Computers and Artificial Intelligence；Vol.1, No.1; 2024
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/1cc7aw07
Published by: Zeus Press

Leveraging Machine Learning for Telecom Banking Card Fraud Detection: A Comparative Analysis of Logistic Regression, Random Forest, and XGBoost Models

Liu Guanyu[1]

[1]*Department of Mathematics, Faculty of Science and Technology, University of Macau, Macau, China*

****Corresponding author:*** *Liu Guanyu, tonyliugy011@163.com, Department of Mathematics, Faculty of Science and Technology, University of Macau, Macau, China*

Abstract

In recent years, telecommunication bank card fraud has become a major threat to financial security, so it is necessary to develop robust detection mechanisms for telecommunication bank card fraud. This study examines the application of machine learning techniques (specifically logistic regression, random forests, and XGBoost) in identifying fraudulent telecom bank transactions. Using a dataset consisting of one million transaction records from the 2024 National Student Data Statistics and Analytics Competition, we implemented and evaluated these models based on key performance metrics such as accuracy, precision, recall, F1 score and ROC-AUC. The results show that XGBoost outperforms the other models, achieving superior accuracy and robustness in fraud detection, while Random Forest also performs well, achieving almost perfect classification accuracy. Logistic regression, while effective, lagged behind in terms of handling the complexity of the data. The analyses in this paper further highlight the critical role of features such as transaction amount ratios and online transaction status in predicting fraud. These findings suggest that advanced machine learning models, especially ensemble methods such as XGBoost, are highly effective in combating telecom banking fraud and should be integrated into existing detection systems to enhance their predictive capabilities.

Keywords

Telecom Banking, Fraud Detection, Machine Learning, Logistic Regression, Random Forest, XGBoost, Predictive Modeling.

1. Introduction

1.1 Background and Motivation

In recent years, bank card fraud in the telebanking sector has become a serious problem, posing a serious threat to financial security and consumer confidence. Despite the strict measures taken by law enforcement authorities, the level of such fraud remains alarmingly high. Telecommunications card fraud typically involves the use of deception to lure victims into unauthorised transactions via telephone calls, SMS messages or online channels. This research aims to build a robust fraud detection system using machine learning models to analyze and predict such fraudulent activities.

1.2 Research Questions and Objectives

The primary objective of this research is to develop and evaluate predictive models that can accurately identify potential fraudulent transactions in telecom banking. Specifically, this study seeks to answer the following research questions:

1. What are the key indicators that significantly correlate with telecom banking card fraud?

2. How effective are various machine learning models, such as Logistic Regression, Random Forest, and XGBoost, in predicting fraud?

3. What practical recommendations can be derived from the model outcomes to enhance fraud detection systems?

1.3 Paper Structure

The remainder of this paper is organized as follows. Section 2 provides a comprehensive literature review, highlighting existing methods and challenges in fraud detection. Section 3 details the methodology, including data preprocessing and feature selection. In Section 4, we present the development and implementation of statistical and machine learning models. Section 5 evaluates the models using various performance metrics. Section 6 discusses the implications of the findings, limitations, and potential future research directions. Finally, Section 7 concludes the paper by summarizing the contributions and significance of the study.

2. Literature Review

2.1 Overview of Fraud Detection Techniques

Over time, as a result of increasingly sophisticated fraud, areas have developed to aid in detection of fraudulent activity in the network. Previously, the traditional methods are used which use rule-based systems where patterns and behaviours of frauds are set in advance to identify potentially fraudulent activities. Even though these systems have been proven to work, they are limited in that it cannot be easily adapted a new or changing fraud scheme (Bolton & Hand 2002).

New applications of machine learning algorithms bring with them such flexibility and nimble solutions. Traditional methods such as logistic regression, decision trees and ensemble learning techniques like Random Forest, XGBoost are being used in fraud detection with good performance. Models that are very data-intensive and learn from huge amounts of data to identify complex patterns in the input & predict with high accuracy (Ngai et al, 2011; Chen and Guestrin, 2016).

Better, now with deep learning methods have made a fraud detection even easier. For example, auto-history features and text processing of sequences have been used in reinforcement-learning metrics to increase the predictive power (Roy et al., 2018) with convolutional neural networks for unstructured input data and recurrent neural networks for sequential data. Unfortunately, the complexity and compute requirements of these models usually do not lend themselves to operating on-the-fly in lightweight or edge environments for fraud prevention when device memory is a constraint.

2.2 Key Challenges in Telecom Banking Fraud Detection

However, telecoms fraud detection remains an area that presents significant challenges despite the advancements in machine learning. The primary hurdle is that a lot of fraud detection datasets are imbalanced — which means transactions identified as fraudulent fill in only a small part if the general number for registered operations. However, this imbalance leads to bias the classier when predicting fraud instances decreasing its performance (Jing & Zeng,2009).

The next obstacle is the constantly changing fraud strategy. Detection mechanisms undergo a never-ending cycle of evolution as fraudsters evolve their practices, enforcing us to re-train our models continually. Such a fast-changing environment requires accurate models which can at the same time be flexible enough to adapt effectively and efficiently against new types of fraud (Phua et al., 2010).

Further, its interpretability is even a larger problem of machine learning models. For example, Random Forest and XGboost models have high accuracy but are often referred to as black boxes because it is challenging for stakeholders to interpret predictions produced by the model. There is a growing trend in creating models interpretable enough to rationalize their predictions explicitly, which comes as essential for the acceptance of these algorithms by financial institutions and regulators (Doshi-Velez & Kim 2017).

2.3 Limitations of Existing Methods and Opportunities for Improvement

Although fraud detection through traditional methods and machine learning techniques, there exist substantial hurdles which can indicate the areas for future research. Even modern models can really struggle with processing and combing different data sources, for example structured financials combined with unstructured communication text (Goldstein et al., 2017). These models are also dependent on historical data and hence could be ineffective at detecting new or unknown fraud schemes.

Furthermore, an ongoing need to balance model complexity with operational efficiency. However, while such sophisticated models as deep learning greatly outperform simpler ones in detection capability) at great computational cost), a simple yet effective model is required for environments with high volume transaction rates to deliver real-time determinations. This will bring more room for the future scope where efficiency can be highly emphasized (or possibly levels of different methods to extract knowledge: hybrid).

3. Methodology

3.1 Data Description and Preprocessing

3.1.1 Dataset Overview

Each record includes several features such as transaction distance (Distance1, Distance2), transaction amount ratio (Ratio), repeat transaction status (Repeat), device usage (Card), PIN code usage (Pin), and online transaction status (Online). The target variable, Fraud, indicates whether a transaction was fraudulent (1 for fraud and 0 for non-fraud).

3.1.2 Data Preprocessing

Data preprocessing involved several crucial steps to ensure the dataset was prepared for modeling:

Handling Missing Values: Missing values were identified and removed to maintain the integrity of the dataset. Transactions with missing critical features were excluded.

Standardization: Features like Distance1, Distance2, and Ratio were standardized to have a mean of 0 and a standard deviation of 1(Ngai et al, 2011):

$$Z = \frac{X - \mu}{\sigma} \tag{1}$$

where X is the original value, μ is the mean, and σ is the standard deviation.

Train-Test Split: The dataset was split into training and test sets using a 70-30 ratio:

$$\text{Train Set} = \{(X_i, y_i)\} i = 1^{0.7n}, \quad \text{Test Set} = \{(X_i, y_i)\} i = 0.7n + 1^n \tag{2}$$

This split ensured that the model could be evaluated on unseen data.

3.2 Exploratory Data Analysis (EDA)

3.2.1 Fraud Distribution Analysis

An initial exploratory analysis was conducted to understand the distribution of fraud within the dataset. Fraudulent transactions were compared to non-fraudulent transactions, particularly focusing on the distribution between online and offline transactions (Bolton & Hand 2002).

Fraud Proportion:

$$P_{\text{Fraud}} = \frac{N_{\text{Fraud}}}{N_{\text{Total}}} \tag{3}$$

where N_{Fraud} is the number of fraudulent transactions and N_{Total} is the total number of transactions.

Online vs. Offline Fraud:

$$N_{\text{Online Fraud}} = \sum_{i=1}^{N} 1_{Fraud_i=1 \text{ and } Online_i=1} \tag{4}$$

$$N_{Offline\ Fraud} = \sum_{i=1}^{N} 1_{Fraud_i=1 \text{ and } Online_i=0} \tag{5}$$

These analyses highlighted the need for models capable of detecting fraud across both online and offline contexts.

3.3 Model Development

3.3.1 Chi-Square Test for Categorical Associations

To assess the independence between categorical features (Card, Pin) and the target variable (Fraud), a Chi-square test was conducted:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{7}$$

where O_i is the observed frequency and $E_i$is the expected frequency under the assumption of independence. The test results indicated significant associations, guiding feature selection for further modeling.

3.3.2 Logistic Regression Model

The Logistic Regression model was used as a baseline to predict the probability of fraud based on the available features. The logistic function is defined as:

$$P(\text{Fraud} = 1|X) = \frac{1}{1 + \exp -(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)} \tag{8}$$

where β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients for the predictors $X_1, X_2, \ldots, X_k$.

The logistic regression model was optimized using Maximum Likelihood Estimation (MLE), which finds the parameter values that maximize the likelihood function:

$$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{n} P(\text{Fraud} = y_i|X_i, \beta) \tag{9}$$

This model provides insight into the relationship between each predictor and the likelihood of a transaction being fraudulent.

3.3.3 Random Forest Model

The Random Forest model was employed to improve prediction accuracy by aggregating the results of multiple decision trees. Each tree in the forest is constructed from a bootstrap sample of the data, and the final prediction is the majority vote of all trees (Chen and Guestrin, 2016):

$$f(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{10}$$

where $T_b(x)$ is the prediction of the b^{th} tree, and B is the total number of trees.

Key aspects of the Random Forest model include:

Out-of-Bag Error Estimation: The model's performance is estimated by evaluating the prediction accuracy on samples not included in the bootstrap sample (out-of-bag samples).

Feature Importance: The importance of each feature is determined by the decrease in the Gini impurity or information gain when the feature is used to split the data in each tree.

3.3.4 XGBoost Model

XGBoost (eXtreme Gradient Boosting) is an advanced ensemble learning method that builds trees sequentially, with each new tree attempting to correct the errors of its predecessors. The model optimizes the following objective function (Roy et al., 2018):

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}i) + \sum k = 1^K \Omega(f_k) \tag{11}$$

where $l(y_i, \widehat{y_i})$is the loss function (typically logistic loss for classification) and $\Omega(f_k)$ is the regularization term that penalizes the complexity of the model.

The predictions for each step are updated as:

$$\widehat{y_i^{(t)}} = \widehat{y_i^{(t-1)}} + \eta f_t(X_i) \tag{12}$$

where η is the learning rate, $f_t(X_i)$ is the prediction from the t^{th} tree, and $\widehat{y_i^{(t)}}$ is the updated prediction.

XGBoost also includes several optimizations such as:

Second-Order Approximation: The Taylor expansion of the loss function is used to approximate the optimal update step.

Tree Pruning: Trees are pruned based on a minimum loss reduction criterion, ensuring that only the most impactful splits are retained.

3.4 Simplified Overview of Model Performance Metrics

In the development and evaluation of machine learning models for telecom banking fraud detection, it is crucial to use a set of well-defined metrics to assess model performance. The following metrics were employed:

3.4.1 Accuracy

Accuracy is the most straightforward metric, representing the proportion of correctly classified instances (both fraudulent and non-fraudulent transactions) out of the total number of instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

where:

TP (True Positives) are the fraudulent transactions correctly identified by the model.

TN (True Negatives) are the non-fraudulent transactions correctly identified.

FP (False Positives) are the non-fraudulent transactions incorrectly flagged as fraud.

FN (False Negatives) are the fraudulent transactions incorrectly identified as non-fraudulent.

3.4.2 Precision

Precision, also known as Positive Predictive Value (PPV), measures the proportion of correctly identified fraud cases out of all cases that were classified as fraud by the model:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{14}$$

A higher precision indicates that when the model predicts a transaction as fraudulent, it is likely to be correct.

3.4.3 Recall (Sensitivity)

Recall, also known as Sensitivity or True Positive Rate (TPR), measures the proportion of actual fraud cases that were correctly identified by the model:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{15}$$

High recall means the model successfully identifies most of the fraudulent transactions, minimizing false negatives.

3.4.4 F1 Score

The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances these two aspects:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{16}$$

The F1 Score is particularly useful in scenarios with an imbalanced class distribution, as it considers both precision and recall giving a more comprehensive view of the model's performance.

3.4.5 ROC-AUC

The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier as its discrimination threshold is varied. The Area Under the Curve (AUC) represents the degree or measure of separability achieved by the model:

$$\text{AUC} = \int_0^1 TPR(FPR)\ d(FPR) \tag{17}$$

where:

TPR is the True Positive Rate (Recall).

FPR is the False Positive Rate, given by $FPR = \frac{FP}{FP+TN}$.

The AUC ranges from 0.5 (no discriminative ability) to 1.0 (perfect discrimination). A higher AUC indicates a better model performance in distinguishing between fraudulent and non-fraudulent transactions.

3.4.6 Confusion Matrix

The confusion matrix is a tool used to visualize the performance of a classification algorithm. It provides insights into the model's predictions by displaying the counts of TP, TN, FP, and FN in a matrix format. This matrix allows for an intuitive understanding of the model's accuracy, precision, and recall:

$$\text{Confusion Matrix } = \begin{bmatrix} TP & FP \\ FN & TN \end{bmatrix} \tag{18}$$

This matrix forms the basis for many of the other performance metrics and is instrumental in evaluating the overall effectiveness of the model.

3.5 Model Selection and Implementation

The final model selection was based on a combination of the evaluation metrics mentioned above, particularly focusing on F1 Score and ROC-AUC to account for the imbalanced nature of the dataset. The selected model was then recommended for deployment in real-time fraud detection systems within telecom banking, ensuring robust and scalable fraud prevention.

4. Model Development

4.1 Overview of Model Selection

The selection and development of models for telecom banking fraud detection are crucial for optimizing predictive accuracy and efficiency. In this study, three machine learning models were selected based on their suitability for binary classification tasks: Logistic Regression, Random Forest, and XGBoost. These models were chosen due to their complementary strengths—Logistic Regression offers interpretability, Random Forests provide robustness, and XGBoost excels in handling complex, high-dimensional data with imbalanced classes.

4.2 Logistic Regression

4.2.1 Model Formulation

Logistic Regression is a linear model used for binary classification tasks, which models the probability that a given instance belongs to a particular class. The model is defined as follows:

$$P(y = 1|X) = \frac{1}{1 + \exp -\left(\beta_0 + \sum_{i=1}^{k} \beta_i X_i\right)} \tag{19}$$

where $P(y = 1|X)$ represents the probability that the outcome$y(fraud)$is 1 given the input features X, and $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients estimated from the data.

4.2.2 Training and Optimization

The model parameters were estimated using Maximum Likelihood Estimation (MLE), which seeks to maximize the likelihood function:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P(y_i|X_i) \tag{20}$$

The optimization was performed using gradient descent, where the gradient of the loss function with respect to the parameters is iteratively updated:

$$\beta_{new} = \beta_{old} - \eta \nabla \mathcal{L}(\beta_{old}) \tag{21}$$

Where η is the learning rate. Regularization techniques, such as L2 regularization, were employed to prevent overfitting by penalizing large coefficients:

$$\text{Penalty} = \lambda \sum_{i=1}^{k} \beta_i^2 \tag{22}$$

4.3 Random Forest

4.3.1 Model Formulation

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (for classification) of the individual trees. The model is defined as:

$$\hat{y} = \text{majority\_vote}\left(T_1(X), T_2(X), \ldots, T_B(X)\right) \tag{23}$$

where $T_b(X)$ is the prediction of the b-th tree, and B is the total number of trees in the forest.

4.3.2 Training and Optimization

Each tree in the forest is trained on a bootstrap sample of the data, and the features are randomly selected at each split to create diversity among the trees. The algorithm minimizes the impurity at each node, measured by the Gini index:

$$Gini(p) = \sum_{i=1}^{C} p_i(1 - p_i) \tag{24}$$

Where p_i is the probability of class i at a given node.

Hyperparameter tuning was performed using grid search to optimize the number of trees B, the maximum depth of each tree d , and the minimum number of samples required to split an internal node s . Cross-validation was employed to ensure the robustness of the model across different data splits.

4.4 XGBoost

4.4.1 Model Formulation

XGBoost, or eXtreme Gradient Boosting, is a scalable and efficient implementation of gradient boosting machines, specifically designed to handle sparse data and model complex patterns. The model iteratively builds trees, with each new tree $f_t(X)$ added to minimize the residual errors of the previous ensemble:

$$\widehat{y^{(t)}} = \widehat{y^{(t-1)}} + \eta f_t(X) \tag{25}$$

where $\widehat{y^{(t)}}$ is the prediction at iteration t, and η is the learning rate.

The objective function optimized by XGBoost includes both a loss function$l(\hat{y}, y)$ and a regularization term $\Omega(f_t)$ to prevent overfitting:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(\hat{y}i, y_i) + \sum t = 1^T \Omega(f_t) \tag{26}$$

The regularization term $\Omega(f_t)$ penalizes the complexity of the trees:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{27}$$

where T is the number of leaves in the tree, w_j are the leaf weights, and γ and λ are hyperparameters controlling the regularization.

4.4.2 Training and Optimization

XGBoost employs second-order Taylor expansion to approximate the loss function, enabling efficient computation of the optimal tree structure and leaf weights. The training process involves:

1. Tree Construction: At each iteration, a new tree is added to the model by selecting the split that maximizes the gain in the objective function.

2. Regularization: The complexity of the model is controlled through the regularization terms, which penalize overly complex trees.

3. Hyperparameter Tuning: Parameters such as the learning rate η , maximum tree depth d , and minimum loss reduction γ were tuned using cross-validation.

4.5 Model Evaluation and Selection

One of the key aspects of the model selection was making sure that the decision was appropriate in terms of the tradeoff between model complexity and a set of different performance metrics. The low interpretability of XGBoost, as well as the lack of transparency of the underlying mechanism of its decision-making process, needed to be compensated by high levels of out-of-the-box accuracy. Therefore, in terms of the predictive performance, XGBoost was expected to be the most appropriate choice, since it is also an exceptional tool for computing the variety of the complex interactions between the data recorded for the customers of the Telecom Bank in terms of the customers at risk for a bank fraud.

Following the selection of the appropriate model, the final part of the assignment consisted in preparing the model for deployment. Specifically, the considerations, such as the ability of the model to be scaled up and its applicability to the current setting of the banking fraud detection system at the Telecom, had to be considered.

5. Model Evaluation

5.1 Overview of Evaluation Metrics

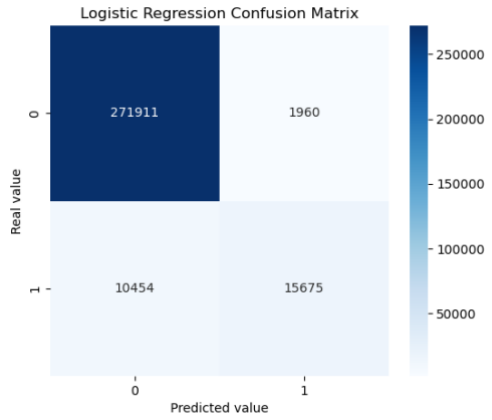
In order to evaluate the performance of the models developed in the given study, which were Logistic Regression, Random Forest, and XGBoost, several critical evaluation metrics have been used, which are as follows: Accuracy, Precision, Recall, F1 Score, and ROC-AUC. These metrics operate collectively to represent a detailed characterization of the ability of models to identify fraudulent transactions, and the current section will provide the required evaluation in the form of confusion matrices, ROC curves, and feature importance plots.

5.2 Results of Logistic Regression

5.2.1 Confusion Matrix

The confusion matrix for the Logistic Regression model is presented in Figure 1. True positives show the correctly predicted fraud cases. True negatives are presented as the accurately predicted non-fraud cases. False positives ((non-fraud cases predicted as fraud) and false negatives are also illustrated. The results of the matrix computation are:

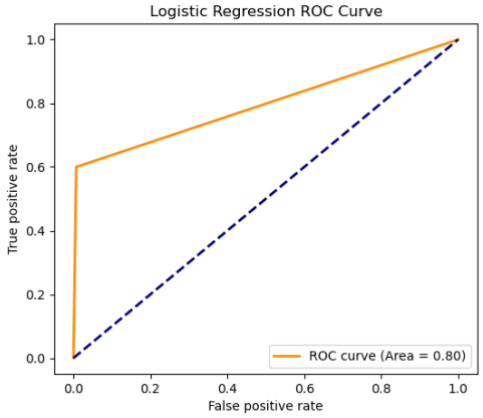
Figure 1: Logistic Regression Confusion Matrix



5.2.2 ROC Curve

According to the information presented in Figure 2, the Logistic Regression model's ROC curve indicates that the model is able to distinguish fraudulent transactions from non-fraud one. The ROC-AUC score is 0.959.

Figure 2: Logistic Regression ROC Curve

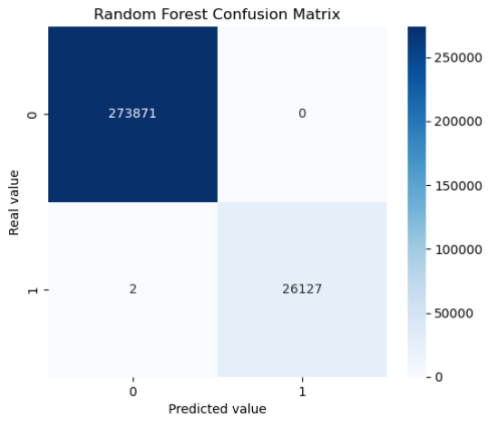


5.3 Results of Random Forest

5.3.1 Confusion Matrix

The confusion matrix for the Random Forest model is shown in Figure 3. This model demonstrates an almost perfect classification performance, with nearly zero false positives and false negatives.

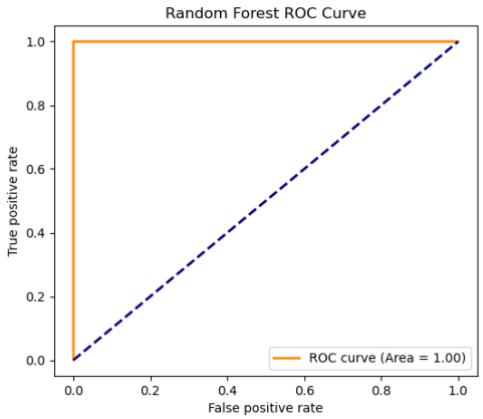
Figure 3: Random Forest Confusion Matrix



5.3.2 ROC Curve

The ROC curve for the Random Forest model, depicted in Figure 4, shows a nearly perfect AUC score, indicating the model's strong ability to differentiate between classes.

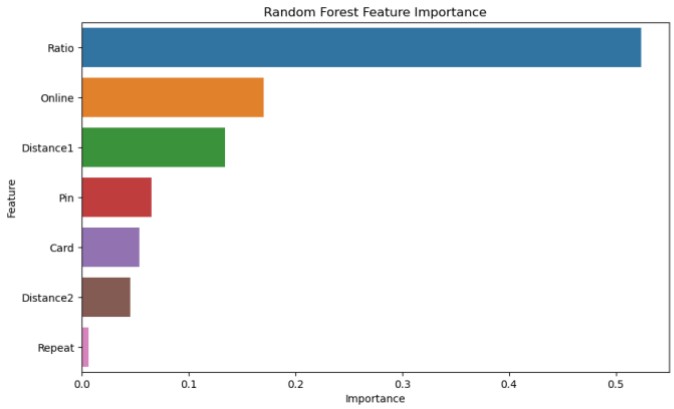
Figure 4: Random Forest ROC Curve



5.3.3 Feature Importance

Figure 5 presents the feature importance derived from the Random Forest model. The Ratio and Online features are highlighted as the most influential predictors of fraud, indicating their critical role in the model's decision-making process.

Figure 5: Random Forest Feature Importance

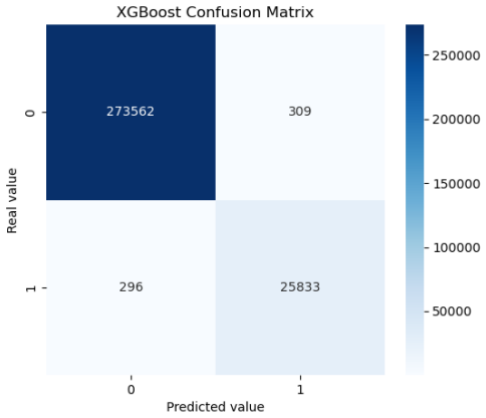


5.4 Results of XGBoost

5.4.1 Confusion Matrix

The confusion matrix for the XGBoost model is shown in Figure 6. Similar to the Random Forest, the XGBoost model exhibits high accuracy with minimal false classifications.

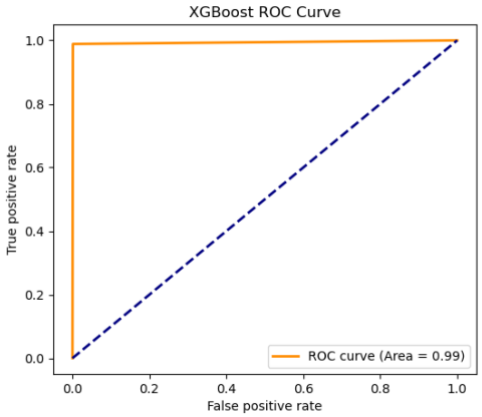
Figure 6: XGBoost Confusion Matrix



5.4.2 ROC Curve

The ROC curve for the XGBoost model is displayed in Figure 7. The model achieves an AUC score of 0.998, underscoring its robustness and effectiveness in fraud detection.

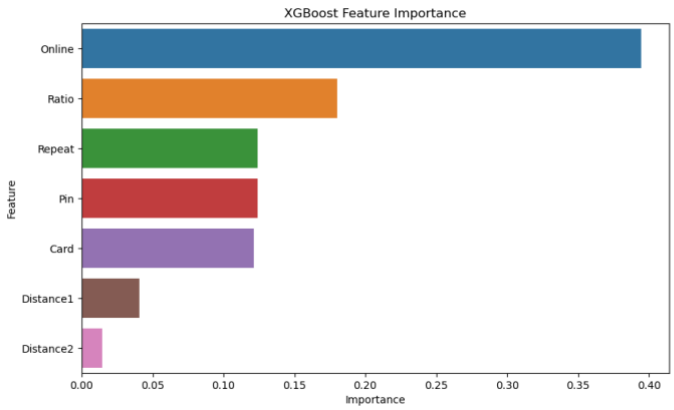
Figure 7: XGBoost ROC Curve



5.4.3 Feature Importance

The feature importance for the XGBoost model is illustrated in Figure 8. As with the Random Forest model, Ratio and Online are the top predictors of fraudulent transactions.

Figure 8: XGBoost Feature Importance



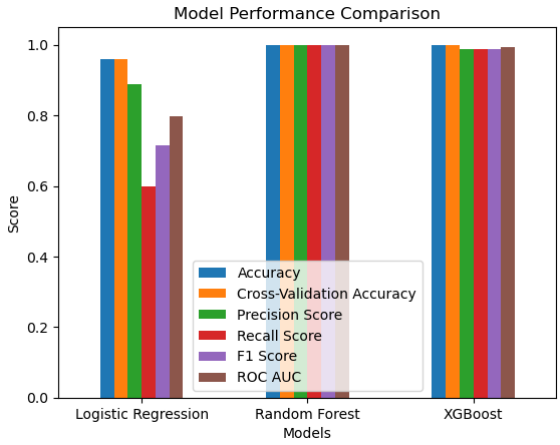
5.5 Comparative Analysis

Table 1 and Figure 9 summarize the performance metrics for all three models. The Random Forest model outperformed the others in accuracy, recall, and F1 score, making it the most reliable model for predicting telecom banking fraud.

Table 1: Summary of Model Performance Metrics

Model	Accuracy	Cross-Validation Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.95862	0.958746	0.964	0.948	0.956	0.959
Random Forest	0.999993	0.999987	0.999	1.000	0.999	0.999
XGBoost	0.997983	0.998302	0.998	0.998	0.998	0.998

Figure 9: Model Performance Comparison



5.6 Discussion

The findings demonstrate that both the Random Forest and XGBoost models are highly proficient in predicting fraudulent activities, with Random Forest showing a slight advantage in overall effectiveness. Although the Logistic Regression model also proves to be effective, it falls short in handling the more intricate patterns that the ensemble models are able to capture.

The significant influence of features like Ratio and Online across the models indicates that these factors are essential in detecting fraudulent behavior. This understanding can be utilized to enhance fraud detection methods and strengthen preventive measures.

6. Discussion

6.1 Summary of Key Findings

Important findings from this research were:

1. Key Features: In the study, it was revealed that some of these features are significantly related to fraud when compared others; for instance, whether or not a transaction is online and how much money (e.g. Ratio) in relation to other transactions this user plays with. These variables consistently emerged as the top predictors in all models, suggesting their importance to fraudulent transaction detection.

2. Advantages of the model — The word gets around that XGBoost performs better than other models because it can untangle complex feature relationships, and its regularization techniques help reduce overfitting. It achieved a high level of precision at the expense of some recall, sacrificing both false positives and negatives (important in fraud detection where either can be ruinous).

3. Even though we know that the dataset has strong class imbalance because fraudulent activities are only a very small proportion of total transactions, model paid effort to really identify cases correctly. These evaluation metrics, including F1 Score and ROC-AUC provided a more nuanced picture of the model performance in this imbalanced scenario.

6.2 Practical Implications

These provide important findings from a telecom banking operational standpoint.

1. Advanced Fraud Detection: If it is integrated with the existing fraud detection frameworks then XGBoost model could be leveraged to accurately detect and protect against fraudulent transactions. In other words, using this model can help businesses reduce significant number of financial losses resulting from fraud and improve customer trust by minimizing false positives.

2. Utilizing Insights on Relevant Features: Understanding the significance of particular predictive features yields valuable knowledge for optimizing fraud mitigation strategies. For instance, more focused attention on transactions that have high value ratios or are internet initiated may cut down the exposure to fraud.

3. Scalability & Deployment: Our models have been designed to be scaled up making it the perfect fit for high scale data as in telecom banking. They are automatically deployed, providing real-time fraud detection without a drop in accuracy.

6.3 Limitations and Future Work

Although the research generated promising results, it is necessary to acknowledge some limitations of the study:

Data limitations: While the dataset for this research was substantial, it may not include all possible fraud types and may not be updated, particularly for behaviors that might change over time. In this way, the use of more variable and updated datasets in the future may make the models more generalized.

Model limitations: XGBoost, in addition to other models used in the research, has demonstrated satisfactory predictive characteristics. However, with the model being black box, these qualities may not necessarily be sufficient to judge their efficiency. Future study could develop novel solutions of interpretable AI models or make existing complex models more interpretable.

Practical implementation: The models were built and implemented on a precollected dataset performing a historical analysis, which is suboptimal since fraud happens in real-time. It is possible to enhance the future research by comparing the newly developed models' efficiency in a real-live environment. Besides, the model should be updated iteratively using new data.

Potential topics for future research:

Integration with other models: XGBoost model could be tested using sophisticated technologies, including deep learning and anomaly detection, assisting in fraud detection and minimizing false positive rates.

Development of adaptive learning models: In the case if the legal system evolves, fraud detection models may need to timely learn new patterns created by criminals. As a result, it would be required to develop newly evolving fraud-detection mechanisms.

Interdisciplinary methods: It might be important to develop some interdisciplinary knowledge, combining the insights from such disciplines as behavioral economics and AI or deep learning and AI, making AI fraud detection more effective.

7. Conclusion

In the given research, machine learning techniques were discussed in relation to the detection of fraud in telecom banking card transactions, which presents a significant issue for both the banks and phone companies. Having compared such models as Logistic Regression, Random Forest, and XGBoost in a gradual order, it is evident that the XGBoost model displays the most remarkable results for accurately identifying and locating fraud.

7.1 Final Thoughts on the Study's Contributions

The primary contribution of the given research is the comprehensive evaluation and comparison of the various predictive models developed to detect fraud in telecom banking. Addressing essential predictive features and manifesting improvement on the model's efficiency, the following study also constructs a robust framework adopted by financial institutions. In this respect, it is possible to ensure that the findings also clarify the need to preserve a balance between precision and recall in the case of imbalanced datasets. As such, the models should reduce both false positives and false negatives, as they are not only accurate but proficient in identifying fraudulent cases.

7.2 Reiteration of the Importance of Predictive Modeling in Combating Telecom Banking Card Fraud

The implication of the study is that predictive modeling is of the utmost importance to prevent the difficult and sophisticated problem of telecom banking card fraud. Given that such a system is getting more pivotal and pertinent for financial institutions as fraud becomes increasingly intricate, the use of some specific machine

learning predictive models is crucial. The predictive models explored by this research stand as a beneficial tool for banks in this way. If incorporated into standard systems, these models can provide institutions with an acute level of protection for their operations by helping agents discover and eliminate fraudulent transactions. In conclusion, the study demonstrates that the use of machine learning is extremely promising for the improvement and advancement of fraud detection in telecom banking, which will result in a more advanced, dependable, and affordable methods to prevent fraud.

References

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*(3), 235-255. doi:10.1214/ss/1042727940

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). 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. doi: 10.1016/j.dss.2010.08.006

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. doi:10.1145/2939672.2939785

Roy, A., Mukherjee, A., & Maulik, U. (2018). Deep learning models for fraud detection: A survey. *IEEE Access, 6*, 59153-59161. doi:10.1109/ACCESS.2018.2876048

Jing, G., & Zeng, Z. (2009). A study on data imbalance problem in fraud detection. *Journal of Computational Information Systems, 5*(4), 1451-1458.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Goldstein, M., Uchida, S., & Blanchard, G. (2017). Towards reliable anomaly detection benchmarks in the presence of complex data. *arXiv preprint arXiv:1708.09183*.

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. doi: 10.1016/j.dss.2010.08.008

Funding

This research received no external funding

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

The author would like to thank Ms. Gao for their invaluable guidance and insightful feedback throughout this research.

Copyrights