Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 17, Issue, 12, pp.35658-35663, December, 2025
DOI: https://doi.org/10.24941/ijcr.49832.12.2025

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

A COMPUTATIONAL ECONOMICS FRAMEWORK FOR RARE-EVENT PREDICTION: EVIDENCE FROM CRYPTOCURRENCY PRICE SHOCKS USING SMOTE AND ENSEMBLE LEARNING

1,*V. Shiva Sankari and 2Dr. R. Kavitha

1Research Scholar, Department of Commerce, Periyar University, Salem – 11, India; 2Assistant Professor, Department of Commerce, Periyar University, Salem – 11, India

ARTICLE INFO

Article History:
Received 27th September, 2025
Received in revised form
28th October, 2025
Accepted 20th November, 2025
Published online 30th December, 2025

Keywords:

Cryptocurrency, Machine Learning, Price Shocks, XGBoost, SMOTE, Predictive Analytics, Financial Risk.

Corresponding author:
Dr. R. Kavitha

ABSTRACT

This study develops a computational economics framework for modeling and predicting rare financial events, using cryptocurrency price shocks as an illustrative case. We treat shock detection as a binary classification problem, where extreme returns are defined by the 90th percentile of absolute log returns. To address the rarity of such events, we integrate the Synthetic Minority Over-sampling Technique (SMOTE) with ensemble learning methods (Random Forest, XGBoost, Light GBM) and benchmark against logistic regression. Using daily Bitcoin data from August 2024 to August 2025, our results show that ensemble models—especially Light GBM—achieve strong predictive performance (AUC = 0.919) and substantially higher recall for shock detection. Feature importance analysis highlights the predictive role of short-term volatility memory and liquidity variation, consistent with theoretical constructs such as EVT, EMH deviations, and the Mixture of Distributions Hypothesis. Beyond crypto currencies, the proposed framework provides a generalizable computational approach to rare-event forecasting in economics, with direct applications to crises, systemic liquidity shocks, credit defaults, and exchange rate collapses. By combining imbalance correction with ensemble learners, this study contributes methodologically to the computational economics literature on tail-risk prediction and systemic stability.

Copyright©2025, Shiva Sankari and Kavitha. 2025. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Citation: V. Shiva Sankari and Dr. R. Kavitha. 2025. "A Computational Economics Framework for Rare-Event Prediction: Evidence from Cryptocurrency Price Shocks Using SMOTE and Ensemble Learning". ***International Journal of Current Research, 17, (12), 35658-35663.***

INTRODUCTION

Cryptocurrency markets have rapidly evolved since Bitcoin's 2009 inception. Analysts note that Bitcoin's academic literature "has grown very fast in recent years," reflecting the intense research interest(Baur & Dimpfl, 2021). Likewise, cryptocurrency's decentralized design has "garnered significant interest from both retail and institutional investors"(*A Comparative Analysis of Statistical and Machine Learning Models for Outlier Detection in Bitcoin Limit Order Books*, n.d.). Proponents cite benefits like cheaper and faster transactions and the elimination of single points of failure(Team, 2024), as well as open global access for anyone with an internet connection. Critics counter that crypto markets exhibit extreme volatility – for example, Bitcoin's price swings are "almost 10 times higher" than those of major currency exchange rates(Baur & Dimpfl, 2021b) – along with pronounced liquidity fluctuations and a "comparatively underdeveloped market microstructure"(*A Comparative Analysis of Statistical and Machine Learning Models for Outlier Detection in Bitcoin Limit Order Books*, n.d.-b). These factors make cryptocurrencies highly sensitive to large price moves. In particular, price shocks – sudden, substantial spikes or drops in returns – pose serious challenges to portfolio risk management, automated trading systems, and regulatory oversight. Indeed, researchers emphasize that accurately modelling and forecasting cryptocurrency volatility is "vital for risk assessment, asset management, and regulatory policy" in these markets (Zhou *et al*., 2025), highlighting the need for effective shock prediction and hedging strategies. Financial researchers have extensively applied time-series and statistical models to forecast asset returns and volatility. For example, early cryptocurrency studies used econometric models like GARCH and ARIMA on Bitcoin returns(Zhou *et al*., 2025). However, because crypto price series tend to be nonstationary, highly nonlinear, and clustered, traditional models often struggle to capture their complex patterns(Zhou *et al*., 2025). In response, many recent studies have turned to machine learning. Advanced AI/ML methods (including neural networks and ensemble techniques) have proven more flexible and adaptive for volatile, non-linear data(Zhang *et al*., 2023). He observed that deep learning models "perform better in prediction tasks than linear and machine learning models in the financial field, especially in the cryptocurrency market". These techniques can ingest large sets of features (e.g. lagged returns, volume, sentiment indices) and uncover hidden signals. Nonetheless, most existing work still targets general price prediction or volatility forecasting rather than specifically identifying rare shock events. In fact, some literature reviews note that the transmission mechanisms of cryptocurrency price shocks remain understudied(Chen, 2025). In other words, few studies explicitly tackle the classification of high-impact, low-frequency jumps in crypto prices. This represents a critical gap in the forecasting literature: most methods are optimized for average-case movements, with relatively little attention to extreme tail events. Several methodological gaps motivate this study. First, very few studies have compared the performance of different ML classifiers (e.g. linear versus ensemble methods) specifically for detecting crypto price shocks, especially under severe class imbalance. Second, it remains unclear which market-based variables (such as past returns, rolling volatility, or trading volume) might serve as early warning indicators of an impending shock. Third, class imbalance is a significant issue: shock events are rare, so naive models tend to predict the non-shock class overwhelmingly. While techniques like Synthetic Minority Over-sampling (SMOTE) are known to help in imbalanced settings (Chawla *et al*., 2002), their use in crypto-shock forecasting has been limited. Addressing these gaps could improve hedging and risk-management tools for crypto portfolios. To fill these gaps, we develop a supervised ML framework to predict rare price shocks in the Bitcoin market using daily data over one year. The process begins by cleaning the raw price series and constructing feature variables, including lagged returns, volatility measures, and volume indicators. A binary shock label is then defined by flagging days when the absolute log-return exceeds the 90th percentile (a high-return threshold). Because shocks are scarce, we apply the Synthetic Minority Over-sampling Technique (SMOTE) to the training set, creating synthetic shock examples to balance the classes(Chawla *et al*., 2002). Finally, four classifiers – Logistic Regression, Random Forest, XGBoost, and LightGBM – are trained on the features. We evaluate their ability to predict shocks using metrics such as accuracy, recall, F1 score, and the area under the ROC curve.

The study is guided by the following research questions

- Can supervised ML models reliably predict rare cryptocurrency price shocks using limited but relevant market data?
- What are the most influential market predictors—such as volatility, trading volume, or lagged returns—of impending Bitcoin price shocks?
- How do different classification algorithms compare in terms of performance, sensitivity to rare events, and robustness under class-imbalanced conditions?

Beyond cryptocurrency-specific applications, this study contributes more broadly to computational economics by offering a rare-event modeling framework applicable to other domains where extreme events play a disproportionate role. Financial crises, exchange rate collapses, systemic liquidity shocks, and credit defaults share the same statistical and computational challenges as cryptocurrency price shocks: nonlinearity, volatility clustering, and severe class imbalance. By integrating machine learning with statistical theories such as EVT, EMH, and MDH, the framework provides a generalizable computational architecture for anticipating tail events in economic systems. In this sense, the contribution extends beyond Bitcoin prediction and offers methodological insights into how computational economics can more effectively address low-frequency, high-impact events across different markets and macroeconomic contexts. The remainder of this paper is structured as follows. Section 2 reviews related work on machine learning applications in financial risk prediction. Section 3 presents the theoretical context for rare-event modeling. Section 4 details the methodology, including data processing, feature engineering, and model development. Section 5 reports and interprets the empirical findings. Section 6 concludes with practical implications and future research directions.

LITERATURE REVIEW

Cryptocurrency price prediction: Predicting cryptocurrency prices is challenging due to extreme volatility and lack of historical structure(John *et al*., 2024). Bouri *et al.* found that cryptocurrencies exhibit far greater volatility than conventional stocks. Machine learning methods are therefore favored. Recent studies show that ensemble and deep learning models (e.g. GRUs, RNNs, LightGBM) outperform traditional approaches for crypto forecasting (Bouteska, 2024). Sun *et al* introduced a LightGBM-based model combining multiple cryptocurrency and economic indicators, reporting higher robustness than other techniques. Incorporating alternative data can also help. Gurrib and Kamalov (2021) used Bitcoin prices plus news sentiment in LDA/SVM models, improving next-day directional accuracy to 58.5% (better than chance). A recent survey highlights emerging architectures as promising directions to further boost prediction accuracy (John *et al*., 2024). Overall, these works suggest that advanced ML algorithms (ensemble, deep nets, gradient boosting) are effective for cryptocurrency price forecasting under high volatility.

Volatility forecasting and shock detection: Accurately modeling volatility and detecting market shocks is critical in finance. Machine learning models have been applied to capture volatility spikes and regime changes. In an adaptive forecasting framework, Sun *et al* demonstrate that their hypernetwork-LSTM model maintains accuracy during extreme conditions by dynamically adjusting to heightened volatility. Similarly, simpler recurrent models like LSTM often outperform more complex transformers in crisis periods: one study reports that during the 2008 financial crisis, an LSTM continued to capture sudden volatility increases while a Transformer failed to adapt(Bouteska, 2024). In systematic evaluations, Mansilla-López *et al.* (2025) review stock market volatility forecasting and find that hybrid ML models (e.g. LSTM+GARCH) can reduce forecast errors by over 10% in terms of MAE/MSE. These results indicate that combining deep learning with traditional volatility models significantly enhances performance, and that adaptive ML architectures can effectively detect and adjust to shocks.

Imbalanced classification in finance: Many financial datasets are highly imbalanced (e.g., fraud detection, credit defaults), requiring specialized ML techniques. Approaches combine sampling with learning to improve minority-class detection. For example:

Weighted oversampling + ensemble: Abedin *et al.* (2022) propose a WSMOTE-ensemble method for small-business credit risk. By generating synthetic defaults and using a bagged ensemble, they improved minority (default) accuracy by 15.16% over baseline methods. They show sampling-based classifiers significantly outperform no-sampling approaches.

GAN-based augmentation: Adiputra *et al.* (2025) benchmark GAN oversampling for multi-class credit scoring. Using methods like WGAN-GP to synthesize minority-class data, the best model (WGAN-GP + Random Forest) achieved accuracy 0.873 and F1-scores 0.806–0.936 across classes. This demonstrates that GAN-generated samples can greatly improve classification on highly imbalanced credit datasets.

Generative autoencoders and VAEs: Tayebi & El Kafhali (2025) develop autoencoder, VAE, and GAN models to augment fraudulent transaction data. Their generative models synthesize new fraud examples, which markedly boost detection performance compared to conventional oversampling (SMOTE/ADASYN). They report superior balanced accuracy and introduce a composite score (BFDS) to evaluate improvements.

Differentiated sampling ensembles: Wang *et al.* (2024) introduce a KSDE algorithm that detects and removes outliers, then creates multiple balanced subsets via varying sampling rates. The weighted ensemble of submodels attains a 12.46% higher true positive rate than prior methods on credit risk data. These works collectively highlight that advanced resampling (weighted SMOTE, GAN/autoencoder augmentation) and ensemble strategies are effective for imbalanced financial classification, substantially improving minority-class recall and overall detection accuracy. Machine learning has become indispensable for financial applications under challenging conditions. For volatile and unpredictable markets like cryptocurrencies, ensemble deep-learning models currently yield the best price forecasts (Sun, 2020). In volatility forecasting, hybrid ML models reduce error and can adapt quickly to shocks (Y. Sun *et al*., 2025). Handling imbalanced data is crucial in finance: techniques such as weighted oversampling, generative augmentation (GAN/VAE), and ensemble learning have all proven successful in improving minority-class classification (Abedin *et al*., 2022). Future research is focusing on integrating emerging architectures (e.g. Transformers, hypernetworks) and novel data sources (news sentiment, on-chain indicators) to further enhance prediction accuracy and robustness in financial ML models (Kamalov, 2021).

Theoretical Framework: The conceptual foundation of this study is grounded in a confluence of statistical theories, financial market hypotheses, and machine learning paradigms that collectively inform the behavior of extreme price movements in cryptocurrency markets. The application of machine learning to predict rare events such as cryptocurrency price shocks is supported by the increasingly complex, noisy, and non-linear structure of digital financial systems—conditions that often diverge from traditional assumptions of normality and market efficiency.

Extreme Value Theory (EVT): Extreme Value Theory (EVT) offers a statistical basis for modeling tail events that exceed high thresholds, making it particularly suitable for identifying rare but impactful phenomena in financial markets. Originally advanced by Tippett (1928) and later developed through the work of Gumbel (1958), Pickands, Balkema, and de Haan (1970s), EVT culminated in widely used models such as the Generalized Pareto Distribution (GPD). In this study, a percentile-based approach is used to define price shocks—specifically, instances where the absolute return exceeds the 90th percentile of all observed returns—consistent with EVT's threshold exceedance framework (Extreme Value Theory," 1989).

Efficient Market Hypothesis (EMH): According to the Efficient Market Hypothesis (Fama, 1970), asset prices in fully efficient markets reflect all available information, rendering prediction theoretically infeasible(Team, 2024). However, the cryptocurrency market frequently deviates from this ideal due to high volatility, information asymmetry, and regulatory fragmentation. These deviations create exploitable patterns such as autocorrelation and volatility clustering, which this study incorporates through lagged returns and rolling volatility measures. Such empirical irregularities challenge the strong-form EMH and justify the use of predictive learning algorithms in this domain.

Mixture of Distributions Hypothesis (MDH): The Mixture of Distributions Hypothesis (Clark, 1973; Tauchen & Pitts, 1983) posits that observed asset returns arise from a mixture of normal distributions, where the variance of each component distribution is determined by the rate of information arrival(Darolles *et al*., 2017). In this context, trading volume and market capitalization are considered proxies for latent information flow and serve as key variables in modeling return heteroskedasticity. These features are retained in the predictive framework to enhance the model's ability to capture the changing volatility structure associated with extreme events.

Liquidity–Volatility Linkage: The Liquidity–Volatility Linkage Hypothesis provides further justification for including volume and capitalization metrics. As argued by Karpoff (1987) and extended by Chordia, Roll, and Subrahmanyam (2000), a positive relationship exists between trading volume and price volatility, driven by common information flow and heightened investor activity. This theoretical relationship supports the hypothesis that increases in market activity often precede price shocks. Accordingly, the study incorporates both trading volume change and market capitalization as predictive features to reflect this dynamic.

Supervised Learning and Imbalanced Classification: The predictive architecture of this study is grounded in supervised learning, where historical data are used to classify future events. The selected models—Logistic Regression (Hosmer & Lemeshow, 2000), Random Forest (Breiman, 2001), XGBoost (Friedman, 2001), and LightGBM (Ke *et al*., 2017)—are well-suited to handle non-linear interactions, feature heterogeneity, and high-dimensional financial datasets. These algorithms are especially relevant in cryptocurrency markets, which exhibit complex structure and high levels of noise. Given the rarity of price shocks, the classification problem is inherently imbalanced. This imbalance can skew learning algorithms toward the majority class and reduce the ability to detect minority events. To address this issue, the study adopts the Synthetic Minority Over-sampling Technique (SMOTE), developed by Chawla *et al*. (2002), which synthetically generates new samples from the minority class by interpolating between nearest neighbors. SMOTE has proven effective in improving minority class recall and overall robustness in rare-event detection.

METHODOLOGY

This study adopts a supervised machine learning framework to forecast extreme price shocks in the Bitcoin market. The methodological design integrates data preprocessing, binary target construction, financial feature engineering, class imbalance correction using SMOTE, and the training and evaluation of multiple predictive models. All implementation was conducted in Python, using libraries such as pandas, scikit-learn, imbalanced-learn, xgboost, and lightgbm.

Data Collection and Preprocessing: Daily historical data for Bitcoin were collected from the CoinGecko platform (https: //www.coingecko.com/en/coins/bitcoin/historical_data), covering the period from August 1, 2024, to August 1, 2025. The dataset includes market capitalization, trading volume, and closing price in USD, commonly used as indicators of market activity and liquidity. To ensure data integrity, the dataset was arranged chronologically, checked for missing values, and cleaned for anomalies. Daily log returns were computed as

$$R_t = \ln(P_t/P_{t-1})$$

where P_t denotes the closing price on day t. This transformation stabilizes variance and expresses relative price changes in additive form. The log returns were used to construct lag features and rolling statistics for improved signal extraction.

Target Variable Construction: To define the binary classification target, a price shock was identified when the absolute return exceeded the 90th percentile of all absolute returns over the sample period. Formally,

$$\text{Shock}_t = \begin{cases} 1, & \text{if } |R_t| > P_{90}(|R|) \\ 0, & \text{otherwise} \end{cases}$$

where $P_{90}(|R|)$ denotes the 90th percentile of absolute log returns. This threshold-based approach enables systematic detection of extreme market fluctuations without reliance on external event labeling.

Feature Engineering: To enhance the model's predictive capabilities, several engineered variables were introduced. These included lagged returns (R_{t-1} and R_{t-2}) to capture short-term memory effects, moving averages over 7-day and 14-day windows to reflect medium-term momentum, and rolling standard deviations over equivalent windows (Volatility7 and Volatility14) to measure price fluctuation intensity. In addition, the percentage change in daily trading volume was included as a proxy for liquidity shifts. All features were normalized using z-score standardization to ensure numerical stability and accelerate convergence during training.

Handling Class Imbalance: The distribution of the target variable revealed significant class imbalance, with shock events comprising less than 10% of the total observations. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was employed exclusively on the training dataset. SMOTE creates synthetic examples of the minority class by interpolating between existing samples, thereby mitigating the model's tendency to overfit the majority class. A comparative analysis was also conducted by training models on the original imbalanced dataset, allowing for before-and-after resampling evaluation to quantify the improvement in detecting rare events.

Model Development and Evaluation: Four supervised classification algorithms were applied to the final dataset: Logistic Regression, Random Forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM). Logistic Regression served as a baseline linear model, while Random Forest and XGBoost, both tree-based ensemble methods, were employed for their robustness in capturing non-linear interactions. LightGBM was included due to its computational efficiency and proven performance in large-scale classification tasks. The dataset was partitioned into 80% training and 20% testing subsets, and SMOTE was applied only to the training set. Each model was trained independently, and performance was evaluated using accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). Empirical results indicated that LightGBM and Random Forest outperformed other classifiers, both achieving an AUC of 0.919 and demonstrating strong recall for the shock class. XGBoost and Logistic Regression also showed competitive AUC scores, though with relatively lower recall. Feature importance analysis, particularly from XGBoost, identified seven-day volatility, volume change, and lagged returns as the most influential predictors. These findings affirm the relevance of short-term volatility memory and liquidity variation in anticipating sudden market disruptions in the cryptocurrency domain.

RESULTS AND DISCUSSION

This section presents the empirical findings from the application of supervised learning models to predict extreme price shocks in the Bitcoin market. The performance of four classifiers—Logistic Regression, Random Forest, XGBoost, and LightGBM—was evaluated on a feature-engineered dataset comprising daily market indicators such as lagged returns, rolling volatility, moving averages, and volume changes. The classification task was designed to distinguish between normal market activity and shock days, defined as instances where the absolute return exceeded the 90th percentile threshold of the return distribution (see Figure 1).

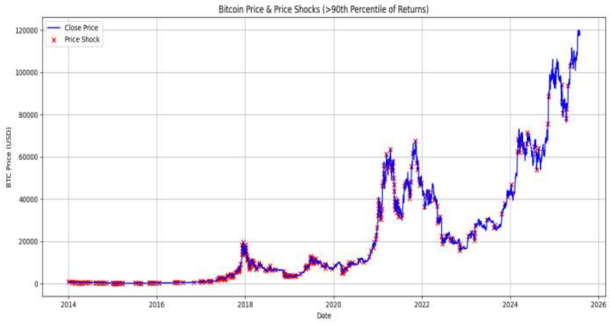

Figure 1. Bitcoin closing prices (USD) from August 2024 to August 2025

Class Imbalance and the Need for Resampling: Initial models trained on the imbalanced dataset (without resampling) exhibited high overall accuracy but performed poorly in detecting shock events. Logistic Regression and Decision Tree models, in particular, achieved high accuracy due to their correct classification of the dominant non-shock class, yet they demonstrated zero recall for the minority class. This limitation reinforced the need for addressing class imbalance in rare-event forecasting. The class distribution before and after SMOTE application is summarized in Table 1, showing a balanced sample post-resampling, which allowed the models to learn more effectively from minority-class patterns.

Table 1. Class Distribution Before and After SMOTE Resampling

Class Label	Before SMOTE (Training Set)	After SMOTE (Training Set)
No Shock (0)	2710	2710
Shock (1)	370	2710
Total	**3080**	**5420**

Model Performance after SMOTE: To improve the models' sensitivity to shocks, the training data were resampled using the Synthetic Minority Over-sampling Technique (SMOTE). After applying SMOTE, all classifiers demonstrated improved ability to detect shocks, as evidenced by significant increases in recall and F1-score for the minority class. Post-resampling classification performance is reported in Table 2, highlighting that Random Forest and LightGBM outperformed the other models in terms of both accuracy and discriminative power. LightGBM achieved the highest Area Under the ROC Curve (AUC = 0.919), followed closely by Random Forest (AUC = 0.919), XGBoost (AUC = 0.906), and Logistic Regression (AUC = 0.901).

Table 2. Classification Metrics for Each Model (After SMOTE Resampling)

Model	Accuracy	Precision (Shock)	Recall (Shock)	F1-Score (Shock)	AUC
Logistic Regression	0.83	0.32	0.81	0.46	0.901
Random Forest	0.90	0.47	0.71	0.57	0.919
XGBoost	0.89	0.46	0.64	0.51	0.906
LightGBM	0.90	0.46	0.73	0.56	0.919

In terms of classification metrics, LightGBM demonstrated strong balance between precision and recall, particularly for the shock class, achieving a recall of 0.73 and an overall accuracy of 90%. XGBoost, although slightly lower in recall, maintained competitive precision and F1-score, highlighting its robustness in modeling noisy, non-linear relationships. Random Forest exhibited similarly high performance, reinforcing the utility of ensemble tree-based methods in financial prediction tasks where signal complexity and volatility clustering are prominent.

ROC Curve Comparison: The ROC curve comparison in Figure 2 further supported the superiority of ensemble methods over the linear baseline. All tree-based models consistently outperformed Logistic Regression across the full range of threshold values, indicating better discriminative ability under varying sensitivity-specificity trade-offs. This visual evidence complements the AUC values reported in **Table 2**, validating the models' robustness. LightGBM and Random Forest showed the highest AUC of 0.919, indicating superior discriminative power.

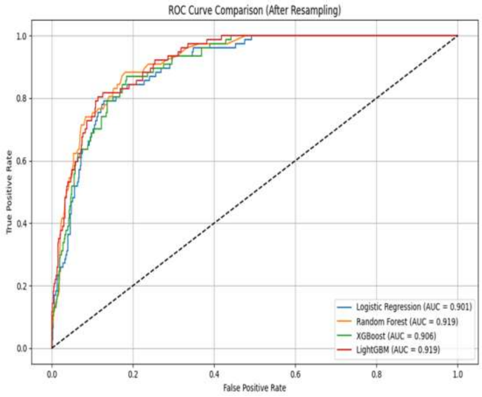

Figure 2. Receiver Operating Characteristic (ROC) curves for Logistic Regression, Random Forest, XGBoost, and LightGBM classifiers after SMOTE resampling

Feature Importance Analysis: Feature importance analysis from XGBoost, visualized in Figure 3, provided further insight into the predictors that contributed most significantly to model performance. The seven-day rolling volatility emerged as the most influential feature, followed by trading volume change, the seven-day moving average of price, and lagged returns. These findings validate the hypothesis that short-term volatility memory, price momentum, and liquidity shifts play critical roles in signalling upcoming extreme market movements. The prominence of volatility-based features also aligns with the Mixture of Distributions Hypothesis and volatility clustering patterns observed in speculative financial markets. Volatility over a 7-day window (Volatility7) was the most influential predictor, followed by volume change and short-term moving averages.

Implications: Overall, the empirical results confirm that the application of resampling techniques such as SMOTE, combined with advanced ensemble models like LightGBM and XGBoost,

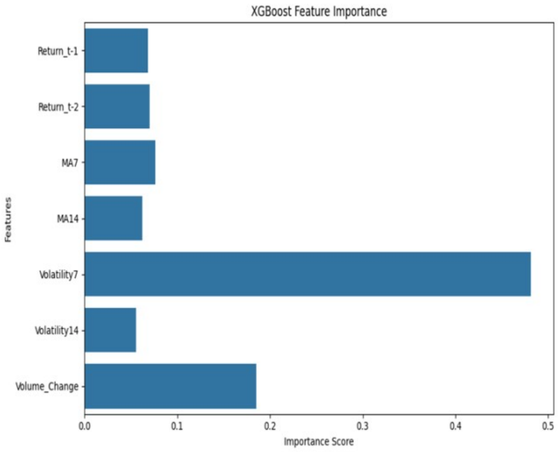

Figure 3. Relative feature importance values from the XGBoost model

significantly enhances the predictive capacity of machine learning systems in rare-event classification problems. These models offer strong potential for early warning systems, risk-sensitive trading algorithms, and volatility forecasting frameworks within the cryptocurrency ecosystem. Future extensions may incorporate deep learning architectures or hybrid models, and explore the integration of exogenous variables such as sentiment indicators or macroeconomic news to further enrich prediction quality.

Operational Use Case and Practical Significance: The predictive framework developed in this study holds substantial operational value for various stakeholders in the cryptocurrency ecosystem, particularly for algorithmic traders, portfolio risk managers, fintech firms, and regulatory bodies. The early identification of extreme price shocks enables more responsive and adaptive decision-making in high-volatility environments, where traditional forecasting tools often fail to detect abrupt market disruptions. From a trading perspective, the proposed ensemble-based classification system can be integrated into real-time risk alert engines, providing early warnings of shock probabilities based on daily input features such as volatility, volume shifts, and lagged returns. This capability is crucial for automated trading systems that must dynamically adjust ass*et al*location, stop-loss parameters, or hedge positions to mitigate exposure to sudden losses. For fintech platforms and crypto exchanges, the framework can serve as the foundation for volatility dashboards or risk-sensitive order routing systems, where liquidity constraints and user trade execution are directly affected by price stability. Additionally, regulators and compliance teams can utilize such a model as part of market surveillance systems to flag anomalous behavior and pre-empt systemic risks triggered by sudden price moves.

Operationally, the use of SMOTE to counteract data imbalance, coupled with efficient ensemble learners like LightGBM and XGBoost, ensures scalability and robustness even in noisy, high-frequency environments. The model's lightweight architecture and reliance on commonly available market indicators make it deployable in cloud-based or edge environments, enabling high-speed inference without the need for excessive computational resources. Importantly, the framework also advances the computational economics literature on rare-event prediction. By combining resampling techniques with ensemble learners, this study demonstrates how data-driven methods can improve the detection of shocks that traditional econometric approaches struggle to capture. These insights are relevant not only for cryptocurrency markets but also for the study of systemic risk in interbank lending, sovereign defaults, and exchange rate crashes. Thus, the model serves as a template for embedding machine learning approaches into economic simulation and forecasting environments where rare, high-impact events are of central concern. In summary, this study bridges methodological innovation with operational feasibility, offering a practical blueprint for deploying predictive analytics in the dynamic and often unpredictable landscape of digital finance.

Conclusion and Future Research Directions

This paper introduced a rare-event prediction framework that integrates imbalance correction with ensemble machine learning, demonstrating its effectiveness in forecasting extreme price shocks in the cryptocurrency market. While Bitcoin provided a suitable testbed due to its high volatility and data availability, the computational architecture extends well beyond digital assets. Rare but impactful events—such as financial crises, exchange rate crashes, sovereign defaults, or systemic liquidity breakdowns—share the same statistical challenges of nonlinearity, volatility clustering, and class imbalance. The framework presented here thus offers a generalizable tool for computational economics, supporting both theoretical exploration and practical applications in risk-sensitive environments. The findings highlight three main contributions. First, they show how resampling strategies (SMOTE) can improve the detection of low-frequency events in economic data. Second, they confirm that ensemble learners outperform linear models when predicting tail outcomes in volatile markets. Third, they bridge computational methods with economic theory (EVT, EMH deviations, MDH, liquidity–volatility linkages), positioning machine learning within a broader theoretical context. For future work, integrating alternative data (e.g., sentiment, news flows, on-chain metrics) and hybrid architectures (deep learning with econometric models) could further enhance rare-event prediction. Moreover, applying the framework to multi-asset settings and systemic simulations would strengthen its role in crisis modeling and policy design. In this sense, the study contributes not only to cryptocurrency research but also to the computational economicsliterature on systemic risk and volatility dynamics, providing a methodological foundation for anticipating high-impact, low-probability events in complex economic systems.

Declarations

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Conflict of Interest: The authors have no relevant financial or non-financial interests to disclose.

Abbreviation	Full Form
SMOTE	Synthetic Minority Over-sampling Technique
EVT	Extreme Value Theory
EMH	Efficient Market Hypothesis
MDH	Mixture of Distributions Hypothesis
GPD	Generalized Pareto Distribution
ML	Machine Learning
AI/ML	Artificial Intelligence / Machine Learning
AUC	Area Under the ROC Curve
ROC	Receiver Operating Characteristic
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
RNN	Recurrent Neural Network
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
ARIMA	Autoregressive Integrated Moving Average
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
WGAN-GP	Wasserstein GAN with Gradient Penalty
BFDS	Balanced Fraud Detection Score
KSDE	Kernel Smoothing Differential Ensemble (from cited work)
MAE	Mean Absolute Error
MSE	Mean Squared Error
USD	United States Dollar
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
LR	Logistic Regression (used contextually)
RF	Random Forest
F1-Score	Harmonic mean of precision and recall
JEL	Journal of Economic Literature (classification codes)

REFERENCES

1. A comparative analysis of statistical and machine learning models for outlier detection in Bitcoin limit order books. (n.d.-a). https://arxiv.org/html/2507.14960v1#:~:text=undermine%20market%20integrity%20,to%20detect%20and%20mitigate%20market

2. A comparative analysis of statistical and machine learning models for outlier detection in Bitcoin limit order books. (n.d.-b). https://arxiv.org/html/2507.14960v1#:~:text=The%20cryptocurrency%20market%20is%20characterised,both%20retail%20and%20institutional%20investors
3. Abedin, M. Z., Guotai, C., Hajek, P., & Zhang, T. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. Complex & Intelligent Systems, 9(4), 3559–3579. https://doi.org/10.1007/s40747-021-00614-4
4. Baur, D. G., & Dimpfl, T. (2021a). The volatility of Bitcoin and its role as a medium of exchange and a store of value. Empirical Economics, 61(5), 2663–2683. https://doi.org/10.1007/s00181-020-01990-5
5. Baur, D. G., & Dimpfl, T. (2021b). The volatility of Bitcoin and its role as a medium of exchange and a store of value. Empirical Economics, 61(5), 2663–2683. https://doi.org/10.1007/s00181-020-01990-5
6. Bouteska, A. &. A. M. Z. &. H. P. &. Y. K. (2024). Cryptocurrency price forecasting – A comparative analysis of ensemble learning and deep learning methods. ideas.repec.org. https://ideas.repec.org/a/eee/finana/v92y2024ics1057521923005719.html#:~:text=and%20deep%20learning%2C%20as%20the,investors%20in%20the%20cryptocurrency%20markets
7. Brauneis, A., & Sahiner, M. (2024). Crypto Volatility Forecasting: Mounting a HAR, sentiment, and machine learning horserace. Asia-Pacific Financial Markets. https://doi.org/10.1007/s10690-024-09510-6
8. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324
9. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
10. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
11. Chen, Z. (2025). From disruption to integration: cryptocurrency prices, financial fluctuations, and macroeconomy. Journal of Risk and Financial Management, 18(7), 360. https://doi.org/10.3390/jrfm18070360
12. Chordia, T., Roll, R., & Subrahmanyam, A. (2000). Commonality in liquidity. Journal of Financial Economics, 56(1), 3–28. https://doi.org/10.1016/s0304-405x(99)00057-4
13. Clark, P. K. (1973). A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. Econometrica, 41(1), 135. https://doi.org/10.2307/1913889
14. Darolles, S., Fol, G. L., & Mero, G. (2017). Mixture of distribution hypothesis: Analyzing daily liquidity frictions and information flows. Journal of Econometrics, 201(2), 367–383. https://doi.org/10.1016/j.jeconom.2017.08.014
15. Extreme value theory. (1989). In Lecture notes in statistics. https://doi.org/10.1007/978-1-4612-3634-4
16. Fama, E. F. (1970). Efficient Capital Markets: A review of theory and Empirical work. The Journal of Finance, 25(2), 383. https://doi.org/10.2307/2325486
17. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5). https://doi.org/10.1214/aos/1013203451
18. Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. https://doi.org/10.1002/0471722146
19. John, D. L., Binnewies, S., & Stantic, B. (2024). Cryptocurrency Price Prediction Algorithms: A Survey and Future Directions. Forecasting, 6(3), 637–671. https://doi.org/10.3390/forecast6030034
20. Kamalov, I. G. &. F. (2021). Predicting bitcoin price movements using sentiment analysis: a machine learning approach. ideas.repec.org. https://ideas.repec.org/a/eme/sefpps/sef-07-2021-0293.html#:~:text=more%20accurate%20in%20predicting%20BTC,specific%20are%20prized%20factors%20in
21. Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. Journal of Financial and Quantitative Analysis, 22(1), 109. https://doi.org/10.2307/2330874
22. Ke, G., 1, Meng, Q., 2, Finley, T., 3, Wang, T., 1, Chen, W., 1, Ma, W., 1, Ye, Q., 1, Microsoft Research, Peking University, & Microsoft Redmond. (2017). LightGBM: a highly efficient gradient boosting decision tree. In 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA [Conference-proceeding].
23. Mansilla-Lopez, J., Mauricio, D., & Narváez, A. (2025). Factors, forecasts, and simulations of volatility in the stock market using machine learning. Journal of Risk and Financial Management, 18(5), 227. https://doi.org/10.3390/jrfm18050227
24. Pickands, J., III. (1975). Statistical inference using extreme order statistics. The Annals of Statistics, 3(1). https://doi.org/10.1214/aos/1176343003
25. Qiu, Z., Kownatzki, C., Scalzo, F., & Cha, E. S. (2025). Historical Perspectives in Volatility Forecasting Methods with Machine Learning. Risks, 13(5), 98. https://doi.org/10.3390/risks13050098
26. Sun, X. &. L. M. &. S. Z. (2020). A novel cryptocurrency price trend forecasting model based on LightGBM. ideas.repec.org. https://ideas.repec.org/a/eee/finlet/v32y2020ics1544612318307918.html#:~:text=Forecasting%20cryptocurrency%20prices%20is%20crucial,can%20effectively%20guide%20investors%20in
27. Sun, Y., Qu, Z., Zhang, T., & Li, X. (2025). Adaptive Ensemble Learning for Financial Time-Series Forecasting: A Hypernetwork-Enhanced Reservoir Computing Framework with Multi-Scale Temporal Modeling. MDPI. https://doi.org/10.3390/axioms14080597
28. Team, C. (2024, May 22). Efficient Markets Hypothesis. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/efficient-markets-hypothesis/
29. Team, I. (2024, June 15). Cryptocurrency explained with pros and cons for investment. Investopedia. https://www.investopedia.com/terms/c/cryptocurrency.asp#:~:text=industries%2C%20including%20finance%20and%20law,and%20use%20in%20criminal%20activities
30. Zhang, J., Cai, K., & Wen, J. (2023). A survey of deep learning applications in cryptocurrency. iScience, 27(1), 108509. https://doi.org/10.1016/j.isci.2023.108509
31. Zhou, Y., Xie, C., Wang, G., Gong, J., & Zhu, Y. (2025). Forecasting cryptocurrency volatility: a novel framework based on the evolving multiscale graph neural network. Financial Innovation, 11(1). https://doi.org/10.1186/s40854-025-00768-x

****