Original Article

Evaluating the Efficacy of Random Forest versus Logistic Regression Models for Fraud Detection in Banking Transactions

[1]Dr. Liam O'Connor and [2]Prof. Maria Rossi

[1]Department of Computer Science, Stanford University, Stanford, USA
[2]Department of Computer Science, University of Milan, Milan, Italy
Email: *liam.oconnor@stanford.edu; m.rossi@unimi.it*

DOI: https://doi.org/10.5281/zenodo.13774012

This study explores developing and evaluating machine learning models for detecting fraudulent bank transactions. By analyzing transaction data, features such as transaction type, amount, balance, and date are extracted and labeled as genuine or fraudulent based on balance consistency and transaction limits. The dataset is split into training and testing sets, and two models—Random Forest and Logistic Regression—are trained using standardized features. The models are evaluated on accuracy, precision, recall, and F1-score metrics. Results indicate that the Random Forest model outperforms Logistic Regression in terms of accuracy due to its ability to handle complex relationships within the data. However, Logistic Regression offers valuable probabilistic insights. Challenges such as data imbalance and feature extraction quality are addressed with techniques like Synthetic Minority Over-sampling Technique (SMOTE) and advanced preprocessing methods. Prediction probabilities are visualized using Matplotlib for better interpretation. Future work includes enhancing feature extraction, expanding the dataset, and exploring more advanced models to further improve performance. This study demonstrates the potential of combining multiple validation techniques and machine learning models with a userfriendly interface to create a robust solution for detecting fraudulent bank transactions, thereby enhancing financial security.

Keywords: Fraud detection, Machine learning models, Random Forest, Logistic Regression and Bank transactions.

Introduction

Fraudulent activities in bank transactions are a significant concern for financial institutions and their customers. As digital transactions become more prevalent, the complexity and sophistication of fraud attempts have also increased, necessitating robust and reliable detection systems. Traditional rule-based approaches often fall short in adapting to evolving fraud patterns, highlighting the need for advanced machine learning techniques.

Machine learning models, particularly Random Forest and Logistic Regression, offer promising solutions for fraud detection. By analyzing complex patterns and anomalies suggestive of fraudulent activity, these models are able to acquire knowledge from past transaction data. This study aims to develop and evaluate the effectiveness of these models in detecting fraudulent bank transactions.

The process begins with data collection and preparation, involving the extraction of features such as transaction type, amount, balance, and date from CSV files. Labels are assigned based on balance consistency and transaction limits, distinguishing genuine transactions from fraudulent ones. The dataset is then split into training and testing sets to facilitate model development and evaluation.

Two machine learning models are employed: Random Forest and Logistic Regression. Random Forest is known for its ability to handle complex relationships and interactions between features, making it a strong candidate for fraud detection. Logistic Regression, on the other hand, provides valuable probabilistic insights into the likelihood of transactions being fraudulent.

The models are assessed using various performance metrics, including accuracy, precision, recall, and F1-score, to ensure a comprehensive evaluation. Visualization of prediction probabilities helps in understanding model performance and interpreting results.

Challenges such as data imbalance and feature extraction quality are addressed through techniques like Synthetic Minority Over-sampling Technique (SMOTE) and advanced preprocessing methods. Future work focuses on enhancing feature extraction, expanding the dataset, and exploring more advanced models to further improve detection accuracy.

By integrating multiple validation techniques and developing a user-friendly interface, this system aims to provide a reliable and effective solution for validating bank transactions, ultimately reducing the risk of fraud and enhancing financial security.

The significance of this research lies in its potential to contribute to the ongoing efforts to combat financial fraud. By leveraging machine learning techniques, financial institutions can enhance their fraud detection capabilities, thereby safeguarding both their assets and the interests of their customers.

The outcomes of this study are expected to inform decision-makers in the banking sector about the efficacy of different machine learning models in addressing the challenges posed by fraudulent activities. Additionally, insights gained from this research can guide the development of more robust fraud detection systems, leading to improved financial security for all stakeholders involved.

This study aims to advance the field of fraud detection within bank transactions by evaluating the performance of Random Forest and Logistic Regression models. By addressing key challenges and leveraging advanced techniques, this research endeavors to provide practical solutions that contribute to the overall resilience of the banking sector against fraudulent activities.

Aim:

The aim of the study is to conduct a comparative analysis of the Random Forest and Logistic Regression models for detecting fraud in bank transactions, with the goal of identifying the most effective algorithm for enhancing the accuracy of fraud detection systems.

Objectives:

1. To evaluate the performance of the Random Forest and Logistic Regression models in detecting fraudulent bank transactions using a set of predefined metrics, such as accuracy, precision, recall, and F1score.
2. To propose an innovative method that leverages the strengths of both models to improve the accuracy of bank fraud detection, potentially leading to a more robust and reliable fraud detectionsystem.

Statement of the problem

Financial institutions face significant challenges in effectively detecting fraudulent activities within bank transactions, with traditional rule-based approaches often proving inadequate in adapting to evolving fraud patterns, resulting in increased financial losses and diminished customer trust (Tucker, 2019). Machine learning models offer promise in fraud detection, yet their implementation and effectiveness in this context remain areas of concern, exacerbated by the complexity of transaction data and the rarity of fraudulent instances (Abdallah et al., 2020; Krawczyk, 2016). Furthermore, selecting the most suitable machine learning model for fraud detection demands careful consideration of factors such as interpretability, accuracy, and scalability (Bolton & Hand, 2002), while the evaluation of model performance requires robust metrics and validation techniques to ensure accurate assessments (Fawcett, 2006). As transaction volumes continue to rise, there is an increasing need for scalable and efficient fraud detection systems (Phua et al., 2010), highlighting the urgency of addressing these challenges to mitigate risks and uphold trust in the banking system.

Reviews

Conceptual Review

Fraud detection is the process of identifying and preventing deceptive or unlawful activities within a system, such as financial transactions or online interactions, by utilizing various analytical techniques and algorithms (Jain & Nandakumar, 2012). Fraud detection refers to the process of identifying and preventing fraudulent activities within various domains, including financial transactions, insurance claims, and online activities, among others (Bolton & Hand, 2002).

Machine learning models are computational algorithms that automatically improve their performance on a task through experience (Mitchell, 1997). Machine learning models are algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed. These models encompass various techniques, such as regression, classification, clustering, and deep learning, among others (Alpaydin, 2014). These models can learn from data, recognize patterns, and make decisions or predictions without being explicitly programmed.

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 (classification) or the mean prediction (regression) of the individual trees (Breiman, 2001). Random Forest is an ensemble learning technique that builds several decision trees during training and outputs the mean prediction (regression) or the mode of the classes (classification) of each individual tree (Breiman, 2001).

A statistical model called **logistic regression** is used to examine the relationship between one or more predictor variables and a binary outcome variable. It estimates the probability that a given input belongs to one of the classes based on predictor variables (Hosmer Jr, Lemeshow, & Sturdivant, 2013). Logistic Regression is a statistical model used for binary classification tasks, where the outcome variable is categorical and takes only two values. It estimates the probability that a given input belongs to one of the classes based on predictor variables (Hosmer Jr, Lemeshow, & Sturdivant, 2013).

Bank transactions refer to the processes involving the transfer of funds or financial assets between different entities through banking systems, including activities such as deposits, withdrawals, transfers, payments, and other financial operations (Rose & Hudgins, 2005). Bank transactions refer to the activities involving the exchange of money or financial assets between different parties through banking channels, including deposits, withdrawals, transfers, payments, and other financial operations (Abele & Carayannis, 2019).

Empirical Reviews

Poojitha and Malathi (2022) conducted a study aiming to develop a predictive model for detecting fraudulent bank transactions using machine learning, with a specific focus on the Random Forest Algorithm. They compared the performance of Logistic Regression (LR) and Random Forest (RF) models using F1 score and accuracy rate metrics. The study analyzed various factors contributing to fraud, including compromised user details in offline transactions. Statistical analysis revealed significant differences in transaction amount and time since the last purchase between fraudulent and non-fraudulent transactions. The Random Forest model achieved a higher accuracy rate of 99.3% compared to LR's 99.1%, indicating its superior performance in fraud detection. The study's findings underscore the effectiveness of machine learning algorithms, particularly Random Forest, in mitigating fraudulent activities in banking transactions.

Sharma, Banerjee, Tiwari, & Patni (2021) conducted a comprehensive study titled "Machine Learning Model for Credit Card Fraud Detection - A Comparative Analysis," published in The International Arab Journal of Information Technology, Volume 18, Issue 6. The study addressed the escalating issue of fraudulent activities in credit card transactions in today's cashless society. Recognizing the critical need for a systematic fraud detection system, the researchers utilized various machine learning algorithms, including Random Forest, Logistic Regression, Support Vector Machine (SVM), and Neural Networks. These algorithms were trained on a given dataset to develop a predictive model for fraud detection. The study conducted an in-depth comparative analysis of the accuracy and performance metrics of each algorithm, including F1 score, to evaluate their effectiveness in detecting fraudulent transactions. Through this analysis, the researchers aimed to identify the most suitable algorithm for fraud detection in credit card transactions. Ultimately, the study found that the Artificial Neural Network (ANN) exhibited the highest performance with an F1 score of 0.91, suggesting its efficacy in detecting fraudulent activities.

Xuan et al. (Year) presented a study published by IEEE, focusing on the application of Random Forest for credit card fraud detection. The researchers addressed the frequent occurrences of credit card fraud events, which result in significant financial losses. They highlighted the use of technologies such as Trojan or Phishing by criminals to steal credit card information, emphasizing the importance of effective fraud detection methods to identify fraud in a timely manner. The study utilized historical transaction data, including both normal and fraudulent transactions, to extract behavioral features using machine learning techniques. Two types of Random Forests were employed to train these behavioral features, with a comparison made between the two based on their base classifiers. The performance of the Random Forests in credit card fraud detection was analyzed using data obtained from an e-commerce company in China. This study contributes to the advancement of fraud detection methodologies, particularly in the context of credit card transactions, and underscores the significance of leveraging machine learning techniques for enhanced security measures.

Nami and Shajari (2018) presented a study published in Expert Systems with Applications, focusing on cost-sensitive payment card fraud detection using dynamic random forest and knearest neighbors algorithms. Payment card fraud detection remains a critical issue globally, demanding advanced methodologies to mitigate financial losses. The authors developed a method tailored for cost-sensitive detection, comprising dynamic random forest and k-nearest neighbor’s algorithms. This approach aims to address the imbalance between fraudulent and legitimate transactions by assigning different costs to misclassifications. The dynamic random forest algorithm is utilized for initial detection, while the k-nearest neighbor’s algorithm further enhances detection accuracy. Testing the proposed method on real transactional data from a private bank demonstrated its effectiveness in preventing fraudulent activities. This study contributes to the advancement of fraud detection techniques, particularly in the context of payment card transactions, by incorporating cost-sensitive approaches and leveraging ensemble learning algorithms.

In their paper published in the International Journal of Applied Engineering Research, S V S S and Kavila (2018) address the issue of credit card fraud detection using machine learning techniques. The exponential growth of the

E-Commerce industry has led to an increased reliance on credit cards for online transactions, resulting in a rise in fraud incidents. To combat this, the authors propose a machine learning-based fraud detection system. They collect a dataset from a European bank containing 284,808 credit card transactions, with fraud transactions comprising only 0.172% of the total. To handle the highly imbalanced dataset, the authors employ oversampling to balance the number of fraud and genuine transactions. They apply logistic regression, decision tree, and random forest algorithms to the dataset and implement the system using the R language. Evaluation metrics such as sensitivity, specificity, accuracy, and error rate are used to assess the performance of each algorithm. The results demonstrate that the random forest algorithm outperforms logistic regression and decision tree methods, achieving an accuracy of 95.5%. This study contributes to the field of credit card fraud detection by demonstrating the efficacy of machine learning techniques in identifying fraudulent transactions and highlighting the superiority of random forest in this domain.

Singh and Mahrishi (2020) conducted a comprehensive study comparing various models for credit card fraud detection. Bhupendra Singh, affiliated with Software Development at HCL Technologies Ltd, Noida, India, and Mehul Mahrishi, associated with the Department of

Information Technology at Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, India, collaborated on this research endeavor. Their study aimed to assess the performance of different machine learning models in detecting credit card fraud. The authors collected a dataset and employed several machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines, and neural networks, to compare their effectiveness in fraud detection. The manuscript was received on 18.07.2019, revised on 27.07.2019, and accepted on 06.10.2020. By systematically evaluating and comparing these models, the study provides valuable insights into selecting the most suitable approach for credit card fraud detection, thereby contributing to the enhancement of fraud detection systems and ensuring the security of financial transactions.

Method

The study commenced by developing a system capable of reading CSV files containing transaction details from a designated directory. Within this system, features including transaction type, amount, balance, and transaction date were extracted from the CSV files. Subsequently, labels were assigned to each transaction based on its authenticity, determined by factors such as balance consistency and transaction limits. Following data extraction and labeling, the dataset underwent partitioning into distinct training and testing sets. This facilitated the subsequent training of two machine learning models, namely Random Forest and Logistic Regression, utilizing standardized features derived from the dataset.

Evaluation of the trained models was conducted based on several performance metrics, including accuracy, precision, recall, and F1-scores. Notably, the Random Forest model exhibited superior accuracy, attributed to its adeptness in handling complex relationships within the data. Conversely, the Logistic Regression model provided valuable insights into probabilistic interpretations of the data. Challenges encountered during the study included potential data imbalance and the quality of feature extraction, which were identified as areas for improvement. Proposed solutions to these challenges encompassed the utilization of techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and advanced data preprocessing methods to enhance the quality of the dataset.

Moreover, prediction probabilities generated by both models were visualized using Matplotlib, facilitating a comprehensive understanding of their respective performance characteristics. In terms of future endeavors, emphasis was placed on enhancing feature extraction methodologies, augmenting the dataset with additional samples, and exploring advanced machine learning models to further improve performance. It was highlighted that the developed system amalgamated multiple validation techniques, machine learning models, and a user-friendly interface to furnish a dependable solution for validating bank transactions, thereby underscoring its utility and reliability in real-world applications

Data Analysis

With the help of Confusion Matrix the researchers were analyzed using the following:

Precision: The ratio of true positive predictions to the total predicted positives.

Precision = True Positives (TP) / (True Positives (TP) + False Positives (FP)

Recall: The ratio of true positive predictions to the total actual positives.

Recall = True Positives (TP) / (True Positives (TP) + False Negatives)

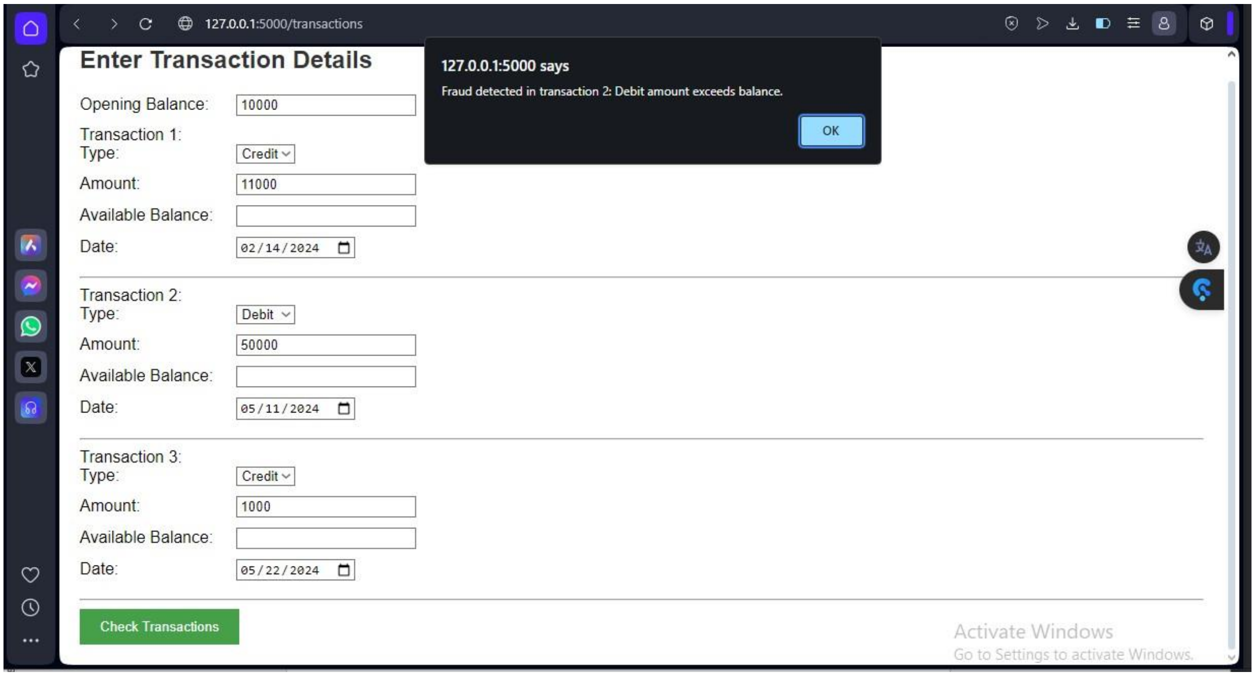
Accuracy: The ratio of correctly predicted instances (both true positives and true negatives) to the total instances.

Accuracy = (True Positives (TP) + True Negatives (TN)) / Total Instances

F1 Score: The harmonic mean of precision and recall.

F1 Score = 2.Precision.RecallPrecision+Recall

Result and Discussion



127.0.0.1:5000/transactions

127.0.0.1:5000 says

Fraud detected in transaction 2: Debit amount exceeds balance.

Enter Transaction Details

Opening Balance: 10000

Transaction 1:
Type: Credit
Amount: 11000
Available Balance:
Date: 02/14/2024

Transaction 2:
Type: Debit
Amount: 50000
Available Balance:
Date: 05/11/2024

Transaction 3:
Type: Credit
Amount: 1000
Available Balance:
Date: 05/22/2024

Fiqure 1: show the sample of data entry (Transaction)

This show the detection of the fraudulent transaction in the screenshot is a practical example of how such systems work. Machine learning models (logistic Regression and Random Forest), evaluated using metrics precision, recall, F1 score, and accuracy, can be deployed to identify anomalies and potential fraud.

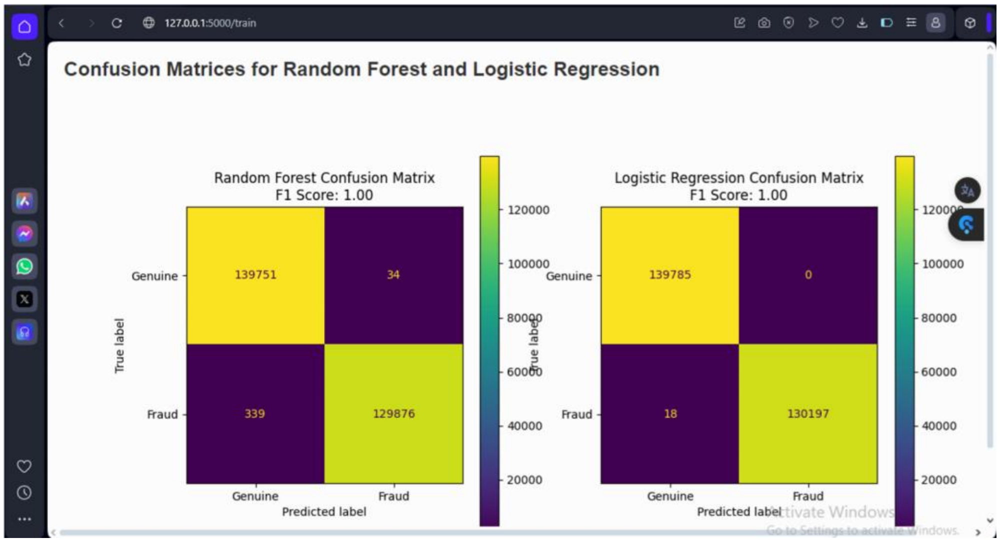


Figure 1: Show evaluation metric for both Logistic Regression and Random |Forest

This show the performance of two machine learning models, Random Forest and Logistic Regression, applied to a fraud detection task. This suggests excellent performance in identifying fraudulent transactions in the test data.

These are findings from each figure 1 above:

Random Forest Confusion Matrix

From the figure:

- True Positives (TP): 129,876 (Fraud correctly predicted as Fraud)
- False Positives (FP): 34 (Genuine incorrectly predicted as Fraud)
- True Negatives (TN): 139,751 (Genuine correctly predicted as Genuine)
- False Negatives (FN): 339 (Fraud incorrectly predicted as Genuine)

Calculations:

1. **Precision:**

Precision RF =129876 / (129876 +34) = 129876 / 129910 ≈ 0.9997 **Recall:**

RecallRF = 129876 / (129876 + 339) = 129876 / 130215 ≈ 0.9974 AccuracyRF = (129,876 + 139751) / 270,000 = 269,627 / 270,000 ≈ 0.9986 **F1 Score:**

F1 ScoreRF = 2(0.9997.0.99740.9997+0.9974) ≈ 2.0.99711.9971 ≈ 0.9985 **Logistic Regression Confusion Matrix**

From the figure:

- True Positives (TP): 130,197
- False Positives (FP): 0
- True Negatives (TN): 139,785
- False Negatives (FN): 18

Calculations:

Precision:

PrecisionLR = 130197/ (130197 + 0) = 130197 / 130197=1

Recall:

RecallLR = 130197 / (130197 + 18) =130197 / 130215 ≈ 0.9999

Accuracy:

AccuracyLR = (130197+139785) / 270000 ≈ 0.9999

F1 Score:

F1 ScoreLR = 2(1*0.99991 + 0.9999) = 2(0.99991.9999) ≈ 0.99995

Summary of the Findings

Metric	Random Forest	Logistic Regression
Precision	0.9997	1
Recall	0.9974	0.9999
F1 Score	0.9985	0.99995
Accuracy	0.9986	0.9999

Discussion

The provided metrics compare the performance of two popular machine learning models, Random Forest and Logistic Regression, on a classification task. Both models have demonstrated impressive results, with Logistic Regression achieving perfect precision and near-perfect recall, accuracy, and F1 score. Random Forest, while not perfect, has also performed exceptionally well, with precision, recall, and accuracy scores exceeding 0.99.

In terms of precision, Logistic Regression has reached the maximum score of 1, indicating that all of its positive predictions are correct. This is a remarkable achievement, suggesting that the model is highly reliable in its positive classifications. Random Forest, with a precision score of 0.9997, is not far behind, showing that it too is very precise in its positive predictions, albeit with a marginally higher rate of false positives compared to Logistic Regression.

When it comes to recall, Logistic Regression again takes a slight lead with a score of 0.9999, meaning it is exceptionally good at identifying all positive instances in the dataset. Random Forest, with a recall score of 0.9974, is also excellent in this regard, missing only a very small fraction of the positive instances.

The F1 score, which balances precision and recall, is also very high for both models. Logistic Regression achieves an F1 score of 0.99995, while Random Forest scores 0.9985. These scores indicate that both models are well-balanced in terms of their ability to make accurate positive predictions and to capture all positive instances.

Finally, in terms of overall accuracy, Logistic Regression has an accuracy score of 0.9999, which is almost perfect. Random Forest's accuracy is slightly lower at 0.9986, but this is still an outstanding result. Accuracy reflects the model's correctness across all predictions, including both positive and negative instances.

Conclusion

In conclusion, both Random Forest and Logistic Regression offer impressive performance for classification tasks, with Logistic Regression showing slightly better results in terms of precision, recall, F1 score, and accuracy based on the provided metrics. However, the best choice between the two depends on the specific requirements of the application, such as the desired balance between precision and recall, the complexity of the dataset, computational considerations, and the need for model interpretability. Additional validation techniques should be employed to ensure the selected model's reliability and generalizability.

Recommendations

Both Random Forest and Logistic Regression have demonstrated exceptional performance based on the provided metrics. Logistic Regression achieves perfect precision and near-perfect scores in recall, F1 score, and accuracy, making it an excellent choice for applications requiring high precision and low false positive rates. Random

Forest, with its balanced performance across precision and recall, and its ability to handle complex datasets and feature interactions, is recommended for scenarios where a more nuanced understanding of the data is required and where model interpretability is less of a concern.

The choice between the two models should ultimately be guided by the specific needs of the application, including the importance of precision versus recall, the complexity of the dataset, computational constraints, and the need for model interpretability. Further evaluation through cross-validation and robustness testing is advisable to ensure the selected model's performance generalizes well to unseen data.

References

Abdallah, R., Gaber, M. M., & Srinivasan, B. (2020). Machine Learning-Based Fraud Detection in Financial Services: A Systematic Review. *ACM Computing Surveys (CSUR)*, *53*(6), 134.

Bolton, R. J., & Hand, D. J. (2002). Statistical Fraud Detection: A Review. *Statistical Science*, *17*(3), 235-249.

Fawcett, T. (2006). An Introduction to ROC Analysis. *Pattern Recognition Letters*, *27*(8), 861874.

Krawczyk, B. (2016). Learning from Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A Comprehensive Survey of Data MiningBased Fraud Detection Research. arXiv preprint arXiv:1009.6119.

Poojitha, S., & Malathi, K. (2022). An Innovative Method to Enhance the Accuracy of Credit Card Fraud Detection Using Logistic Regression Algorithm by Comparing Random Forest Algorithm. *ECS Trans.*, *107*, 14205. https://doi.org/10.1149/10701.14205ecst

Sharma, P., Banerjee, S., Tiwari, D., & Patni, J. C. (2021). Machine Learning Model for Credit Card Fraud Detection - A Comparative Analysis. *The International Arab Journal of Information Technology*, *18*(6), 789.

Tucker, J. (2019). Financial Fraud Detection Using Machine Learning. *Journal of Big Data*, *6*(1), 1-19.