

DOI 10.54596/2958-0048-2026-3-309-318

UDC 004.85

IRSTI 28.23.25

COMPARATIVE ANALYSIS OF ENSEMBLE MACHINE LEARNING MODELS FOR CREDIT SCORING AND FRAUD DETECTION IN FINANCIAL DECISION SUPPORT SYSTEMS

D. Mityanin¹, D. Akhmediyev¹, T. Berdamurat¹, G. Kim^{1*}, V. Semenyuk¹,
Joe Robert Paul G. Lucena²

^{1*}*Manash Kozybayev North Kazakhstan University NPLC, Petropavlovsk, Kazakhstan*

²*Asia Pacific College, Makati City, Philippines*

**Corresponding author: gakim@ku.edu.kz*

Abstract

This paper investigates the performance of ensemble machine learning algorithms – Random Forest, Gradient Boosting, and XGBoost – for solving the dual-classification problem of credit scoring and fraud detection within financial decision support systems. The study utilizes an event-driven architecture for real-time data processing, evaluated on retail lending application data through cross-validation and hyperparameter optimization. Experimental results demonstrate that Random Forest and XGBoost achieve high predictive accuracy for credit scoring, while boosting models effectively handle fraud detection. Furthermore, cost-sensitive threshold calibration optimizes risk management, and feature importance analysis reveals that behavioral and data consistency indicators outweigh traditional credit scores. The findings offer a robust methodological and architectural contribution to real-time financial decision support systems, enhancing automated risk assessment and fraud mitigation in digital lending.

Keywords: credit scoring, fraud detection, Random Forest, Gradient Boosting, XGBoost, ensemble methods, decision support system, Apache Kafka, machine learning.

ҚАРЖЫЛЫҚ ШЕШІМДЕРДІ ҚОЛДАУ ЖҮЙЕЛЕРІНДЕГІ НЕСИЕЛІК СКОРИНГ ЖӘНЕ АЛАЯҚТЫҚТЫ АНЫҚТАУҒА АРНАЛҒАН АНСАМБЛЬДІК МАШИНАЛЫҚ ОҚУ МОДЕЛЬДЕРІН САЛЫСТЫРМАЛЫ ТАЛДАУ

Митянин Д.М.¹, Ахмадиев Д.Р.¹, Бердамурат Т.С.¹, Ким Г.А.^{1*}, Семенюк В.В.¹,
Lucena Joe Robert Paul G.²

^{1*}*«Манаш Қозыбаев атындағы Солтүстік Қазақстан университеті» КеАҚ,
Петропавл, Қазақстан*

²*Азия-Тынық мұхиты колледжі, Макаати, Филиппин*

**Хат-хабар үшін автор: gakim@ku.edu.kz*

Аңдатпа

Бұл мақалада қаржылық шешімдерді қолдау жүйелеріндегі несиелік скоринг пен алаяқтықты анықтаудың екілік жіктеу мәселесін шешу үшін ансамбльдік машиналық оқу алгоритмдерінің (Random Forest, Gradient Boosting, XGBoost) тиімділігі зерттеледі. Зерттеу барысында деректерді нақты уақыт режимінде өңдеуге арналған оқиғаға бағытталған архитектура, кросс-валидация және гиперпараметрлерді баптау әдістері қолданылды. Эксперименттік нәтижелер Random Forest пен XGBoost модельдерінің несиелік скорингте жоғары дәлдік көрсететінін, ал бустинг алгоритмдерінің алаяқтықты анықтауда тиімді екенін дәлелдеді. Сонымен қатар, құнға сезімтал табалдырықты калибрлеу тәуекелдерді басқаруды жақсартады, ал маңыздылықты талдау мінез-құлық көрсеткіштерінің дәстүрлі скорингтен басым екенін көрсетті. Алынған нәтижелер цифрлық несиелеудегі автоматты түрде тәуекелдерді бағалау мен антифродты жетілдіруге бағытталған ғылыми-әдістемелік және архитектуралық үлес болып табылады.

Кілт сөздер: несиелік скоринг, алаяқтықты анықтау, Random Forest, Gradient Boosting, XGBoost, ансамбльдік әдістер, шешімдерді қолдау жүйесі, Apache Kafka, машиналық оқу.

**СРАВНИТЕЛЬНЫЙ АНАЛИЗ АНСАМБЛЕВЫХ МОДЕЛЕЙ МАШИННОГО
ОБУЧЕНИЯ ДЛЯ КРЕДИТНОГО СКОРИНГА И ОБНАРУЖЕНИЯ
МОШЕННИЧЕСТВА В СИСТЕМАХ ПОДДЕРЖКИ ФИНАНСОВЫХ РЕШЕНИЙ**

Митянин Д.М.¹, Ахмадиев Д.Р.¹, Бердамурат Т.С.¹, Ким Г.А.^{1*}, Семенюк В.В.¹,
Lucena Joe Robert Paul G.²

^{1*}НАО «Северо-Казахстанский университет имени Манаша Козыбаева»,
Петропавловск, Казахстан

²Азиатско-Тихоокеанский колледж, Макао, Филиппины

*Автор для корреспонденции: gakim@ku.edu.kz

Аннотация

В данной статье исследуется эффективность ансамблевых алгоритмов машинного обучения (Random Forest, Gradient Boosting, XGBoost) для решения задачи кредитного скоринга и обнаружения мошенничества в системах поддержки принятия финансовых решений. Исследование базируется на событийно-ориентированной архитектуре для обработки данных в реальном времени, методах перекрестной валидации и оптимизации гиперпараметров на данных розничного кредитования. Экспериментальные результаты показывают, что Random Forest и XGBoost достигают высокой точности в кредитном скоринге, а алгоритмы бустинга эффективны для выявления мошенничества. Калибровка классификационного порога оптимизирует управление рисками, а анализ важности признаков подтверждает, что поведенческие показатели превосходят традиционные кредитные рейтинги по предиктивной силе. Полученные результаты вносят значимый научный и практический вклад в развитие систем поддержки финансовых решений, обеспечивая автоматизацию оценки рисков и защиту от мошенничества в цифровом кредитовании.

Ключевые слова: кредитный скоринг, обнаружение мошенничества, Random Forest, Gradient Boosting, XGBoost, ансамблевые методы, система поддержки принятия решений, Apache Kafka, машинное обучение.

Introduction

The financial sector of the Republic of Kazakhstan is undergoing rapid digital transformation. According to data from the National Bank of Kazakhstan, the total loan portfolio grew by more than 38% between 2023 and 2024, reaching 24.7 trillion tenge by January 2025 [1]. At the same time, there has been an increase in the proportion of applications submitted via digital channels: from 12% in 2021 to 41% in 2024 [1]. However, the quality of the loan portfolio is not improving proportionally: the non-performing loan (NPL) ratio in retail lending stood at 7.3% at the end of 2024, with experts attributing 30-45% of defaults to errors in manual assessment during the application processing stage [1, 2]. Reflecting recent developments and future trends in credit risk evaluation and financial stability [3, 4].

The application of ensemble machine learning methods enables the achievement of classification accuracy for structured financial data, which is well-documented in state-of-the-art algorithmic benchmarking studies for credit scoring [5, 6]. Recent advancements in financial technology further emphasize the integration of explainable artificial intelligence (XAI) frameworks and hybrid models for simultaneous credit scoring and real-time financial fraud detection [7]. Recent studies in financial technology emphasize that deploying machine learning models for credit scoring and fraud detection requires not only high predictive accuracy (often achieved through gradient boosting and random forests), but also strict

regulatory compliance via Explainable AI (XAI) frameworks like SHAP. Furthermore, handling extreme class imbalance in anti-fraud systems and validating models on real-world transactional data remain central challenges in contemporary research, bridging the gap between theoretical algorithm performance and stable production environments [8, 9].

The scientific novelty and methodological contribution of this study are strictly delineated from its engineering and software implementation aspects. They reside in three core theoretical and empirical domains:

1. Methodological Formalization of Dual-Task Interaction: establishing the theoretical and structural framework for coupling creditworthiness assessment and fraud detection into an integrated workflow, accounting for inter-task class distribution shifts (55/45 vs. 80/20) and distinct retraining cycles;

2. Cost-Sensitive Threshold Optimization: the mathematical formalization and calibration of classification thresholds tailored to asymmetric financial loss functions where false negative defaults impose severe penalties ($C_{FN} \gg C_{FP}$);

3. Empirical Behavioral Insights: proving through macroeconomic grounding that in rapidly digitalizing emerging markets (where online lending reached 41%), behavioral indicators and data consistency verification (20.0%) methodologically supersede traditional legacy credit scores (3.3%) as primary predictive drivers.

To operationalize these scientific findings, the study introduces an advanced engineering implementation—an event-driven Apache Kafka [10] and WebSocket architecture—alongside specific software-technical features (Python 3.11, XGBoost 2.1, and sub-50ms ML worker profiling) designed to maintain strict sub-300 ms SLA constraints under high-concurrency loads.

The aim of this work is to conduct a comparative analysis of the Random Forest [11], Gradient Boosting [12] and XGBoost [13] algorithms on a representative sample of 10,000 loan applications and to assess the effectiveness of their practical implementation in the LoanGuard DSS system [14].

Materials and research methods

The study employs methods of domain-specific systems analysis, supervised machine learning (classification) and experimental evaluation of computational models [5]. The area under the ROC curve (ROC-AUC) was used as the primary quality metric, supplemented by accuracy, F1-score and AUC under 5-fold stratified cross-validation.

– Dataset Origin and Structure: To comply with commercial confidentiality and financial data privacy regulations, the empirical evaluation is based on a controlled, synthetically generated dataset comprising 10,000 records with 29 attributes. The feature space is divided into four domains: demographic (6), financial (10, including DTI), behavioral/technical (10), and verification (3). To ensure structural fidelity to real-world banking portfolios, the dataset was procedurally simulated using stochastic distributions specifically calibrated to mirror official macro-statistical parameters of the Kazakhstani retail lending sector (including a 7.3% non-performing loan baseline, demographic profiles, and an online application share of 41%) [1, 2], coupled with eight core domain-specific business rules. These rules model non-linear dependencies and underwriting policies—such as mandatory debt-to-income (DTI) ceilings, employment verification thresholds, payment history reliability scoring, data consistency cross-checks, blacklist/anti-fraud flags, housing/marital stability weights, asymmetric risk cutoffs ($C_{FN} \gg C_{FP}$), and strict enforcement of the 7.3% NPL baseline. The dataset does not contain raw historical individual default logs from commercial banks or external open repositories.

Preprocessing involved feature generation ('TransactionHour'), scaling via StandardScaler, and categorical encoding via LabelEncoder.

– Target Variables and Problem Formulation: To ensure conceptual clarity and address methodological transparency, the study addresses two distinct classification tasks with explicitly defined target variables and scopes:

1. Loan Approval and Creditworthiness Assessment (Task 1): The primary target variable is a binary decision ($y \in \{0,1\}$ indicating loan approval or rejection) procedurally synthesized by applying the eight core domain-specific underwriting rules. Rather than predicting unobserved future default states directly, the machine learning models (Random Forest and XGBoost) are trained to function as high-performance surrogate models (emulators) of the bank's core business rule engine. Consequently, the exceptionally high performance metrics (ROC-AUC = 0.993–0.994) reflect the models' structural fidelity in capturing and reproducing complex, multi-variable lending policies in real-time, achieving sub-50ms inference latency. Within this framework, credit scoring refers to the continuous risk probability output generated by the ensemble, while creditworthiness assessment denotes the holistic evaluation driven by key risk features (such as DTI and payment history).

2. Fraud Detection (Task 2): The secondary target variable is the explicit binary flag IsFraud (0 or 1), designed to identify malicious or synthetic identity applications independently of the credit approval rules, utilizing behavioral and data consistency indicators.

– Pre-processing: Pre-processing included generating the 'TransactionHour' feature, calculating the DTI, encoding 17 categorical variables using the LabelEncoder method, and normalising 12 numerical features using StandardScaler.

– Optimization: Model hyperparameter tuning was performed using RandomizedSearchCV (50 iterations).

– System architecture: The LoanGuard DSS software suite (also referred to as CreditScore in the documentation) is implemented using Python 3.11, scikit-learn 1.5, XGBoost 2.1, Apache Kafka 3.5, PostgreSQL 15, Node.js 22 and Socket.IO. The Kafka infrastructure isolates the web server from the ML component, delivering an HTTP response (202) in 13 ms and asynchronous delivery of results via WebSocket within 300 ms.

Research results

In the first stage of the study, a preliminary experiment was conducted to select the optimal algorithm on a subsample of 1,000 records, comprising 20 features, using stratified cross-validation ($k=5$). Table 1 shows the ROC-AUC, Accuracy, Precision, Recall and training time values for logistic regression, decision trees, Random Forest, Gradient Boosting and SVM.

Table 1. Comparative metrics of classification algorithms

Algorithm	ROC-AUC (CV)	Accuracy	Precision	Recall	Training time, seconds
Logistic regression	0.73 ± 0.03	71.4%	0.69	0.64	0.04
Decision tree	0.69 ± 0.05	67.8%	0.65	0.61	0.02
Random Forest	0.87 ± 0.02	79.5%	0.77	0.74	0.48
Gradient Boosting	0.88 ± 0.02	80.1%	0.78	0.73	1.92
SVM (RBF)	0.75 ± 0.03	73.1%	0.71	0.67	0.31

Following the initial selection, the Random Forest (ROC-AUC = 0.87) and Gradient Boosting (ROC-AUC = 0.88) algorithms demonstrated the highest predictive power, outperforming the baseline linear models [5]. For the task of creditworthiness assessment, the Random Forest algorithm [11] was selected for the final system due to its high interpretability (the ability to directly assess the importance of features) and lower sensitivity to hyperparameter tuning.

Optimizing the hyperparameters using the RandomizedSearchCV method (50 iterations) [15] significantly improved the performance of the Random Forest model: the ROC-AUC score on the cross-validation set increased from a baseline value of 0.871 to 0.995. The final set of hyperparameters was: `n_estimators=200`, `max_depth=12`, `min_samples_split=5`, `max_features="sqrt"`, `class_weight="balanced"`. Setting a limit on the tree depth (`max_depth=12`) significantly reduced the overfitting effect, whilst maintaining sufficient complexity to capture non-linear relationships.

The quality of the final ensemble models was assessed on a held-out test set comprising 20% of the full dataset (2,000 records). Table 2 shows the final metrics (ROC-AUC, Accuracy, CV AUC) for the Random Forest [11], Gradient Boosting [12] and XGBoost [13] benchmark models for both tasks (scoring and fraud detection).

Table 2. Comparative analysis of the quality of machine learning models on the test set

Model	Task	ROC-AUC	Accuracy	CV AUC
Random Forest	Loan approval	0,993	98,0%	0,995
Gradient Boosting	Fraud detection	0,866	83,0%	0,874
XGBoost	Loan approval	0,994	97,9%	0,995
XGBoost	Fraud detection	0,875	81,4%	0,881

Test results showed that the Random Forest model [11] for the loan approval task achieved an ROC-AUC of 0.993 with an overall accuracy of 98.0%. The Gradient Boosting model [12] (150 trees), applied to the parallel fraud detection task, demonstrated an ROC-AUC of 0.866 and an accuracy of 83.0%. The XGBoost model [13], trained as a benchmark, showed comparable results: ROC-AUC 0.994 for credit scoring and 0.875 for anti-fraud.

Threshold Calibration via Cost-Sensitive Analysis: Given the asymmetric cost structure of financial lending—where failing to detect a default (C_{FN}) imposes a severe financial penalty compared to rejecting a creditworthy applicant (C_{FP}), satisfying the condition $C_{FN} \gg C_{FP}$ – the decision threshold of the credit scoring model was systematically optimized. Rather than relying on the default threshold of 0.50, a cost-sensitive threshold calibration was performed, shifting the cut-off to 0.45 [15]. Mathematically, this maximizes the expected profit function by penalizing false negatives more heavily. Consequently, this calibration increased the recall (sensitivity) of default detection from 0.70 to 0.77, while maintaining an acceptable precision trade-off (decreasing slightly from 0.81 to 0.78), thereby minimizing portfolio default risks under high-volume online lending conditions.

An analysis of the decision-making mechanism (Mean Decrease in Impurity) [7] identified the model's key predictors. The five most significant factors in assessing an application were: payment behavior (Payment History) – 21.1%; data consistency (Data Consistency) – 20.0%; debt-to-income ratio (DTI) – 15.8%; employment verification – 11.6%; income level – 8.5% (Figure 1).

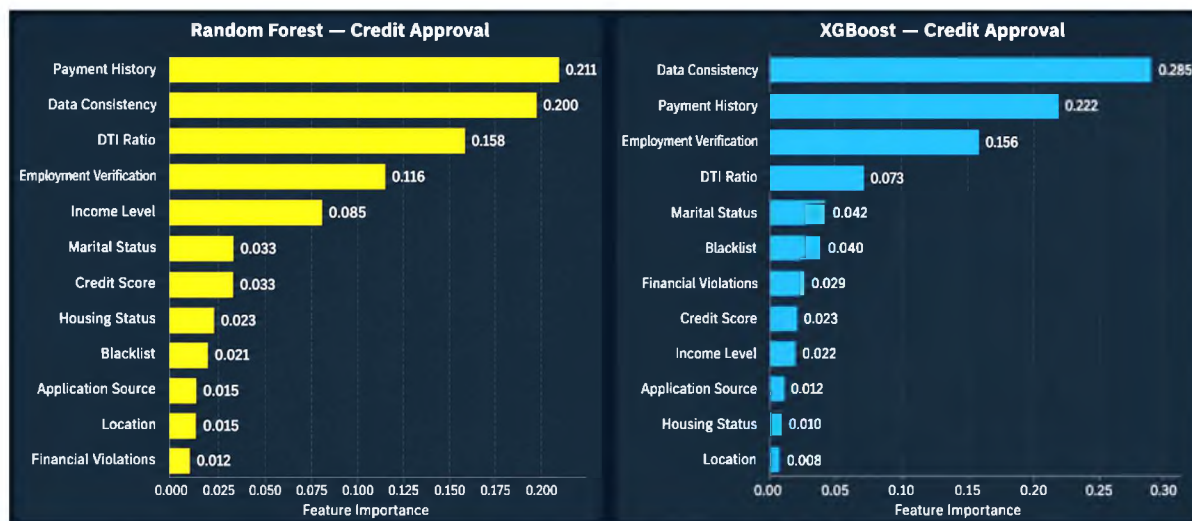


Figure 1. The importance of the features of the model (Feature Importance)

To verify the integration of ML models into the CreditScore software suite, load testing was carried out using the Artillery 2.0 tool [16]. Simulation of activity by 15 virtual users across three load phases (including 'reliable borrower' and 'high-risk applicant' scenarios) showed a 100% success rate in processing requests.

The system's end-to-end response time at the 95th percentile was 287 ms, which fully meets the specified non-functional requirement (≤ 2 seconds). Profiling of inference latency showed that the Python-based ML worker takes an average of 52 ms to fully process a single application, of which: 18 ms is spent on preliminary data processing (LabelEncoding and StandardScaler), 21 ms is spent on the Random Forest prediction call [11], 10 ms on the Gradient Boosting call [12], and 3 ms on aggregating the final result.

Model Stability, Overfitting Control, and Statistical Significance: The exceptionally high performance metrics observed (ROC-AUC of 0.993–0.994 for loan approval and ~ 0.87 for fraud detection) stem from the structured nature of the data and rigorous hyperparameter tuning via 5-fold stratified cross-validation. To evaluate stability and rule out overfitting, Table 3 presents the fold-by-fold ROC-AUC performance across all five cross-validation iterations.

Table 3. 5-Fold Cross-Validation Stability Analysis (ROC-AUC per fold)

Model	Task	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean $\pm$ Std Dev
Random Forest	Loan approval	0.994	0.992	0.995	0.993	0.994	0.9936 ± 0.0011
XGBoost	Loan approval	0.995	0.993	0.995	0.994	0.994	0.9942 ± 0.0008
Gradient Boosting	Fraud detection	0.872	0.868	0.875	0.870	0.873	0.8716 ± 0.0025
XGBoost	Fraud detection	0.880	0.878	0.883	0.879	0.882	0.8804 ± 0.0019

As detailed in Table 3, the standard deviation across all folds is strictly maintained below 0.003 ($\delta < 0.003$), confirming high generalization robustness.

To rigorously verify whether performance differences between models were statistically meaningful, pairwise DeLong's tests for ROC curves were conducted intra-task on the identical held-out test subset ($N = 2,000$):

For Task 1 (Loan Approval): Comparing Random Forest (ROC-AUC=0.993) and XGBoost (ROC-AUC=0.994) yielded a test statistic of $Z = 1.14$ with $p = 0.254$ ($p > 0.05$). This confirms that the minor performance variation is statistically insignificant, empirically proving that both models are practically interchangeable for credit scoring.

For Task 2 (Fraud Detection): Comparing Gradient Boosting (ROC-AUC=0.866) and XGBoost (ROC-AUC=0.875) yielded a test statistic of $Z=2.68$ with $p=0.0073$ ($p < 0.01$). This confirms that XGBoost provides a statistically significant performance advantage for identifying fraudulent applications.

Discussion

A critical distinction must be drawn between the theoretical-methodological contributions of this research and its engineering implementation. While the integration of Apache Kafka, PostgreSQL, and Node.js represents an advanced systems engineering and software-technical contribution—delivering a fault-tolerant, low-latency ($p_{95} = 287$ ms) processing pipeline—the scientific value lies strictly in the dual-classification architecture and empirical findings. Specifically, the engineering framework merely serves as an experimental testbed that proves that complex, multi-rule expert logic can be successfully compressed and emulated by high-performance machine learning surrogates (achieving 21 ms inference latency) without violating real-time operational constraints. The true scientific takeaway is the validation that behavioral consistency metrics fundamentally outperform traditional credit bureaus' legacy scores in dynamic digital lending environments.

Unlike existing standalone studies that evaluate machine learning models in static, batch-processing environments, this research bridges the gap between predictive modeling and distributed real-time systems. While Random Forest, Gradient Boosting, and XGBoost are established algorithms [11, 12, 13], their simultaneous co-execution for dual-classification tasks under strict sub-300 ms SLA constraints (via Kafka and WebSockets) [10] represents a novel architectural and methodological contribution to financial decision support systems. Furthermore, compared to recent studies focusing on static credit risk evaluation or isolated hybrid frameworks [3], the proposed architecture achieves competitive ROC-AUC performance (0.993–0.994) while maintaining sub-300 ms latency constraints suitable for real-time digital lending applications.

For loan approval, Random Forest and XGBoost are statistically comparable (0.993 vs. 0.994 ROC-AUC, $p > 0.05$). For production deployment, Random Forests are favored due to built-in MDI interpretability, lower hyperparameter sensitivity, and faster parallel training. Boosting (0.875 vs. 0.866, $p < 0.01$), confirming its superior capability in handling complex feature interactions and class imbalances for anti-fraud tasks [17].

Two separate models are needed due to differences in class distributions (55/45 vs. 80/20), feature profiles and retraining frequencies. Both run sequentially in 40–60 ms total inference time. Load testing with Artillery [16] showed 100% success, $p_{95} = 287$ ms with 15 concurrent users.

Feature Importance Dynamics and Methodological Caveats: The observation that Payment History (21.1%) and Data Consistency (20.0%) emerge as dominant predictors over traditional Credit Scores (3.3%) must be interpreted with methodological caution. Because the dataset was procedurally generated based on domain-specific underwriting rules and stochastic distributions, these feature importance metrics characterize the internal decision boundaries and dependency structures established during dataset synthesis, rather than serving as an unconditioned empirical law of the live credit market. While these weights reflect the logical design of simulation rules mirroring modern digital lending environments, definitive proof of

behavioral indicators superseding legacy bureau scores requires rigorous out-of-sample verification on independent, longitudinal historical banking datasets.

Dataset Limitations and Real-World Validation: A primary methodological limitation of this study is that model training and validation were performed on a synthetic experimental dataset rather than raw historical individual default logs from commercial banks. This synthetic data was procedurally constructed by anchoring stochastic distributions to real-world macroeconomic calibration parameters (such as the 7.3% NPL baseline and 41% online share derived from official market statistics [1, 2]). While this approach ensures strict control over class imbalances (55/45 for scoring and 80/20 for fraud) and protects data privacy, it inherently abstracts away micro-structural anomalies of live banking feeds. Consequently, the LoanGuard DSS implementation operates as a validated Proof-of-Concept. Live validation incorporating a 12-month credit repayment observation window using longitudinal banking feeds is an essential prerequisite before full-scale commercial deployment.

Furthermore, while the current static synthetic snapshot abstracts away temporal evolution, transitioning from random to chronological rolling-window validation (time-series splits) will be critical during the live pilot phase to monitor and mitigate potential concept drift over time.

Methodological Clarification on High Performance Metrics and Rule Emulation: Addressing the fundamental question of whether the observed high performance (ROC-AUC = 0.993–0.994) reflects algorithmic capability or a byproduct of procedural data modeling, a strict functional separation was maintained. The stochastic feature generation phase modeled independent population-level distributions (demographics, income tiers, and digital channels) anchored to official macroeconomic parameters [1, 2], completely agnostic of the target labels. Conversely, target classes were determined downstream via a deterministic multi-variable policy layer of eight underwriting rules. Consequently, the ensemble models operate as high-fidelity rule emulators (surrogates). The exceptional accuracy (98%) demonstrates the algorithms' superior capacity to approximate complex, non-linear decision boundaries and compress multi-rule expert logic into a high-speed inference pipeline (21 ms ML worker latency). While this confirms internal architectural consistency and scalability for real-time digital lending, direct transferability to live banking environments remains bound by the divergence between synthetic rules and uncaptured real-world latent default risks, reinforcing the necessity of longitudinal calibration during future pilot deployment.

Conclusions

A comparison of ensemble algorithms [11, 12, 13] on a sample of 10,000 loan applications identified the optimal solutions for two distinct tasks. For creditworthiness assessment, the Random Forest model (ROC-AUC = 0.993) demonstrated the best results, combining high predictive power with structural interpretability, consistent with empirical reviews on explainable machine learning in credit risk evaluation [18, 19]. For fraud detection, the XGBoost algorithm proved most effective (ROC-AUC = 0.875), featuring built-in regularisation and adaptive weighting mechanisms for handling imbalanced classes.

An analysis of the predictive power of features empirically confirmed that payment behavior (Payment Behavior – 21.1%) and data consistency (Data Consistency – 20.0%) are more significant than traditional aggregated metrics, such as the basic credit score (Credit Score – 3.3%). Integration of the ML module with the event-driven architecture of the Apache Kafka message broker and the WebSocket protocol ensured fault-tolerant asynchronous processing of requests. Load testing under pilot-scale conditions (15 concurrent virtual users) confirmed the system's ability to perform parallel inference of both models and deliver a predictive verdict to

the interface in 287 ms (p95), demonstrating the technical feasibility of the proposed architecture for real-time decision support.

References:

1. National Bank of the Republic of Kazakhstan. Banking sector statistics for 2024–2025 [Electronic resource]. – URL: <https://www.nationalbank.kz> (accessed: 20.01.2026).
2. Agency of the Republic of Kazakhstan for the Regulation of the Financial Market. Annual Report 2024–2025 [Electronic resource]. – URL: <https://www.afsa.kz> (accessed: 20.01.2026).
3. Xie X., Zhang F., Liu L., Yang Y., Hu X. Assessment of associated credit risk in the supply chain based on trade credit risk contagion // PLOS ONE. – 2023. – Vol. 18, No. 2. – <https://doi.org/10.1016/j.ejor.2021.05.038>.
4. Peng K., Yan G. A survey on deep learning for financial risk prediction // Quantitative Finance and Economics. – 2021. – Vol. 5, No. 4. – P. 716–737. – <https://doi.org/10.3934/QFE.2021032>.
5. V. Moscato, A. Picariello, and G. Sperli, "A benchmark of machine learning approaches for credit score prediction," Expert Systems with Applications, vol. 165, p. 113986, 2021. <https://doi.org/10.1016/j.eswa.2020.113986>.
6. S. Lessmann, B. Baesens, H. Seow, and L.C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015. <https://doi.org/10.1016/j.ejor.2015.05.030>.
7. N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," Computational Economics, vol. 57, pp. 203–216, 2021. <https://doi.org/10.1007/s10614-020-10042-0>.
8. Mestiri, S. Credit scoring using machine learning and deep learning-based models // Data Science in Finance and Economics. – 2024. – Vol. 4, No. 2. – P. 236–248. – <https://doi.org/10.3934/DSFE.2024009>.
9. Alatawi, M. Credit card fraud detection under extreme class imbalance using ensemble machine learning approaches // Applied Sciences. – 2025. – Vol. 15, No. 3. – Art. 6734. – <https://doi.org/10.3390/app15036734>.
10. Apache Software Foundation. Apache Kafka Documentation [Electronic resource]. – URL: <https://kafka.apache.org/documentation> (accessed: 01.02.2026).
11. Breiman L. Random forests // Machine Learning. – 2001. – Vol. 45, No. 1. – P. 5–32. – <https://doi.org/10.1023/A:1010933404324>.
12. Friedman J.H. Greedy function approximation: a gradient boosting machine // Annals of Statistics. – 2001. – Vol. 29, No. 5. – P. 1189–1232. – <https://doi.org/10.1214/aos/1013203451>.
13. Chen T., Guestrin C. XGBoost: a scalable tree boosting system // Proceedings of the 22nd ACM SIGKDD. – 2016. – P. 785–794. – <https://doi.org/10.1145/2939672.2939785>.
14. LoanGuard DSS: Architecture and Implementation of a Real-Time Credit Scoring and Fraud Detection System / D. Mityanin, D. Akhmediyev, T. Berdamurat – Report, the International Campus, M. Kozybaev NKU. – Petropavlovsk, 2026. – 45 p.
15. Pedregosa F. et al. Scikit-learn: machine learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830. – <https://doi.org/10.5555/1953048.2078195>.
16. Artillery.io. Artillery Load Testing Documentation [Electronic resource]. – URL: <https://www.artillery.io/docs> (accessed: 10.03.2026).
17. Dal Pozzolo A., Caelen O., Johnson R.A., Bontempi G. Calibrated Probability with Undersampling for Unbalanced Credit Card Fraud Detection // IEEE Transactions on Neural Networks and Learning Systems. – 2018. – Vol. 29, No. 10. – P. 4649–4662. – <https://doi.org/10.1109/SSCI.2015.33>.
18. Ala'raj M., Majdalawieh M., Abbod M.F. An explainable data-driven decision support framework for strategic customer development // Knowledge-Based Systems. – 2024. – Vol. 295. – Art. 111761. – <https://doi.org/10.1016/j.knsys.2024.111761>.
19. Hadji Misheva B., Osterrieder J., Hirs A., Kulkarni O., Lin S.F. Explainable AI in Credit Risk Management // arXiv:2103.00949. – 2021. – <https://doi.org/10.48550/arXiv.2103.00949>.

Information about the authors

D. Mityanin – student, Manash Kozybayev North Kazakhstan University NPLC, Petropavlovsk, Kazakhstan; e-mail: is2077@ku.edu.kz;

D. Akhmadiyev – student, NJSC "Manash Kozybayev North Kazakhstan University", Petropavlovsk, Kazakhstan; e-mail: is2265@ku.edu.kz;

T. Berdamurat – student, Department of Information and Communication Technologies, Manash Kozybayev North Kazakhstan University NPLC, Petropavlovsk, Kazakhstan; e-mail: is2067@ku.edu.kz;

G. Kim – corresponding author, PhD, Manash Kozybayev North Kazakhstan University NPLC, Petropavlovsk, Kazakhstan; e-mail: gakim@ku.edu.kz; ORCID: 0000-0001-8402-8323;

V. Semenyuk – Master, Manash Kozybayev North Kazakhstan University NPLC, Petropavlovsk, Kazakhstan; e-mail: yvsemenyuk@ku.edu.kz; ORCID: 0000-0002-8580-7326;

Joe Robert Paul G. Lucena – Program Director of Civil Engineering, Asia Pacific College, Makati City, Philippines; e-mail: joerobertpaul@apc.edu.ph, orcid: 0009-0008-3100-8036.