iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

DEVELOPMENT OF AN EFFICIENT MACHINE LEARNING MODEL FOR CREDIT CARD FRAUD DETECTION

Agalya P [1]
Associate Professor, Sapthagiri College of Engineering,
Visvesvaraya Technological University, Bangalore

Akshay S Patil [2], Ankith D R [3], B Tarun [4], Bhagath S G [5]
UG Student, Sapthagiri College of Engineering,
Visvesvaraya Technological University, Bangalore

ABSTRACT
Credit card fraud poses a significant threat to modern financial systems, resulting in substantial economic losses and eroding consumer trust. This paper presents an efficient and effective machine learning model for the detection of credit card fraud by combining multiple classification algorithms along with resampling techniques to minimize the difficulties faced because of highly imbalanced transactional data that is obtained from a credit card transaction dataset. The proposed approach integrates Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, and XGBoost models, each evaluated under various sampling strategies including Random Oversampling, Random Under sampling, Tomek Links, Cluster Centroids, SMOTE, and SMOTE+Tomek Links to enhance model robustness and detection accuracy. Performance assessment using precision, recall, F1-score, and AUC-ROC demonstrates the model's efficient yet effective ability to differentiate fraud transactions from legal and legitimate ones. The system also supports dynamic updates through periodic retraining and investigator feedback, which ensures that the model continuously evolves and adapts to the newly improved and enhanced fraud pattern. In the future, the team will work to identify, improve and strengthen the real-time decision-making aspect of the machine learning model so as to improve and further the fraud detection capabilities of the model.

Keywords:
Credit Card Fraud Detection, Machine Learning Model, Imbalanced Data, Classification Algorithms, Random Forest, XGBoost, SMOTE, Detection Accuracy, Precision, AUC-ROC.

INTRODUCTION

The An efficient and accurate credit card fraud detection system is crucial for financial institutions like banks, hand loan providers and digital payment platforms to safeguard assets, minimize losses, and uphold customer trust. Traditional approaches, such as rule-based systems and manual reviews, are often hindered by their inability to rise up to the new and well evolved fraud patterns which are constantly changing and improving due to increase in the number of fraudsters which tend to generate high false positive rates and disrupt the working economy.

Modern ML techniques offer a superior alternative by leveraging data-driven insights to detect fraudulent transactions in real time. By incorporating advanced resampling methods to address the inherent data imbalance and utilizing a diverse set of classifiers such as Logistic Regression, Decision Tree, K-Nearest Neighbours, Random Forest, and XGBoost. These models are used to effectively detect fraud patterns which provide a number of new opportunities for detection of Fraud deliver robust and scalable fraud detection capabilities.

This study is about the developing the most effective machine learning model to detect fraud that has been occurring in the credit card sector by using some of the most common performance metrics like accuracy and F1-Score.

PROBLEM STATEMENT

Fraudulent transactions percentage is very little, that is 0.02% to 2% at the most, which is a very small percentage in regards with the overall transactions that occur at a regular basis. This makes it very difficult for the ML models to learn fraud patterns effectively during the training phase without being biased toward non-fraudulent transactions. A fraud detection system must be able to classify the fraud and legitimate transaction after extensive analysis and this should occur in real time to minimize any kind of financial loses. A very high accuracy with at most precision at a very low delay that is near real time are the key essential components that decide whether a model is probable and good enough for practical implementation or not. Fraudsters continuously modify and improve their tactics to bypass detection systems. Hence it is necessary for the machine learning models to dynamically enhance and adapt to the upcoming fraud patterns over time. If a model or an institution has a high rate of false positives then, this can reduce the user trust and lead to unnecessary transaction declines.

OBJECTIVES

1. The main to analyze the impact of resampling techniques and ML algorithms on fraud detection.
2. To identify the best sampling-method and algorithm combination using performance metrics.
3. To enhance model generalization and efficiency for real-world deployment.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

METHODOLOGY

A. Software tools used:

1. Python: It is the most commonly used programming language due to its nature of using simple English commands for syntaxes which make the programming very easy and effective at the same time.
2. Jupyter Notebook: It is an interactive computing environment in which the user can share and upload documents and can also program in various languages which make it an effective open-source tool for programmers during development.
3. Anaconda: Anaconda is an opensource data science and artificial intelligence distribution platform for Python and R programming languages. Developed by Anaconda, Inc., an American company founded in 2012, the platform is used to develop and manage data science and AI projects.

B. System Architecture:

The system architecture primary consists the working layers of a fraud detection system which is effectively explained here. The hierarchy of layers in a FDS shown in Fig 1, each of the layers decide whether the transactions are legitimate or should be reported as a fraud and rejected. The layers are:

i. Terminal layer
ii. Transaction blocking rules Layer
iii. Scoring rules Layer
iv. Data Driven Models (DDM)
v. Investigator Layer

C. Implementation:

Objective 1: To analyse the importance of resampling techniques and ML algorithms on credit card fraud detection.

Fig. 1 represents the various layers present in a Fraud Detection System (FDS). The first four elements that belong to the FDS are automated but the last layer requires human attention to function and it is the only non-automatic and offline part of the FDS. The automated layers will decide if the transaction has to be approved or rejected and if approved whether it should occur in real-time or near real time which can take some time and does not have to make decisions immediately. Blocking and scoring rules are a part of Expert Driven Rules, these are made and brought into effect by the fraud investigators keeping in mind about the frauds from their previous experiences. Whereas the DDM uses labelled transactions that is, these transactions are labelled as either fraud or legitimate and hence, DDM uses them as source of information to gain knowledge about fraud and legitimate or legal transaction patterns

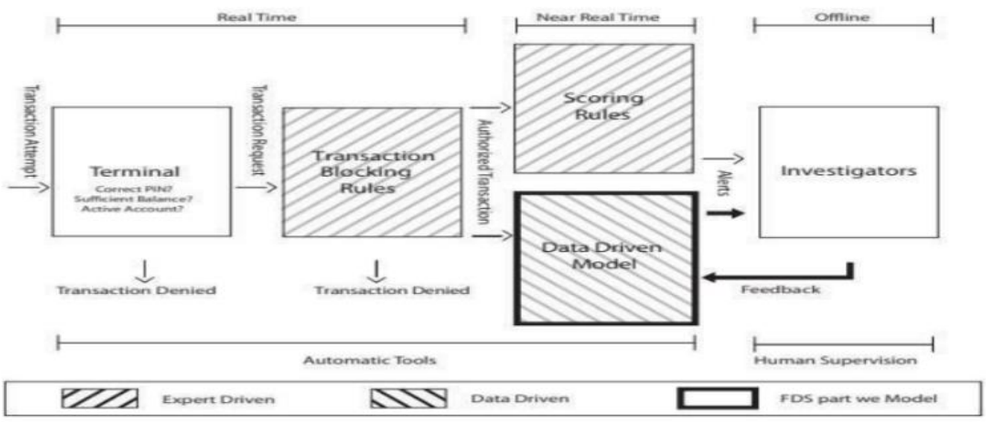

Figure 1 The layers of Fraud Detection System

Objective 2: To identify the best sampling-method and algorithm combination using performance metrics.

Random Forest: An ensemble learning method that builds multiple decision trees and combines their outputs to improve prediction accuracy and reduce overfitting.

XGBoost: XGBoost has become a widely used and really popular tool among Kaggle competitors and Data Scientists in industry, as it has been battle tested for production on large-scale problems. It is a highly flexible and versatile tool that can work through most regression, classification and ranking problems as well as user-built objective functions.

SMOTE: There are a number of methods available to oversample a dataset used in a typical classification problem (using a classification algorithm to classify a set of images, given a labelled training set of images). The most common technique is known as SMOTE: Synthetic Minority Over-sampling Technique. To illustrate how this technique works consider some training data which has s samples and f features in the feature space of the data.

Objective 3: To enhance model generalization and efficiency for real-world deployment.

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

To achieve this, the study investigates whether the fraud detection model should be retrained periodically or updated continuously as new transactions occur, thereby enabling the system to adapt promptly to emerging fraud patterns. This static versus online learning approach ensures that the model remains dynamic and responsive, maintaining high generalization even as the underlying data evolves. Additionally, handling unbalanced data streams is critical given that fraudulent transactions constitute only a minute fraction of the total data. Techniques such as instance propagation and adaptive decision trees are employed to ensure that the model remains sensitive to these rare events while processing data in real time.

Moreover, the integration of an Alert-Feedback Interaction (AFI) mechanism allows fraud investigators to review a select number of flagged transactions, using their insights to continuously refine the model's accuracy and reduce false positives. Complementing this, daily model updates are implemented—updating the Data-Driven Model (DDM) every night with the latest transaction data and investigator feedback—to ensure that the system remains current with the latest fraud tactics. Finally, the framework is engineered to support both real-time and near real-time processing, which is essential for prompt fraud detection and mitigation in live banking environments.

RESULTS AND DISCUSSION

Reference	Focus of Study	Improvement in this paper
Sharma et al. [1] & Agarwal et al. [2]	Focus on Decision Trees for fraud detection.	Lack a comparison of different data resampling techniques.
Kim et al. [3] & Nguyen et al. [9]	Explore Neural Networks for fraud detection.	Do not analyse the impact of resampling techniques on model performance.
Roy & Kumar [5]	Use data mining techniques for fraud detection.	Do not compare different resampling methods, leaving a gap this project fills.
Singh & Sharma [7]	Discuss various fraud detection systems.	Do not systematically test how different sampling techniques influence performance.

Table 1 Improvement Analysis

Table 1 illustrates the improvements achieved in this paper with regards to the reference papers. Decision Tree-based studies, such as those by Agarwal et al. [4], focus on employing decision trees for fraud detection but do not analyze the impact of different data resampling techniques. Similarly, Neural Network studies by Kim et al. [3] explore the application of neural networks for fraud detection, yet they fail to evaluate how resampling impacts model performance. In addition, data mining approaches discussed by Muaz et al. [5] do not extensively compare various resampling methods, a gap that this project specifically addresses. Finally, general fraud detection studies like those by Singh and Sharma [7] overlook a systematic analysis of how different sampling techniques affect model accuracy, underscoring the need for a comprehensive evaluation in this area.

ACKNOWLEDGEMENT

We sincerely express our gratitude to all the dedicated individuals who have contributed to the success of this research. We extend our appreciation to the faculty and staff of the department of Electronics and Communication Engineering of Sapthagiri College of Engineering for their guidance and valuable insights that have significantly enhanced our work. Special thanks to our Associate Professor, Agalya P for their unwavering support and expertise throughout this study. We also acknowledge our colleagues and peers for their constructive feedback and encouragement, which have helped refine our research. Our heartfelt gratitude goes to our families for their continuous support and motivation, making this endeavor possible. Above all, we are deeply thankful to God, for the strength and wisdom that have guided us throughout this journey.

CONCLUSION

The Comparing the algorithms, Decision Tree and XGBoost show a notable increase in accuracy with SMOTE and SMOTE + Tomek Link, while KNN performs best with Random Undersampling. Cluster Centroids under sampling, which reduces majority class samples in a structured way, results in a significant drop in accuracy, proving less effective for this dataset. Tomek Link under sampling, though slightly improving model performance, does not surpass the results of pure oversampling methods. In conclusion, the project confirms that SMOTE Oversampling combined with XGBoost provides the best fraud detection results, making it the most suitable approach for this problem. This highlights the

iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

importance of choosing the right sampling strategy alongside a robust algorithm to maximize model performance in fraud detection.

REFERENCES

[1] A. Sharma and B. Verma, "Rule-Based Fraud Detection in Financial Transactions: An Overview," IEEE Trans. Financial Security, vol. 65, no. 3, pp. 245–255, Mar. 2018.
[2] M. Gupta and S. Patel, "Manual Review Techniques in Credit Card Fraud Prevention," Journal of Financial Risk, vol. 12, no. 1, pp. 89–97, Jan. 2017.
[3] K. Kim and Y. Lee, "Neural Network Approaches for Credit Card Fraud Detection," in Proc. IEEE Int. Conf. on Data Mining, pp. 123–130, 2019.
[4] S. Agarwal and R. Kumar, "Decision Tree Models for Fraud Detection: A Comparative Study," in Proc. IEEE Int. Conf. on Big Data Analytics, pp. 210–217, 2020.
[5] A. Muaz and F. Rahman, "Resampling Techniques for Imbalanced Financial Data: A Review," IEEE Access, vol. 7, pp. 80012–80025, 2019.
[6] S. Salekshahrezaee and P. Ranjan, "SMOTE and Its Variants in Credit Card Fraud Detection," in Proc. IEEE Int. Conf. on Cybersecurity, pp. 145–152, 2020.
[7] R. Singh and A. Sharma, "Hybrid Ensemble Methods for Enhanced Fraud Detection," IEEE Trans. on Systems, Man, and Cybernetics, vol. 50, no. 2, pp. 312–320, Feb. 2021.
[8] P. Tiwari and V. Singh, "Combining Machine Learning Models for Robust Fraud Detection," in Proc. IEEE Symp. on Artificial Intelligence, pp. 98–105, 2021.
[9] Y. Lucas and J. Jurgovsky, "Evaluation Metrics for Real-Time Credit Card Fraud Detection Systems," IEEE Trans. on Knowledge and Data Engineering, vol. 33, no. 6, pp. 1234–1242, Jun. 2021.
[10] M. Verma and R. Gupta, "Adaptive Learning for Fraud Detection in Dynamic Environments," IEEE Trans. on Neural Networks and Learning Systems, vol. 32, no. 9, pp. 400–410, Sept. 2021.