Exposure to default: estimation for a credit card portfolio

Bambino-Contreras, Carlos and Morales-Oñate, Víctor

Banco Solidario, Universidad San Francisco de Quito, Escuela Superior Politécnica de Chimborazo, Universidad Técnica de Ambato

December 2021

Exposición al *default*: estimación para un portafolio de tarjeta de crédito

Carlos Bambino-Contreras[1], Víctor Morales-Oñate[1,2,3,4]

[1]Banco Solidario, División de Riesgos, Analítica de Datos, Quito, Ecuador

[2]Universidad San Francisco de Quito, Colegio de Administración y Economía, Quito, Ecuador

[3]Escuela Superior Politécnica de Chimborazo, Data Science Research Group - CIDED, Riobamba, Ecuador

[4]Universidad Técnica de Ambato, Territorial Development, Business and Innovation Research Group - DeTEI, Ambato, Ecuador

Diciembre 2021

Resumen

Este trabajo estima la exposición al riesgo de la cartera de clientes de un Banco ecuatoriano sin hacer uso del factor de conversión de crédito, mecanismo habitual empleado en la literatura de estimación de pérdidas y sugerido por el Comité de Basilea. Para lograr este objetivo se ha identificado la distribución de probabilidad de esta variable (exposición al default) para poder emplearla en un contexto de modelos lineales generalizados. Los resultados muestran que se puede usar el modelo para realizar predicciones basadas en supuestos más cercanos a la realidad del comportamiento de los clientes en función de las variables utilizadas en la regresión.

Keywords— Pérdida esperada, Riesgo de crédito, Exposición al riesgo, Modelos lineales generalizados, Distribución Gamma, Aprendizaje Automático

Abstract

This work estimates the exposure at default of a credit card portfolio of an Ecuadorian bank without using the credit conversion factor, a common mechanism used in the expected loss distribution estimation literature and suggested by the Basel Committee. To achieve this goal, the probability distribution of this variable (exposure at default) has been identified so that it can be used in the context of generalized linear models. The results show that the model can be used to make predictions based on assumptions closer to the reality of customer behavior based on the variables used in the regression.

Keywords— Expected loss, Credit risk, Exposure at default, Generalized linear models, Gamma Distribution, Machine Learning.

1. INTRODUCCIÓN

El negocio de intermediación financiera es una de las principales, sino la principal actividad de la industria bancaria, generadora de una parte importante de sus beneficios y como consecuencia generadora también de una parte importante de sus riesgos.

Este proceso de transformación de activos y pasivos o, en otras palabras, este proceso de intercambio (compra - venta) de riesgos en el que participan las entidades financieras está sujeto a una variedad de riesgos financieros y operativos. Uno de ellos el riesgo de crédito o de contraparte (counterparty risk), inherente a la gestión de carteras que tienen cuentas pendientes de cobro.

Si bien la palabra riesgo puede tener una connotación negativa, es posible usar una de sus tantas explicaciones etimológicas que a la vez nos sugiere la forma del cómo se debe abordarlo. García y Sánchez. (2005) señalan que el vocablo riesgo proviene del latín risicare *atreverse o transitar por un sendero peligroso*. Es en este sentido que las instituciones financieras deben abordar la gestión de riesgo de crédito, porque el negocio bancario supone precisamente esto: arriesgarse o atreverse a entregar dinero a un deudor sabiendo que existe el peligro de no pago, pero con el objetivo de administrarlo de tal forma que se pueda obtener una rentabilidad generadora de valor para los dueños del capital o accionistas.

La teoría moderna del portafolio y por ende la actual administración del riesgo crediticio tiene como meta que su gestión permita buscar la rentabilidad que se adecue al nivel de pérdidas esperadas que se esté dispuesto a asumir. Lo que se traduce en que un cliente con una mayor probabilidad de impago no se reduce a una negación sino a una correcta determinación del nivel de riesgo y por ende del precio (tasa de interés activa) para que se pueda obtener la rentabilidad esperada que compense el riesgo de crédito asumido. En este contexto, la medición de este tipo de riesgo se vuelve relevante, no sólo en su comprensión sino en su medición, desarrollando para esto varias metodologías de estimación de pérdidas.

Hace más de 20 años que el Banco JP Morgan publicó en su documento técnico Riskmetrics el concepto de Valor en Riesgo (VaR, por sus siglas en inglés), un modelo estadístico que permite medir cuantitativamente la pérdida máxima que puede experimentar una entidad financiera, con un nivel de confianza dado y en un horizonte temporal determinado (Phelan, 1997). Desde ahí hasta la actualidad la medición del riesgo de crédito no ha tenido descanso en la búsqueda de mejoras permanentes de las técnicas empleadas en su estimación, porque el riesgo no se puede eliminar, pero sí se debe y puede administrar.

Una adecuada gestión permitirá optimizar el binomio rentabilidad-riesgo dado que existe una relación directa entre el nivel de riesgo que la entidad está dispuesta a asumir y el potencial de beneficios que se podrán generar (Elizondo y Altman, 2004).

El resultado final del análisis de riesgo de crédito es poder obtener el nivel de pérdidas de capital que una entidad puede alcanzar producto del incumplimiento (*default*) de sus prestatarios. Este incumplimiento no es otra cosa que el deterioro progresivo observado en los activos de la institución y que se terminará traduciendo en lo que hoy se conoce como pérdida esperada (PE).

La PE de una cartera de créditos nos indica el monto de capital que se podría perder en un horizonte dado, como resultado de la exposición al riesgo de incumplimiento de pago y representa el costo de hacer el negocio crediticio. Estas pérdidas deberían estar cubiertas por las reservas preventivas o provisiones (Arias et al., 2006).

Pero, ¿es tan necesaria la medición y el control de riesgo de crédito en la vida real? La historia reciente nos responde que sí. En 2007, el mercado de créditos hipotecarios en EE.UU. dejó en evidencia la importancia de una adecuada administración y medición del riesgo. Las hipotecas subprime o hipotecas basura eran créditos otorgados a deudores con muy poca o ninguna liquidez ni solvencia para cubrir estas obligaciones y en donde las altas tasas de interés no compensaban el riesgo de crédito que se estaba asumiendo Majid y Kassim. (2006). El análisis *post mortem* de estas operaciones crediticias dejó en evidencia que los préstamos se daban a personas con trabajos precarios lo que incrementaba el riesgo de no pago, pero mientras el mercado hipotecario estaba al alza disfrazaba con su crecimiento el peligro asumido y latente. La única preocupación era lograr empaquetar la deuda en títulos que pudieran ser ofrecidos a inversores con un gran apetito por el riesgo y que buscaban altos retornos.

Se puede apreciar que la utilidad de la gestión y cuantificación del riesgo de crédito es

evidente. Esta cuantificación, considerando los enfoques más avanzados entregados por el Comité de Supervisión Bancaria de Basilea, debe realizarse con base en los conceptos de frecuencia y severidad de las pérdidas. Las pérdidas por este riesgo se dividen en esperadas y en no esperadas o inesperadas, las primeras relacionadas con los requerimientos de provisiones por incobrabilidad, mientras que las segundas se asocian con el requerimiento de capital regulatorio mínimo por riesgo crediticio.

Basilea propone a las instituciones financieras elegir entre dos métodos para la medición del riesgo de crédito:

- Método Estándar. - este enfoque plantea que los bancos realicen sus mediciones de riesgo de crédito empleando calificaciones o ratings provistos por externos, los cuales deben ser empresas calificadoras de riesgo con prestigio internacional. Para la estimación de los activos ponderados por riesgo de crédito se aplicará a los saldos de cartera netos de provisiones específicas, un coeficiente de ponderación en función a esta calificación de riesgo.

- Método Avanzado basado en ratings internos (IRB por sus siglas en inglés). - este enfoque a diferencia del anterior incorpora nuevos términos para la cuantificación del riesgo crediticio. La estimación de pérdidas, basada en modelos internos, deberá calcular los componentes del riesgo de una determinada cartera u operación de crédito, los cuales son : a) la probabilidad de incumplimiento o default, b) la exposición en caso de incumplimiento y c) la pérdida dado el incumplimiento.

Este enfoque avanzado o de ratings internos propuesto por Basilea para la medición del riesgo de crédito se expone en el siguiente apartado.

1.1. Basilea y la estimación de pérdidas para la administración de Riesgo de Crédito

En su documento técnico de julio de 2005 *An Explanatory Note on the Basel II IRB Risk Weight Functions*, el Comité de Supervisión Bancaria de Basilea explica la pérdida tanto desde una perspectiva *top-down*, es decir desde una visión de portafolio, como desde una *bottom-up* o desde sus componentes (Joseph, 2005).

Cuánto capital mantener es la clave de este ejercicio de estimación de pérdidas, pues las entidades financieras en su búsqueda de optimizar el binomio rentabilidad-riesgo tienen un incentivo para minimizar el capital que requieren porque al hacerlo se liberan recursos que pueden destinarse a inversiones rentables y aquí es cuando surge el *trade-off* entre riesgo y retorno, pues cuanto menos capital tienen las instituciones, mayor es la probabilidad de no ser capaz de responder a sus obligaciones porque las pérdidas experimentadas en un año fiscal no pueden ser cubiertas por la utilidad (beneficio) más el capital disponible, lo que puede conducir a una quiebra por insolvencia.

Basilea II propone, para determinar cuánto capital debe tener un banco, en su enfoque IRB (*Internal Rating Based*) centrarse en la frecuencia de las insolvencias bancarias producto de las pérdidas crediticias que los supervisores bancarios están dispuestos a aceptar. Mediante el uso de modelos estocásticos sobre un portafolio de créditos es posible estimar la máxima pérdida que se puede soportar con un nivel de confianza dado o en su defecto cuál sería ese nivel de pérdida que superará el capital de la entidad con una probabilidad pequeña y predefinida.

El número exacto de incumplimientos en un determinado año, el monto exacto adeudado al momento del incumplimiento o la tasa de pérdida real son variables aleatorias y las entidades bancarias no los pueden conocer de antemano, pero pueden estimar su promedio. El enfoque IRB de Basilea se fundamenta sobre estos tres parámetros de riesgo, los cuales se definen a continuación:

- Probabilidad de incumplimiento (PI o PD) por grado de calificación que indica el porcentaje promedio de deudores que incumplirán en este grado de calificación en el transcurso del año. La probabilidad de incumplimiento es función de una ratio de riesgo, un intervalo de tiempo y del momento específico del tiempo en donde se evaluará el evento de default o incumplimiento.

- Exposición en caso de incumplimiento (EAD) que proporciona una estimación del saldo pendiente de pago (montos recibidos más los posibles retiros que se puedan realizar a las líneas de crédito aprobadas) en caso de incumplimiento del prestatario. En otras palabras, su estimación debe considerar no sólo la deuda directa que mantiene el deudor, sino la exposición potencial de las operaciones contingentes que pueden volverse cartera en el futuro.

- Pérdida en caso de incumplimiento (LGD) que es el porcentaje de exposición que el banco podría perder en caso de incumplimiento del prestatario. La pérdida en caso de incumplimiento o severidad busca medir la pérdida que sufriría el acreedor después de haber realizado todas las gestiones posibles para lograr recuperar los créditos impagos.

Con estos tres factores la pérdida esperada en montos monetarios se puede escribir como:

$$\text{Pérdida Esperada} = PD \times EAD \times LGD.$$

1.2. Saldo expuesto al *default* en el contexto de Basilea y su tratamiento para un portafolio de tarjetas de crédito

La exposición o el saldo expuesto al incumplimiento o al default no es otra cosa que el importe de la deuda que está pendiente de pago al momento en que el prestatario cae en incumplimiento.

Por lo general, al momento del incumplimiento, suele ocurrir que la exposición coincide con el saldo impago de la operación crediticia, pero esto no es una regla absoluta. En productos con límites explícitos como las tarjetas de crédito, hay una proporción de este límite que no está en uso hoy pero que puede usarse durante el tiempo que el deudor tarda en caer en emphdefault. Por tanto, el cálculo de la exposición debe incorporar no solo la parte del límite o cupo de la tarjeta de crédito que se está usando sino también el potencial incremento de saldo que pudiera generarse desde una fecha de referencia hasta el momento en que se declare impago.

Como consecuencia de esto, la EAD se obtiene como la suma del riesgo asumido de la operación más un porcentaje del riesgo no dispuesto respecto del límite disponible por el cliente. EAD = Saldo utilizado del límite o cupo disponible + Porcentaje del Saldo no utilizado del límite o cupo disponible al momento del incumplimiento.

Este porcentaje se calcula a partir del *Credit Conversión Factor* (CCF) y en la literatura es común que la estimación de la EAD se reduzca a encontrar este Factor de Conversión. Se puede definir al CCF como el porcentaje sobre el saldo no dispuesto o no utilizado que se espera vaya a emplearse antes o hasta que se produzca el incumplimiento (Thomas at al., 2007).

Para la estimación de la EAD en el caso de las tarjetas de crédito, objeto de este estudio, Basilea II y III sugieren el uso de datos históricos que permitan evaluar el CCF, de tal forma que a partir del comportamiento reciente de los clientes se pueda, aunque parcialmente, conocer cuál será el saldo expuesto al default.

Si bien el acuerdo de Basilea no define como mandatorio el cálculo del CCF para estimar la exposición al incumplimiento, es una referencia recurrente a lo largo de la documentación generada por el Comité, esta aproximación hace que la precisión en la estimación de la EAD recaiga sobre la calidad de los modelos para el CCF.

No obstante, los modelos para estimar el *Credit Conversion Factor* han enfrentado importantes retos estadísticos dado que la distribución de esta variable no se ajusta a las distribuciones

estadísticas estándar. Según Tong et al. (2016) en su artículo de 2016 *Exposure at Default models with and without the credit conversion factor*, las distribuciones del CCF tienden a ser bimodales con una función de probabilidad en cero y otra en uno, entre ambas una distribución relativamente plana (ver Figura 1).

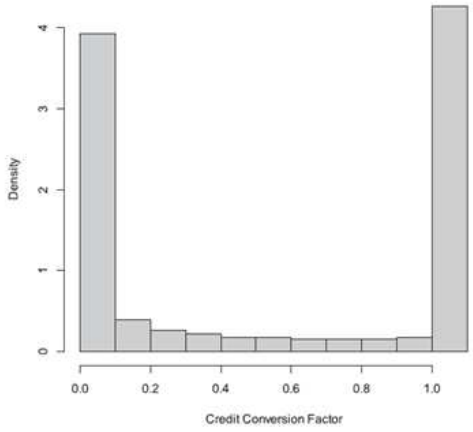

Figura 1: Característica de la distribución del Credit Conversion Factor después de haber realizado un truncamiento de la función (Tong et al., 2013).

Ante esta dificultad y siguiendo el trabajo realizado por Tong et al. (2013), en este trabajo se estima directamente la exposición al *default* y se ignorar la habitualidad de calcular el CCF.

Para lograr esto, la EAD se modeliza como una variable de respuesta continua empleando un modelo de regresión lineal generalizado usando la distribución Gamma (Dobson y Barnett, 2018), esto implica que, en primer lugar y dado que la cantidad de saldo expuesto mayor a cero tiene una distribución sesgada a la derecha, se utilizará una distribución gamma. Además, se usa un modelo Regresión Spline Adaptativa Multivariante como método alternativo el enfoque tradicional paramétrico de estimación. Posterior a esto, la probabilidad de (no)ocurrencia de una EAD con valor cero se estudiará con un modelo logístico.

Finalmente, el desempeño de estos modelo es contrastado contra el saldo expuesto observado al momento del default del grupo de prueba y también contra el saldo expuesto observado de un grupo distinto de clientes en un periodo diferente al empleado para el grupo de construcción y prueba del modelo. Este trabajo utiliza un conjunto de datos reales de una entidad financiera ecuatoriana que tiene un portafolio de tarjetas de crédito.

2. METODOLOGÍA

Tanto desde la perspectiva de la gestión del riesgo como desde el punto de vista regulatorio, la estimación del riesgo de crédito es de gran relevancia. Este se basa en el análisis de las pérdidas esperadas (EL):

$$EL = PD(\,\%) \times LGD(\,\%) \times EAD(\$)$$

donde la probabilidad de incumplimiento (PD), la severidad (LGD) y la exposición del activo (EAD) juegan un papel fundamental.

En particular, Basilea II y III recomiendan que para la estimación de la exposición se use la proporción de la cantidad actual no utilizada que probablemente se utilizará en el momento del incumplimiento, o factor de conversión del crédito (Vytautas, 2008). No obstante, esta recomendación no es vinculante y trabajos como Tong et al. (2016) y Taplin at al. (2007) han mostrado que la estimación directa de la exposición puede ser realizada sin necesidad de la utilización del factor de conversión.

Un objetivo primordial de este trabajo modelizar la EAD directamente de la distribución observada de la EAD. En esta línea, la estimación directa de la EAD ha sido abordada usando el modelo con variable dependiente Gamma por su capacidad de trabajar con datos de naturaleza positivamente sesgada (Tong et al., 2013, 2016).

Además de la estimación del modelo lineal generalizado para modelizar la EAD, este trabajo evalúa empíricamente métodos estadísticos alternativos y realiza un *benchmark* de los resultados obtenidos. El primer método consiste en elegir la distribución de probabilidad univariada que genere el mejor ajuste de la distribución observada de la EAD y luego usar esa distribución en el contexto de regresión. La elección de la distribución paramétrica se basa en estadísticos de bondad de ajuste como Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling (Stephens, 1974).

El segundo enfoque se basa en técnicas de aprendizaje automático o *machine learning* (ML), específicamente, el modelo MARS. Trabajos como Tanoue et al. (2020) han usado exitósamente ténicas de ML en el contexto de riesgo de crédito.

2.1. Datos

El conjunto de datos consiste en más de 200 mil observaciones de tarjeta habientes de un banco ecuatoriano. Se ha trabajado sobre dos periodos de observación, el primero es desde diciembre de 2017 hasta noviembre de 2018 con un total de 212.796 registros. Este grupo de cuentas sirve para la construcción del modelo y para obtener el grupo de prueba contra el cual se contrasta el resultado del modelo con la EAD observada.

El segundo periodo de observación es desde diciembre de 2018 hasta noviembre de 2019 con 221.213 tarjeta habientes y es empleado únicamente para contrastar la precisión del modelo obtenido para el cálculo de la EAD con un grupo de clientes en un periodo distinto al de la población de construcción y prueba.

Junto con la muestra de clientes se obtienen o calculan variables explicativas que son empleadas en los modelos que se usan para la estimación de la EAD y su posterior validación. Este conjunto de variables se puede obtener tanto al momento del default (t_d) como a la fecha de referencia o de corte de la información (t_r), en este caso esta fecha es diciembre de 2017.

Este trabajo presenta un esquema de modelización basado datos de entrenamiento, prueba y validación. Este es un esquema que se usa ampliamente en la modelización de aprendiza automático y ciencia de datos (Yadav y Shukla, 2016). La siguiente lista muestra las variables que se consideran en los modelos para la estimación de la EAD:

- Límite o Cupo de la TC o Tarjeta de crédito - $L(t_r)$: Cupo asignado a la tarjeta a la fecha de corte.
- Cupo utilizado - $E(t_r)$: Cupo utilizado a la fecha de corte.
- Cupo no utilizado - $L(t_r) - E(t_r)$: Cupo de la tarjeta menos el cupo utilizado a la fecha de corte.
- Porcentaje de uso de la TC - $E(t_r)/L(t_r)$ Cupo utilizado a la fecha de corte dividido para el cupo asignado a la tarjeta a la fecha de corte.
- Tiempo al default - $t_d - t_r$: Fecha de caída al incumplimiento menos la fecha de corte.
- Segmento de riesgo - $R(t_r)$: Clasificación de riesgo medida por el score de comportamiento, a la fecha de corte.
- Promedio de días de atraso: Número promedio de días de no pago medidos a los 3, 6, 9 y 12 meses anteriores a la fecha de corte.

- Porcentaje de no uso de la TC - $(L(t_r) - E(t_r))/L(t_r)$: Porcentaje de no uso a la fecha de corte dividido para el cupo o límite de la tarjeta a la fecha de corte.
- Cupo utilizado al default - $E(t_d)$: Cupo utilizado al momento del default o incumplimiento.
- Límite o Cupo de la Tarjeta de crédito (TC) al default - $L(t_d)$ Cupo o límite de la tarjeta de crédito al momento del incumplimiento.
- Incremento de cupo: Variable binaria que indica 1 si tuvo un incremento de cupo durante los 12 meses anteriores a la fecha de corte o 0 en caso contrario.
- Cambio absoluto en el cupo utilizado: Cambio en valor en el cupo utilizado calculado como la diferencia entre el Cupo utilizado a la fecha de corte menos el cupo utilizado 3, 6, 9 y 12 meses antes de la fecha de corte.
- Cambio relativo en el cupo utilizado: Cambio relativo en el cupo utilizado calculado como la diferencia entre el Cupo utilizado a la fecha de corte menos el cupo utilizado 3, 6, 9 y 12 meses antes de la fecha de corte dividido para el cupo utilizado a la fecha de corte.
- Segmento RFM: Clasificación de RFM (recencia, frecuencia y monto) a la fecha de corte.

2.2. Modelo Logit

Uno de los principales motivos de la amplia utilización del modelo logit es el hecho de que puede estimar modelos cuando la variable dependiente es binaria, es decir, toma valores cero o uno. Esta característica hace que el modelo de regresión lineal múltiple clásico no sea viable en este tipo de variables dependientes debido a que sus predicciones pueden tener valores negativos, así como valores superiores a uno.

Formalmente, la regresión logística o logit puede formularse como:

$$y = \begin{cases} 1 & \beta_1 + \beta_2 X_1 + \cdots + \beta_k X_k + u > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Suponga un modelo con una variable explicativa x, entonces

$$t = \beta_0 + \beta_1 x$$

Ahora la función logística ($\sigma(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}}$) se puede expresar como

$$p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}$$

Donde $p(x)$ representa la probabilidad de que $y = 1$.

El modelo logit pertenece a la familia de los modelos lineales generalizados. Para observar este particular se define la inversa de la función logística:

$$g(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x,$$

Lo que indica que esta función define un modelo lineal a partir de la función logística de enlace.

2.3. Modelo Gamma

Sea y_i la EAD observada para el i-ésimo cliente, $i = 1, \ldots, n$ (por simplicidad se omite el índice i en adelante); x denota la matriz de covariables observados para cada cliente. La función de densidad para y es mixta:

$$f(y) = \begin{cases} \pi & \text{si y=0} \\ (1-\pi)g(y) & \text{si y>0} \end{cases}$$

donde $g(y)$ es la función de densidad de una distribución continua y π es la probabilidad de la EAD en cero, en este caso se usa un modelo lineal generalizado para variable dependiente con distribución Gamma.

2.4. Mars

MARS, o Regresión Spline Adaptativa Multivariante (*Multivariate adaptive regression spline*) es una forma de regresión introducida por Jerome Friedman Friedman. (1991). MARS es una técnica de regresión no paramétrica y puede ser vista como una extensión de los modelos lineales que automáticamente identifica no linealidades e interacciones entre variables. El término *MARS* está protegido por derechos de autor y pertenece a *Salford Systems*. Para evitar violentar esos derechos, las implementaciones abiertas de MARS se suelen llamar *Earth* (El paquete earth en R (Milborrow, 2021)).

MARS es ideal para usuarios que prefieren obtener resultados similares a la regresión tradicional mientras capturan no linealidades e interacciones necesarias. Revela patrones importantes en los datos que otras técnicas suelen fallar en revelar. También construye su modelo uniendo pedazos de líneas rectas que mantienen su propia pendiente. Esto permite que se detecte cualquier patrón en los datos. Se puede utilizar para cuando se tiene variables de respuesta cuantitativa y cualitativa. MARS realiza, automátcamente y con gran velocidad: selección de variables, transformación de variables, detección de interacciones, testeo. Para detalles formales del modelo ver Yaseen et al. (2018).

3. RESULTADOS

La Figura 3 muestra el ajuste de diferentes distribuciones de probabilidad para la exposición. Se puede apreciar que se tiene una variable con una fuerte asimetría positiva. Esta característica de la distribución descarta opciones de modelización más tradicionales como cuando se asume normalidad en la variable de respuesta, ampliamente estudiadas en textos de econometría clásica (Gujarati, 2012).

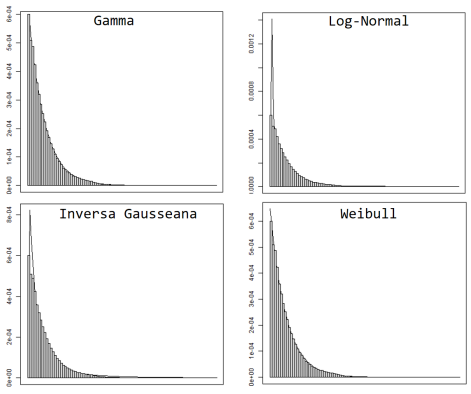

Figura 2: Distribuciones continuas candidatas para valores mayores a cero en la exposición.

Una de las tareas centrales en el análisis de datos paramétrico responde a la elección del modelo que mejor se ajusta a los datos analizados (Forbes et al., 2011). Usando el paquete Ricci (2005), las distribuciones de probabilidad candidatas para el ajuste de la exposición fueron: Gamma, Inversa Gausseana, Log-normal y Weibull y se muestran en la Figura 3. La Figura 3 muestra el histograma de la variable analizada (EAD), así como las densidades estimadas para cada una de las candidatas. Se usa el criterio de Kolmogorov-Smirnov para evaluar la bondad de ajuste de donde se obtuvieron, respectivamente: 0,00533, 0,37897, 0,07263 y 0,00716. Este criterio indica que el mejor ajuste es el de la distribución Gamma.

Una vez identificada la distribución de probabilidad de la variable dependiente, se procede a la estrategia de modelización basado en dos etapas. La primera etapa consiste en estimar un modelo para valores de exposición igual a cero. Para este objetivo se usa un modelo logit. La segunda etapa consiste en la modelización de la exposición para valores mayores a cero usando la distribución gamma como supuesto distribucional de la variable dependiente y el modelo MARS. Los resultados del modelo logit se muestran en la Tabla 1.

Asumiendo un nivel de significancia de 5 %, casi todas las variables son significativas y se puede apreciar que algunos segmentos no son significativos pero la variable en conjunto si lo es. Este modelo nos permite identificar a los clientes cuya predicción de exposición sea igual a cero. Es decir, ante un nuevo conjunto de información, el primer modelo en estimarse es el logit. Los clientes que tengan una probabilidad mayor al umbral estimado se consideran con predicción igual a cero.

El umbral elegido para determinar la probabilidad de que la exposición sea igual a cero se realiza mediante la estimación de la curva ROC (*Receiver Operating Characteristic*) que indica el balance que existe entre la especificidad y la sensibilidad para cada punto de corte. Se elige el punto más cercano al par ordenado (0,1) porque estos valores indican una mejor clasificación. Cabe indicar que el área bajo la curva de este modelo es 94 %, que es mejor mientras más cerca de 100 % se encuentre. En los datos de prueba del modelo logit se obtiene una sensibilidad del 88 % y una especificidad de 93 %.

Tabla 1: Resultados del modelo logístico para la estimación de EAD = 0

	Estimate	Std.	Error z	Pr(>—z—)
(Intercept)	-5.334	0.216	-24.751	0.000
Rfm	-0.436	0.034	-12.981	0.000
Timetodefault	-0.015	0.003	-4.647	0.000
SegmentoTDC_Operacional1.	0.633	0.353	1.794	0.073
SegmentoTDC_Operacional2.	0.332	0.236	1.405	0.160
SegmentoTDC_Operacional3.	-0.211	0.186	-1.131	0.258
SegmentoTDC_Operacional4.	-0.019	1.047	-0.018	0.986
SegmentoTDC_Operacional5.	0.395	0.221	1.791	0.073
SegmentoTDC_Operacional6.	0.458	0.186	2.464	0.014
Absolute_change_drawn3	0.000	0.000	2.378	0.017
Relative_change_drawn3	-6.785	0.322	-21.102	0.000
Relative_change_drawn6	1.196	0.263	4.557	0.000

En la segunda fase de modelización se estima un modelo de regresión basado en modelos lineales generalizados cuyos resultados se muestran la Tabla 5 del Anexo A. Se puede apreciar que todas las variables de este modelo son significativas. Asmismo, la Tabla 4 del Anexo B muestra los coeficientes ajustados del modelo MARS.

La estimación comparada con lo observado en los datos de prueba del modelo gamma y MARS se presenta en la Figura 3. Se puede apreciar que la predicción tiene una distribución

similar a la observada en ambos casos.

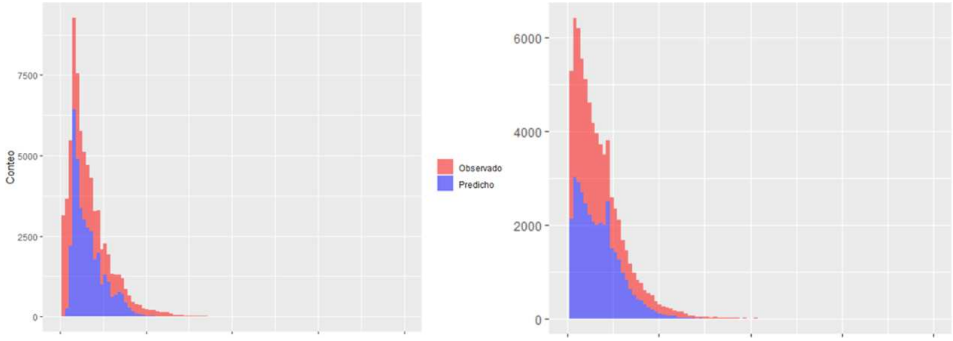

Figura 3: Predicción de EAD vs Observada en datos de prueba. Izquierda: Gamma. Derecha: MARS

Tabla 2: Resultados de medidas del rendimiento de los modelos estimados (Periodo 1)

	Gamma	Mars	Total (Gamma)	Total (Mars)
Pearson	0.555	0.806	0.578	0.808
Spearman	0.630	0.807	0.652	0.813
MAE	873.561	620.928	849.477	612.086
RMSE	1278.665	905.019	1255.909	900.391

Tabla 3: Resultados de medidas del rendimiento de los modelos estimados (Periodo 2)

	Gamma	Mars	Total (Gamma)	Total (Mars)
Pearson	0.533	0.810	0.563	0.813
Spearman	0.595	0.818	0.637	0.820
MAE	949.442	641.077	908.430	625.217
RMSE	1380.786	952.507	1338.925	935.376

Las Tablas 2 y 3 muestra cuatro medidas de rendimiento usadas para evaluar modelos cuya variable dependiente es cuantitativa (Burges et al., 2006). La correlación de Pearson y la correlación de Spearman indican mejores rendimientos cuanto más alta es. El MAE y raíz de MSE son mejores cuanto más bajo es el valor. Como es de esperarse, se puede apreciar que las medidas de rendimiento del modelo presentan rendimientos más bajos en la medida en que se trabaja con valores no observados. Por ejemplo, la correlación de Spearman tiene un valor inicial de 0,630 en la fase de prueba del periodo 1 (Tabla 2). Pero en el periodo 2 (Tabla 3) es de 0,595, muestra una ligera disminución.

Es notorio que el uso del Modelo MARS supera considerablemente al modelo Gamma. Debe también destacarse que el modelo combinado (columnas *Total* de las tablas 2 y 3) muestran mejores valores de rendimiento que únicamente los modelos parciales que no toman en cuenta EAD = 0. Esto indica que el uso del modelo logit para predecir la exposición igual a cero aporta al rendimiento de las predicciones.

4. CONCLUSIONES

En los modelos de estimación de pérdidas por riesgo de crédito minorista (*retail credit risk*) la estimación de la probabilidad de *default* ha sido permanentemente el principal foco de investigación y la literatura es extensa no sólo en los documentos científicos sino también en los libros académicos que abordan este componente de la pérdida esperada. En cambio, la Exposición al *Default* (EAD) o Saldo Expuesto al Incumplimiento tanto en la industria financiera como en la literatura académica ha sido una de las áreas más débiles de medición y modelización por lo que este trabajo representa un aporte en el esfuerzo de medición del EAD.

La correcta medición de la EAD en un portafolio de tarjetas de crédito permite a la institución financiera reducir el riesgo de subestimación de pérdidas esperadas e inesperadas y por tanto tener un mejor control en la optimización de la evaluación del desempeño de la cartera, el rendimiento sobre el capital ajustado por riesgo, las decisiones relativas a la operación, el análisis de rentabilidad, así como a la toma de decisiones respecto a la estructura de capital.

Se ha mostrado a lo largo del trabajo que el primer desafío es la identificación de la distribución de probabilidad de la variable dependiente. Para lograr este objetivo se plantearon diferentes distribuciones candidatas y se evaluó la bondad de ajuste entre la variable observada y la distribución teórica propuesta.

Se ha demostrado que es posible la estimación de la EAD omitiendo el cálculo del Factor de Conversión de Crédito. Para este objetivo se ha empleado una regresión con modelos lineales generalizados dada una distribución Gamma en la variable dependiente y el Modelo MARS, resultando este último muy superior al primero.

Los resultados empíricos obtenidos con datos de entrenamiento han sido validados con dos muestras distintas, unos datos de prueba sobre el mismo periodo de la información de entrenamiento y unos datos de validación que corresponden a un periodo de tiempo diferente a los de entrenamiento y prueba, encontrando resultados estables en ambos casos, lo que significa que se puede implementar un modelo para la estimación de la EAD combinando los resultados obtenidos para cuando EAD = 0 y EAD¿0.

Este trabajo puede ser un insumo importante para la estimación más fina de las pérdidas en la cartera de clientes de tarjeta de crédito en Ecuador y puede ser utilizado por la institución de tal manera que permita cumplir con la normativa vigente, así como tener resultados más precisos desde el punto de vista estadístico.

REFERENCIAS

Arias-Serna, M. A., Guzmán-Aguilar, D. S., Valdez-Betancur, D. (2021). Sistema de información para la cuantificación de pérdidas esperadas: Una aplicación en las entidades del sector solidario colombiano. *Revista Ibérica de Sistemas y Tecnologías de Información, (E39)*, 444-460.

Burges, C., Ragno, R., & Le, Q. (2006). Learning to rank with nonsmooth cost functions. *Advances in neural information processing systems*, 193-200.

Dobson, A. J., & Barnett, A. G. (2018). *An introduction to generalized linear models*. Nueva York: CRC press.

Elizondo, A., & Altman, E. I. (2004). *Medición integral del riesgo de crédito*. México: Limusa.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions*. Nueva York: John Wiley Sons.

Friedman, J. H. (1991). Estimating functions of mixed ordinal and categorical variables using adaptive splines. *Stanford Univ CA Lab for Computational Statistics*.

García Sánchez, M., Sánchez Barradas, C. (2005). Riesgo de crédito en México: aplicación del modelo CreditMetrics. México: *Departamento de Contaduría y Finanzas. Escuela de Negocios, Universidad de las Américas Puebla.* Recuperado el, 17.

Gujarati, D. (2012). *Econometrics by example*. Nueva York: Macmillan.

Joseph, M. P. (2005). A PD validation framework for Basel II internal ratings-based systems. *Credit Scoring and Credit Control IV*.

Shabri Abd Majid, M. and Hj Kassim, S. (2009), Impact of the 2007 US financial crisis on the emerging equity markets, *International Journal of Emerging Markets*, Vol. 4 No. 4, pp. 341-357. https://doi.org/10.1108/17468800910991241

Stephen Milborrow. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. (2021). earth: Multivariate Adaptive Regression Splines. *R package version 5.3.1.* https://CRAN.R-project.org/package=earth

Phelan, M. J. (1997). Probability and statistics applied to the practice of financial risk management: The case of JP Morgan's RiskMetrics. *Journal of Financial Services Research, 12*(2), 175-200.

Ricci, V. (2005). Fitting distributions with R. *Contributed Documentation available on CRAN, 96*.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association, 69*(347), 730-737.

Tanoue, Y., Yamashita, S., & Nagahata, H. (2020). Comparison study of two-step LGD estimation model with probability machines. *Risk Management, 22*(3), 155-177.

Taplin, R., To, H. M., & Hee, J. (2007). Modeling exposure at default, credit conversion factors and the Basel II accord. *Journal of Credit Risk, 3*(2), 75-84.

Thomas, L., Crook, J., Edelman, D. (2017). *Credit scoring and its applications*. Society for industrial and Applied Mathematics.

Tong, E. N., Mues, C., & Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting, 29*(4), 548-562.

Tong, E. N., Mues, C., Brown, I., & Thomas, L. C. (2016). Exposure at default models with and without the credit conversion factor. European Journal of Operational Research, 252(3), 910-920.

Vytautas, V. (2008). Estimating EAD for retail exposures for Basel II purposes. *Journal of Credit Risk, 4*(1), 79-110.

Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *2016 IEEE 6th International conference on advanced computing (IACC)*, 78-83.

Yaseen, Z. M., Deo, R. C., Hilal, A., Abd, A. M., Bueno, L. C., Salcedo-Sanz, S., Nehdi, M. L. (2018). Predicting compressive strength of lightweight foamed concrete using extreme learning machine model. *Advances in Engineering Software, 115*, 112-125.

Anexo A. GLM: Gamma

Resultados del modelo lineal generalizado con distribución gamma para la estimación del EAD mayor a cero, ver Tabla 4.

Tabla 4: Resultados del modelo lineal generalizado con distribución gamma para la estimación del EAD mayor a cero

	Estimate	Std.	Error z	Pr(>—z—)
(Intercept)	0.002	0.000	107.661	0.000
Rfm	0.000	0.000	-85.855	0.000
Undrawnamount	0.000	0.000	5.115	0.000
Average_days_delinquent_3M	0.000	0.000	-15.294	0.000
Timetodefault	0.000	0.000	-12.283	0.000
SegmentoTDC_Operacional1.	-0.001	0.000	-48.630	0.000
SegmentoTDC_Operacional2.	-0.001	0.000	-47.393	0.000
SegmentoTDC_Operacional3.	-0.001	0.000	-36.441	0.000
SegmentoTDC_Operacional4.	-0.001	0.000	-38.050	0.000
SegmentoTDC_Operacional5.	-0.001	0.000	-38.693	0.000
SegmentoTDC_Operacional6.	-0.001	0.000	-27.299	0.000
Absolute_change_drawn3	0.000	0.000	-23.964	0.000
Relative_change_drawn3	0.000	0.000	4.971	0.000
Relative_change_drawn6	0.000	0.000	-1.888	0.059

Anexo B. Mars

Resultados del modelo lineal generalizado con distribución gamma para la estimación del EAD mayor a cero, ver Tabla 5.

Tabla 5: Resultados del modelo Mars para la estimación del EAD mayor a cero

Variables	Coeficientes
(Intercept)	-1443.5478
h(Rfm-7)	212.3675
h(Absolute_change_drawn3- -31.41)	130.7348
h(103.34-Absolute_change_drawn3)	-110.1749
h(Absolute_change_drawn3-103.34)	-14.7207
h(0.313667-Relative_change_drawn3)	20932.3912
h(Relative_change_drawn3-0.313667)	-13814.0467
h(Absolute_change_drawn3-31.41)*	
h(Relative_change_drawn3-1.47706)	102.5315
h(Absolute_change_drawn3-31.41)*	
h(1.47706-Relative_change_drawn3)	-101.2973
h(-10.26-Absolute_change_drawn3)*	
h(0.313667-Relative_change_drawn3)	168.1839
h(103.34-Absolute_change_drawn3)*	
h(Relative_change_drawn3-0.350793)	186.9744
h(103.34-Absolute_change_drawn3)*	
h(-0.350793-Relative_change_drawn3)	-170.7754
h(103.34-Absolute_change_drawn3)*	
h(Relative_change_drawn3-0)	-994.0103
h(103.34-Absolute_change_drawn3)*	
h(Relative_change_drawn3-0.10797)	99.5814
h(103.34-Absolute_change_drawn3)*	
h(Relative_change_drawn3-0.0303594)	512.9692
h(103.34-Absolute_change_drawn3)*	
h(Relative_change_drawn3-0.0417451)	298.6189
h(Absolute_change_drawn3-103.34)*	
h(Relative_change_drawn3-0.356966)	-102.5351
h(Absolute_change_drawn3-103.34)*	
h(0.356966-Relative_change_drawn3)	116.9814