Modeling and Forecasting the Probability of Crypto-Exchange Closures: A Forecast Combination Approach

Magomedov, Said and Fantazzini, Dean

2025

Modeling and Forecasting the Probability of Crypto-Exchange Closures: A Forecast Combination Approach

Said Magomedov[*] Dean Fantazzini[†]

Abstract

The popularity of cryptocurrency exchanges has surged in recent years, accompanied by the proliferation of new digital platforms and tokens. However, the issue of credit risk and the reliability of crypto exchanges remain critical, highlighting the need for indicators to assess the safety of investing through these platforms. This study examines a unique, hand-collected dataset of 228 cryptocurrency exchanges operating between April 2011 and May 2024. Using various machine learning algorithms, we identify the key factors contributing to exchange shutdowns, with trading volume, exchange lifespan, and cybersecurity scores emerging as the most significant predictors. Since individual machine learning models often capture distinct data characteristics and exhibit varying error patterns, we employ a forecast combination approach by aggregating multiple predictive distributions. Specifically, we evaluate several specifications of the generalized linear pool (GLP), beta-transformed linear pool (BLP), and beta-mixture combination (BMC). Our findings reveal that the beta-transformed linear pool and the beta-mixture combination achieve the best performances, improving forecast accuracy by approximately 4.1% based on a robust H-measure, which effectively addresses the challenges of misclassification in imbalanced datasets.

Keywords: forecast combination; exchange; bitcoin; crypto assets; cryptocurrencies; credit risk; bankruptcy; default probability.

JEL classification: C35; C51; C53; C58; G12; G17; G32; G33.

Journal of Risk and Financial Management, forthcoming

[*] Higher School of Economics, International College of Economics and Finance, Moscow, (Russia).

[†] Moscow School of Economics, Moscow State University, Leninskie Gory, 1, Building 61, 119992, Moscow, Russia. Fax: +7 4955105256 . Phone: +7 4955105267 . E-mail: dean.fantazzini@gmail.com .

This is the working paper version of the paper *Modeling and Forecasting the Probability of Crypto-Exchange Closures: A Forecast Combination Approach*, forthcoming in the Journal of Risk and Financial Management.

1 Introduction

Over the past decade, the global financial system has undergone significant changes, one of the most prominent being the emergence of Bitcoin in 2009. Bitcoin, like other cryptocurrencies, "*is a digital or virtual currency secured by cryptography, which makes it nearly impossible to counterfeit or double-spend*".[1] In contrast to fiat currencies, Bitcoin is not a "real" currency in the physical sense and cannot be embodied in coins or banknotes.

The motivation behind this digital asset, as proposed by [Nakamoto, 2008], was to optimize internet commerce, which traditionally relied on intermediaries such as financial institutions. These intermediaries charge fees for their services — transaction costs — and impose restrictions on operations.

The foundation of cryptocurrency lies in a unique protocol and algorithm that establish its functionality, operational rules, and distinctive features. Key aspects include decentralization, consensus mechanisms, fraud protection, and anonymity. Central to these innovations is the "blockchain", a continuous chain of blocks that record all transactions and associated data for a given cryptocurrency. Blockchain technology ensures transparency, reliability, and data immutability (transactions cannot be reversed), making it a cornerstone of the cryptocurrency ecosystem.

Cryptocurrency exchanges, online platforms facilitating the buying, selling, and trading of digital assets, also play a critical role in this ecosystem. These exchanges can be categorized into centralized platforms, where operations are managed by the exchange itself, and decentralized platforms, which enable peer-to-peer trading without intermediaries. Exchanges differ in security levels, liquidity, regional accessibility, supported cryptocurrencies, and transaction fees.

As of May 22, 2024, there are approximately 14,000 cryptocurrencies and over 1,000 crypto-exchanges worldwide.[2] However, not all exchanges have a long lifespan. The rapid growth of the cryptocurrency market, with a total market capitalization of $2.6 trillion[2] has attracted substantial capital but also exposed vulnerabilities. Many exchanges fail to adapt to this evolving landscape, often struggling to create secure and trustworthy trading environments. Consequently, crypto-exchanges remain prone to risks such as bankruptcy and cyber-attacks, which opportunistic fraudsters exploit to target inexperienced users. Some notable instances of crypto-exchange hacks include:[3]

1. Ronin Network: $625 million (March 2022),

2. Poly Network: $611 million (August 2021),

3. FTX: $600 million (November 2022).

[1] Definition by the world-leading source of financial content Investopedia

[2] Data from Coingecko

[3] Investopedia research, December 2, 2023.

For additional information on cyber-attacks targeting crypto-exchanges, refer to the SlowMist Hacked web blog.

[Lee and Milunovich, 2023] emphasize that such events not only result in irreversible financial losses and breaches of personal data but also cause price shocks and heightened market volatility. In some cases, these incidents lead to the collapse of exchanges. For example, [Moore et al., 2018] found that nearly half of the 80 Bitcoin exchanges operating before 2015 had ceased operations.

Selecting a trustworthy and secure exchange is, therefore, a crucial concern for cryptocurrency market participants. According to [Fantazzini and Calabrese, 2021] and [Milunovich and Lee, 2022], factors strongly associated with the survival of crypto-exchanges include:

- 24-hour trading volume,
- Exchange lifetime,
- Cybersecurity measures,
- Number of supported cryptocurrencies,
- Presence of a public developer team.

Additionally, the variety of trading pairs offered on an exchange may also play a role. A broader selection of trading pairs provides users with opportunities to explore diverse trading strategies and reflects increased support from projects offering their tokens for trading.

This study aims to forecast the risk of crypto-exchange closures (referred to as defaults) using the factors outlined above, employing a combination approach. By leveraging credit scoring models and state-of-the-art machine learning techniques, this research introduces an expanded set of explanatory variables and an innovative approach to combining predictive distributions. Advanced methods such as the generalized linear pool (GLP), beta-transformed linear pool (BLP), and beta mixture combination (BMC) are employed, with model performance evaluated using the robust H-measure metric. This metric is particularly suitable for imbalanced datasets, where the number of operational exchanges significantly outweighs closed ones.

The hypotheses driving this research are:

1. A forecast combination approach yields superior statistical metrics and enhances forecast accuracy compared to individual algorithms.
2. The probability of crypto-exchanges remaining operational is significantly influenced by their lifetime, daily trading volume, and cybersecurity scores.

2 Literature Review

The existing literature relevant to this study can be divided into two primary areas: research on the cryptocurrency market and its associated risks, and studies focused on technical and empirical methodologies, including forecasting and model combination approaches.

The first body of literature centers on the cryptocurrency market itself. As mentioned earlier, [Nakamoto, 2008] identified the challenges of traditional financial systems, including centralization, reliance on third-party intermediaries, and a lack of confidentiality. To address these issues, Nakamoto introduced the conceptual framework for Bitcoin, including blockchain technology and the mechanisms of mining and consensus. This foundational work remains central to understanding and advancing cryptocurrency and blockchain technology.

[Moore et al., 2018] studied 80 exchanges operating before 2015 and found that 38 closures were primarily due to security breaches rather than fraudulent activities. They also noted that exchanges with higher trading volumes were less likely to shut down. Similarly, [Schueffel and Groeneweg, 2019] provided a framework for evaluating cryptocurrency exchanges in the absence of centralization or regulatory oversight. Their credit scoring model assessed 34 factors, grouped into four categories: user experience, fees and costs, trustworthiness, and support.

The conceptual groundwork for this research is built on three key studies: [Fantazzini and Calabrese, 2021], [Milunovich and Lee, 2022], and [Lee and Milunovich, 2023]. These studies applied various machine learning techniques, such as logistic regression, decision trees, random forests, and support vector machines, to analyze the factors influencing cryptocurrency exchange closures. Across all three studies, random forest models demonstrated the highest predictive accuracy. Key determinants identified include trading volume, exchange age, cybersecurity measures, cryptocurrency variety, and the presence of a public development team.

The second body of literature is related to the technical and empirical methodologies employed in this study. [James et al., 2013] provided a comprehensive guide to statistical learning techniques in R, covering essential methods such as regression analysis, classification, resampling, shrinkage approaches, tree-based methods, support vector machines, clustering, and advanced neural networks. This resource underpins the technical implementation of our research.

[Lahiri and Yang, 2013] offered a systematic review of forecasting binary outcomes, distinguishing between probability and point forecasts generated by regression models. They highlighted the potential improvements achievable through combination and bootstrap methods, which are particularly relevant to our research. The aggregation of predictive distributions, a cornerstone of this study, finds its theoretical basis in the work of [Gneiting and Ranjan, 2013]. Their study introduced linear and non-linear combination approaches, such as generalized, spread-adjusted, and beta-transformed linear pools, emphasizing their adaptability in capturing dispersion patterns in underlying distributions. Their application to forecasting S&P 500 returns demonstrated the efficacy of these methods compared to traditional approaches.

[Lahiri et al., 2015] further developed the beta-transformed linear pool in forecasting probabilistic outcomes. Their approach involved three steps: selecting forecasts using the Kuiper Skill Score (KSS), testing forecast significance, and combining forecasts via beta-transformed linear pools. This methodology significantly improved forecast accuracy across various horizons compared to individual and average forecasts.

Comparative analyses of opinion pools, including linear, harmonic, logarithmic, and beta mixture combinations, were conducted by [Casarin et al., 2016] and [Wattanachit et al., 2023]. Their studies evaluated these methods using real data, such as the S&P 500 log returns and the US seasonal influenza data, respectively, showcasing the performance of these combinations in practical applications.

Machine learning applications in default risk analysis have also been explored extensively. [Fonseca and Lopes, 2017] and [Bracke et al., 2019] provided insights into the use of machine learning models for assessing default risk, while [Nabipour et al., 2020] compared machine learning and deep learning methods for forecasting stock market trends, treating them as binary outcomes.

The issue of evaluating classifier performance under imbalanced data conditions, as relevant to this study, was addressed by [Hand, 2009]. Hand critiqued the AUC metric, arguing that it depends on classifier-specific weight distributions. As an alternative, he proposed the H-measure, which uses a beta distribution as a weighting function for misclassification costs. [Hand and Anagnostopoulos, 2014] further refined this measure, providing optimal beta distribution parameters tailored to class imbalances.

[Lee and Yu, 2021] provide a comprehensive analysis of traditional statistical methods applied to credit risk assessment, including discriminant analysis, factor analysis, logistic regression, and the KMV-Merton model. Their work offers valuable insights into the effectiveness of these techniques in evaluating creditworthiness and predicting default probabilities. While [Lee and Yu, 2021] employed classical statistical methods to address credit risk, our study builds upon and extends this foundation by applying machine learning ensemble techniques to the unique challenges of cryptocurrency markets. Classical methods often rely on strict parametric assumptions and linear relationships, which may limit their ability to capture the complex, non-linear dynamics inherent in cryptocurrency exchanges. In contrast, the machine learning methods we employ, including Random Forest, Categorical Boosting, and advanced forecast combination techniques like the Beta Linear Pool (BLP) and Beta Mixture Combination (BMC), excel in environments characterized by high volatility and heterogeneity. These methods effectively aggregate diverse models, balancing bias and variance to deliver robust predictions. By addressing the specific challenges of predicting exchange closures in the volatile cryptocurrency domain, our work provides a complementary yet distinct contribution to the literature, demonstrating the potential of flexible, data-driven approaches in financial risk modeling.

3 Materials and Methods

3.1 Machine Learning Techniques

The first step of our empirical analysis involves applying individual machine learning algorithms to generate out-of-sample forecasts and identify the most significant features. Here, a feature refers to an individual measurable property or variable used as an input for the predictive models (e.g., trading volume, lifetime of the exchange). An out-of-sample forecast is a prediction generated by the model using data not included in the training set, ensuring an unbiased evaluation of the model's performance. Below, we outline the classifiers used in the study ([Lahiri and Yang, 2013], [James et al., 2013]).

Overview of Methodological Approach. Machine learning methods employed in this study are well-suited for the prediction of default probabilities and the identification of risk factors. These methods span from simple probabilistic models to advanced ensemble techniques, allowing for both interpretability and high accuracy. The selected classifiers include Na"ive Bayes, Logistic Regression, Support Vector Machines, Categorical Boosting, and Random Forest. Each method is designed to handle the specific challenges posed by our dataset, such as class imbalance and categorical features. These algorithms are extensively used in default probability estimation ([Fonseca and Lopes, 2017], [Bracke et al., 2019]) and are particularly well-suited for handling categorical features.

Probabilistic and Linear Classifiers (Credit Scoring models). *Naive Bayes*, one of the simplest and fastest classification algorithms, is based on Bayes' theorem:

$$P(y \mid X) = \frac{P(y) \cdot P(X \mid y)}{P(X)},$$

where $X = (x_1, x_2, \ldots, x_n)$ represents n conditionally independent features. To estimate the posterior probability $P(y \mid X)$, the algorithm finds the argument that maximizes the numerator (as $P(X)$ is constant across values of y):

$$\hat{y} = \underset{y}{\operatorname{argmax}}\ P(y) \prod_{i=1}^{n} P(x_i \mid y).$$

where $\hat{y}$ represents the predicted class for a given instance, that is the class y that maximizes the posterior probability.

Logistic Regression serves as a benchmark for modeling the probability of default due to its intuitive simplicity and relatively high accuracy. It applies a sigmoid transformation to the linear combination of features, yielding the estimated probability of the positive class:

$$\hat{p} = \frac{1}{1 + e^{-X\beta}},$$

where X is an $n \times (k+1)$ matrix of regressors, and β is a $(k+1)$-dimensional vector of coefficients, including the intercept.

Support Vector Machines (SVMs) find the optimal hyperplane that maximizes the margin between two classes. By transforming data into a higher-dimensional space using a kernel function, SVM ensures linear separability. The decision boundary is defined by the support vectors, i.e., the data points closest to the hyperplane. Given the small sample size, we opted for a basic linear SVM specification to avoid overfitting.

Ensemble Methods. Ensemble methods combine predictions from multiple models to enhance accuracy. *Categorical Boosting (CatBoost)* is part of the gradient boosting family, designed to improve prediction accuracy by sequentially adding decision trees. Each tree corrects the errors made by the previous ones (*boosting*), creating an "ensemble of decision trees." The iterative process can be described as:

$$\begin{cases} \hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x), \\ \quad r_i \leftarrow r_i - \lambda \hat{f}^b(x), \end{cases}$$

where $\hat{f}(x)$ and r_i represent the aggregated predictions and residuals, respectively. The index $b = 1, \ldots, B$ refers to the b-th decision tree, and B is the total number of trees, while $\lambda > 0$ denotes the learning rate. The loss function is minimized using a gradient-based optimization algorithm. Among gradient boosting methods, CatBoost is distinguished by its effective handling of categorical features via ordered encoding.

Random Forest constructs multiple decision trees on bootstrapped subsets of the original dataset (sampling with replacement). It decorrelates these trees by selecting a random subset of $m \approx \sqrt{k}$ predictors at each split. This randomized approach reduces the risk of overfitting and enhances generalization. For classification tasks, predictions are aggregated based on a majority vote, while for regression tasks, they are averaged.

3.2 Forecast Combination Approach

Before implementing the forecast combination methods, we briefly define each approach. According to [Gneiting and Ranjan, 2013], the combination formula $G(\cdot)$ is defined based on the predictive cumulative distribution functions $F_i(\cdot) \in \mathcal{F}$, where $i = 1, \ldots, k$, and k represents the number of previously estimated base models (in our case, $k = 5$). The formula is expressed as:

$$G : \mathcal{F}^k = \underbrace{\mathcal{F} \times \cdots \times \mathcal{F}}_{k \text{ times}} \to \mathcal{F}, \quad (F_1, \ldots, F_k) \mapsto G\left(F_1, \ldots, F_k\right).$$

The family of combination approaches is defined as $\mathcal{G} = \{G_\theta : \theta \in \Theta\}$, where G_θ represents the aggregated predictive distribution.

1. *Generalized Linear Pool (GLP):*

$$G(y) = h^{-1}\left(\sum_{i=1}^{k} w_i\, h(F_i(y))\right), \quad \sum_{i=1}^{k} w_i = 1,$$

where $h(\cdot)$ is a continuous and strictly monotonic link function. Examples include:

- Linear Pool: $h(x) = x$;
- Harmonic Pool: $h(x) = 1/x$;
- Logarithmic Pool: $h(x) = \ln(x)$;
- Normal Pool: $h(x) = \Phi^{-1}(x)$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard N(0,1) normal distribution.

2. *Beta-transformed Linear Pool (BLP):*

$$G_{\alpha,\beta}(y) = B_{\alpha,\beta}\left(\sum_{i=1}^{k} w_i F_i(y)\right),$$

where $B_{\alpha,\beta}(\cdot)$ is the CDF of the beta distribution with shape parameters $\alpha > 0$ and $\beta > 0$.

3. *Beta Mixture Combination (BMC):*

$$G_{\theta_m,\alpha_m,\beta_m}(y) = \sum_{j=1}^{m} \theta_j G_{\alpha_j,\beta_j}(y) = \sum_{j=1}^{m} \theta_j B_{\alpha_j,\beta_j}\left(\sum_{i=1}^{k} w_{ji} F_i(y)\right), \quad \sum_{j=1}^{m} \theta_j = 1,$$

where m is the number of beta components, and θ_j are the weights of each component.

Since this study addresses a binary classification problem, the task of aggregating predictive distributions simplifies to combining probability forecasts for an observation belonging to the positive class. Thus, for each approach, we optimize the linear combination of predicted probabilities for cryptocurrency exchange closures along with the beta distribution parameters to maximize the binary log-likelihood function ([Wattanachit et al., 2023]):

$$\ln \mathcal{L} = \sum_{i=1}^{228} \Big[y_i \ln \hat{G}_i(\hat{y}_i; \alpha, \beta, w, \theta) + (1 - y_i) \ln \Big(1 - \hat{G}_i(\hat{y}_i; \alpha, \beta, w, \theta) \Big) \Big] \longrightarrow \max_{\alpha,\beta,w,\theta}. \tag{1}$$

where 228 is the number of exchanges in our dataset (more below in the Data section). This non-linear constrained optimization problem is solved numerically using the Sequential Least Squares Quadratic Programming (SLSQP) algorithm, implemented in libraries such as `SciPy` and `pyslsqp` in Python.

3.3 Evaluation Metrics for Binary Classification

3.3.1 The Confusion Matrix and associated metrics

The performance of a binary classification model is typically assessed using the confusion matrix and related metrics, which provide insights into its predictive capability.

The confusion matrix summarizes the model's predictions:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

From this, we can derive the following commonly used metrics:

- Sensitivity (Recall or True Positive Rate): Reflects the ability to identify positive cases:

$$\text{Sensitivity} = \frac{TP}{TP + FN}.$$

- Precision: Measures the proportion of true positive predictions among all positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

- F1-Score: Harmonic mean of precision and recall, balancing false positives and false negatives:

$$\text{F1-Score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}.$$

Sensitivity, Precision, and the F1-score are widely used metrics for evaluating classification models; however, they are inherently threshold-dependent, meaning their values change based on the chosen probability cutoff. This threshold dependency can lead to biased evaluations, especially when comparing models across varying thresholds or in applications where selecting the optimal threshold is challenging. As an alternative, threshold-independent measures such as the Area Under the receiver operating characteristic Curve (AUC) and the H-measure provide a more robust evaluation of a model's overall discriminatory power, independent of any specific threshold. Additionally, model selection can benefit from loss-based metrics like the [Brier, 1950]'s score, which quantifies the accuracy of probabilistic forecasts, and robust statistical frameworks such as the Model Confidence Set (MCS) procedure by [Hansen et al., 2011], which identifies models that are statistically indistinguishable from the best-performing one. These approaches ensure a more comprehensive and reliable model comparison.

The Receiver Operating Characteristic (ROC) curve is a graphical representation of a classifier's performance across different decision thresholds. It plots the True Positive Rate (TPR), also known as sensitivity, against the False Positive Rate (FPR), also known as 1-specificity, defined as:

$$\mathrm{FPR} = \frac{FP}{FP + TN}, \quad \mathrm{TPR} = \frac{TP}{TP + FN}.$$

where, TP represents the number of true positives, FP the false positives, TN the true negatives, and FN the false negatives. Each point on the ROC curve corresponds to a specific threshold, with the curve illustrating the trade-off between sensitivity (the ability to correctly identify positive cases) and specificity (the ability to correctly identify negative cases). An ideal classifier has a curve that closely approaches the top-left corner, which corresponds to both high sensitivity and specificity. The closer the ROC curve is to this point, the better the model's overall performance.

The Area Under the ROC Curve (AUC) proposed by [Metz, 1978], [Metz and Kronman, 1980], and [Hanley and McNeil, 1982] quantifies the overall performance of the model, summarizing the ROC curve into a single value:

$$\mathrm{AUC} = \int_0^1 TPR\, d(FPR).$$

AUC values range from 0.5 (random guessing) to 1.0 (perfect classification). A higher AUC indicates better discriminative ability of the model, see [Sammut and Webb, 2011], pp. 869-875, and references therein for more details.

These metrics are particularly valuable in evaluating models that predict financial risks, such as defaults or closures, where false negatives (missed detections) can be costly. Sensitivity is often prioritized when the consequences of failing to predict positive cases outweigh those of false alarms. Meanwhile, the AUC provides a measure of model performance independent of specific thresholds.

3.3.2 The H-Measure

[Hand, 2009] highlighted several limitations of the Area Under the ROC Curve (AUC) metric. While AUC is widely used to evaluate classifier performance, it has significant drawbacks:

1. *Aggregation over thresholds*: The Area Under the Curve (AUC) of the ROC curve provides an aggregated measure of a classifier's performance across all possible thresholds. It reflects the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. However, it is important to note that when ROC curves intersect, one classifier may outperform another at certain thresholds, while the opposite is true elsewhere. This implies that AUC may fail to provide a definitive comparison in applications where certain thresholds are more critical. In such cases, it may be more appropriate to evaluate classifiers based on performance metrics at specific thresholds of interest, depending on the application's requirements, or use an alternative robust measure.

2. *Lack of Focus on Specific Regions*: In many real-world applications, specific regions of the ROC curve are more relevant. For instance, in financial risk analysis, minimizing false positives may be particularly important, and AUC does not emphasize performance in such critical regions.

To address these issues, [Hand, 2009] proposed the H-measure. This metric incorporates application-specific cost considerations and prioritizes classifier performance in the most relevant areas of the ROC curve. Below, we outline its key components and formulation.

The H-measure begins by identifying the optimal probability threshold $T(c)$ that minimizes the weighted loss for a given severity ratio c:

$$T(c) = \underset{t}{\operatorname{argmin}} \left\{ c\pi_0(1 - F_0(t)) + (1 - c)\pi_1 F_1(t) \right\}, \quad \text{where:} \tag{2}$$

- $c = \frac{c_0}{c_0 + c_1} = \left(1 + \frac{c_1}{c_0}\right)^{-1}$ is derived from the severity ratio $\frac{c_1}{c_0}$, which specifies the relative costs of misclassification for the two classes c_i $(i \in \{0, 1\})$;
- π_i is the prior probability of class i, i.e. its true share in the whole sample;
- $b = c_0 + c_1$ is a redundant scaling factor excluded from minimization;
- $F_i(t)$ is the cumulative distribution function (CDF) of scores for class i.

The loss function for a given threshold t is defined as:

$$Q(t; b, c) = b\left[c\pi_0(1 - F_0(t)) + (1 - c)\pi_1 F_1(t)\right]. \tag{3}$$

The general loss is then calculated by substituting the optimal threshold $T(c)$ from (2) into (3), weighting it using a severity distribution $u(c)$, and integrating over all possible severity ratios:

$$L_{\alpha^*,\beta^*} = \int Q\left(T(c); b, c\right) u_{\alpha^*,\beta^*}(c)\, dc. \tag{4}$$

Here, $u_{\alpha^*,\beta^*}(c)$ is the probability density function (PDF) of a Beta distribution with parameters $\alpha^* = \pi_1 + 1$ and $\beta^* = \pi_0 + 1$ [Hand and Anagnostopoulos, 2014]:

$$u_{\alpha^*,\beta^*}(c) = \frac{c^{\alpha^*-1}(1-c)^{\beta^*-1}}{B(\alpha^*, \beta^*)},$$

where $B(\alpha^*, \beta^*) = \dfrac{\Gamma(\alpha^*)\Gamma(\beta^*)}{\Gamma(\alpha^* + \beta^*)}$ is the Beta function and $\Gamma(\cdot)$ is the Gamma function.

The H-measure is defined as the normalized ratio of the general loss to the maximum possible loss, which occurs when the two class score distributions are indistinguishable (e.g., diagonal ROC curve with AUC = 0.5):

$$H = 1 - \frac{L_{\alpha^*,\beta^*}}{L_{\max}}, \tag{5}$$

where $L_{\max}$ is computed as:

$$L_{\max} = \pi_0 \int_0^{\pi_1} c\, u_{\alpha^*,\beta^*}(c)\, dc + \pi_1 \int_{\pi_1}^{1} (1-c)\, u_{\alpha^*,\beta^*}(c)\, dc.$$

The H-measure provides a more application-specific evaluation of classifier performance by incorporating misclassification costs and class imbalance. This is particularly beneficial in domains like finance, where imbalanced datasets and the high cost of certain errors (e.g., false positives in fraud detection) are common. By focusing on critical regions of the ROC curve, the H-measure addresses limitations of AUC and offers a more nuanced assessment of predictive models.

3.3.3 The Model Confidence Set (MCS) Procedure

The Model Confidence Set (MCS) procedure, proposed by [Hansen et al., 2011], is a statistical method used to compare and select forecasting models. Unlike traditional model selection techniques that focus solely on identifying a single "best" model, the MCS procedure identifies a set of models that are statistically indistinguishable from the best model at a given confidence level.

The MCS procedure is based on iterative hypothesis testing to eliminate inferior models. The process begins with an initial set of candidate models $\mathcal{M}_0$ of size m_0. For binary classification, the models are evaluated using a loss function, such as the *Brier score*:

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2, \tag{6}$$

where $\hat{p}_i$ is the predicted probability for observation i, $y_i \in \{0,1\}$ is the true class label, and n is the sample size. The MCS identifies the set of models $\mathcal{M}^*$ that are not significantly worse than the best model in $\mathcal{M}_0$. This is achieved by testing the null hypothesis that the expected performance of all models in $\mathcal{M}_0$ is equal:

$$H_0 : E[d_{ij}] = 0 \quad \forall\, i, j \in \mathcal{M}_0, \tag{7}$$

where $d_{ij} = L_i - L_j$ represents the pairwise difference in loss between models i and j.

The test statistic measures the relative performance of models using the loss differences d_{ij}. Two commonly used test statistics are the *Range Statistic* $R = \max_{i,j\in\mathcal{M}_0}|\bar{d}_{ij}|$, where $\bar{d}_{ij}$ is the sample mean of d_{ij}, and the *T-Statistic* $T = \max_{i\in\mathcal{M}_0}\frac{\bar{d}_{i+}}{\sqrt{\text{Var}(\bar{d}_{i+})}}$, where $\bar{d}_{i+}$ is the average loss difference of model i relative to others.

Models that fail the test are removed iteratively until the null hypothesis can no longer be rejected at the specified confidence level α. The resulting set of models $\mathcal{M}^* \subseteq \mathcal{M}_0$ contains models that are statistically indistinguishable from the best model. This ensures robustness in model evaluation, as the MCS accounts for uncertainty and avoids over-reliance on a single "best" model, particularly in small samples or when models perform similarly.

In our case, the Brier score serves as the loss function to evaluate the predictive performance of models in $\mathcal{M}_0$. By applying the MCS procedure, we can identify a subset of models that perform equivalently well in terms of probabilistic forecasts for binary outcomes. This is particularly useful in financial applications, where robustness and interpretability are critical, and small performance differences can have significant practical implications.

4 Results

4.1 Data

The target variable of this study is a binary indicator of whether an exchange is closed or active:

$$\textit{closed} = \begin{cases} 1, & \text{if closed,} \\ 0, & \text{if active.} \end{cases}$$

The explanatory variables used in the analysis are as follows:

(a) **Binary variables:**

1. *decentralized*: whether the exchange is decentralized;
2. *wire_transfer*: availability of fund deposits via bank transfer;
3. *credit_card*: availability of payment via credit or debit card;
4. *public_team*: presence of a publicly available Senior Leadership team profile;
5. *pen_test*: evidence of penetration tests assessing security resilience;
6. *proof_of_funds*: disclosure of reserve holdings by the exchange;
7. *bug_bounty*: existence of a bug bounty program incentivizing ethical hackers to identify vulnerabilities;
8. *hacked*: history of a security breach at the exchange.

(b) **Quantitative variables:**

9. *lifetime*: time in months from the exchange's foundation to its closure, or to May 2024 if still active;
10. *coins_traded*: number of cryptocurrencies available for trading;
11. *pairs_traded*: number of trading pairs offered by the exchange;
12. *cer_score*: cybersecurity score assigned by the CER platform;
13. *mozilla_score*: website security score provided by Mozilla Observatory;
14. *volume_mln*: daily trading volume (in million USD).

The dataset was manually compiled using information from various sources, including CoinMarketCap, Coingecko, CryptoWisser, BitDegree, CER.live, Mozilla Observatory, and SlowMist Hacked platforms as of May 15, 2024. For closed exchanges, additional information was obtained using the WayBack Machine, which provides archived versions of websites.

The final dataset consists of 228 exchanges, exactly one-third of which are closed. The full list of the analyzed crypto-exchanges can be found in Table 4 in the Appendix. Descriptive statistics for all variables are provided in Table 1. For example, the variable "decentralized" indicates whether an exchange operates in a decentralized manner (1) or is centralized (0). The mean value of 0.04 indicates that only 4% of the exchanges analyzed are decentralized, reflecting the dominance of centralized exchanges in the cryptocurrency market. This result is consistent with industry trends, where centralized exchanges typically offer higher trading volumes and user accessibility, despite the decentralized nature of blockchain technology, see [Fantazzini and Calabrese, 2021] and [Milunovich and Lee, 2022] and references therein. Given this imbalance, the results of our analysis primarily apply to centralized exchanges. Furthermore, as shown in Figure 2 below, the 'decentralized' variable was the least important predictor for both the CatBoost and Random Forest models, indicating limited relevance for predicting exchange closures in our study — a result consistent with the findings of [Fantazzini and Calabrese, 2021].

	mean	std	min	25%	50%	75%	max
closed	0.33	0.47	0	0	0	1	1
decentralized	0.04	0.184	0	0	0	0	1
wire_transfer	0.68	0.468	0	0	1	1	1
credit_card	0.53	0.5	0	0	1	1	1
lifetime	67.82	35.612	5	39	67	84.75	154
coins_traded	170.81	292.36	1	19.75	62.5	200.25	2424
pairs_traded	254.58	448.1	1	28.5	98.5	262.25	3452
public_team	0.71	0.45	0	0	1	1	1
cer_score	4.92	2.46	0.76	2.7	4.29	7.31	10
pen_test	0.30	0.46	0	0	0	1	1
proof_of_funds	0.49	0.5	0	0	0	1	1
bug_bounty	0.41	0.493	0	0	0	1	1
mozilla_score	43.25	27.51	0	25.	47.5	70	110
hacked	0.28	0.45	0	0	0	1	1
volume_mln	361.21	1248.9	0	1.575	31	256.25	17

Table 1: Full sample descriptive statistics for the analyzed crypto-exchange dataset.

To account for the significant variability in the means and standard deviations of quantitative features, we applied min-max scaling to normalize these variables to a range of $[0, 1]$:

$$X_i^* = \frac{X_i - X_{min}}{X_{max} - X_{min}} \in [0, 1].$$

This preprocessing step ensures comparability across features and facilitates the implementation of logistic regression with regularization, which will be employed later in this study.

4.2 Empirical Analysis: Machine Learning models

To evaluate the performance of our models, we employed out-of-sample predictions computed using the Leave-One-Out Cross-Validation (LOOCV) technique. This approach is particularly suited for small datasets, such as ours, which consists of 228 exchanges. LOOCV works by iteratively training the model on all observations except one, then using the excluded observation for testing. This process is repeated for each observation in the dataset, resulting in a comprehensive assessment of the model's predictive ability.

We deliberately avoided conducting any in-sample analysis and focused exclusively on out-of-sample forecasting. The primary reason for this decision was to mitigate the risk of overfitting, which is a significant concern when working with small datasets. In-sample evaluation could lead to overly optimistic performance metrics, as the model would be assessed on data it has already seen. By contrast, out-of-sample evaluation ensures a more realistic estimate of the model's ability to generalize to unseen data, which is essential for deriving meaningful insights in practical applications.

The use of LOOCV allowed us to maximize the utilization of the limited data available while maintaining the integrity of the evaluation process. By training the model on nearly the entire dataset for each iteration, LOOCV provides robust predictions without the need to set aside a separate validation set, which would have further reduced the sample size available for training. This makes LOOCV a natural choice for empirical studies involving small samples, such as this one.

Figure 1 depicts the receiver operating characteristic (ROC) curves for the predictions made by our five Machine Learning models. These curves illustrate the relationship between the true positive rate (TPR) and the false positive rate (FPR) across different classification thresholds t, where $\hat{y} = \mathbf{1}\{\hat{p} \geq t\}$. The ROC curves provide a visual representation of how well each model balances sensitivity (TPR) and specificity (1 - FPR) as the threshold t is varied. For example, points closer to the top-left corner represent better performance, with higher sensitivity and specificity. The multiple intersections of these curves highlight the limitations of comparing models solely based on the area under the ROC curve (AUC) metric, as discussed by [Hand, 2009]. Such intersections suggest that one model may outperform another at certain thresholds while underperforming at others. This reinforces the need to consider additional robust evaluation metrics, such as the H-measure, to accurately assess model performance in real-world applications.

Table 2 reports key performance metrics for the five ML models: AUC, H-measure, F1-score, Brier Score, and their inclusion in the Model Confidence Set (MCS). These metrics provide a comprehensive evaluation of classification accuracy, calibration, and robustness. Notably, CatBoost and Random Forest achieved the highest performance, as evidenced by their superior H-measure values (0.614 and 0.621, respectively), lowest Brier Scores (0.103 and 0.102, respectively), and their inclusion in the MCS. The MCS procedure, conducted at a 95% confidence level with the Brier Score as the loss function, identified these two models as statistically indistinguishable in terms of predictive ability.

	AUC	F1-score	Brier Score	H	MCS
Naive Bayes	0.841	0.748	0.162	0.523	No
Logistic Regression	0.878	0.775	0.124	0.553	No
SVC	0.857	0.715	0.132	0.527	No
CatBoost	0.914	0.769	0.103	0.614	**Yes**
Random Forest	0.921	0.696	0.102	0.621	**Yes**

Table 2: Performance comparison of Machine Learning models (Naive Bayes, Logistic Regression, SVC, CatBoost, and Random Forest) based on AUC, F1-score, Brier Score, H-measure, and inclusion in the Model Confidence Set (MCS).

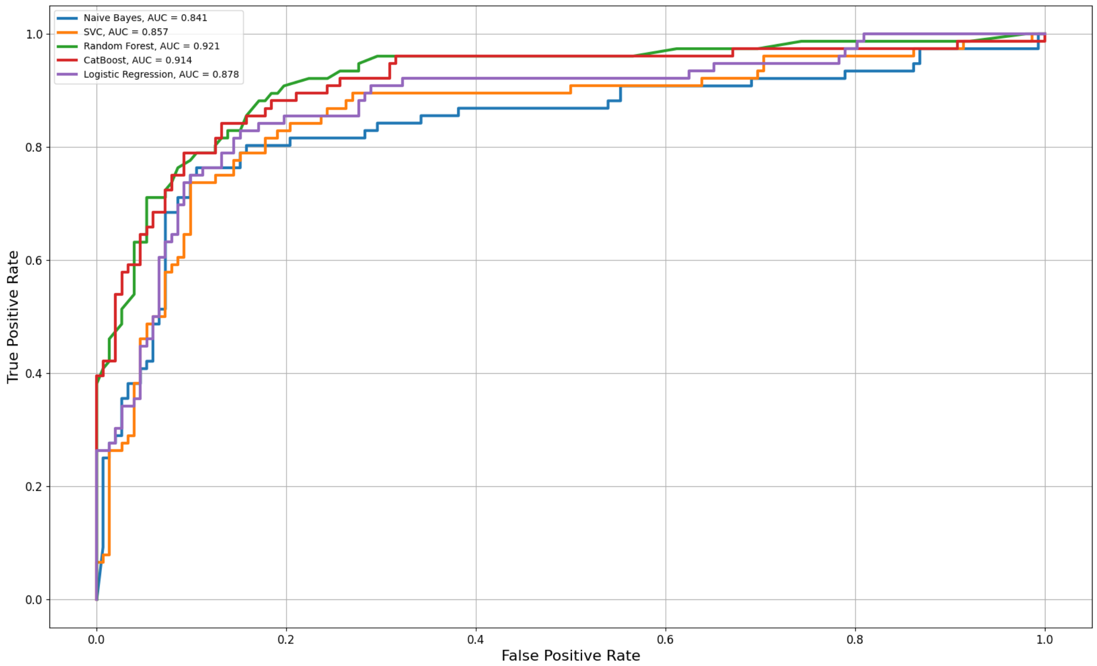

Figure 1: Receiver Operating Characteristic (ROC) curves for the five Machine Learning models. The x-axis represents the false positive rate (FPR), and the y-axis represents the true positive rate (TPR). The multiple intersections of these curves demonstrate that AUC alone may not capture the nuanced differences in performance, necessitating further evaluation metrics.

In the context of Machine Learning algorithms, the term *feature importance* refers to the contribution of each input variable to the predictive performance of the model. For tree-based algorithms like CatBoost and Random Forest, feature importance is typically measured by how often a feature is used to split data points across decision trees and the degree to which it reduces prediction error (e.g., Gini impurity or entropy). Features with higher importance scores have a greater influence on the model's predictions. This interpretability is particularly valuable in financial applications, as it allows researchers to identify the key drivers of the target variable and gain insights into underlying patterns.

Figure 2 visualizes the feature importances for the two best ML models, CatBoost and Random Forest. The most influential features are the exchange's lifetime and daily trading volume, both of which are consistently ranked at the top. Additionally, the CER security score and Mozilla Observatory security score are among the top five features. These findings align with our second hypothesis, discussed in the Introduction, that the likelihood of a cryptocurrency exchange remaining operational is significantly influenced by its operational history, market activity, and security measures.

The prominence of lifetime and trading volume underscores the critical role of long-term trust and liquidity in sustaining exchanges. Security metrics, such as the CER and Mozilla scores, further highlight the importance of robust cybersecurity practices in preventing potential vulnerabilities that could lead to exchange closure. These results not only validate our hypotheses but also offer practical insights for industry stakeholders aiming to assess the viability and resilience of cryptocurrency exchanges.

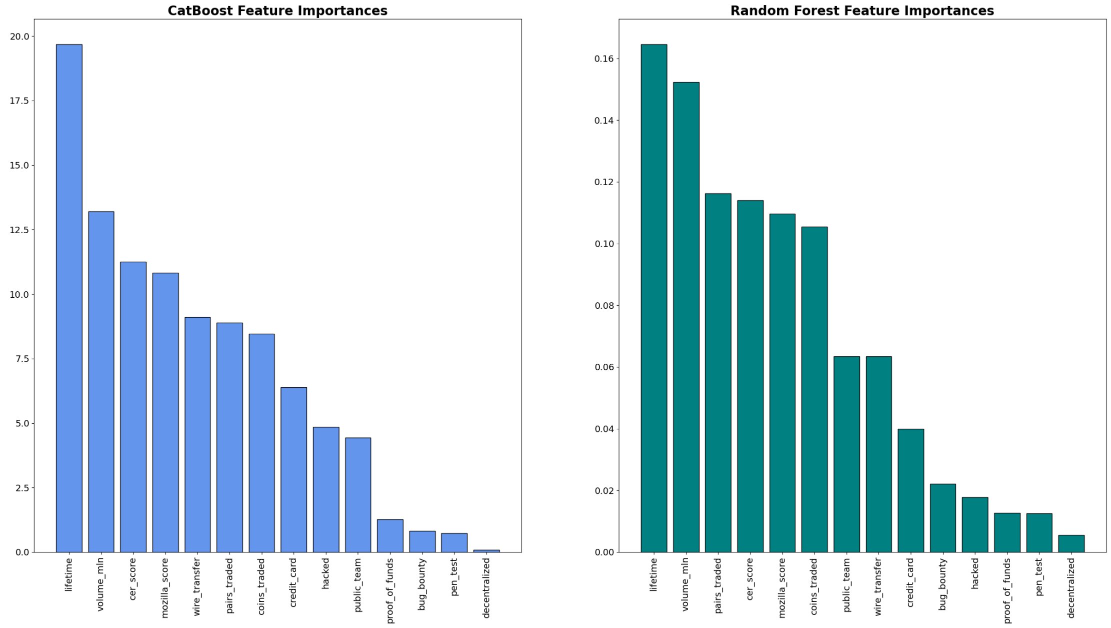

Figure 2: Feature Importances for the Two Best ML Models (CatBoost and Random Forest)

4.3 Empirical Analysis: Forecast Combination Approach

To further enhance predictive performance, we investigated whether combining forecasts from multiple models could outperform the best individual algorithm. Forecast combination methods are well-known for improving accuracy by leveraging the strengths of different models and mitigating their weaknesses. In this subsection, we compare several combination approaches to the baseline Random Forest model, which was previously identified as the best-performing base algorithm, performing better in 3 out of 4 forecasting metrics (see Table 2).

	AUC	F1-score	Brier Score	H	MCS
Random Forest	0.921	0.775	0.102	0.621	No
Linear Pool	0.922	0.772	0.100	0.632	No
Harmonic Pool	0.901	0.757	0.105	0.612	No
Logarithmic Pool	0.919	0.755	0.100	0.631	No
N(0,1) Pool	0.921	0.772	0.100	0.631	No
BLP	0.924	0.767	0.099	0.647	**Yes**
BMC(2)	0.924	0.767	0.099	0.647	**Yes**
BMC(3)	0.924	0.767	0.099	0.647	**Yes**

Table 3: Performance metrics for forecast combination methods (for example, linear pool, harmonic pool, and Beta Mixture Combination) compared to the Random Forest baseline, evaluated using AUC, F1-score, Brier Score, H-measure, and MCS inclusion.

Table 3 reports the comparative performance metrics for the Random Forest model and various forecast combination approaches, including the linear pool, harmonic pool, logarithmic pool, Normal N(0,1) pool, beta-transformed linear pool (BLP), and Beta Mixture Combination (BMC) with two or three beta components. Several observations emerge from this analysis:

1. *Performance of Combination Methods*:
   - Both the BLP and BMC methods (with 2 and 3 components) achieved the highest performance metrics, with an AUC of 0.924, F1-score of 0.767, Brier Score of 0.099, and H-measure of 0.647. These represent an improvement in the H-measure by 4.1% compared to Random Forest (H-measure = 0.621) and a reduction in the Brier Score by 2.8% (from 0.102 to 0.099). The inclusion of the BLP and BMC models in the Model Confidence Set (MCS) at a 95% significance level, coupled with the exclusion of all other models, confirms that their improvements in forecasting performance are statistically significant..
   - Simpler combination methods, such as the linear pool, also improved performance relative to Random Forest, achieving an H-measure of 0.632 (an increase of 1.8%) and a Brier Score of 0.100 (a reduction of 2.0%). However, these improvements are less pronounced compared to the BLP and BMC methods.
2. *Bias-Variance Tradeoff*:
   - The harmonic pool and logarithmic pool exhibited slightly lower performance than the Random Forest baseline, with H-measures of 0.612 and 0.631, respectively. Additionally, the harmonic pool had the highest Brier Score of 0.105, indicating a poorer calibration of probabilities. This suggests that overly simplistic or rigid pooling strategies may fail to capitalize on the diversity of forecasts effectively.
   - In contrast, the BLP and BMC methods demonstrated a better balance between bias and variance, achieving the lowest Brier Score of 0.099 and the highest H-measure of 0.647, indicating robust and well-calibrated forecasts. This supports the hypothesis that more flexible combination techniques can effectively harness the strengths of individual models without introducing excessive variance.
3. *Validation of the Forecast Combination Hypothesis*: The superior performance of the BLP and BMC methods provides strong evidence in support of our first hypothesis: combining forecasts enhances accuracy compared to relying on a single model. The BLP and BMC methods not only achieved the highest AUC and H-measure values but also consistently outperformed simpler pooling methods in terms of calibration and overall predictive ability. These results are particularly valuable in contexts like ours, where high-stakes decisions require robust and well-calibrated predictions.

In conclusion, the results demonstrate the value of forecast combination approaches in improving model performance. The BLP and BMC methods are particularly effective, leveraging the strengths of individual models while maintaining robustness and avoiding overfitting. This underscores the importance of considering ensemble techniques, especially in scenarios with complex relationships and high uncertainty, such as predicting the closure of cryptocurrency exchanges.

5 Discussion and Conclusions

This study set out to address two key hypotheses: (1) it is possible to improve the accuracy of probabilistic forecasts through ensemble methods, and (2) the probability of closure of cryptocurrency exchanges is significantly influenced by their lifetime, daily trading volume, and cybersecurity scores. Both hypotheses were successfully confirmed, yielding the following key results:

1. The application of ensemble methods, particularly the Beta-Transformed Linear Pool (BLP) and Beta Mixture Combination (BMC), resulted in a significant improvement in forecast quality. These methods increased the robust H-measure by over 4% and reduced the Brier Score by 2.8% compared to the already highly accurate Random Forest classifier. This demonstrates the value of combining forecasts to achieve superior predictive performance.

2. The analysis of feature importance revealed that the lifetime of a crypto-exchange and its daily trading volume together account for over 30% of feature importance. When security-related features such as CER and Mozilla security scores are included, this proportion exceeds 50%. These findings strongly support the hypothesis that operational longevity, trading activity, and robust security measures are critical factors in determining the survival of cryptocurrency exchanges.

The novelty of this research lies in its focus on a unique, manually collected dataset of 228 cryptocurrency exchanges, offering up-to-date insights into a rapidly evolving industry. The application of modern statistical methods, including state-of-the-art machine learning algorithms and advanced ensemble forecasting techniques, further distinguishes this study from prior work in the field.

Our results not only provide practical tools for evaluating the probability of default for cryptocurrency exchanges but also contribute to the broader understanding of risk factors in this nascent and volatile sector. The development of reliable and accurate probability-of-default models will remain an essential area of inquiry as the cryptocurrency market continues to expand and mature.

Limitations of the Study

Despite the valuable contributions of this research, several limitations must be acknowledged:

- **Sample Size:** The dataset includes 228 exchanges, which, while sufficient for initial analysis, limits the generalizability of the findings. A larger sample size would enable the use of more sophisticated validation techniques, such as a train-validate-test split, and provide more robust estimates of model performance.

- **Data Quality and Availability:** The manually collected dataset relies on multiple external sources, which may introduce biases or inconsistencies. Furthermore, historical data for closed exchanges often depended on archived websites, which could lack accuracy or completeness.

- **Model Complexity:** While ensemble methods like BLP and BMC showed significant improvements, the study avoided overly complex models to mitigate the risk of overfitting given the small sample size. This decision may have excluded some advanced techniques that could perform better with larger datasets.

- **Dynamic Factors:** The crypto market evolves rapidly, with new factors such as regulatory changes, technological innovations, and macroeconomic conditions influencing exchange closures. Our static dataset does not fully capture these dynamic effects, potentially limiting the predictive power of the models in changing environments.

Future Research Directions

Building on the findings and limitations of this study, several avenues for future research are worth exploring:

- **Expanding the Dataset:** Incorporating additional exchanges and updating the dataset with more recent closures and newly established platforms would provide a more comprehensive view of the market. A larger sample size would also enable the application of deep learning techniques and more complex ensemble methods.

- **Dynamic Modeling:** Future studies could investigate time-dependent models to capture the evolving nature of the cryptocurrency market. Approaches such as dynamic survival models or recurrent neural networks could provide insights into how risks change over time.

- **Alternative Feature Engineering:** While this study focused on operational and security-related features, future work could explore additional predictors, such as user sentiment analysis from social media, blockchain activity data, or regulatory announcements.

- **Explainability and Interpretability:** As machine learning models become increasingly complex, incorporating methods to enhance model interpretability (e.g., SHAP or LIME, see [Lundberg and

Lee, 2017] and [Ribeiro et al., 2016]) could make the results more actionable for stakeholders.

- **Scenario Analysis and Stress Testing:** Developing models that can evaluate the impact of extreme events, such as major hacks or regulatory crackdowns, would provide valuable insights for risk management in the crypto sector.

Concluding Remarks

This study has demonstrated the potential of ensemble methods and machine learning algorithms to significantly improve the accuracy of default predictions for cryptocurrency exchanges. The findings have practical implications for multiple stakeholders:

- **For Investors**: By identifying the key factors that influence exchange survival—such as operational longevity, trading volume, and security features—this research provides a data-driven framework to assess the risks associated with specific exchanges. Investors can use these insights to make informed decisions about where to allocate their funds, mitigating potential losses from exchange closures.
- **For Exchange Operators**: The results highlight the importance of robust security measures and sustained trading activity in maintaining operational longevity. Exchange operators can leverage these findings to prioritize cybersecurity investments and strategies to increase trading volume, thereby improving their chances of long-term success.
- **For Regulators**: The study offers a foundation for developing regulatory frameworks aimed at enhancing market stability. By focusing on the key risk factors identified in this research, regulators can create guidelines that promote transparency, security, and sustainability within the cryptocurrency market.

In addition to its practical contributions, this research also advances the academic understanding of risk assessment in the nascent and rapidly evolving cryptocurrency sector. By leveraging state-of-the-art ensemble methods such as the Beta-Transformed Linear Pool (BLP) and Beta Mixture Combination (BMC), the study demonstrates the value of combining probabilistic forecasts to achieve superior predictive performance. The robust improvement in forecast quality—reflected by a 4% increase in the H-measure and a 2.8% reduction in the Brier Score compared to the Random Forest classifier—sets a benchmark for future research in this area.

Finally, this study underscores the importance of addressing the limitations and challenges associated with data quality and market dynamics. The proposed avenues for future research, such as expanding the dataset, incorporating dynamic modeling techniques, and exploring additional predictive features, provide a roadmap for advancing the field further. As the cryptocurrency market continues to mature, ongoing research will be critical to developing tools and strategies that can adapt to its evolving risks and opportunities.

By combining methodological rigor with practical relevance, this study contributes to the growing body of literature on risk assessment and predictive modeling in the cryptocurrency sector. The findings serve as a call to action for researchers, practitioners, and policymakers to work collaboratively in addressing the challenges and seizing the opportunities presented by this dynamic and transformative market.

References

Philippe Bracke, Anupam Datta, Carsten Jung, and Shayak Sen. Machine learning explainability in finance: an application to default risk analysis. Technical report, Bank of England, working paper n. 816, 2019.

Glenn Brier. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Roberto Casarin, Giulia Mantoan, and Francesco Ravazzolo. Bayesian calibration of generalized pools of predictive distributions. *Econometrics*, 4(1):17, 2016.

Dean Fantazzini and Raffaella Calabrese. Crypto Exchanges and Credit Risk: Modeling and Forecasting the Probability of Closure. *Journal of Risk and Financial Management*, 14(11):516, 2021.

Pedro G Fonseca and Hugo D Lopes. Calibration of machine learning classifiers for probability of default modelling. *arXiv preprint arXiv:1710.08901*, 2017.

Tilmann Gneiting and Roopesh Ranjan. Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782, 2013.

David J Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123, 2009.

David J Hand and Christoforos Anagnostopoulos. A better beta for the h measure of classification performance. *Pattern Recognition Letters*, 40:41–46, 2014.

James Hanley and Barbara McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

Peter Hansen, Asger Lunde, and James Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

Kajal Lahiri and Liu Yang. Forecasting binary outcomes. In *Handbook of economic forecasting*, volume 2, pages 1025–1106. Elsevier, 2013.

Kajal Lahiri, Huaming Peng, and Yongchen Zhao. Testing the value of probability forecasts for calibrated combining. *International journal of forecasting*, 31(1):113–129, 2015.

Cheng Few Lee and Hai-Chin Yu. Application of discriminant analysis, factor analysis, logistic regression, and KMV-Merton model in credit risk analysis. In *Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning*, pages 4313–4348. World Scientific, 2021.

Seung Ah Lee and George Milunovich. Digital exchange attributes and the risk of closure. *Blockchain: Research and Applications*, 4(2):100131, 2023.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

Charles Metz. Basic principles of ROC analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

Charles Metz and Helen Kronman. Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology*, 22(3):218–243, 1980.

George Milunovich and Seung Ah Lee. Cryptocurrency exchanges: Predicting which markets will remain active. *Journal of forecasting*, 41(5):945–955, 2022.

Tyler Moore, Nicolas Christin, and Janos Szurdi. Revisiting the risks of bitcoin currency exchange closure. *ACM Transactions on Internet Technology*, 18(4):1–18, 2018.

Mojtaba Nabipour, Pooyan Nayyeri, Hamed Jabani, S Shahab, and Amir Mosavi. Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. *Ieee Access*, 8: 150199–150212, 2020.

Satoshi Nakamoto. A peer to peer electronic cash system. Technical report, 2008. URL `https://bitcoin.org/bitcoin.pdf`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Claude Sammut and Geoffrey Webb. *Encyclopedia of machine learning*. Springer, 2011.

Patrick Schueffel and Nikolaj Groeneweg. Evaluating crypto exchanges in the absence of governmental frameworks-a multiple criteria scoring model. *Available at SSRN 3432798*, 2019.

Nutcha Wattanachit, Evan L Ray, Thomas C McAndrew, and Nicholas G Reich. Comparison of combination methods to create calibrated ensemble forecasts for seasonal influenza in the us. *Statistics in Medicine*, 42(26):4696–4712, 2023.

Appendix

3xbit	6x	Aax	ABCC
Abucoins	AlphaX	AlterDice	Altilly
Altsbit	AscendEx	Azbit	B2BX
Backpack	bancor	BHEX (HBTC)	bibox
Biconomy	BigOne	BiKi	Bilaxy
binance	BingX	Bit2Me	Bitazza
Bitbank	BitBNS	Bitcastle	Bitci TR
Bitcointry	BitDelta	Bitexen	Bitfinex
bitFlyer	BitForex	Bitfront (Bitbox)	Bitget
BitGrail	Bithumb	BITKER	Bitkub
Bitlish	Bitlo	BitMart	BitMesh
BitMex	BitoPro	Bitrue	bitso
BitStamp	Bitsten	BitStorage	Bittrex
Bitunix	Bitvavo	BitVenus	BKEX
Bleutrade	Blockchain.com (The PIT)	Blofin	BTCbear
BTCEX	BtcTurk	BTSE	Bullish
Bybit	BYDFi	C-CEX	C-Patex
Catex	Chainrift	ChaoEX	Chilebit.net
CITEX	Cobinhood	Coinbase	CoinBene
Coinchangex	Coincheck	CoinCorner (Coinfloor)	CoinDeal
Coineal	CoinEgg	CoinEx	CoinFalcon
Coinhub	CoinJar	CoinLim	Coinlist
Coinmetro	Coinnest	Coinone	Coinrate
Coins.ph	Coinsbit	Coinstore	Coinsuper
CoinTiger	CoinTR Pro	CoinW	CPDAX
CredoEx	Cryptal	Crypto Dao	Crypto.com
CryptoBridge DEX	Cryptology	CryTrEx	Currency.com
Dcoin	Deepcoin	Deribit	Dex-Trade
DigiFinex	Emirex	Exmo	Fairdesk
Fastex	FatBTC	Fcoin	Fisco
FMFW.io	Foxbit	FTX	Gate.io
GDAC	Gemini	GMO Japan	GokuMarket
GoPax	Hashkey	HB.top	HBUS
HitBTC	Hoo.com	Hotbit	Hotcoin
HTX (Huobi)	iCE3	ICOCryptex	Icrypex
Independent Reserve	Indodax	Instant Bitex	IQFinex
itBit	Kanga	KickEx	KoinBX
Koinpark	Korbit	Kraken	KuCoin
Kuna	LakeBTC	LATOKEN	Lbank
LCX	LEOxChange	Liquid	Livecoin
LocalTrade	Lukki	Luno (BitX)	Max Maicoin
Mercado Bitcoin	Mercatox	MEXC	Narkasa
Neraex	Nicehash	NLexch	Nominex
Nonkyc.io	OceanEx	Okcoin	OKX (OKEx)
One Trading (Bitpanda)	OPNX	OrangeX	OTCBTC
P2B	Paribu	Phemex	Pionex
PointPay	Poloniex	ProBit	Purcow
QMall	Shortex	Sistemkoin	Slex
Sparkdex	SpectroCoin (Bankera)	STEX	StormGain
Tapbit	TheRockTrading	Thodex (Koineks)	Tidex
Tokenize	TokensNet	TokoCrypto	Tokpie
Toobit	TopBTC	Trade Satoshi	Tux Exchange
Txbit	Unichange	Upbit	VALR
Vbitex	Vebitcoin	VirWox	WazirX
Websea	WEEX	WhiteBIT	WOO X
Worldcore	XeggeX	XT.com	YoBit
Zaif	ZebPay	ZG.top	zondacrypto (BitBay)

Table 4: List of Analyzed Crypto-exchanges