Development of Financial Relations

Spyridon D. LAMPROPOULOS,
Georgios L. THANASAS,
Georgia N. KONTOGEORGA

FRAUD DETECTION IN BANKING TRANSACTIONS WITH THE USE OF ARTIFICIAL INTELLIGENCE AND ANONYMIZED DATA

Abstract

This paper examines whether AI machine-learning classifiers trained on anonymized bank transaction data can effectively predict fraudulent transactions. The study tests H1: at least one classifier's area under the ROC curve (AUC) > 0.50 against H0: the best classifier's AUC ≤ 0. 50. Using an anonymized dataset from a U.S.-based commercial bank, we assess an extensive set of classifiers, including tree-based ensembles, probabilistic, distance-based, linear and margin-based learners and a neural network using Orange Data Mining Software. The models were evaluated with stratified 10-fold cross-validation. Multiple models achieved AUC > 0.50, with tree-boosting methods providing the strongest balance between detecting fraud and limiting false alarms. Linear baselines and distance-

Lampropoulos Spyridon D., PhD, Adjunct Assistant Professor, Department of Tourism Management, University of Patras, Patras, Greece. ORCID: 0009-0003-6701-9427 Email: spyridonlampropoulos@upatras.gr
Thanasas Georgios L., PhD, Associate Professor, Department of Management Science and Technology, University of Patras, Patras, Greece. ORCID: 0000-0002-7893-9363 Email: thanasasgeo@upatras.gr
Kontogeorga Georgia N., PhD, Auditor, Hellenic Court of Audit, Athens, Greece; Affiliated Researcher, University of Paris 1 Panthéon-Sorbonne, Paris, France. ORCID: 0000-0002-9830-324X Email: kont_georgia@yahoo.gr

based methods were weak, while SVM produced high recall with operationally costly false positives. Overall, results support H1 and are inconsistent with H0. The study offers a transparent, bank-ready benchmark on anonymized, production-plausible features, and the framework is readily replicable for threshold tuning and governance in financial institutions.

Key Words:

artificial intelligence, bank fraud, banking transactions, CatBoost, financial transaction analytics, machine learning classification, XGBoost.

JEL: G21, C45, C52, C55, M42.

1 table, 1 figure, 32 references.

Introduction

Artificial Intelligence has affected and transformed finance and accounting when it comes to data capture, validation, analysis, and anomaly detection on a large scale. In reporting and assurance settings enabled by AI, data extraction, data analysis and validation can be effectively and efficiently coordinated end to end, thereby eliminating manual errors, enhancing precision and auditability. Machine Learning classifiers have been used in the banking sector to detect fraud at transaction levels where the fraud patterns remain nonlinear, sparse and fast paced changing (Bolton & Hand, 2002; Ngai et al., 2011).

The study aims to examine AI systems trained on an anonymized dataset provided by a U.S.-based commercial bank (institution anonymized); we evaluate whether AI/ML classifiers can predict fraudulent transactions. To operationalize the research objective of our study, we resorted to an ample suite of models (Random Forest, XGBoost, CatBoost, SVM, Neural Networks) subjected to stratified cross-validation, evaluated the discrimination and error-balance metrics (AUC, F1, MCC) in most suited for imbalanced fraud problems (Chicco & Jurman, 2020; Breiman, 2001; Chen & Guestrin, 2016; Cortes & Vapnik, 1995; Heaton, 2018).

Our research aim is to assess how effectively AI/ML classifiers can detect fraudulent banking transactions under realistic conditions by using only anonymized production-plausible features.

So, the main hypothesises are the following:

H1 (Main): AI machine learning classifiers trained on anonymized bank-transaction data can effectively predict whether a transaction is fraudulent or not (AUC > 0.50).

H0 (Null): AI machine learning classifiers trained on anonymized bank-transaction data do not predict fraud effectively (top model performance below (AUC ≤ 0.50).

This paper: (i) provides transparent dataset cases for evaluation across complementary model types, (ii) threshold-dependent operating behavior (AUC, precision/recall, MCC, confusion matrices) and (iii) situates empirical findings within the fraud analytics literature to find out when and why some particular AI models perform better than linear or distance based baselines against bank data (Bolton & Hand, 2002; Ngai et al., 2011; Chen & Guestrin, 2016).

The rest of the paper is organized as follows. Section 2 presents the literature review. Section 3 describes the dataset and variables (all anonymized), as well as analyses the methodology followed by the evaluation design. Section 4 presents results and section 5 with the practical implementation. Section 6 discusses the results and implications for bank fraud monitoring and avenues for future work. Section 7 concludes the paper.

Literature Review and Problem Statement

Across the financial and accounting verticals, AI has been widely adopted to enhance efficiency and accuracy, yet outcomes are contingent on organizational preparedness, quality of data and governance. As per reviews of AI adoption, firms experience benefits when robust data pipelines and clear oversight are present, while those without these factors face poor performance and trust (Cubric, 2020; Petkov, 2020). This context matters for banking fraud. Even strong algorithms will underperform if data, controls or monitoring are weak. Governance and data readiness are not side notes as they set the ceiling for any model we train on bank transactions.

Decades of research have shown that transaction fraud has three properties that complicate prediction such as: (i) severe class imbalance (far fewer frauds than legitimate payments), (ii) changing patterns over time as criminals adapt and (iii) costs that are asymmetric (missing a fraud is expensive, but too many false alarms are costly too) (Bolton & Hand, 2002; Ngai et al., 2011). Because of this, fraud detection has been framed as a supervised learning problem (classifying each transaction) alongside anomaly-detection ideas (flagging unusual behavior) (Bolton & Hand, 2002; Ngai et al., 2011; Bulatova et al., 2019; Kuryliak et al., 2025).

A non-linear tree ensemble is especially effective when dealing with structured bank data. A random forest (Breiman, 2001) captures interactions without heavy feature engineering and gradient boosting methods (e.g., XGBoost) iteratively correct errors (Chen & Guestrin, 2016). With CatBoost, categorical variables are handled, and overfitting is reduced through ordered boosting which is particularly useful when features include device, channel and transaction-type fields. Empirical studies that add temporal or relational signals (sequences or networks across cards and merchants) also find that flexible, non-linear learners outperform simple linear or distance-based baselines (van Vlasselaer et al., 2015; Jurgovsky et al., 2018).

Due to fraud's rarity and cost asymmetry, accuracy alone may be misleading. The literature recommends: (i) discrimination metrics that are insensitive to thresholds (AUC), (ii) precision / recall analysis that focuses on the minority (fraud) class and (iii) balanced metrics that use all four cells of the confusion matrix such as MCC (Fawcett, 2006; Saito & Rehmsmeier, 2015; Chicco & Jurman, 2020). In addition, it encourages making the costs explicit in training (cost-sensitive learning) and in setting decision thresholds (Bahnsen et al., 2013; Bahnsen et al., 2015). In addition, Pozzolo et al. (2018) have warned that evaluation should not be based solely on static, random partitions but must also consider real-world conditions (e.g., temporal splits, concept drift, verification delays).

Together, these findings provide a clear theoretical foundation for our design. We trained a diverse set of AI/ML classifiers on anonymous bank transaction data and evaluated them with AUC (threshold-insensitive discrimination), precision/recall (minority-class focus), and MCC (balanced correctness) under stratified cross-validation.

AI helps to identify fraud and abnormal financial activity, which increases precision in financial reporting. These include financial statements that are distorted by fraudulent transactions, compromised audit integrity and material misstatement (Wells, 2020). Fraud detection systems based on machine learning can detect anomalies faster and more accurately than manual review alone (Ngai et al., 2011). Artificial intelligence-driven fraud prediction models, such as Random Forests, Gradient Boosting and Neural Networks, have been proven to be effective in detecting high-risk behavior in banking environments, thereby enhancing internal control frameworks and protecting financial reporting accuracy (Ryman-Tubb et al., 2018). Therefore, evaluating fraud detection algorithms directly contributes to understanding how artificial intelligence can enhance the reliability and credibility of financial processes.

Methodology and Theoretical Background

The study uses an anonymized transactional dataset provided by a commercial bank based in the United States (U.S.) that wishes to remain unnamed. The dataset was provided to the authors under confidentiality restrictions. Therefore, the raw transaction-level data cannot be publicly shared. Prior to access, any personally identifying information (PII) was removed. This study focuses solely on transaction-level fraud detection and is intended for research use only.

Target (binary): *Is_Fraud (class «1» treated as the positive/target class in Orange).*

Predictor features: *Transaction_Amount, Transaction_Type, Transaction_Time, Device_Used, Account_Age, Credit_Score, Previous_Fraud*.

Typical banking data include transaction amount, time stamp (datetime), device/channel, transaction type, account tenure and prior-fraud flag (binary).

Featuring transaction descriptors, channel indicators, customer/account context and simple behavioral history, the feature set follows established practice in transaction-level fraud analytics. Transaction_Amount and Transaction_Time capture the magnitude and temporal patterns of purchases that often distinguish fraudulent from legitimate ones (Whitrow et al., 2009; Jurgovsky et al., 2018). Various payment channels and access vectors (e.g., card-present versus remote, ATM versus online) are repeatedly shown to carry distinct risk profiles (Bolton & Hand, 2002; Ngai et al., 2011). While Account_Age reflects lifecycle effects (newer accounts tend to carry a higher risk than older ones), Credit_Score provides an overall measure of credit risk that is intertwined with fraud propensity in operational settings (Ngai et al., 2011; Bhattacharyya et al., 2001). As a final note, Previous_Fraud encodes a minimal behavior history: prior confirmed fraud has been shown to provide a strong operational signal and is commonly used in bank rulesets and models.

This set of variables is production-plausible (no PII), aligns with what banks can lawfully use and reflects the three core dimensions the literature links with fraud and more precise amount & time dynamics, channel & devices, maturity & quality of accounts and adverse past events (Bolton & Hand, 2002; Ngai et al., 2011).

In Orange's feature statistics, there were no missing values across features. The transaction timestamps span eight weeks starting in January and class «1» indicates fraud. Model training without imputation and appropriate evaluation schemes for class imbalances are supported by these checks.

All experiments were performed in were conducted in Orange Data Mining software, a visual analytics environment for machine learning and evaluation workflows (Demšar et al., 2013). The learner panel covered the following:

- **Linear baselines:** Ridge Regression and Lasso Regression (kept intentionally as linear reference points to test linear separability and provide a conservative baseline).
- **Distance-based:** k-Nearest Neighbors (kNN).
- **Margin-based:** Support Vector Machine (SVM).
- **Probabilistic:** Naive Bayes.
- **Ensembles:** Random Forest, AdaBoost.
- **Gradient Boosting:** XGBoost, CatBoost.
- **Neural model:** Neural Network.
- **Online linear:** Stochastic Gradient Descent (SGD).

Several studies have shown that tree ensembles and boosting methods perform well on structured banking data (Breiman, 2001; Chen & Guestrin, 2016), while linear models provide interpretable, conservative baselines (Hoerl & Kennard, 1970; Tibshirani, 1996). The use of SVMs, kNNs, Naive Bayes, and neural networks can complement margin, instance-based, probabilistic, and deep learning approaches (Cortes & Vapnik, 1995; Cover & Hart, 1967; Mitchell, 1997; Heaton, 2018).

An evaluation of the effectiveness of a variety of machine learning models was conducted in this study. The chosen models represent four major categories of predictive learning methods commonly applied in fraud analytics: (i) distance-based learning (kNN), (ii) tree-based ensemble learning (Random Forest, XGBoost, CatBoost, AdaBoost), (iii) probabilistic learning (Naive Bayes), and (iv) linear and nonlinear discriminative learning (SVM, Logistic/Ridge/Lasso Regression, Stochastic Gradient Descent, Neural Networks).

According to previous literature, fraud patterns are often nonlinear, sparse and dynamic, making it necessary to compare simple interpretable models with more complex AI-based ensemble models (Ngai et al., 2011). Particularly, tree-based gradient boosting methods perform well on structured financial data due to their ability to learn intricate feature interactions and class boundaries (Chen & Guestrin, 2016).

The advanced AI models were also compared against classical models such as Logistic Regression, Naive Bayes and KNN, which allowed any observed improvement to be directly attributed to the advanced AI models rather than dataset bias. According to this approach, the study's primary objective is to demonstrate the value of artificial intelligence in modern fraud detection systems by comparing AI-driven predictive algorithms with traditional statistical and rule-based methods.

Evaluation used the Test & Score widget in Orange with the following settings:

- **Protocol:** stratified 10-fold cross-validation (the «stratified» flag in Orange enforces class-proportion preservation per fold when feasible) to obtain reliable, low-variance estimates under class imbalance.
- **Target class:** set to the positive class «1» in the widget (important for per-class metrics and downstream PR/ROC tools).
- **Outputs for diagnostics**: per-model scores and per-model confusion matrices (via Orange's Confusion Matrix widget) to inspect error patterns (false positives vs. false negatives).

We evaluate effectiveness at the level of cross-validated discrimination. Our primary criterion is AUC.

Operationally, we interpret «effective» as mean cross-validated AUC exceeding 0.50 for at least one model, assessed alongside complementary diagnostics (precision/recall, F1, MCC, and confusion matrices) to rule out degenerate operating points. This rule is pre-specified and does not identify any model a priori.

Orange's Test & Score reports standard classification metrics. Below are the metrics used and their official meanings in Orange:

- **AUC (Area under ROC):** probability that the classifier ranks a randomly chosen positive instance higher than a random negative; reported by Test & Score and analysed further in the ROC Analysis widget.
- **CA (Classification Accuracy):** proportion of correctly classified instances across all classes.
- **Precision:** proportion of true positives among instances predicted as positive.
- **Recall (Sensitivity):** proportion of true positives among all actual positives.
- **F1-score:** weighted harmonic mean of precision and recall.
- **MCC (Matthews Correlation Coefficient):** balanced correlation-style index using all four cells of the confusion matrix.
- **Confusion Matrix (diagnostic):** tabulates predicted vs. actual classes to visualize false positives/negatives and class errors.

Because fraud is rare and costs are asymmetric, accuracy alone can mislead. Using AUC, precision/recall & F1, and MCC aligns with best practice in imbalanced classification (Fawcett, 2006; Saito & Rehmsmeier, 2015; Chicco & Jurman, 2020).

Research Results

A stratified 10-fold cross-validation method was used to assess the predictive performance of multiple machine learning models with the target variable Is_Fraud. Among the evaluation metrics were AUC (Fawcett, 2006), accuracy, precision, recall (Saito & Rehmsmeier, 2015), F1-score and Matthews Correlation Coefficient (MCC) which, since it incorporates all the components of the confusion matrix, provides an accurate measure of classification performance under class imbalance (Chicco & Jurman, 2020).

Table 1

Model Evaluation Summary

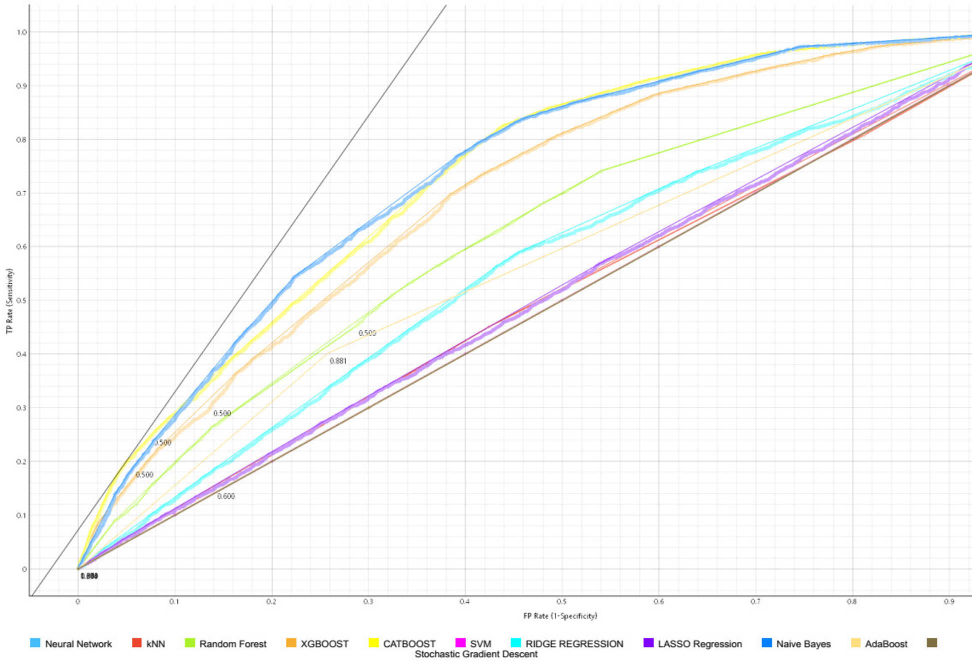
Model	AUC	Accuracy	F1	Precision	Recall	MCC
CatBoost	*0.737*	*0.733*	*0.345*	*0.566*	*0.248*	*0.236*
Naive Bayes	0.740	0.729	0.283	0.567	0.188	0.202
Ridge Regression	0.567	0.716	0.000	0.000	0.000	0.000
Lasso Regression	0.516	0.716	0.000	0.000	0.000	0.000
XGBoost	0.703	0.704	0.368	0.467	0.303	0.193
Random Forest	0.642	0.690	0.338	0.428	0.279	0.152
kNN	0.510	0.657	0.198	0.294	0.149	0.010
AdaBoost	0.572	0.646	0.391	0.382	0.401	0.142
Neural Network	0.500	0.543	0.332	0.284	0.400	0.000
Stochastic Gradient Descent	0.500	0.500	0.362	0.284	0.500	0.000
SVM	0.499	0.329	0.431	0.284	0.897	0.001

Source: calculated by the authors.

ROC curves for all classifiers and fraud as the positive class are shown in Figure 1. Naive Bayes and CatBoost create the upper envelope, with XGBoost close behind and Random Forest trailing, similar to where AUCs ranking is in the Model Evaluation Summary. kNN and AdaBoost, these baseline models are positioned closer to the 45° diagonal, while linear models and the untuned neural network follow the diagonal, respectively, which is consistent with AUC > 0.50. While Naive Bayes has a slightly higher AUC than CatBoost, it has a lower MCC due to smaller recall and a less favourable operating balance, which is identified in its curvature and indicated in the confusion matrices. These ROC profiles are in line with our previous results and validate the integration of precision–recall/MCC with ROC for imbalanced problems (Fawcett, 2006; Saito & Rehmsmeier, 2015).

Figure 1

ROC Classifiers Models Curves



Source: calculated by the authors.

CatBoost (Best Performing Model)

CatBoost displayed the best-balanced performance (AUC = 0.737, MCC = 0.236). The confusion matrix revealed 704 fraud cases that were detected (True Positives), 2,134 that were missing (False Negatives) and 540 that were classified as false positives (False Positives). The model was able to identify fraudulent cases effectively while keeping false positives controlled, making it the most reliable and operationally achievable model. CatBoost's ordered boosting and native categorical encoding helps mitigate overfitting when working with structured financial data, while its built-in native categorical encoding improves generalization.

Naive Bayes

Naive Bayes obtained high precision (0.567) but has very low recall (0.188). The confusion matrix indicated 534 detected frauds (TP), 2,304 missed (FN) & 408 false alarms (FP), suggesting a cautious fraud detection process. Its strong conditional independence assumption is due to this behavior, rarely aligns with financial fraud dynamics (Mitchell, 1997).

Ridge and Lasso Regression

Ridge and Lasso both predicted all transactions as non-fraud, obtaining F1 = 0 and MCC = 0. The confusion matrices gave 0 TP, 2,838 FN and 0 FP and 7,162 TN. These linear models do not recognize nonlinear fraud patterns that may occur in financial behavior data (Hoerl & Kennard, 1970; Tibshirani, 1996).

XGBoost

The trade-off between detection and false alarms was better in XGBoost (AUC = 0.703, MCC = 0.193). The confusion matrix displays 861 detected fraud cases (TP), 1,977 missed (FN) and 982 false alarms (FP), which is a better trade-off than Random Forest. This is in line with gradient boosting's ability to incrementally eliminate errors (Chen & Guestrin, 2016).

Random Forest

Moderate effectiveness was obtained with Random Forest (AUC = 0.642, MCC = 0.152). The confusion matrix predicted 779 fraud cases (TP), 2,059 fraud cases missed (FN) and 1,061 false alarms on legitimate transactions (FP). This is consistent with a conservative model design in which fewer false positives are picked rather than detecting fraud (Breiman, 2001).

k-Nearest Neighbors (kNN)

kNN showed a very low recall (0.149) and low general discrimination (MCC = 0.010). The confusion matrix of 422 fraud cases correctly identified (TP), 2,416 misclassified as legitimate (FN) and 1,011 legitimate transactions wrongly flagged (FP). It demonstrates strong bias toward the majority non-fraud class. This is expected when distance-based similarity does not reflect behavioral fraud patterns (Cover & Hart, 1967).

AdaBoost

AdaBoost yielded performance at the medium range (F1 = 0.391). The confusion matrix revealed 1,138 detected fraud cases (TP), 1,700 missed (FN) & 1,843 false positive alerts (FP) which indicated improved fraud sensitivity against false alarms (Freund & Schapire, 1997) but lower precision.

Neural Network

The discriminative ability of the neural network was weak (AUC = 0.500, MCC = 0.000). With 1,136 correctly detected fraudulent transactions (TP), 1,702 fraud cases missed (FN), and 2,864 legitimate transactions incorrectly flagged (FP), the confusion matrix indicates that this is quite true. This is in line with poorly trained neural-models on tabular-based data (Heaton, 2018).

Stochastic Gradient Descent

SGD showed poor stable classification (AUC = 0.500, MCC = 0.000) in classifying the data indicating difficulty in forming decision boundaries in this dataset (Bottou, 2010).

Support Vector Machine (SVM)

A high recall (0.897) was achieved by the SVM model which had 2,545 detected fraud cases (TP) but 6,418 false alarms (FP) and 293 misses (FN) while precision was very poor (0.284) (Saito & Rehmsmeier, 2015). This imbalance makes SVM impractical for real-world fraud screening, where false alerts are costly (Cortes & Vapnik, 1995).

CatBoost was proven to be the strongest model, providing the best balance of fraud detection capability and false-positive control, making it the most suitable candidate for real-world financial fraud monitoring and intervention systems. Because multiple models achieved AUC values (Fawcett, 2006) materially above 0.50 (e.g., Naive Bayes = 0.740, CatBoost = 0.737), the evidence supports H1 and is inconsistent with H0 on this dataset. We identify CatBoost as «best overall» based on balanced performance (top MCC = 0.236) alongside competitive AUC. So, our main H1 is accepted and H0 is rejected.

Practical Implementation

The empirical results can be interpreted as the workflow for a commercial fraud-screening workflow for banks. Firstly, reliable discriminative and balanced error behavior models (represented by AUC, paired with precision/recall, and MCC) are suitable for use as a first-line transaction scoring layer and generate a fraud risk score for each transaction. Secondly, banks can establish and periodically recalibrate decision thresholds to account for operational capacity and asymmetric costs (e.g., the cost of missed fraud versus the cost of investigating a false alarm) (He & Garcia, 2009; Bahnsen et al., 2015). Thirdly the model should be embedded within governance controls: (i) routine performance monitoring to detect degradation as fraud patterns evolve (Pozzolo et al., 2018), (ii) scheduled retraining using recent labeled outcomes when available and (iii) clear escalation protocols so that high-risk cases receive timely review while low-risk cases are processed normally. Although the data are anonymized, real-world bank deployment still requires documentation, validation, monitoring and auditability under model-risk management expectations (Division of Banking Supervision and Regulation, 2011), supported by strong risk data governance and reporting practices (Basel Committee on Banking Supervision, 2013).

Discussion

We studied four families of general classifiers commonly employed in fraud analytics in this study: (i) distance-based (k-nearest neighbors), (ii) tree-based ensembles (Random Forest, XGBoost, CatBoost, AdaBoost), (iii) probabilistic (Naive Bayes), and (iv) linear and margin-based discriminative learners (Ridge/Lasso, Stochastic Gradient Descent, Support Vector Machines), as well as (v) a neural network. Earlier literature has indicated that fraud patterns observed for transactions are non-linear and evolve over time, which motivates us to investigate comparing flexible ensemble methods with simple baselines (Ngai et al., 2011). Gradient-boosted trees are frequently effective on structured banking data because they capture interactions and complex decision boundaries (Chen & Guestrin, 2016).

These results are consistent with H1 and inconsistent with H0, since multiple classifiers demonstrate discrimination well above AUC > 0.5 on this dataset.

Among families of models, tree-based boosting provided the highest discrimination and error balances in anonymized bank-transaction data. The largest AUC came from Naive Bayes and CatBoost (AUC ≈ 0.74 for the first and the second model), with CatBoost yielding the highest MCC (≈ 0.24) and the best error representation. On its confusion matrix 704 true positives (TP), 2,134 false negatives (FN) and 540 false positives (FP) indicated its cautious stance on fraud flags with restrained false alarms (Chicco & Jurman, 2020).

While Naive Bayes exhibits somewhat higher AUC than CatBoost, its very low recall results in significantly less MCC, reflecting a less balanced overall operating balance on imbalanced data (Chicco & Jurman, 2020). That’s why CatBoost is placed in the first place.

XGBoost (AUC ≈ 0.70, MCC ≈ 0.19) obtained similar results, 861 TP, 1,977 FN and 982 FP, in line with the expected results obtained by gradient boosting on structured financial dataset (Chen & Guestrin, 2016). Random Forest (AUC ≈ 0.64; MCC ≈ 0.15) worked in a moderate manner with 779 TP, 2,059 FN and 1,061 FP and it was able to understand interactions, though with a more conservative boundary (Breiman, 2001).

Naive Bayes with a high AUC, demonstrated high precision yet a low recall (precision ≈ 0.57; recall ≈ 0.19) with 534 TP, 2,304 FN and 408 FP. This is a stereotypical example of its conditional-independence assumption with complex, dependent fraud features (Mitchell, 1997).

Compared to the two baselines with different values, SVM had a significant recall (≈ 0.90) at an operationally prohibitive false positive rate (2,545 TP, 293 FN but 6,418 FP), which reflects known sensitivity to the class imbalance and thresholding (Cortes & Vapnik, 1995). kNN exhibited low recall (≈ 0.15) with 422 TP, 2,416 FN, 1,011 FP, in which recall had little impact, when distance does not correspond to behavior similarity (Cover & Hart, 1967). Neural Network and SGD are at the rejection level of AUC (= 0.50) (Hoerl & Kennard, 1970; Tibshirani, 1996; Heaton, 2018; Bottou, 2010), and Ridge/Lasso collapsed into majority class (0 TP, 2,838 FN, 0 FP), indicating the non-linearity of fraud patterns.

In the case of categorical channels, devices and heterogeneous behaviors of bank transactions, the regularized tree boosting produced the most reliable operating results, whereas linear or distance-based baselines detected fraud too little or gave false alarms.

Because fraud is rare and the costs are asymmetric, accuracy alone can be misleading. For this approach AUC, precision/recall, F1 and MCC are applicable in this context (Fawcett, 2006; Saito & Rehmsmeier, 2015; Chicco & Jurman, 2020). The patterns we notice are rather common. Naive Bayes offers high precision (few false alarms) but does not notice many frauds. SVM attains very high recall but very high false-positive rates. Boosted trees present a real balance that can be adjusted by changing the decision thresholds according to bank's costing requirements (Chen & Guestrin, 2016). Alarm thresholds for banks are aligned to asymmetric costs. Post-hoc thresholding or cost-sensitive training may fine-tune the precision–recall tradeoff (He & Garcia, 2009; Bahnsen et al., 2015).

Conclusions

We evaluated a wide range of AI/ML classifiers against anonymized bank-transaction dataset using stratified, 10-fold cross-validation and evaluation metrics that are optimal for class-imbalanced fraud detection (AUC, precision/recall, F1, MCC). Tree-boosting methods provided the best performing operating profiles, CatBoost achieved AUC = 0.737 and MCC = 0.236 with a more subdued false-alarm rate (TP = 704, FN = 2,134, FP = 540). XGBoost also had good performance (AUC = 0.703, MCC = 0.193, TP = 861, FN = 1,977, FP = 982). Naive Bayes obtained the highest AUC (0.740) with minimal recall (0.188) (TP = 534, FN = 2,304, FP = 408) which corresponds to the conservative nature of ignoring many frauds. From the Ridge/Lasso linear baselines, we see that it collapsed to the majority class (TP = 0, FN = 2,838, FP = 0) and SVM pushed recall very high (0.897) at the cost of 6,418 false alarms.

In other words, these results confirm H1 («AI machine learning classifiers trained on anonymized bank-transaction data can effectively predict whether a transaction is fraudulent or not (AUC > 0.50) and are inconsistent with H0 (top model performance below (AUC ≤ 0.50).The results corroborate previous findings that gradient-boosted trees generally perform well with structured financial data and that MCC is a strong summary in the presence of imbalance (Breiman 2001; Chen and Guestrin, 2016; Chicco and Jurman, 2020).

On this dataset, boosted trees (CatBoost /XGBoost) provide a workable balance between detecting fraud and controlling alert volume. For example, the confusion pattern of CatBoost (TP = 704; FP = 540) reflects fewer unnecessary escalations compared to the extremely high-recall like the very high false-positive profile of SVM (FP = 6,418). Banks may choose tree-boosting models as their first-line detectors, and tune thresholds according to their cost ratios (the relative cost of each false alert and missed fraud would incur). This strategy is in line with best practice in imbalanced classification, where thresholds or cost-sensitive learning are tailored to business restrictions (He & Garcia, 2009; Bahnsen et al., 2015).

Model selection is just one factor. Good data pipelines, monitoring, and oversight are important for successful deployment, serving as bottlenecks for performance in banking applications (Cubric, 2020). Institutions are advised to implement: (i) periodic threshold calibration to current fraud pressure, (ii) drift monitoring and re-training cycles to deal with new patterns and (iii) clear escalation pathways to ensure that analysts are focusing on the most valuable alerts (Bolton & Hand, 2002; Pozzolo et al., 2018).

With this in mind, one of the strongest supporting pieces of evidence for modern boosted trees with post-hoc thresholding has emerged as a feasible, high-value baseline for banks to enhance preventive controls and mitigate losses whilst keeping analyst workload manageable (Chen & Guestrin, 2016).

Coming now to limitations of the study, the transactions cover only the first six weeks of 2025 (early January to mid-February). Results could change due to seasonal impacts or new fraud incidents. The dataset originates from a single U.S.-based bank, so external validity would benefit from a multi-institution replication. On this limited dataset, there were only transactional, and basic account features available. Network and sequence features (merchant–card graphs, session dynamics) would usually further increase detection (van Vlasselaer et al., 2015; Jurgovsky et al., 2018).

Robustness should be tested for in future work. Time-ordered validation and rolling or online updates could quantify how well models cope with the fact that fraud patterns constantly evolve (Pozzolo et al., 2018). Second, by training and calibrating models with example-dependent costs and bank-specific loss matrices would help in basing predictions on operational economics as well as on statistical fit (Bahnsen et al., 2015). Finally, external validation on different time frames, geographies, and institutions is warranted to determine generalizability and uncover possible dataset-specific biases.

This study offers a clear, practice-oriented benchmark for bank fraud detection on anonymized transaction features. It (i) evaluates diverse model predictors side-by-side under stratified cross-validation and (ii) demonstrates, on anonymized production-plausible features, that modern tree-boosting achieves effective discrimination while linear and distance-based baselines are less reliable. The result is a transparent reference point for both economists and risk managers that can replicate, audit and extend when designing cost-awareness and when it comes to building explainable banking fraud-monitoring systems.

Data availability declaration

The transaction-level dataset analyzed in this study was obtained from a banking institution under confidentiality restrictions. Although the data are anonymized and contain no personally identifiable information, the raw records remain proprietary and cannot be deposited in a public repository or included in an appendix. To support transparency and methodological replicability, the paper reports the full feature set, the evaluation protocol, and complete model performance outputs. Readers may contact the corresponding author (Dr. Spyridon D. Lampropoulos, spyridonlampropoulos@upatras.gr) regarding questions about the dataset and the study design, subject to the applicable confidentiality constraints.

Ethical considerations declaration

The dataset was provided in anonymized form and does not include direct identifiers. Use of the dataset is restricted to research purposes under confidentiality constraints, and no attempt was made to re-identify any individual or entity.

References

Bahnsen, A. C., Aouada, D., & Ottersten, B. (2015). Example-dependent cost-sensitive decision trees. *Expert Systems with Applications, 42*(19), 6609–6619. https://doi.org/10.1016/j.eswa.2015.04.042

Bahnsen, A. C., Stojanovic, A., Aouada, D., & Ottersten, B. (2013). Cost sensitive credit card fraud detection using Bayes Minimum Risk. In *2013 12th International Conference on Machine Learning and Applications* (pp. 333–338). https://doi.org/10.1109/icmla.2013.68

Basel Committee on Banking Supervision. (2013, January). *Principles for effective risk data aggregation and risk reporting* (BCBS Working paper No 239). Bank for International Settlements. https://www.bis.org/publ/bcbs239.pdf

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.008

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235–249. http://www.jstor.org/stable/3182781

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier & G. Saporta (Eds.), *Proceedings of COMPSTAT'2010* (pp. 177–186). Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2604-3_16

Breiman, L. (2001) Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Bulatova, O., Kuryliak, V., Savelyev, Y., Zakharova, O., & Sachenko, S. (2019, September). Modeling the multi-dimensional indicators of regional integration processes [Conference presentation abstract] (pp. 1024–1029). In *2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, Metz, France. https://doi.org/10.1109/IDAACS.2019.8924430

Chen, T., & Guestrin, C. (2016, August 13-17). XGBoost: A scalable tree boosting system (pp. 785–794). In *2016 KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), Article 6. https://doi.org/10.1186/s12864-019-6413-7

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://dx.doi.org/10.1007/BF00994018

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory, 13*(1), 21–27. https://isl.stanford.edu/~cover/papers/transIT/0021cove.pdf

Cubric, M. (2020). Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study. *Technology in Society, 62*, Article 101257. https://doi.org/10.1016/j.techsoc.2020.101257

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., & Zupan, B. (2013). Orange: Data mining toolbox in Python. *Journal of Machine Learning Research, 14*, 2349–2353. https://www.jmlr.org/papers/v14/demsar13a.html

Division of Banking Supervision and Regulation. (2011, April 4). *SR 11-7: Guidance on model risk management* (Supervision and Regulation letter). Board of Governors of the Federal Reserve System, Washington, D.C. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heaton, J. (2018). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. *Genetic Programming and Evolvable Machines, 19*(1–2), 305–307. https://doi.org/10.1007/s10710-017-9314-z

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Jurgovsky, J., Granitzer, M., Ziegler, K., Calabretto, S., Portier, P.-E., He-Guelton, L., & Caelen, O. (2018). Sequence classification for credit-card fraud detection. *Expert Systems with Applications, 100*, 234–245. https://doi.org/10.1016/j.eswa.2018.01.037

Kuryliak, V., Lyzun, M., Hayda, Y., Lishchynskyy, I., & Ukhova, N. (2025). Cross-correlation analysis of dynamic interdependencies between socioeconomic development and the demand for higher education in Ukraine. *Journal of European Economy*, *24*(3), 467–485. https://doi.org/10.35774/jee2025.03.467

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill. https://www.cs.cmu.edu/~tom/mlbook.html

Ngai, E. W. T, Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

Petkov, R. (2020). Artificial intelligence (AI) and the accounting function – A revisit and a new perspective for developing framework. *Journal of Emerging Technologies in Accounting, 17*(1), 99–105. https://doi.org/10.2308/jeta-52648

Pozzolo, A. D., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems, 29*(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. *Engineering Applications of Artificial Intelligence, 76*, 130–157. https://doi.org/10.1016/j.engappai.2018.07.008

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE, 10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems, 75*, 38–48. https://doi.org/10.1016/j.dss.2015.04.013

Wells, J. T. (2020). *Principles of fraud examination* (6th ed.). Wiley.

Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery, 18*(1), 30–55. https://doi.org/10.1007/s10618-008-0116-z

Received: September 18, 2025.
Reviewed: October 27, 2025.
Accepted: December 3, 2025.