, observamos que el punto máximo conjunto de la especificidad y la sensibilidad se da cuando el *cutoff* es de 711, bajo esta premisa encontramos la siguiente matriz de confusión:

		Predicción	
		Buenos	Malos
Real	Buenos	7980	3462
	Malos	1168	3084

Precisión Global	**70.51%**
Error	**29.49%**
Precisión Positiva (Sensibilidad)	**72.51%**
Precisión Negativa (Especificidad)	**69.85%**

Asertividad Positiva	**47.11%**
FP/(FN+FP)	**25.22%**

La cuál indica que se pierden 4 puntos porcentuales de precisión global, y se ganan cerca de 36 puntos porcentuales de Especificidad respecto del original y de 2,831 cuentas que teníamos en misclassification se reducen a 1168, sin embargo, éste no es el mínimo de FP, ese punto se da, cuando el *cutoff* (aplicando la fórmula del $CutOffMax_2$) se estipula en 786 que da como resultado la siguiente matriz de confusión:

		Predicción	
		Buenos	Malos
Real	Buenos	6643	4799
	Malos	717	3535

Precisión Global	**64.85%**
Error	**35.15%**
Precisión Positiva (Sensibilidad)	**83.11%**
Precisión Negativa (Especificidad)	**58.14%**
Asertividad Positiva	**42.42%**
FP/(FN+FP)	**12.00%**

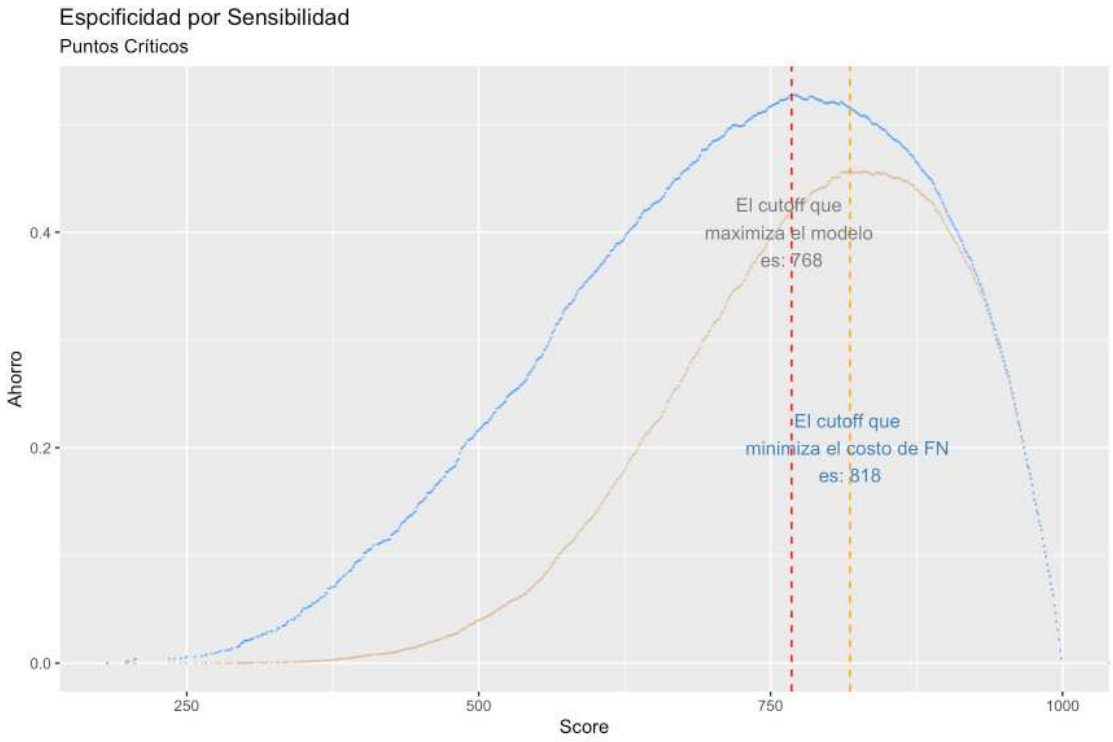
Lo que indica que hemos llegado el número mínimo de FP con el costo de disminuir casi 12 puntos porcentuales, pero con un 717 cuentas catalogadas en *missclasification*, reduciendo al 4% del 18.3% que originalmente teníamos y que representan una pérdida neta en el sentido de que hacemos pasar créditos como buenos, que al final serán malos. Este punto únicamente nos servirá como conexión para realizar el ajuste a la Capacidad de pago, la cual se modulará bajo reglas asociadas a este *cutoff*. En el caso del modelo No HIT, al buscar el *CutoffMax* y el $CutOffMax_2$ a través de las fórmulas propuestas evaluados en todos los puntos del score, encontramos los siguientes resultados:

Score	VP	FP	VN	FN	Sens	Esp	CutoffMax	CutoffMax2
850	4474	286	1740	3346	57.21%	85.88%	49.14%	45.27%
849	4496	287	1739	3324	57.49%	85.83%	49.35%	45.43%
848	4516	290	1736	3304	57.75%	85.69%	49.48%	45.49%
847	4531	293	1733	3289	57.94%	85.54%	49.56%	45.51%
846	4545	300	1726	3275	58.12%	85.19%	49.51%	45.36%
845	4558	301	1725	3262	58.29%	85.14%	49.63%	45.43%
844	4575	303	1723	3245	58.50%	85.04%	49.75%	45.51%
843	4586	307	1719	3234	58.64%	84.85%	49.76%	45.44%
842	4606	309	1717	3214	58.90%	84.75%	49.92%	45.54%
841	4617	311	1715	3203	59.04%	84.65%	49.98%	45.55%
840	4639	315	1711	3181	59.32%	84.45%	50.10%	45.58%
839	4659	325	1701	3161	59.58%	83.96%	50.02%	45.36%
838	4667	328	1698	3153	59.68%	83.81%	50.02%	45.31%
837	4681	331	1695	3139	59.86%	83.66%	50.08%	45.30%
836	4700	333	1693	3120	60.10%	83.56%	50.22%	45.38%
835	4732	337	1689	3088	60.51%	83.37%	50.45%	45.48%
834	4746	339	1687	3074	60.69%	83.27%	50.54%	45.52%
833	4760	341	1685	3060	60.87%	83.17%	50.62%	45.55%

832	4766	341	1685	3054	60.95%	83.17%	50.69%	45.60%
831	4785	344	1682	3035	61.19%	83.02%	50.80%	45.63%
830	4792	345	1681	3028	61.28%	82.97%	50.84%	45.64%
829	4794	349	1677	3026	61.30%	82.77%	50.74%	45.50%
828	4804	351	1675	3016	61.43%	82.68%	50.79%	45.49%
827	4812	352	1674	3008	61.53%	82.63%	50.84%	45.52%
826	4820	352	1674	3000	61.64%	82.63%	50.93%	45.58%
825	4841	357	1669	2979	61.91%	82.38%	51.00%	45.54%
824	4858	359	1667	2962	62.12%	82.28%	51.11%	45.59%
823	4870	362	1664	2950	62.28%	82.13%	51.15%	45.56%
822	4893	366	1660	2927	62.57%	81.93%	51.27%	45.57%
821	4906	368	1658	2914	62.74%	81.84%	51.34%	45.58%
820	4917	371	1655	2903	62.88%	81.69%	51.36%	45.54%
819	4933	372	1654	2887	63.08%	81.64%	51.50%	45.62%
818	4949	374	1652	2871	63.29%	81.54%	51.60%	45.66%
817	4959	377	1649	2861	63.41%	81.39%	51.61%	45.60%
816	4972	379	1647	2848	63.58%	81.29%	51.69%	45.62%
815	4986	384	1642	2834	63.76%	81.05%	51.67%	45.51%
814	5007	386	1640	2813	64.03%	80.95%	51.83%	45.58%
813	5021	389	1637	2799	64.21%	80.80%	51.88%	45.55%
812	5036	391	1635	2784	64.40%	80.70%	51.97%	45.57%
811	5052	392	1634	2768	64.60%	80.65%	52.10%	45.64%
810	5068	399	1627	2752	64.81%	80.31%	52.04%	45.45%
809	5081	410	1616	2739	64.97%	79.76%	51.83%	45.08%
808	5094	411	1615	2726	65.14%	79.71%	51.93%	45.12%
807	5097	412	1614	2723	65.18%	79.66%	51.92%	45.10%
806	5105	413	1613	2715	65.28%	79.62%	51.97%	45.11%
805	5119	415	1611	2701	65.46%	79.52%	52.05%	45.12%
804	5130	417	1609	2690	65.60%	79.42%	52.10%	45.11%
803	5138	421	1605	2682	65.70%	79.22%	52.05%	44.99%
802	5152	424	1602	2668	65.88%	79.07%	52.09%	44.95%
801	5161	430	1596	2659	66.00%	78.78%	51.99%	44.75%
800	5182	434	1592	2638	66.27%	78.58%	52.07%	44.71%
799	5193	441	1585	2627	66.41%	78.23%	51.95%	44.48%
798	5223	449	1577	2597	66.79%	77.84%	51.99%	44.32%
797	5246	456	1570	2574	67.08%	77.49%	51.99%	44.16%
796	5255	458	1568	2565	67.20%	77.39%	52.01%	44.13%
795	5264	459	1567	2556	67.31%	77.34%	52.06%	44.14%
794	5287	461	1565	2533	67.61%	77.25%	52.22%	44.18%
793	5295	463	1563	2525	67.71%	77.15%	52.24%	44.14%
792	5306	468	1558	2514	67.85%	76.90%	52.18%	43.99%
791	5333	471	1555	2487	68.20%	76.75%	52.34%	44.01%
790	5343	475	1551	2477	68.32%	76.55%	52.31%	43.89%
789	5353	476	1550	2467	68.45%	76.51%	52.37%	43.90%
788	5369	479	1547	2451	68.66%	76.36%	52.42%	43.85%
787	5397	484	1542	2423	69.02%	76.11%	52.53%	43.78%
786	5413	487	1539	2407	69.22%	75.96%	52.58%	43.73%
785	5429	491	1535	2391	69.42%	75.77%	52.60%	43.64%
784	5442	495	1531	2378	69.59%	75.57%	52.59%	43.53%
783	5457	500	1526	2363	69.78%	75.32%	52.56%	43.38%
782	5472	510	1516	2348	69.97%	74.83%	52.36%	43.02%
781	5482	514	1512	2338	70.10%	74.63%	52.32%	42.89%

780	5495	518	1508	2325	70.27%	74.43%	52.30%	42.77%
779	5507	519	1507	2313	70.42%	74.38%	52.38%	42.78%
778	5518	523	1503	2302	70.56%	74.19%	52.35%	42.66%
777	5528	524	1502	2292	70.69%	74.14%	52.41%	42.66%
776	5539	524	1502	2281	70.83%	74.14%	52.51%	42.70%
775	5550	525	1501	2270	70.97%	74.09%	52.58%	42.70%
774	5562	528	1498	2258	71.13%	73.94%	52.59%	42.62%
773	5577	528	1498	2243	71.32%	73.94%	52.73%	42.68%
772	5589	531	1495	2231	71.47%	73.79%	52.74%	42.60%
771	5604	536	1490	2216	71.66%	73.54%	52.70%	42.44%
770	5615	539	1487	2205	71.80%	73.40%	52.70%	42.35%
769	5630	541	1485	2190	71.99%	73.30%	52.77%	42.32%
768	5635	542	1484	2185	72.06%	73.25%	52.78%	42.29%
767	5654	549	1477	2166	72.30%	72.90%	52.71%	42.05%
766	5663	554	1472	2157	72.42%	72.66%	52.61%	41.86%
765	5675	562	1464	2145	72.57%	72.26%	52.44%	41.55%
764	5685	566	1460	2135	72.70%	72.06%	52.39%	41.41%
763	5690	568	1458	2130	72.76%	71.96%	52.36%	41.34%
762	5699	572	1454	2121	72.88%	71.77%	52.30%	41.19%
761	5709	576	1450	2111	73.01%	71.57%	52.25%	41.05%
760	5734	582	1444	2086	73.32%	71.27%	52.26%	40.86%
759	5748	584	1442	2072	73.50%	71.17%	52.32%	40.81%
758	5758	589	1437	2062	73.63%	70.93%	52.23%	40.62%
757	5764	595	1431	2056	73.71%	70.63%	52.06%	40.38%
756	5779	602	1424	2041	73.90%	70.29%	51.94%	40.11%
755	5792	604	1422	2028	74.07%	70.19%	51.99%	40.06%
754	5797	607	1419	2023	74.13%	70.04%	51.92%	39.94%
753	5802	611	1415	2018	74.19%	69.84%	51.82%	39.78%
752	5812	614	1412	2008	74.32%	69.69%	51.80%	39.67%
751	5816	616	1410	2004	74.37%	69.60%	51.76%	39.59%
750	5824	620	1406	1996	74.48%	69.40%	51.68%	39.44%

Cabe mencionar que por simplicidad sólo revisamos 100 puntos del *score*, (de la puntuación 750 al 850) y vemos que los puntos críticos *CutoffMax* y el $CutOffMax_2$ se encuentran en el punto 768 y en el 818 respectivamente. Al verlo de manera gráfica, encontramos los siguiente:



Los detalles de la matriz de confusión con el *CutoffMax*, es el siguiente:

	Predicción	
	Buenos	Malos
Buenos	5635	2185
Malos	542	1484

Precisión Global	**72.06%**
Error	**27.94%**
(Sensibilidad)	**73.25%**
(Especificidad)	**72.06%**
Asertividad Positiva	**40.45%**
Asertividad Negativa	**91.23%**
Falsos Negativos	**26.75%**
Falsos Positivos	**28%**

Por lo que en *CutoffMax* hemos encontrado el punto donde la especificidad y la sensibilidad es máxima, es decir, que alcanzamos a predecir lo máximo conjuntamente de clientes "Buenos" y "Malos", sin embargo, el costo de missclasification sigue siendo alto, ya que aún no se reduce al mínimo el costo de clasificar a los malos como "Buenos". Sin embargo, este se encuentra en el punto 818 aplicando la fórmula del $CutoffMax_2$, que al aplicarlo en la matriz de confusión, encontramos el siguiente resultado:

	Predicción	
	Buenos	Malos
Buenos	4949	2871
Malos	374	1652

Precisión Global	**67.29%**

Error	**32.71%**
(Sensibilidad)	**81.54%**
(Especificidad)	**63.29%**
Asertividad Positiva	**36.52%**
Asertividad Negativa	**92.97%**
Falsos Negativos	**18.46%**
Falsos Positivos	**37%**

Por lo que en este cutoff se cumple el máximo de especificidad y sensibilidad, sujeto a la restricción de obtener el mínimo de verdaderos negativos (VN), ya que es el costo más grande que puede tener una institución financiera por clasificar a sus prospectos de manera errónea.

10. Ajuste por Capacidad de Pago: La inclusión financiera

Este trabajo se ha puntualizado en demostrar la importancia que tienen los scores de crédito, así como los puntos de corte, ya que estos son la principal línea de fuego para que un cliente tenga acceso al crédito, la mayoría de las instituciones financieras optan por no aceptar a los clientes que están por debajo de ese *cutoff*, cual sea que hayan decidido, que en muchas ocasiones, sólo se realiza a través del Bad Rate asumida en deciles, percentiles, etc.

Es por ello que en la sección anterior presentamos una metodología más adecuada para elegir esa línea de fuego (*cutoff*) sin embargo, nuestra solución para el control de la mora, no consiste en excluir a todos los prospectos que estén por encima de cierta probabilidad de impago o por debajo de cierto cutoff, la solución consiste en medidas más inteligentes, ya que como mencionábamos anteriormente, los modelos sufren de errores que pueden representar oportunidades de negocio muy grandes desde el punto de vista de oferta, mientras que desde el punto de vista de la demanda de crédito, consiste en la exclusión financiera de facto.

De manera que dicha solución es modular el *exposure*, dada una probabilidad de impago, es decir que, de los Ingresos y Egresos reportados por el cliente podemos agregarle un componente que llamaremos "Castigo", el cual está asociado al riesgo para determinar el *exposure* que queremos asumir.

En las secciones anteriores mencionábamos que la forma más fácil de determinar la capacidad de pago es a través del remanente existente entre Ingresos Totales Mensuales y Egresos Totales Mensuales, al cual le añadiremos la parte del "Castigo", el cuál representará la confianza sobre esta capacidad de pago dada la probabilidad de impago, es decir, que se calculará de la siguiente forma:

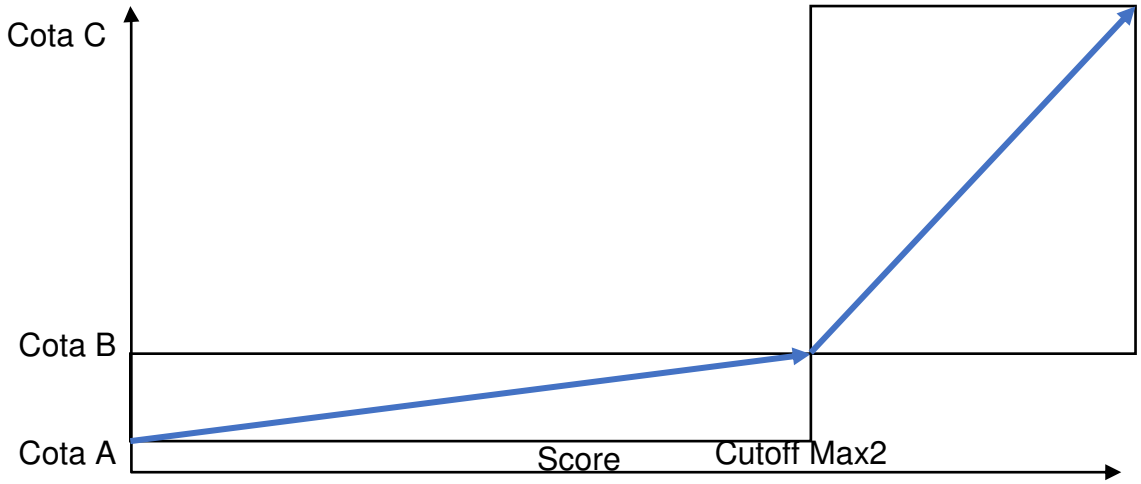
$$Capacidad\ de\ pagoM = (IngresosTotalesM - Egresos\ TotalesM) * Castigo$$

$$Castigo(Score) = \begin{cases} Score \geq CutoffMax_2 \ \therefore Cota_B + (Cota_C - Cota_B) * 1 - \left(\frac{Score - CutoffMax_2}{1000 - CutoffMax_2}\right) \\ Score < CutoffMax_2 \ \therefore Cota_A + (Cota_B - Cota_A) * 1 - \left(\frac{CutoffMax_2 - Score}{CutoffMax_2 - 0}\right) \end{cases}$$

Donde las cotas estarán en función de lo que la institución de crédito esté dispuesta a asumir, pero mostros para efecto de este trabajo nosotros utilizamos las siguientes:

Cota A	0.1
Cota B	0.3
Cota C	1.0

Es decir, que cumplan con una función de ajuste por riesgo que gráficamente se ve de la siguiente forma.



Lo que se sugiere es que, todo lo que esté por debajo del *Cutoff*, el *exposure* sea muy conservador y sea directamente proporcional a las cotas definidas las cuales están diseñadas para que ningún prospecto con alto riesgo, supere el 30% de la capacidad de pago original y así, evitar pérdidas que se pueden generar si nosotros no realizáramos este ajuste. A continuación, mostraremos la fórmula con la que hemos calculado el monto base del préstamo, cabe mencionar que utilizamos la frecuencia semanal para este ejemplo

$$Factor = \frac{días\ mes}{Frecuencia} = \frac{30.4166}{7} = 4.33$$

$$Monto\ Base = \left(\frac{Capacidad\ de\ pagoM}{Factor}\right) * 52\ semanas$$
$$= \left(\frac{Capacidad\ de\ pagoM}{4.33}\right) * 52\ semanas$$

Al realizar este ejercicio, tuvimos una condición a priori y era, que sí la capacidad de pago mensual era menor a $230.00, lo considerara como "Insuficiencia en Capacidad de Pago", puesto a que el crédito no calificaría para el crédito mínimo por política de $3,000.00. De manera que se obtuvo el monto del crédito en toda la base de datos asumiendo que los ingresos y egresos son certeros.

Modelo HIT Ajustado por riesgo

Rangos Score	Casos	Total Buenos*	Total Malos*	Total*	Ticket Promedio	Bad Rate	AcumB*	AcumM*	AcumT*	Riesgo Asumido
j.)Mayor950	1529	$23.76	$0.59	$24.35	$15,925.44	2%	$23.76	$0.59	$24.35	2%
i)924-959	1596	$21.85	$1.03	$22.88	$14,337.72	5%	$45.61	$1.62	$47.23	3%
h)884-923	1571	$17.51	$1.90	$19.41	$12,356.46	10%	$63.12	$3.53	$66.65	5%
g)832-883	1560	$14.60	$2.32	$16.92	$10,844.87	14%	$77.71	$5.85	$83.56	7%
f)767-831	1588	$12.05	$3.10	$15.15	$9,539.04	20%	$89.76	$8.95	$98.71	9%
e)701-766	1565	$7.76	$2.87	$10.62	$6,786.58	27%	$97.52	$11.81	$109.33	11%
d)627-700	1575	$6.44	$3.33	$9.77	$6,202.54	34%	$103.96	$15.14	$119.10	13%
c)547-626	1557	$4.95	$3.48	$8.43	$5,412.97	41%	$108.91	$18.62	$127.53	15%
b)448-546	1582	$3.95	$3.76	$7.70	$4,868.52	49%	$112.86	$22.38	$135.23	17%
a)0-447	1571	$2.50	$4.67	$7.16	$4,558.24	65%	$115.35	$27.04	$142.39	19%

*Cifras en millones de pesos

En la tabla podemos observar el monto desembolsado que tendríamos por el uso de la aplicación de la fórmula de Capacidad de Pago y el *exposure* que tendríamos en clientes "Buenos" y "Malos", así como el número de casos y el total de exposición por banda de score. El resultado derivado de esta operación, fue un capital por $142.39 mdp, de los cuales habría un *Bad Rate*[42] de 27%, sin embargo un Riesgo Asumido[43] de 19%. Sí analizamos el monto original que desembolsó la institución de Crédito, obtenemos la siguiente información:

Modelo HIT sin ajustar/Cantidades originales

Rangos Score	Casos	Total Buenos*	Total Malos*	Total*	Ticket Promedio	Bad Rate	AcumB*	AcumM*	AcumT*	Riesgo Asumido
j.)Mayor950	1529	$11.71	$0.38	$12.09	$15,925.44	3%	$11.71	$0.38	$12.09	3%
i)924-959	1596	$11.18	$0.66	$11.84	$14,337.72	6%	$22.90	$1.04	$23.94	4%
h)884-923	1571	$10.16	$1.18	$11.34	$12,356.46	10%	$33.05	$2.23	$35.28	6%
g)832-883	1560	$9.87	$1.64	$11.51	$10,844.87	14%	$42.92	$3.87	$46.79	8%
f)767-831	1588	$9.33	$2.61	$11.94	$9,539.04	22%	$52.26	$6.47	$58.73	11%
e)701-766	1565	$8.27	$3.16	$11.43	$6,786.58	28%	$60.53	$9.64	$70.17	14%
d)627-700	1575	$7.08	$3.98	$11.06	$6,202.54	36%	$67.60	$13.62	$81.22	17%
c)547-626	1557	$6.06	$4.61	$10.67	$5,412.97	43%	$73.66	$18.23	$91.89	20%
b)448-546	1582	$5.21	$5.40	$10.61	$4,868.52	51%	$78.87	$23.63	$102.50	23%
a)0-447	1571	$3.55	$6.88	$10.43	$4,558.24	66%	$82.42	$30.51	$112.93	27%

*Cifras en millones de pesos

Analizamos de la misma manera que el riesgo asumido desde la primer banda es mayor, teniendo un riesgo asumido 8 puntos por encima del modelo de capacidad de pago.

[42] Considérese como el Monto de Casos Malos/ Monto de Casos Totales
[43] Considérese el Monto Acumulado de Malos/ Monto acumulado total

En el modelo NO HIT, al aplicar las mismas reglas y supuestos de los ajustes a la capacidad por riesgo, tenemos los siguientes resultados:

Modelo NO HIT Ajustado por riesgo

Rangos Score	Casos	Total Buenos*	Total Malos*	Total*	TicketP	Bad Rate	AcumB*	AcumM*	AcumT*	Riesgo Asumido
j.)Mayor981	987	$13.92	$0.19	$14.11	$14,297.87	1%	$13.92	$0.19	$14.11	1%
i)961-981	954	$12.59	$0.42	$13.01	$13,633.12	3%	$26.51	$0.61	$27.12	2%
h)931-960	998	$11.83	$0.46	$12.29	$12,317.64	4%	$38.34	$1.08	$39.41	3%
g)892-930	994	$9.68	$0.91	$10.59	$10,651.91	9%	$48.02	$1.98	$50.00	4%
f)842-891	982	$7.13	$0.99	$8.12	$8,272.91	12%	$55.15	$2.97	$58.12	5%
e)786-841	985	$5.43	$1.11	$6.54	$6,641.62	17%	$60.58	$4.08	$64.67	6%
d)719-785	989	$4.54	$1.58	$6.12	$6,187.06	26%	$65.12	$5.66	$70.78	8%
c)637-718	983	$3.74	$1.70	$5.44	$5,532.05	31%	$68.86	$7.36	$76.22	10%
b)529-636	986	$2.85	$2.10	$4.95	$5,023.33	42%	$71.72	$9.46	$81.18	12%
a)0-528	988	$2.19	$2.55	$4.74	$4,794.53	54%	$73.90	$12.01	$85.91	14%

Vemos que existe una relación entre el ticket promedio por banda y el *Bad Rate*, cuando, observamos el riesgo asumido, concluimos que bajo la aplicación de nuestras reglas, llega a un 14% de riesgo que podría materializarse en caso que los clientes llegasen a incumplir. Si contrastamos estos resultados con los montos originalmente desembolsados, encontramos lo siguiente:

Modelo NO HIT sin ajustar/Cantidades originales

Rangos Score	Casos	Total Buenos*	Total Malos*	Total*	TicketP	Bad Rate	AcumB*	AcumM*	AcumT*	Riesgo Asumido
j.)Mayor981	987	$6.82	$0.12	$6.94	$7,032.02	1%	$6.82	$0.12	$6.94	2%
i)961-981	954	$6.25	$0.24	$6.49	$6,802.45	3%	$13.07	$0.36	$13.43	3%
h)931-960	998	$6.64	$0.33	$6.97	$6,983.64	4%	$19.71	$0.69	$20.40	3%
g)892-930	994	$6.30	$0.63	$6.93	$6,970.76	9%	$26.01	$1.32	$27.33	5%
f)842-891	982	$5.74	$0.94	$6.67	$6,795.47	12%	$31.74	$2.26	$34.00	7%
e)786-841	985	$5.45	$1.25	$6.70	$6,804.02	17%	$37.20	$3.51	$40.70	9%
d)719-785	989	$4.72	$1.77	$6.49	$6,559.67	26%	$41.92	$5.27	$47.19	11%
c)637-718	983	$4.33	$2.23	$6.56	$6,669.32	31%	$46.25	$7.50	$53.75	14%
b)529-636	986	$3.75	$2.87	$6.62	$6,713.70	42%	$49.99	$10.37	$60.37	17%
a)0-528	988	$3.17	$3.80	$6.97	$7,051.28	54%	$53.17	$14.17	$67.33	21%

Encontramos que el riesgo asumido, es mucho mayor (21% en su totalidad), que en el modelo ajustado por capacidad de pago, debido a que no se cuidan los montos de pérdida, tal es así, que en el monto promedio, no se observa discriminación alguna por el riesgo que conllevan las operaciones. Por lo que la hipótesis en los dos modelos se cumple.

11. El ajuste por Tasa de Interés: el precio justo del Microcrédito

En el secciones anteriores analizamos la estructura de la tasa de interés y mencionábamos que dependía directamente del precio del fondeo, el nivel de reservas y de los gastos operativos derivados de la colocación y administración del crédito, es por ello que esta sección proponemos una fórmula retomando estos aspectos y que tienen que ver básicamente con el Bad Rate de la Institución, el cuál es el acercamiento más cercano del nivel de riesgo de crédito, esta expresión que se propone se puede tomar de manera general o por banda de nivel del riesgo

$$Tasa = \left(\frac{1 + Fondeo + GOperativos}{1 - BadRate_i}\right) * \overline{T}_i$$

donde,
Fondeo = Tasa de interés de fondeo cobrada a la institución por el uso del Capital (Expresada en porcentaje)
GOperativos = Gastos operativos y administrativos con relación al crédito (Expresada en porcentaje)
$BadRate_i$ = Bad Rate evaluado en la *í*-esima banda
$\overline{T}_i$= Ticket Promedio evaluado en la íesima banda

La cuál asegura que la tasa de interés soportará el pago por el fondeo, el gasto operativo y el capital desembolsado de los créditos buenos y malos, es decir que a dicha tasa, no habrá pérdida alguna, considerando sólo los efectos del *exposure* en capital desembolsado, es decir, no contempla pérdidas por el aprovisionamiento por interés vencido que pueda tener un crédito. Esto no lo tomamos en cuenta porque estaríamos entrando en un *loop*, en donde necesitaríamos cubrir pérdidas por un interés futuro, el cual, por su naturaleza, necesitaría más requerimiento, lo que estribaría en cantidades exponenciales e insuficientes.
Para efectos de este ejercicio tomaremos los siguientes valores para Fondeo y Gastos Operativos.

Fondeo	10%
Gastos Operativos	5%

Una vez con esta información y aplicando el ajuste por tasa de interés en el Modelo Hit llegamos a los siguientes resultados:

Rangos	TicketP	TicketPromAcum	Bad Rate	Bad Rate Acum	Tasa int Banda	Tasa Acumulada	Tasaint Banda IVA	Tasa Acumulada IVA
j.)Mayor950	$15,925.44	$15,925.44	2%	2.4%	17.9%	17.9%	20.7%	20.7%
i)924-959	$14,337.72	$15,114.56	5%	3.4%	20.4%	19.1%	23.7%	22.1%
h)884-923	$12,356.46	$14,191.87	10%	5.3%	27.5%	21.4%	31.9%	24.9%
g)832-883	$10,844.87	$13,357.26	14%	7.0%	33.3%	23.7%	38.6%	27.4%
f)767-831	$9,539.04	$12,584.27	20%	9.1%	44.6%	26.5%	51.7%	30.7%
e)701-766	$6,786.58	$11,619.94	27%	10.8%	57.5%	28.9%	66.7%	33.6%
d)627-700	$6,202.54	$10,843.14	34%	12.7%	74.4%	31.7%	86.3%	36.8%
c)547-626	$5,412.97	$10,168.97	41%	14.6%	95.9%	34.7%	111.3%	40.2%
b)448-546	$4,868.52	$9,575.23	49%	16.5%	124.5%	37.8%	144.4%	43.8%
a)0-447	$4,558.24	$9,073.02	65%	19.0%	230.1%	42.0%	266.9%	48.7%

Observamos que la tasa mínima dados esos costos, al menos en la banda j con un bad rate del 2.4% da como resultado el cobro de una tasa de 17.9% pura, ya con IVA, aumenta a un 20.7%, cuando analizamos el nivel más alto de riesgo en la banda a, observamos que el mínimo requerimiento con un Bad Rate de 65%, al menos deberíamos ser el tener una tasa de 266.9% para cubrir las pérdidas por Bad Rate sólo en esa Banda, sin embargo consideramos que este es un precio muy alto, por lo que para este tipo de escenarios se sugiere que se tome la tasa global, es decir, la acumulada, en donde considera que todos los agentes inmersos pagan de manera conjunta por el riesgo que se tiene por caer en default, que en este modelo específico sería de 42% y 48.7% con IVA incluido.

Siguiendo con esta visión en comparar el modelo Capacidad de Pago contra las cantidades originales con los que se otorgaron los créditos, y aplicando la tasa mínima que debería tener los clientes para al menos cubrir la pérdida en Capital Desembolsado, los resultados se comportan de la siguiente manera:

Rangos	TicketP	TicketPromAcum	Bad Rate	Bad Rate Acum	Tasa int Banda	Tasa Acumulada	Tasaint Banda IVA	Tasa Acumulada IVA
j.)Mayor950	$7,908.49	$7,908.49	3%	3.1%	18.7%	18.7%	21.7%	21.7%
i)924-959	$7,421.48	$7,659.77	6%	4.3%	21.8%	20.2%	25.3%	23.5%
h)884-923	$7,218.67	$7,512.20	10%	6.3%	28.4%	22.7%	33.0%	26.4%
g)832-883	$7,381.30	$7,479.56	14%	8.3%	34.1%	25.4%	39.6%	29.4%
f)767-831	$7,518.24	$7,487.39	22%	11.0%	47.1%	29.2%	54.6%	33.9%
e)701-766	$7,306.37	$7,457.28	28%	13.7%	59.0%	33.3%	68.4%	38.6%
d)627-700	$7,019.06	$7,394.45	36%	16.8%	79.7%	38.2%	92.4%	44.3%
c)547-626	$6,854.24	$7,327.38	43%	19.8%	102.6%	43.5%	119.0%	50.4%
b)448-546	$6,705.80	$7,257.75	51%	23.1%	134.1%	49.5%	155.6%	57.4%
a)0-447	$6,639.28	$7,195.84	66%	27.0%	238.0%	57.6%	276.1%	66.8%

En esta tabla vemos que el Ticket promedio por banda no hace mucha distinción según su nivel de riesgo, ya que en comparación con el modelo Capacidad de Pago se muestra claramente esta relación, esto estriba en que la tasa mínima de interés por no haber ajustado el *exposure a priori* sea mucho más alta que, cuando ajustamos por riesgo. En la siguiente tabla mostramos el desglose de lo que esperamos tener en dinero, dado este riesgo y tasa.

	Capacidad Pago	Original
(+)Monto en Riesgo Total	$ 142,392,000	$ 112,931,510
Tasa Mínima	42.0%	57.6%
(+)Costo Fondeo	$ 14,239,200	$ 11,293,151
(+)Goperativos	$ 7,119,600	$ 5,646,576
(+)Mínimo recuperar (c+d)	$ 163,750,800	$ 129,871,237
(+)Monto Buenos (a)	$ 115,350,000	$ 82,421,294
(+)Costo Fondeo Buenos	$ 11,535,000	$ 8,242,129
(+)Goperativos Buenos	$ 5,767,500	$ 4,121,065
Interés probable a generar (b)	$ 48,400,802	$ 47,449,939
(=)Mínimo Recuperar Buenos (c)	$ 132,652,500	$ 94,784,488
(+)Monto Malos	$ 27,042,000	$ 30,510,216
(+)Costo Fondeo Malos	$ 2,704,200	$ 3,051,022
(+)Goperativos Malos	$ 1,352,100	$ 1,525,511
(=)Mínimo a recuperar Malos (d)	$ 31,098,300	$ 35,086,748
Recuperación Total (a+b)	$ 163,750,802	$ 129,871,233
Saldo (c+d) - (a+b)	-$ 2.32	$ 3.54

Lo que mostramos en primera instancia es que, la fórmula considera como pérdida absoluta aquélla proporción que consideramos como malo, es por eso que cuando realizamos el desglose por “Buenos” y “Malos”, no tiene tasa de interés y el riesgo que debemos asumir es por el total de los desembolsos que fueron “Malos” más los costos *a priori* en los cuáles incurrieron, añadiéndole los costos de créditos “Buenos”, los cuáles absorberán a través de la tasa, las pérdidas por el total de los créditos.

Al aplicar la misma metodología para el modelo No Hit, con los montos propuestos por el ajuste en capacidad de pago, encontramos que:

Rangos Renov	TicketP	TicketPromAcum	Bad Rate	Bad Rate Acum	Tasa int Banda	Tasa Acumulada	Tasaint Banda IVA	Tasa Acumulada IVA

j.)Mayor981	$14,297.87	$14,297.87	1%	1.3%	16.6%	16.6%	19.2%	19.2%
i)961-981	$13,633.12	$13,971.15	3%	2.3%	18.8%	17.7%	21.9%	20.5%
h)931-960	$12,317.64	$13,409.66	4%	2.7%	19.5%	18.2%	22.6%	21.1%
g)892-930	$10,651.91	$12,712.69	9%	4.0%	25.8%	19.8%	29.9%	22.9%
f)842-891	$8,272.91	$11,825.64	12%	5.1%	31.0%	21.2%	35.9%	24.6%
e)786-841	$6,641.62	$10,960.17	17%	6.3%	38.4%	22.7%	44.6%	26.4%
d)719-785	$6,187.06	$10,274.93	26%	8.0%	55.0%	25.0%	63.8%	29.0%
c)637-718	$5,532.05	$9,682.67	31%	9.7%	67.3%	27.3%	78.0%	31.7%
b)529-636	$5,023.33	$9,164.03	42%	11.7%	99.7%	30.2%	115.7%	35.0%
a)0-528	$4,794.53	$8,725.57	54%	14.0%	149.2%	33.7%	173.1%	39.1%

La tasa mínima de interés sin incluir, el pago del IVA, de manera acumulada, es decir, tomando en cuenta todos los niveles de riesgo y la posible pérdida, encontramos que la tasa a la que se debería otorgar los créditos de manera conjunta, para no generar pérdidas, asciende a 33.7%, cuando en la última banda, por sí sola y en donde el Bad Rate es de 54%, la tasa llega a ser de 149.2%. Continuando esta idea comparativa, nos remitimos a los resultados por los montos otorgados originalmente, encontrando lo siguiente:

Rangos Renov	TicketP	TicketPromAcum	Bad Rate	Bad Rate Acum	Tasa int Banda	Tasa Acumulada	Tasaint Banda IVA	Tasa Acumulada IVA
j.)Mayor981	$7,032.02	$7,032.02	2%	2%	17.0%	17.0%	19.7%	19.7%
i)961-981	$6,802.45	$6,919.19	4%	3%	19.4%	18.1%	22.5%	21.0%
h)931-960	$6,983.64	$6,941.07	5%	3%	20.7%	19.0%	24.1%	22.1%
g)892-930	$6,970.76	$6,948.58	9%	5%	26.6%	20.8%	30.8%	24.2%
f)842-891	$6,795.47	$6,917.99	14%	7%	33.8%	23.2%	39.2%	26.9%
e)786-841	$6,804.02	$6,898.96	19%	9%	41.4%	25.8%	48.0%	30.0%
d)719-785	$6,559.67	$6,850.25	27%	11%	58.0%	29.5%	67.3%	34.2%
c)637-718	$6,669.32	$6,827.66	34%	14%	74.1%	33.6%	86.0%	39.0%
b)529-636	$6,713.70	$6,814.97	43%	17%	103.2%	38.9%	119.7%	45.1%
a)0-528	$7,051.28	$6,838.69	54%	21%	152.6%	45.6%	177.0%	52.9%

Encontramos que para los requerimientos dada la pérdida, por no haber ajustado la capacidad de pago, asciende a 45.6% en tasa de interés acumulada, es decir, 14.9 puntos porcentuales encima del escenario de capacidad de pago., adicionalmente, vemos que en la última banda, el requerimiento sube de 149.2% a 177%, lo que representa un incremento de tasa de interés, de 27.8%. Al realizar el análisis financiero, obtenemos lo siguiente:

	Capacidad Pago	Original
(+)Monto en Riesgo Total	$ 85,912,000	$ 67,333,689
Tasa Mínima	33.7%	45.6%
(+)Costo Fondeo	$ 8,591,200	$ 6,733,369
(+)Goperativos	$ 4,295,600	$ 3,366,684
(+)Mínimo recuperar (c+d)	$ 98,798,800	$ 77,433,742
(+)Monto Buenos (a)	$ 73,901,000	$ 53,166,248.00
(+)Costo Fondeo Buenos	$ 7,390,100	$ 5,316,625
(+)Goperativos Buenos	$ 3,695,050	$ 2,658,312
Interés probable a generar (b)	$ 24,897,801	$ 24,267,495
(=)Mínimo Recuperar Buenos (c)	$ 84,986,150	$ 61,141,185
(+)Monto Malos	$ 12,011,000	$ 14,167,441.00
(+)Costo Fondeo Malos	$ 1,201,100	$ 1,416,744
(+)Goperativos Malos	$ 600,550	$ 708,372
(=)Mínimo a recuperar Malos (d)	$ 13,812,650	$ 16,292,557
Recuperación Total (a+b)	$ 98,798,801	$ 77,433,743
Saldo (c+d) - (a+b)	-$ 1.16	-$ 0.30

Vemos que el monto de malos a recuperar es mucho mayor en el modelo original, pese a que la base, es mucho menor, además, si tratamos de hacer una equivalencia, tomando como base la tasa mínima, en el caso de los montos originalmente desembolsados, tendríamos que quitar al menos dos bandas, es decir, todos los clientes que estén por debajo de 636 puntos en el score para tener una tasa acumulada del 33.6%, similar a la que por todos los créditos pero con ajuste en capacidad de pago, llega a 33.7%, sin embargo, esto, gracias a la mitigación del riesgo de crédito a través del *exposure*, lo que permite tener menos pérdidas, por lo que estos se pueden incluir, lo que permite reafirmar nuestra tesis de inclusión, es decir, sí se pueden otorgar créditos a clientes que por modelos paramétricos parezcan de "Alto Riesgo" y que por ende sean excluidos

12. Conclusiones (Proyección de la Cartera)

Se realizaronn cuatro simulaciones de la evolución Cartera con los resultados de los *Scorings* Reactivos, obteniendo los posibles ingresos y días de atraso, aplicando los ajustes por Capacidad de Pago y contrastándolo con la simulación del *exposure* originalmente desembolsado. Dichas simulaciones están sujetas a normatividad regulatoria mexicana vigente, por ejemplo, se realiza el reconocimiento de Cartera de Crédito conforme al Anexo 33, Apartado B-6 "Cartera de Crédito" de las Disposiciones de Carácter General Aplicables a Instituciones de Crédito (CNBV, 2020) y adicionalmente se reconocen dentro de esos pronósticos las Estimaciones Preventivas por Riesgo de Crédito de acuerdo con el Anexo D[44] de las Disposiciones de Carácter General Aplicables a Entidades de Ahorro y Crédito Popular (CNBV, 2020), esto con el fin de darle todo el realismo posible a nuestras proyecciones. Dentro de estas simulaciones, al comparar información original y la obtenida con la metodología propuesta a través del ajuste de riesgo en la Capacidad de Pago deducimos que se puede obtener el mismo nivel de rentabilidad, con una tasa de interés y la *mora asumida*[45] es mucho menor, en el caso del modelo HIT hubo una reducción de 800 puntos base de Mora y en modelo NO HIT hubo una disminución de 700 puntos base, asimismo se concluyó lo siguiente:

1.) No es necesario excluir a personas del sistema financiero personas con un alta probabilidad de incumplimiento ya que se puede reducir la *mora asumida* y aun así mantener un margen similar de ganancias incluso bajando la tasa de interés; siempre y cuando se realice el ajuste de riesgo por capacidad de pago;
2.) Tomando como referencia la utilidad de los diferentes escenarios en donde existe la misma tasa de interés pero con la diferencia que hubo un ajuste por capacidad de pago se observa que conduce una utilidad mayor con una tasa menor.
3.) El aumento en tasa de interés, ante coberturas por altos incumplimientos no representa necesariamente una mayor utilidad, incluso puede ser perjudicial, ya que como observamos en la sección 2, esto afecta directamente a las Estimaciones

[44] La calificación de Cartera se realiza bajo la metodología del Anexo D puesto que sólo se cuenta con la información de días de atraso.
[45] El monto en dinero de prospectos catalogados como "Malos" dividido entre el *exposure* total de los prospectos.

Preventivas por Riesgo de Crédito el cual afecta directamente los Estados de Resultados de la Institución.

4.) La concepción del ajuste de Capacidad de Pago no se puede llevar a cabo si no se tiene la selección del mejor algoritmo dada una base de Datos, esto representa que para llegar a estos resultados tenemos que:
    a) Establecer una metodología adecuada de transformación de de datos,
    b) Se tienen que probar las *n* iteraciones con diferentes técnicas estadísticas.
    c) Buscar el *cutoff* que minimice el número de Falsos Positivos y a la par se encuentre el máximo de precisión global, esto únicamente es posible a través la fórmula que proporcionamos en este documento.
5.) Se demuestra que se puede realizar el otorgamiento de crédito a segmentos como Microcrédito incluso a perfiles que prescinden de historiales de créditos robustos (No Hit) y aún así, ser rentables, con una tasa justa y con menor mora.
6.) Dados los sucesos del SARS-COVID19, en medio de la crisis Financiera-Económica, en donde la Banca se protege a través de tomar riesgos más conservadores, se considera que el conjunto de propuestas dadas en este documento engloban una estrategia integral adecuada para la administración del Riesgo de Crédito en cualquier Institución Financiera en México de tal manera que se puedan enfocar los recursos que se tienen usando a la ciencia de datos como principal eje, y de reorientar exclusiones definitivas del sistema financiero prospectos que tienen necesidad de crédito y no son atendidos por la banca tradicional.

Una vez dados estos créditos, presentamos un Estado de Resultados a través de la simulación de cartera considerando los montos originales y los ajustados por capacidad de pago, asignándoles la tasa mínima que calculamos para cada uno de los modelos, para corroborar que la tasa objetivo en cada uno pueden cubrir al menos los gastos operativos y financieros que de ellos deriven, calculando la posible utilidad o pérdida, para ello nuestra simulación de cartera, la realizamos a través de la plataforma R bajo los siguientes supuestos:

1.) Todos los créditos son otorgados a un plazo de 52 semanas (12 Meses) e inician al mismo tiempo, por lo que la tasa mínima, así como el valor pagaré (Capital más Interés e IVA), se dividirá sobre ese tiempo.
2.) El devengo del crédito se divide en las 52 semanas que tiene de plazo.
3.) El tipo de tasa de interés cobrada es sobre “Saldos Globales”, lo que permite que el devengo de capital e interés sea el mismo en todas las amortizaciones.
4.) No existe devengo por cargos moratorios, ni alguna consideración extra por el impago de un crédito.
5.) Los créditos si mantienen una clasificación de “Malos”, a través de una semilla aleatoria binaria se le asignará un único punto de inicio de impago (que representan el impago de más de 45 días definidos en este documento) dentro de los 12 meses y de ahí el crédito se comenzará a deteriorar en días de atraso de manera constante, es decir, que no tendrá posibilidad de cobranza y no podrá bajar el nivel en días de atraso. (La semilla aleatoria es igual para cada uno de los escenarios para evitar que se preste a diferentes escenarios)

6.) Los créditos dentro de la clasificación de "Buenos", asumiremos que pagan de manera constante y el devengo y recuperación del crédito se da de forma ininterrumpida
7.) Las Estimaciones Preventivas de Riesgo de Crédito (EPRC) se realizaran, bajo la metodología del Anexo D de las Disposiciones de Carácter General aplicables a Entidades de Ahorro y Crédito Popular con base en los días de atraso a cierre de mes y sobre el saldo insoluto que de la operación derive (Esto se obtendrá de los días de atraso derivado del punto 5 de este numeral)
8.) El interés devengado no cobrado se sujeta a lo dispuesto Anexo B-6 "Cartera de Crédito", emitida por la CNBV y visto en secciones anteriores de este documento donde se menciona que:
    a.) "Se deberá suspender la acumulación de los intereses devengados de las operaciones crediticias, en el momento en que el saldo insoluto del crédito sea considerado como vencido"
9.) El fondeo (Costo por Interés) será del 10% sobre el monto total desembolsado dividido en los 12 meses que dure la recuperación del crédito.
10.) No se castiga ningún crédito contable o fiscalmente, lo que permite llevar reservas integras y cartera incobrable al final del Estado de Resultados.
11.) Los impuestos que consideramos dentro del estado de resultados, engloba únicamente el IVA causado por el interés devengado cobrado y no cobrado del crédito.

Adicionalmente a esto, se consideraron distintos escenarios, para cada uno de los modelos.

1.) Escenario Original: Contempla la inclusión de todos los créditos "Buenos" y "Malos" en donde se le asignan los montos originales y la Tasa Acumulada total que derive por el total de operaciones originales.
2.) Escenario *Optimize*: Contempla la inclusión de todos los créditos "Buenos" y "Malos" en donde se le asignan los montos ajustados por capacidad de pago y la Tasa Acumulada total que derive por el total de operaciones con *exposure* ajustado.
3.) Escenario Mix1: Contempla la inclusión de todos los créditos "Buenos" y "Malos" en donde se le asignan los montos ajustados por capacidad de pago y la Tasa Acumulada total que derive por la tasa mínima del modelo sin ajuste por capacidad de pago (66.8% para el modelo HIT y 52.9% para el modelo NO HIT). Esto con el propósito de materializar las utilidades pérdidas derivadas de una tasa no adecuada de interés
4.) Escenario Mix2: Contempla la inclusión de todos los créditos "Buenos" y "Malos" en donde se le asignan los montos originales y la Tasa Acumulada total que derive por la tasa mínima del modelo con ajuste por capacidad de pago (48.7% para el modelo HIT y 39.1% para el modelo NO HIT). Esto con el propósito de materializar las utilidades pérdidas derivadas de una tasa no adecuada de interés.

Al aplicarlo en el Modelo HIT, obtuvimos los siguientes resultados:

Modelo HIT: Escenario Optimize				
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
Cartera	$110,293,808	$79,735,462	$50,740,885	$14,435,923
Ingreso por Interés	$4,416,833	$4,200,546	$3,979,724	$3,723,138

Costo por interés	$1,095,323	$1,095,323	$1,095,323	$1,095,323
Ingreso por interés Acum	$13,468,385	$26,290,651	$38,456,761	$53,690,304
Costo por interés Acum	$3,285,969	$6,571,938	$9,857,908	$14,239,200
Margen financiero	$3,321,510	$3,105,223	$2,884,400	$2,627,815
Margen financiero acum	$10,182,416	$19,718,713	$28,598,853	$39,451,104
Estimaciones Preventivas	$2,573,614	$7,958,357	$12,689,415	$16,590,360
Gasto Reserva	$795,077	$1,905,950	$1,462,548	$706,915
Utilidad Bruta Ejercicio	$7,608,801	$11,760,355	$15,909,438	$22,860,744
Impuestos	$961,731	$1,037,882	$1,115,623	$1,214,461
Impuestos Acumulados	$2,810,951	$5,848,797	$9,118,076	$13,834,644
Utilidad Neta	$4,797,851	$5,911,558	$6,791,362	$9,026,099
Modelo HIT: Escenario Original				
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
Cartera	$87,714,661	$64,257,544	$42,610,022	$16,369,162
Ingreso por Interés	$4,728,609	$4,383,578	$4,040,022	$3,649,995
Costo por interés	$868,704	$868,704	$868,704	$868,704
Ingreso por interés Acum	$14,517,340	$28,021,602	$40,484,081	$55,590,943
Costo por interés Acum	$2,606,112	$5,212,224	$7,818,335	$11,293,151
Margen financiero	$3,859,905	$3,514,874	$3,171,318	$2,781,291
Margen financiero acum	$11,911,228	$22,809,379	$32,665,745	$44,297,792
Estimaciones Preventivas	$2,571,905	$9,026,584	$15,014,053	$20,002,091
Gasto Reserva	$963,375	$2,312,438	$1,870,232	$941,265
Utilidad Bruta Ejercicio	$9,339,324	$13,782,794	$17,651,692	$24,295,701
Impuestos	$1,016,940	$1,137,658	$1,260,745	$1,409,973
Impuestos Acumulados	$2,937,867	$6,228,555	$9,889,270	$15,317,150
Utilidad Neta	$6,401,457	$7,554,239	$7,762,422	$8,978,551
Modelo HIT: Escenario Mix1				
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
Cartera	$110,293,808	$79,735,462	$50,740,885	$14,435,923
Ingreso por Interés	$6,059,995	$5,763,244	$5,460,271	$5,108,230
Costo por interés	$1,095,323	$1,095,323	$1,095,323	$1,095,323
Ingreso por interés Acum	$18,478,929	$36,071,369	$52,763,547	$73,664,313
Costo por interés Acum	$3,285,969	$6,571,938	$9,857,908	$14,239,200
Margen financiero	$4,964,672	$4,667,921	$4,364,948	$4,012,907
Margen financiero acum	$15,192,960	$29,499,431	$42,905,639	$59,425,113
Estimaciones Preventivas	$2,652,501	$8,304,236	$13,347,285	$17,638,519
Gasto Reserva	$837,432	$2,005,224	$1,575,853	$805,363
Utilidad Bruta Ejercicio	$12,540,459	$21,195,195	$29,558,354	$41,786,595
Impuestos	$1,254,319	$1,358,800	$1,465,463	$1,601,070
Impuestos Acumulados	$3,661,095	$7,633,497	$11,923,428	$18,133,877
Utilidad Neta	$8,879,364	$13,561,698	$17,634,926	$23,652,718
Modelo HIT: Escenario Mix2				
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
Cartera	$87,714,661	$64,257,544	$42,610,022	$16,369,162
Ingreso por Interés	$3,446,451	$3,194,975	$2,944,574	$2,660,302
Costo por interés	$868,704	$868,704	$868,704	$868,704
Ingreso por interés Acum	$10,580,977	$20,423,571	$29,506,860	$40,517,511
Costo por interés Acum	$2,606,112	$5,212,224	$7,818,335	$11,293,151
Margen financiero	$2,577,747	$2,326,271	$2,075,870	$1,791,599

Margen financiero acum	$7,974,865	$15,211,347	$21,688,525	$29,224,360
Estimaciones Preventivas	$2,490,720	$8,645,058	$14,269,108	$18,819,257
Gasto Reserva	$916,412	$2,199,210	$1,738,180	$831,421
Utilidad Bruta Ejercicio	$5,484,145	$6,566,289	$7,419,417	$10,405,103
Impuestos	$778,886	$866,871	$956,583	$1,065,348
Impuestos Acumulados	$2,254,330	$4,765,815	$7,546,993	$11,653,860
Utilidad Neta	$3,229,815	$1,800,475	-$127,577	-$1,248,756

Y para el modelo NO HIT, obtenemos los siguientes resultados:

Modelo NO HIT: Escenario Optimize				
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
Cartera	$66,445,269	$47,646,769	$29,550,692	$6,465,192
Ingreso por Interés	$2,159,642	$2,086,274	$2,004,120	$1,915,215
Costo por interés	$660,862	$660,862	$660,862	$660,862
Inrgeso por interés Acum	$6,558,488	$12,891,843	$18,988,547	$26,766,225
Costo por interés Acum	$1,982,585	$3,965,169	$5,947,754	$8,591,200
Margen financiero	$1,498,780	$1,425,412	$1,343,259	$1,254,354
Margen financiero acum	$4,575,903	$8,926,674	$13,040,793	$18,175,025
Estimaciones Preventivas	$1,350,045	$3,676,711	$5,590,550	$7,204,316
Gasto Reserva	$352,377	$798,537	$596,096	$271,817
Utilidad Bruta Ejercicio	$3,225,858	$5,249,963	$7,450,244	$10,970,709
Impuestos	$481,335	$507,478	$536,039	$570,111
Impuestos Acumulados	$1,416,069	$2,912,554	$4,491,028	$6,724,136
Utilidad Neta	$1,809,789	$2,337,409	$2,959,216	$4,246,573
Modelo NO HIT: Escenario Original				
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
Cartera	$52,232,572	$37,922,974	$24,411,788	$7,622,103
Ingreso por Interés	$2,255,317	$2,139,618	$2,012,960	$1,866,730
Costo por interés	$517,951	$517,951	$517,951	$517,951
Inrgeso por interés Acum	$6,892,837	$13,424,389	$19,591,509	$27,255,086
Costo por interés Acum	$1,553,854	$3,107,709	$4,661,563	$6,733,369
Margen financiero	$1,737,365	$1,621,667	$1,495,009	$1,348,779
Margen financiero acum	$5,338,983	$10,316,680	$14,929,946	$20,521,717
Estimaciones Preventivas	$1,386,996	$4,367,801	$6,773,322	$8,880,066
Gasto Reserva	$476,577	$1,002,898	$731,245	$386,004
Utilidad Bruta Ejercicio	$3,951,987	$5,948,879	$8,156,624	$11,641,651
Impuestos	$493,085	$534,902	$578,787	$633,530
Impuestos Acumulados	$1,433,525	$2,996,773	$4,686,591	$7,142,119
Utilidad Neta	$2,518,462	$2,952,107	$3,470,033	$4,499,532
Modelo NO HIT: Escenario Mix1				
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
Cartera	$66,445,269	$47,646,769	$29,550,692	$6,465,192
Ingreso por Interés	$2,925,904	$2,826,504	$2,715,201	$2,594,752
Costo por interés	$660,862	$660,862	$660,862	$660,862

Inrgeso por interés Acum	$8,885,502	$17,465,993	$25,725,866	$36,263,139
Costo por interés Acum	$1,982,585	$3,965,169	$5,947,754	$8,591,200
Margen financiero	$2,265,042	$2,165,642	$2,054,340	$1,933,891
Margen financiero acum	$6,902,917	$13,500,824	$19,778,112	$27,671,939
Estimaciones Preventivas	$1,380,791	$3,799,702	$5,817,394	$7,565,630
Gasto Reserva	$367,846	$832,122	$634,858	$304,530
Utilidad Bruta Ejercicio	$5,522,126	$9,701,122	$13,960,718	$20,106,309
Impuestos	$614,600	$650,019	$688,714	$734,874
Impuestos Acumulados	$1,805,953	$3,720,855	$5,746,837	$8,622,205
Utilidad Neta	$3,716,172	$5,980,266	$8,213,881	$11,484,104
Modelo NO HIT: Escenario Mix2				
	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
Cartera	$52,232,572	$37,922,974	$24,411,788	$7,622,103
Ingreso por Interés	$1,664,674	$1,579,276	$1,485,788	$1,377,854
Costo por interés	$517,951	$517,951	$517,951	$517,951
Inrgeso por interés Acum	$5,087,680	$9,908,690	$14,460,710	$20,117,281
Costo por interés Acum	$1,553,854	$3,107,709	$4,661,563	$6,733,369
Margen financiero	$1,146,723	$1,061,324	$967,837	$859,903
Margen financiero acum	$3,533,825	$6,800,982	$9,799,147	$13,383,912
Estimaciones Preventivas	$1,353,169	$4,223,498	$6,509,505	$8,460,969
Gasto Reserva	$457,588	$963,112	$687,406	$347,115
Utilidad Bruta Ejercicio	$2,180,657	$2,577,484	$3,289,643	$4,922,943
Impuestos	$385,655	$416,520	$448,912	$489,319
Impuestos Acumulados	$1,123,211	$2,342,171	$3,654,555	$5,553,820
Utilidad Neta	$1,057,446	$235,313	-$364,912	-$630,878

Con lo que concluimos los siguientes puntos:

7.) No es necesario excluir a personas del sistema financiero personas con un alta probabilidad de incumplimiento y mantener un margen similar de ganancias incluso bajando la tasa de interés; siempre y cuando se realice el ajuste de riesgo por capacidad de pago, esto se ve claramente cuando hacemos el comparativo entre el escenario Original y el escenario *Optimize*:

	HIT			NO HIT		
	Ajuste por Capacidad de Pago	Tasa Mínima Fórmula Propuesta	Utilidad	Ajuste por Capacidad de Pago	Tasa Mínima Fórmula Propuesta	Utilidad
Optimize (a)	SI	48.70%	$9,026,099	SI	39.10%	$4,246,573
Original (b)	NO	66.80%	$8,978,551	NO	52.90%	$4,499,532
Mix1 (c)	SI	66.80%	$23,652,718	SI	52.90%	$11,484,104
Mix2 (d)	NO	48.70%	-$1,248,756	NO	39.10%	-$630,878

8.) Tomando como referencia la utilidad del escenario Original y Mix1 en donde existe la misma tasa de interés pero con la diferencia que hubo un ajuste por capacidad de

pago en los créditos entre los escenarios, podemos observar que conduce una utilidad menor, esto sucede porque los recursos no se canalizan de manera eficiente y no hay una estrategia consistente de asignación de *exposure* por operación.

9.) También podemos concluir que aumentos en tasa de interés, ante coberturas por altos incumplimientos no representa necesariamente una mayor utilidad, incluso puede ser perjudicial, ya que como observamos en la sección 2, esto afecta directamente a las Estimaciones Preventivas por Riesgo de Crédito el cual considera la provisión por interés devengado no cobrado, el cual se detiene hasta que el crédito es consignado como "Vencido", por lo que genera requerimientos mayores monetarios, lo cual afecta brutalmente los Estados de Resultados.

10.) Aunado a esto, existe un beneficio colateral y es que, al bajar la tasa de interés a nivel agregado, la Institución Financiera se vuelve escalable y más atractiva al público demandante de crédito.

11.) La concepción del ajuste de Capacidad de Pago no se puede llevar a cabo si no se tiene la selección del mejor algoritmo dada una base de Datos, esto representa que para llegar a estos resultados tenemos que:

d) Establecer una metodología adecuada de transformación de de datos, ya que está sumamente ligado con el producto de modelación, en nuestro caso, el tratamiento de datos que más le brindó estabilidad a los modelos y a las diferentes técnicas fue a través de la metodología WOE propuesta por (Siddiqi, 2005).

e) Se tienen que probar las *n* iteraciones con diferentes técnicas estadísticas. En nuestro caso al poner a competir diversos algoritmos de ciencias de Datos a través del K-Fold Cross Validation, nos demostró que la técnica más adecuada para esta base de Datos para los dos modelos fue la Técnica Logit.

f) Una vez obtenido el mejor modelo, la búsqueda del *cutoff* que minimice el número de Falsos Positivos en la Matriz de confusión, ya que estos representan a los "Malos" que pasan como "Buenos" y que son los que en realidad materializan la pérdida de una Institución Financiera, de manera que dado ese riesgo aceptado se encuentre el mínimo y se busque el máximo de precisión global, esto únicamente es posible a través la fórmula que proporcionamos en este documento.

12.) Como se demostró a lo largo de este documento, se puede realizar el otorgamiento de crédito a segmentos como Microcrédito incluso a perfiles que prescinden de historiales de créditos robustos (NO HIT) y aún así, ser rentables, con una tasa justa y con menor mora, lo que apoya la tesis de Azevedo, Lafortune *et al*. (2019). Por lo que se rompe con la concepción tradicionalista bancaria de otorgar créditos a clientes con 'historia' y que incentiva el despliegue de la economía.

13.) Dados los sucesos del SARS-COVID19, en medio de la crisis Financiera-Económica, en donde la Banca se protege a través de tomar riesgos más conservadores, se considera que el conjunto de propuestas dadas en este documento engloban una estrategia integral adecuada para la administración del Riesgo de Crédito en cualquier Institución Financiera en México de tal manera que se puedan enfocar los recursos que se tienen usando a la ciencia de datos como principal eje, y de reorientar exclusiones definitivas del sistema financiero prospectos que tienen necesidad de crédito y no son atendidos por la banca tradicional.

Por último, es importante mencionar que el trabajo se puede extender en varias direcciones, por ejemplo, modelar factores de riesgo o fuentes de incertidumbre con el movimiento browniano o proceso de Wiener como en Venegas-Martínez (2008).

Bibliografía

Abanto, P. A. R., & Chávez, R. S. (s. f.). *Estructura del mercado de créditos y tasas de interés: Una aproximación al segmento de las microfinanzas*. 19.

Achieving the Sustainable Development Goals. (s. f.-a). Recuperado 21 de abril de 2020, de https://www.cgap.org/research/publication/achieving-sustainable-development-goals

Achieving the Sustainable Development Goals. (s. f.-b). Recuperado 16 de marzo de 2020, de https://www.cgap.org/research/publication/achieving-sustainable-development-goals

Albarrán Lozano, I., y Alonso González, P., Fundación MAPFRE, e Instituto de Ciencias del Seguro. (2010). *Métodos estocásticos de estimación de las provisiones técnicas en el marco de solvencia II*. Fundación Mapfre.

Amoroso, C., Noemi, D., y Martínez Jara, M. (2009). *Metodología para la evaluación de la capacidad de pago de un sujeto de crédito de consumo*. http://www.dspace.espol.edu.ec/handle/123456789/2110

Amoroso, D. N. C., y Jara, I. M. M. (s. f.). Metodología para la evaluación de la capacidad de pago de un sujeto de crédito de consumo.

Angelucci, M., Karlan, D., and Zinman, J. (2015). Microcredit Impacts: Evidence from a Randomized Microcredit Program Placement Experiment by Compartamos Banco. *American Economic Journal: Applied Economics*, *7*(1), 151-182. https://doi.org/10.1257/app.20130537

Balmaseda Pérez, B. I., Necoechea Hasfield, L., Balmaseda Pérez, B. I., y Necoechea Hasfield, L. (2013). Metodología de estimación del número de clientes del sistema bancario en México. *El trimestre económico*, *80*(320), 943-963.

Banco de Pagos Internacionales. (2010). *Actividades de microfinanciación y los Principios Básicos para una supervisión bancaria eficaz.* (Basilea, Suiza).

Barona, C. V. B., & Villegas, I. M. U. (s. f.). Determinantes de la probabilidad de morosidad en la cartera de microcrédito.

BBVA Financial Report. (2010). *Probabilidad de incumplimiento (PD)*. https://accionistaseinversores.bbva.com/microsites/informes2010/es/Gestiondelriesgo/ProbabilidaddeincumplimientoPD.html

Barraza, J. E. (2015). *Modelando time to default sensible al contexto sistémico en carteras de consumo* [Text, Universidad de Buenos Aires. Facultad de Ciencias Económicas.]. http://bibliotecadigital.econ.uba.ar/?c=tpos&a=d&d=1502-0868_BarrazaJE

Beck, T., Levine, R., and Loayza, N. (2000). Finance and the sources of growth. *Journal of Financial Economics*, *58*(1), 261-300. https://doi.org/10.1016/S0304-405X(00)00072-6

Bejarano, L., y Fernando, V. (2015). *Modelo de suficiencia de capital utilizando una medida de concentración para la cartera de crédito comercial de una institución financiera*. http://bibdigital.epn.edu.ec/handle/15000/12555

Beltrán Pascual, M. (2015). *Diseño e implementación de un nuevo clasificador de préstamos bancarios a través de la minería de datos*. http://e-spacio.uned.es/fez/view/tesisuned:CiencEcoEmp-Mbeltran

Benchmarking de las microfinanzas en México 2015-2016: Un informe del sector (p. 142). (2016). [Documento]. ProDesarrollo: Finanzas y Microempresa A.C. https://www.microfinancegateway.org/es/library/benchmarking-de-las-microfinanzas-en-m%C3%A9xico-2015-2016-un-informe-del-sector

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152. https://doi.org/10.1145/130385.130401

Bruhn, M., & Love, I. (2014). The Real Impact of Improved Access to Finance: Evidence from Mexico. *Journal of Finance*, *69*(3), 1347-1376.

Cáceres, F. U., y Palacios, Y. A. (2017). Análisis de supervivencia como alternativa metodológica para estimar probabilidades de incumplimiento de los deudores de créditos corporativos y a grandes empresas en el Perú. *Industrial Data*, *20*(1), 7-16. https://doi.org/10.15381/idata.v20i1.13486

Cálculo del valor en riesgo operacional mediante redes bayesianas para una empresa financiera. (2016). *Contaduría y Administración*, *61*(1), 176-201. https://doi.org/10.1016/j.cya.2015.09.009

Carlos, J., & García-Cespedes, J.-C. (2018). *Nuevas técnicas de medición del riesgo de crédito*.

Castillo, F., & León, J. (2019). *"SOBRE EL APETITO POR EL RIESGO EN EL MERCADO DE MICROCRÉDITOS: EL CASO DE LAS CAJAS MUNICIPALES"*. 26.

Castillo Polanco, L. Alfredo. (2013). Determinantes del desempeño de la tecnología del microcrédito individual. *Investigación económica*, *72*(285), 115-140.

CGAP-Occasional-Paper-Microcredit-Interest-Rates-Nov-2002-Spanish.pdf. (s. f.). Recuperado 19 de junio de 2018, de https://www.cgap.org/sites/default/files/CGAP-Occasional-Paper-Microcredit-Interest-Rates-Nov-2002-Spanish.pdf

Chen, G. (s. f.). *Gráfico 1: Cuatro países que recientemente tuvieron crisis de pago en el sector microfinanciero*. 20.

Chiapa, C. (s. f.). *Diagnóstico de las Políticas Públicas de Microcrédito del Gobierno Federal*. 263.

Chong, A., y Schroth, E. (s. f.). *Cajas municipales, microcrédito y pobreza en el Perú*. 39.

Clavellina Miller, J. L. (2013). Crédito bancario y crecimiento económico en México. *Economía Informa*, *378*, 14-36. https://doi.org/10.1016/S0185-0849(13)71306-9

Clavellina Miller, J. L., & Domínguez Rivas, M. I. (2020). Implicaciones económicas de la pandemia por COVID-19 y opciones de política. http://bibliodigitalibd.senado.gob.mx/handle/123456789/4829

Clavellina-Miller, J. L. (2013, febrero 28). *(PDF) Crédito bancario y crecimiento económico en México*. https://www.researchgate.net/publication/276378768_Credito_bancario_y_crecimiento_economico_en_Mexico

CNBV. (2006).*Disposiciones de carácter general aplicables a las entidades de ahorro y crédito popular, organismos de integración, sociedades financieras comunitarias y organismos de integración financiera rural, a que se refiere la ley de ahorro y crédito popular*.

CNBV. (2018). *Disposiciones de Carácter General aplicables a las Instituciones de Crédito.*

Collins, M. C., Harvey, K. D., & Nigro, P. J. (2002). The Influence of Bureau Scores, Customized Scores and Judgmental Review on the Bank Underwriting Decision-Making Process. *Journal of Real Estate Research*, *24*(2), 129-152.

Dellien, H., and Schreiner, M. (s. f.). *Credit Scoring, Banks, and Microfinance: Balancing High-Tech with High-Touch*. 16.

Donou-Adonsou, F., and Sylwester, K. (2016). Financial development and poverty reduction in developing countries: New evidence from banks and microfinance institutions. *Review of Development Finance*, *6*(1), 82-90. https://doi.org/10.1016/j.rdf.2016.06.002

Embríz, F. A., Diez-Canedo, J. M., y Aranda, A. R. (s. f.). *Implantación del Modelo CyRCE:* 45.

Espin-García, O., y Rodríguez-Caballero, C. V. (2013). Metodología para un scoring de clientes sin referencias crediticias. *Cuadernos de Economía*, *32*(59), 139-165.

Esquivel, H. (s. f.). *Medición del efecto de las microfinanzas en México*. 19.

Esquivel, R. S. (2017). *Microfinanzas. Resultados financieros y sociales: México y Perú*. *10*(27), 22.

Fernández, D. (s. f.). *Suficiencia del capital y previsiones de la banca uruguaya por su exposición al sector industrial*. Banco Central de Uruguay.

Fernández, D., y Netto, R. S. (s. f.). *Valor en riesgo de las carteras de préstamos bancarios*. 43.

Figueroa, M. (2006). *Minería de datos aplicada a Credit Scoring*. http://repositorio.usfq.edu.ec/handle/23000/547

Financial development and poverty reduction in developing countries: New evidence from banks and microfinance institutions | Elsevier Enhanced Reader. (s. f.). https://doi.org/10.1016/j.rdf.2016.06.002

G20-Policy-Guide-Digitisation-and-Informality.pdf. (s. f.). Recuperado 16 de mayo de 2020, de http://www.oecd.org/g20/G20-Policy-Guide-Digitisation-and-Informality.pdf

Gómez Jacinto, L. G. (2008). Información Asimétrica: Selección Adversa y Riesgo Moral. *Actualidad Empresarial*, *170*.

Gómez-González, J. E., Morales-Acevedo, P., Pineda, F., & Zamudio-Gómez, N. E. (2007). Estimación de matrices de transición de la calidad de cartera comercial de las entidades financieras colombianas.

Hernández Corrales, L., Cerón, M., Ángel, L., y Benavides, J. (2005). Desarrollo de una metodología propia de análisis de crédito empresarial en una entidad financiera. *Estudios Gerenciales*, *21*(97), 129-165.

Grimshaw, S. D., & Alexander, W. P. (s. f.). Markov Chain Models for Delinquency: Transition Matrix Estimation and Forecasting. 42.

Guerrero, M., & Rolando, D. (2009). Optimización del capital económico mediante la diversificación de una cartera de crédito: Caso práctico para una Institución Financiera. http://bibdigital.epn.edu.ec/handle/15000/8199

Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring. ACM Transactions on the Web (TWEB), 10. https://doi.org/10.1145/2996465

Hadzimustafa, S., & Cipusheva, H. (2012). Microfinancing as a Poverty Reduction Tool (SSRN Scholarly Paper ID 2368787). Social Science Research Network. https://papers.ssrn.com/abstract=2368787

HR Ratings. (2016). *Microfinancieras en México. Análisis Sectorial*.

INEGI. (2018). *Encuesta Nacional de Inclusión Financiera*.

Ivanciuc, O. (2007). Applications of Support Vector Machines in Chemistry. En Reviews in Computational Chemistry (pp. 291-400). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470116449.ch6

Jung, T., & Strohhecker, J. (2009). Risk-adjusted pricing strategies for the corporate loans business: Do they really create value? System Dynamics Review, 25(4), 251-279. https://doi.org/10.1002/sdr.429

Lara Rubio, J. (2010a). *La gestión del riesgo de crédito en las instituciones de microfinanzas*. Editorial de la Universidad de Granada.

Lara Rubio, J. (2010b). *La gestión del riesgo de crédito en las instituciones de microfinanzas*. Editorial de la Universidad de Granada.

Leyshon, A., & Thrift, N. (1995). Geographies of Financial Exclusion: Financial Abandonment in Britain and the United States. Transactions of the Institute of British Geographers, 20(3), 312-341. JSTOR. https://doi.org/10.2307/622654

López, R. A. C. (s. f.). *La importancia de áreas de administración de riesgos en las entidades de ahorro y crédito popular*. 18.

Maldonado, D., y Pazminño, M. (2008). Nuevas herramientas para la Administración del Riesgo Crediticio: El caso de una Cartera Crediticia Ecuatoriana. *Cuestiones Económicas*, *24*(2).

Marimo, M. (2015). *Survival Analysis of Bank Loans and Credit Risk Prognosis*. University of the Witwatersrand.

Martínez Sánchez, J. F. (2012). *Riesgo Operacional en el Mercado de Dinero*. Instituto Politécnico Nacional.

Martínez Sánchez, J. F., y Venegas Martínez, F. (2013). Riesgo operacional en el proceso de pago del Procampo Un enfoque bayesiano. *Contaduría y Administración*, *58*(2), 221-259. https://doi.org/10.1016/S0186-1042(13)71216-6

Martínez-Sánchez, J. F., y Venegas-Martínez, F. (2013). Riesgo operacional en el proceso de liquidación del mercado mexicano de valores: Un enfoque bayesiano. *Investigación Económica*, *72*(286), 101-138.

Marulanda Consultores en colaboración con DAI. (2011). *Estudio. Microfinanzas en México*.

Méndez, M. C. (s. f.). *Modelización estadística con Redes Neuronales. Aplicaciones a la Hidrología, Aerobiología y Modelización de Procesos*. 161.

Mermelstein, D. (2006). *Defaults en carteras hipotecarias, macroeconomía y arreglos institucionales: Más allá de los modelos de Credit-Scoring tradicionales.* Universidad de Buenos Aires. Facultad de Ciencias Económicas.

Mermelstein, D. A. (s. f.). *Credit scoring: del uso tradicional, al pricing ajustado por riesgo*. 5.

Method 8000c - determinative chromatographic separations. (2003). 66.

Microfinanzas en México: Evolución hacia el impacto social. (2015). [Actualidad]. PRONAFIM. https://www.microfinancegateway.org/es/library/microfinanzas-en-m%C3%A9xico-evoluci%C3%B3n-hacia-el-impacto-social

Miled, K. B. H., & Rejeb, J. E. B. (2015). Microfinance and Poverty Reduction: A Review and Synthesis of Empirical Evidence. *Procedia - Social and Behavioral Sciences*, *195*, 705-712. https://doi.org/10.1016/j.sbspro.2015.06.339

Nicolo, G., Jalal, A. M., & Boyd, M. (s. f.). Bank Risk-Taking and Competition Revisited: New Theory and New Evidence. IMF. Recuperado 26 de julio de 2020, de https://www.imf.org/en/Publications/WP/Issues/2016/12/31/Bank-Risk-Taking-and-Competition-Revisited-New-Theory-and-New-Evidence-20126

Nieto Murillo, S., Pérez Salvador, B. R., & Soriano Flores, J. F. (2011). Crédito al Consumo: *La Estadística aplicada. Actuarios Trabajando: Revista Mexicana de investigación actuarial aplicada., 6, 109.*

Niu, B., Ren, J., & Li, X. (2019). Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending. Information, 10, 397. https://doi.org/10.3390/info10120397Ospina Cardona, C. (2015). *Modelo avanzado para administrar la cartera crediticia en la empresa*. http://repositorio.utp.edu.co/dspace/handle/11059/6054

Ospina Cardona, C. (2015). Modelo avanzado para administrar la cartera crediticia en la empresa. http://repositorio.utp.edu.co/dspace/handle/11059/6054

Otero, M., Rhyne, E., & Martínez Vázquez, S. (1998). El nuevo mundo de las finanzas microempresariales: Estructuración de instituciones financieras sanas para los pobres. *Plaza y Valdés : Servicios de Apoyo Local al Desarrollo de Base en México*.

Perossa, M. L., y Gigler, S. (2015). Modelos microfinancieros latinoamericanos: Una experiencia para la inclusión social y el desarrollo. *Cooperativismo & Desarrollo*, *23*(106). https://doi.org/10.16925/co.v23i106.1124

Pineda, G. L. A., y Gómez, J. G. M. (2010). *Modelos de pérdidas agregadas (LDA) y de la teoría del valor extremo para cuantificar el riesgo operativo teoría y aplicaciones*. *2010*, 134.

Raccanello, K., y Roldán-Bravo, G. (2014). Instituciones microfinancieras y cajas de ahorro en Santo Tomás Hueyotlipan, Puebla. *Economía Sociedad y Territorio*. https://doi.org/10.22136/est00201435

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied regression analysis: A research tool* (2nd ed). Springer.

Rayo Cantón, S., Lara Rubio, J., y Camino Blasco, D. (2010). Un Modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea II. *Journal of Economics, Finance and Administrative Science*, *15*(28), 89-124.

Rayo Cantón, S., Lara-Rubio, J., y Blasco, D. (2010). Un Modelo de Credit Scoring para instituciones de microfinanzas en el marco de Basilea II. *Journal of Economics, Finance and Administrative Science*, *15*, 89-124.

Rentería, V., Pilar, V. D., Valencia, Z., y Luis, J. (2012). *Cálculo de la probabilidad de default para una cartera de créditos vehiculares*. http://www.dspace.espol.edu.ec/handle/123456789/24961

Rodríguez Vázquez, V. P., & Hernández Vaquero, J. (s. f.). Matriz de probabilidad de transición de microcréditos: el caso de una microfinanciera mexicana. *Estudios Económicos. El Colegio de México*, *28*(1), 39-77.

Rozo, G., Andrea, E., Montoya, R., & Enrique, L. (2014). Pricing de Crédito en la Banca Comercial Colombiana. *reponame:Repositorio Colegio de Estudio Superiores de Administración (CESA)*. http://repository.cesa.edu.co/handle/10726/1226

Salamanca, T., y Edwin, S. (2014). Macro credit scoring como propuesta para cuantificar el riesgo de crédito. *Investigación y Desarrollo*, *2*(14), 42-63.

Sandiás, A. R., López, S. F., y González, L. O. (1999). Estimación de la capacidad de endeudamiento del proyecto: Propuesta de un modelo de cobertura temporal. *La gestión de la diversidad : XIII Congreso Nacional, IX Congreso Hispano-Francés, Logroño (La Rioja), 16, 17 y 18 de junio, 1999, Vol. 1, 1999, ISBN 84-95301-10-5, págs. 761-772*, 761-772. https://dialnet.unirioja.es/servlet/articulo;jsessionid=C0B4E2DB83FD6E6150B7698F7F741EA5.dialnet01?codigo=565104

Schreiner, M. (2003). *Scoring: The next breakthrough in microcredit ?* World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail

Seijas-Giménez, M. N., Vivel-Búa, M., Lado-Sestayo, R., y Fernández-López, S. (2017). Vista de La evaluación del riesgo de crédito en las instituciones de microfinanzas. *COMPENDIUM*, *4*(9), 36-52.

Siddiqi, N. (2005). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Sas Inst.

Simbaña, M., y Rodrigo, R. (2013). *Modelación e implementación del control de riesgo operativo del área de riesgos mediante la simulación de Montecarlo y Redes Bayesianas para el caso de la cooperativa de ahorro y crédito Cotocollao*. http://bibdigital.epn.edu.ec/handle/15000/8067

Soto Esquivel, R. (2017). Microfinanzas: Resultados financieros y sociales: México y Perú. *Ola Financiera*, *10*(27). http://dx.doi.org/10.22201/fe.18701442e.2017.27.61005

Támara-Ayús, A., Aristizábal, R., y Velásquez, E. (2012). Matrices de transición en el análisis del riesgo crediticio como elemento fundamental en el cálculo de la pérdida esperada en una institución financiera colombiana. *Revista Ingenierías Universidad de Medellín. Vol. 11, (20), 2012, pp.15-120*. http://repository.eafit.edu.co/handle/10784/7632

Téllez Cabrera, M. R. (2010). *Medición del riesgo en crédito: Implementación y cálculo del VaR y el CVaR en tres modelos de incumplimiento.* Universidad Autónoma Metropolitana.

The CFPB Office of Research. (2015, mayo). Data point: Credit invisibles. Consumer Financial Protection Bureau. https://www.consumerfinance.gov/data-research/research-reports/data-point-credit-invisibles/

Thomas, L., Crook, J., & Edelman, D. (2017). *Credit Scoring and Its Applications, Second Edition*. SIAM.

Valencia, G. A. D. (2010). Las imperfecciones del mercado de créditos, la restricción crediticia y los créditos alternativos. Revista CIFE: Lecturas de Economía Social, 12(17), 103-134.

Velásquez, A., E, R., Ayús, T., Lenin, A., & Velásquez Ceballos, E. (2010). *Modelación de riesgo crediticio como elemento fundamental en el cálculo de la pérdida esperada en una institución financiera*. http://repository.eafit.edu.co/handle/10784/1187

Venegas-Martínez, F. (2008). Riesgos financieros y económicos: Productos derivados decisiones económicas bajo incertidumbre. Segunda edición. Editorial Cengage Learning. México.

Vergara, D., y Antonio, M. (2010). Método de Segmentación Utilizando Análisis de Supervivencia. *Repositorio Académico - Universidad de Chile*. http://www.tesis.uchile.cl/tesis/uchile/2010/cf-duarte_mv/html/index-frames.html

Vigano', L. (1993). A credit scoring model for development banks: An African case study. *Savings And Development*, *17*(4), 441-482.

Villafani-lbarnegaray, M., y Gónzales-Vega, C. (2007). Tasas de interés y desempeño diferenciado de cartera de las entidades de microfinanzas ante múltiples shocks sistémicos: ¿Se cumple el teorema de Stiglitz y Weiss en las microfinanzas bolivianas? *Revista Latinoamericana de Desarrollo Económico*, *8*, 11-52.

Vogelgesang, U. (2003). *Microfinance in Times of Crisis: The Efects of Competition, Rising Indebtedness, and Economic Crisis on Repayment Behaviour. 50*.

Yang, A. (2013). *Research on the Pricing Formula on Small Loan for RCC in China*. International Academic Workshop on Social Science (IAW-SC-13). https://doi.org/10.2991/iaw-sc.2013.104

Anexo I. Índice Matemático

1. Formula de regresión. $y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + e$
2. Modelo lineal de probabilidad $y \cong \frac{G}{(G+B)}$
3. Regresión logística $y \cong e^{\frac{G}{B}}$
4. Distancia a la predeterminada $\frac{A-D}{\sigma_A}$
5. Modelización de la probabilidad lineal $P(Good)_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + e_i$
6. Error estándar $s_e = \sqrt{\frac{\sum\left(\hat{Y}_i - Y_i\right)^2}{n-1-k}}$
7. Coeficiente de determinación $R^2 = 1 - \frac{\sum\left(\hat{Y}_i - Y_i\right)^2}{\sum\left(\bar{Y} - Y_i\right)^2}$
8. Regresión de Logit $In\left(\frac{p(Good)}{1-p(Good)}\right) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$

9. Estadística de divergencia $D^2 = \frac{\left(\pi_G - \pi_B\right)^2}{\frac{\left(\sigma_G^2 + \sigma_B^2\right)}{2}}$

10. El peso de la evidencia (WOE) $W_i = Ln\left(\left(\frac{N_i}{\sum N}\right)\middle|\left(\frac{P_i}{\sum P}\right)\right)$

11. Valor de la información (Information Value) $F = \sum_{i=1}^{n}\left[\left(\frac{N_i}{\sum N} - \frac{P_i}{\sum P}\right) \times WoE_i\right]$

12. Estabilidad de la población $F = \sum_{i=1}^{n}\left[\left(\frac{O_i}{\sum O} - \frac{E_i}{\sum E}\right) \times Ln\left(\frac{O_i}{\sum O} / \frac{E_i}{\sum E}\right)\right]$

13. La estadística del KS $D_{ks} = \max\left\{abs\left(cpY - cpX\right)\right\}$

14. Correlación de Pearson $r = \frac{N\sum XY - \sum X \sum Y}{N\sqrt{\left(\sum X^2 - \left(\sum X\right)^2\right)\left(\sum Y^2 - \left(\sum Y\right)^2\right)}}$

15. La correlación del orden de rango de Spearman $r_s = 1 - 6\frac{\sum\left(x_R - y_R\right)^2}{N^3 - N}$

16. Coeficiente de Gini $D = 1 - \sum_{i=1}^{n}\left(\left(cpY_i - cpY_{i-1}\right)\left(cpX_i + cpX_{i-1}\right)\right)$

17. AUROC $C_{P,N} = \Pr\left[S_{TP} < S_{TN}\right] + \frac{1}{2}\Pr\left[S_{TP} = S_{TN}\right]$

18. Chi-cuadrada de Pearson $\chi^2 = \sum_{i=1}^{n}\left(\frac{\left(O_i - E_i\right)^2}{E_i}\right)$

19. IHH $IHH = \sum_{i=1}^{N} s^2$

Anexo II. Especificación Query's R (Cross Validation)

```
library(FactoMineR)
library(factoextra)
library(dplyr)
library(RColorBrewer)
library(pROC)
library(ggplot2)
```

```
library(ggpubr)
library(corrplot)
library(extrafont)
library(caret)
library(e1071)
library(randomForest)
library(ROCR)
library(rpart)
library(ada)
library(e1071)
library(xgboost)
library(naivebayes)
library(caret)
library(kknn)
library(nnet)
library(janitor)

LOCVLogit2 <- Base %>%
select(Buenos,WOEEdad,WOEGiro,WOEMunicipio,WOETipoVivienda,
                    WOEEstado,WOEEstadoCivil,WOEConsultas,WOESexo)

LOCVLogit2$Buenos <- factor(LOCVLogit2$Buenos ,ordered = FALSE)
LOCVLogit2 <- as.data.frame(LOCVLogit2)
n <- dim(LOCVLogit2)[1]
deteccion.no.svm<-rep(0,5)
deteccion.no.knn<-rep(0,5)
deteccion.no.arbol<-rep(0,5)
deteccion.no.bosque<-rep(0,5)
deteccion.no.potenciacion<-rep(0,5)
deteccion.no.red<-rep(0,5)
deteccion.no.Bayes<-rep(0,5)
deteccion.no.boost<-rep(0,5)


deteccion.error.svm<-rep(0,5)
deteccion.error.knn<-rep(0,5)
deteccion.error.arbol<-rep(0,5)
deteccion.error.bosque<-rep(0,5)
deteccion.error.potenciacion<-rep(0,5)
deteccion.error.red<-rep(0,5)
deteccion.error.bayes<-rep(0,5)
deteccion.error.boost<-rep(0,5)


# Validación cruzada 5 veces
for(i in 1:5) {
  grupos <- createFolds(1:n,10) # Crea los 10 grupos
```

```
  no.svm<-0
  no.knn<-0
  no.arbol<-0
  no.bosque<-0
  no.potenciacion<-0
  no.red<-0
  no.Bayes<-0
  no.boost<-0

  error.svm<-0
  error.knn<-0
  error.arbol<-0
  error.bosque<-0
  error.potenciacion<-0
  error.red<-0
  error.bayes<-0
  error.boost<-0

  # Este ciclo es el que hace "cross-validation" (validación cruzada) con 10 grupos (Folds)
  for(k in 1:10) {
    muestra <- grupos[[k]] # Por ser una lista requiere de doble paréntesis
    ttesting <- LOCVLogit2[muestra,]
    taprendizaje <- LOCVLogit2[-muestra,]

    modelo <- svm(Buenos~., data=taprendizaje,kernel ="radial")
    prediccion <- predict(modelo,ttesting[,-which(names(LOCVLogit2) %in%
c("Buenos"))])
    Actual<-ttesting[,which(names(LOCVLogit2) %in%  c("Buenos"))]
    MC<-table(Actual,prediccion)
    # Detección de los Sí detectados
    no.svm<-no.svm+MC[2,2]
    error.svm<-error.svm+(1-(sum(diag(MC)))/sum(MC))*100



    modelo <- train.kknn(Buenos~.,data=taprendizaje,kmax=7)
    prediccion<-predict(modelo,ttesting[,-which(names(LOCVLogit2) %in%
c("Buenos"))])
    Actual<-ttesting[,which(names(LOCVLogit2) %in%  c("Buenos"))]
    MC<-table(Actual,prediccion)
    # Detección de los Sí detectados
    no.knn<-no.knn+MC[2,2]
    error.knn<-error.knn+(1-(sum(diag(MC)))/sum(MC))*100


    modelo = rpart(Buenos~.,data=taprendizaje)
```

```
   prediccion <- predict(modelo, ttesting[,-which(names(LOCVLogit2) %in%
c("Buenos"))], type='class')
   Actual<-ttesting[,which(names(LOCVLogit2) %in%  c("Buenos"))]
   MC<-table(Actual,prediccion)
   # Detección de los Sí detectados
   no.arbol<-no.arbol+MC[2,2]
   error.arbol<-error.arbol+(1-(sum(diag(MC)))/sum(MC))*100

   modelo<-randomForest(Buenos~.,data=taprendizaje,importance=TRUE)
   prediccion<-predict(modelo, ttesting[,-which(names(LOCVLogit2) %in%
c("Buenos"))])
   Actual<-ttesting[,which(names(LOCVLogit2) %in%  c("Buenos"))]
   MC<-table(Actual,prediccion)
   # Detección de los Sí detectados
   no.bosque<-no.bosque+MC[2,2]
   error.bosque<-error.bosque+(1-(sum(diag(MC)))/sum(MC))*100

   modelo<-ada(Buenos~.,data=taprendizaje,iter=60,nu=1,type="real")
   prediccion<-predict(modelo, ttesting[,-which(names(LOCVLogit2) %in%
c("Buenos"))])
   Actual<-ttesting[,which(names(LOCVLogit2) %in%  c("Buenos"))]
   MC<-table(Actual,prediccion)
   # Detección de los Sí detectados
   no.potenciacion<-no.potenciacion+MC[2,2]
   error.potenciacion<-error.potenciacion+(1-(sum(diag(MC)))/sum(MC))*100

   modelo<-nnet(Buenos~.,data=taprendizaje,size=5,rang=0.1,decay=5e-
4,maxit=100,trace=FALSE)
   prediccion<-predict(modelo, ttesting[,-which(names(LOCVLogit2) %in%
c("Buenos"))],type = "class")
   Actual<-ttesting[,which(names(LOCVLogit2) %in%  c("Buenos"))]
   MC<-table(Actual,prediccion)
   correc<- ifelse ( dim(MC)[2]==2, MC[2,2], 0)
   #Detección de los Sí detectados
   no.red<-no.red+correc
   error.red<-error.red+(1-(sum(diag(MC)))/sum(MC))*100

   modelo<-naiveBayes(Buenos~.,data=taprendizaje)
   prediccion<-predict(modelo, ttesting[,-which(names(LOCVLogit2) %in%
c("Buenos"))])
   MC<-table(ttesting[,which(names(LOCVLogit2) %in%  c("Buenos"))]  ,prediccion)
   # Detección de los Sí detectados
   no.Bayes<-no.Bayes+MC[2,2]
   error.bayes<-error.bayes+(1-(sum(diag(MC)))/sum(MC))*100

   taprendizajeboost <- taprendizaje
   ttestingboost <- ttesting
```

```
   valor.variable.predecir <- ttestingboost$Buenos
   taprendizajeboost[] <- lapply(taprendizajeboost, as.numeric)
   ttestingboost[] <- lapply(ttestingboost, as.numeric)

   # Recodifica la variable a predecir pues solo trabaja con "1" y "0"
   taprendizajeboost$Buenos <- as.numeric(ifelse(taprendizajeboost$Buenos == 2, 1, 0))
   ttestingboost$Buenos <- as.numeric(ifelse(ttestingboost$Buenos == 2, 1, 0))

   taprendizajeboost <- xgb.DMatrix(data = data.matrix(taprendizajeboost[,-
which(names(taprendizajeboost) %in%  c("Buenos"))]),label =
data.matrix(taprendizajeboost$Buenos))

   ttestingboost <- xgb.DMatrix(data = data.matrix(ttestingboost[,-
which(names(ttestingboost) %in%  c("Buenos"))]),label =
data.matrix(ttestingboost$Buenos))


   parametros <- list(booster = "gbtree", objective = "binary:logistic", eta=0.3,
                      gamma=0, max_depth=6, min_child_weight=1, subsample=1,
colsample_bytree=1)
   modelo_boost <- xgb.train (params = parametros, data = taprendizajeboost, nrounds =
79,
                           watchlist = list(train=taprendizajeboost, test=ttestingboost),
print_every_n = 10,
                           early_stop_round = 10, maximize = F , eval_metric = "error")

   prediccion_boost<-predict(modelo_boost,ttestingboost)

   prediccion_boost<-ifelse (prediccion_boost > 0.5, 1, 0)

   MC <- table(prediccion_boost, valor.variable.predecir)
   no.boost<-no.boost+MC[2,2]
   error.boost<-error.boost+(1-(sum(diag(MC)))/sum(MC))*100



 }
 deteccion.no.svm[i]<-no.svm
 deteccion.no.knn[i]<-no.knn
 deteccion.no.arbol[i]<-no.arbol
 deteccion.no.bosque[i]<-no.bosque
 deteccion.no.potenciacion[i]<-no.potenciacion
 deteccion.no.red[i]<-no.red
 deteccion.no.Bayes[i]<-no.Bayes
 deteccion.no.boost[i]<-no.boost
```

```
  deteccion.error.svm[i]<-error.svm/10
  deteccion.error.knn[i]<-error.knn/10
  deteccion.error.arbol[i]<-error.arbol/10
  deteccion.error.bosque[i]<-error.bosque/10
  deteccion.error.potenciacion[i]<-error.potenciacion/10
  deteccion.error.red[i]<-error.red/10
  deteccion.error.bayes[i]<-error.bayes/10
  deteccion.error.boost[i]<-error.boost/10

}
```