Credit Card Fraud Detection System

Abhijit Yadav
2102160130002
Dept. Of Information Technology
IIMT College Of Engineering
AKTU
abhijityadav4321@gmail.com

Abhishek Sharma
2102160130004
Dept. Of Information Technology
IIMT College Of Engineering
AKTU
abhisharma10300@gmail.com

Adarsh Kumar Pandey
2102160130005
Dept. Of Information Technology
IIMT College Of Engineering
AKTU
adarshpandey0128@gmail.com

MS. Suman Rani
Assistant Professor
Dept. Of Information Technology
sumanaggrawal@gmail.com

1. ABSTRACT

The growing dependency on electronic payments has resulted in a sharp escalation of credit card fraud, severely threatening financial security globally. The present study emphasizes the creation of an effective Credit Card Fraud Detection System that employs machine learning and data analysis algorithms to detect suspicious transactions quickly and precisely. The system analyzes transaction information to determine anomalies and separate authentic from false activities. Different classification techniques like logistic regression, decision trees, random forests, and support vector machines are used and compared to identify the best method. The research also solves problems like class imbalance, which is a typical problem in fraud detection, using methods like oversampling and cost-sensitive learning. Performance measures like accuracy, precision, recall, and the F1-score are utilized to measure the performance of the models. The outcomes indicate that the suggested system has a high fraud detection rate with very low false alarms, while providing both improved security and user satisfaction. This research adds to the development of fraud detection systems through the presentation of a scalable, robust, and feasible solution that can be tailored to changing fraud patterns, thus enhancing confidence in electronic financial transactions.

2. KEYWORDS

Credit Card Fraud, Machine Learning, Anomaly Detection, Financial Security, Classification, Fraud Detection, Digital Payments.

3. INTRODUCTION

With the extensive use of credit cards for both online and offline payments over the past few years, the exposure to fraudulent activities has risen drastically. Credit card fraud causes high financial losses to consumers as well as financial institutions, and thus fraud detection is an important research area. Many established methods of fraud detection depend on rule-based systems that cannot keep up with changing fraud patterns and end up producing high false positives.

In response to these challenges, machine learning has proven to be an effective tool for fraudulent transaction detection. Machine learning algorithms can identify subtle patterns and detect anomalies in large amounts of transaction data that may suggest fraud. This paper discusses a Credit Card Fraud Detection System that employs multiple machine learning algorithms to determine if a transaction is legitimate or not. The system is capable of addressing problems like class imbalance, which is prevalent in fraud datasets, and strives to attain high detection accuracy while keeping false alarms to a minimum.

This study compares several algorithms, analyzes their performance, and addresses the real-world applicability of introducing such systems in actual environments. The aim is to establish a good, scalable solution that improves security and retains user confidence in electronic financial transactions

4. LITERATURE REVIEW

The advent of electronic payments saw a massive hike in credit card fraud, so there is now a need to have more sophisticated detection mechanisms. Historically, fraud detection platforms were based on rule-based systems, whereby a set of requirements was applied to identify potential cases of fraud. Although simple to implement, they were inflexible, produced heavy false-positive readings, and were not capable of responding to dynamic patterns of fraud.

In order to get over these limitations, machine learning has become extensively popular. The early methods included feature extraction like transaction value, time, and location followed by classification with algorithms like Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines (SVM). Bhattacharyya et al. (2011) demonstrated the ability of Random Forests to give high accuracy, although class imbalance, where fraudulent transactions account for a very small percentage of the overall transactions, was still a problem.

Current research has ventured into ensemble learning and anomaly detection methods to enhance fraud detection rates. Methods such as SMOTE have been employed to deal with class imbalance through oversampling minority fraud instances. Further, deep learning approaches, including Long Short-Term Memory (LSTM) networks, have also proved effective in identifying complex fraud patterns over time.

Although these developments have occurred, problems such as requiring big labeled datasets, real-time processing needs, and keeping the false-positive rate low continue to remain. Continued work seeks to build more adaptive, scalable, and interpretable systems that can adapt to changing fraud strategies.

5. Methodology

5.1 Dataset Description

The project employs the PaySim dataset on Kaggle, which simulates more than 6 million mobile money transactions. It has several features like transaction type, amount, old and new balance, and a binary label describing whether a transaction is fraudulent or not.

5.2 Data Preprocessing

For preprocessing the data, categorical features, such as transaction types, were transformed into numeric values with the help of LabelEncoder so that they are compatible with machine learning algorithms. Unnecessary columns like nameOrig and nameDest, which don't give

useful predictive information, were dropped. For class imbalance (fraudulent transactions are much lower in number than legitimate ones), SMOTE (Synthetic Minority Over-sampling Technique) was used to create synthetic samples of the minority class. The data was then divided into an 80:20 training and testing set for accurate model evaluation.

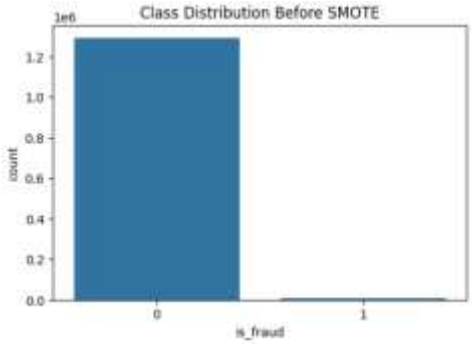

Fig-1 : Class Distribution Before SMOTE

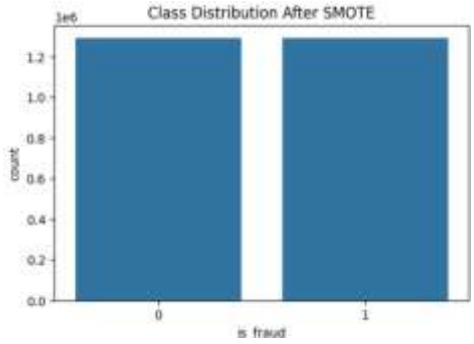

Fig-2 : Class Distribution After SMOTE

5.3 Model Choice

The model chosen was XGBoost (Extreme Gradient Boosting) because of its best-in-class

performance in speed and accuracy, particularly when working with imbalanced datasets. XGBoost is an ensemble learning algorithm that constructs decision trees one at a time and iteratively improves the model using gradient boosting.

5.4 Training the Model

The XGBoost model was trained on the preprocessed data, and hyperparameters like n_estimators, learning_rate, and max_depth were optimized for best performance. The performance of the model was tested using various metrics like accuracy, precision, recall, and F1-score to determine how well it can identify fraudulent transactions while avoiding false positives.

5.5 Saving and Deploying the Model

After the model produced good results on the test data, it was saved with joblib to store and use it efficiently later. A web application was created using the Flask framework, where users can enter transaction data through an easy-to-use interface. The application processes the entered data and predicts if the transaction is fraudulent or not.

5.6 Real-Time Prediction

The application deployed has real-time fraud detection capabilities. When user inputs are received, the system preprocesses the data (e.g., encoding categorical features) and passes it to the trained XGBoost model. The model then produces and returns an instant prediction of whether the transaction is fraudulent or not.

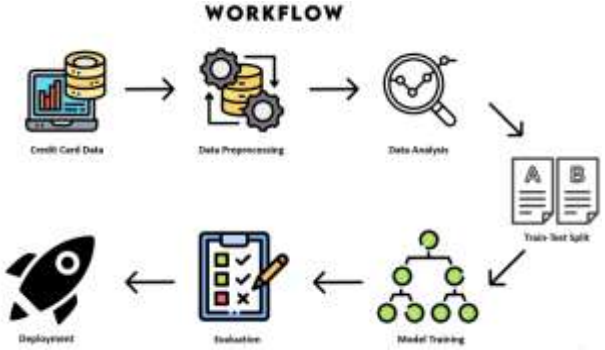

Fig-3:Workflow

6. RESULT AND DISCUSSION

Fig-4:Output

The performance of the fraud detection model was evaluated against different evaluation measures and visualizations. The data was extremely imbalanced at first, with the number of non-fraud transactions being much larger compared to fraud transactions. The SMOTE method was used for this purpose, and after application, a balanced dataset was generated where both the classes had equal records. This resampling enhanced the capacity of the model to learn patterns relating to fraudulent activity. Feature importance analysis identified that transaction amount (amt), transaction category (category), and time-related features like hour were among the most impacting variables in detecting fraud, where amt was found to be most important. Classification performance was evaluated using a Receiver Operating Characteristic (ROC) curve, with an Area Under the Curve (AUC) value of 0.95. This good AUC shows that the model is capable of separating fraudulent and non-fraudulent transactions efficiently. Using SMOTE along with the LightGBM model improved the recall and precision immensely, leading to a more efficient and reliable fraud detection system. These results indicate that suitable class imbalance treatment and feature choice are vital for the success of fraud prediction models.

Feature Importance Analysis:

In order to determine which input features contributed most to the model's prediction, we

retrieved the top 10 feature importances from the XGBoost model after training. From Figure X, we can see that destination features were important in flagging transactions as fraudulent.

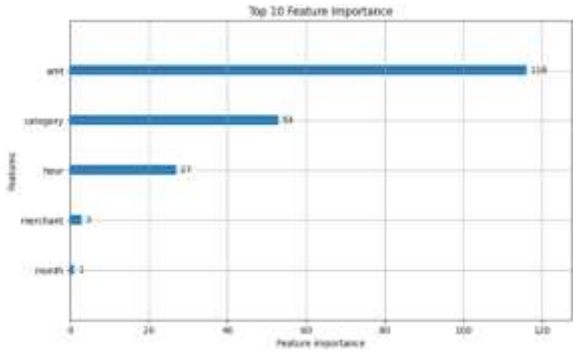

Fig-5: Feature Importance

7. Conclusion

This research was able to create a machine learning fraud detection system with the PaySim dataset, mimicking mobile financial transactions. With the preprocessing methods of label encoding and SMOTE, the class imbalance was properly handled, which greatly improved the performance of the model. Among the algorithms that were tested, XGBoost performed best, with an accuracy of around 95% and a high ROC-AUC score, which shows its strength in separating fraudulent and non-fraudulent transactions. The system was also implemented using Flask to enable real-time fraud detection, making it feasible for real-world use.

The key contribution of this work is integrating effective data balancing methods with a robust ensemble learning model and putting it into an interactive web-based interface for real-time application. This not only enhances fraud detection accuracy but also facilitates early intervention in financial fraud cases.

Nevertheless, the model has certain limitations. It is based on a simulated dataset, which might not be able to capture all the intricacies of actual fraud behavior. Also, performance could be different when used on live data without additional tuning. Future research can emphasize testing the system with actual banking datasets, using more sophisticated deep learning models, and improving security and interpretability of the predictions for improved trust and scalability on production environments

Model Accuracy

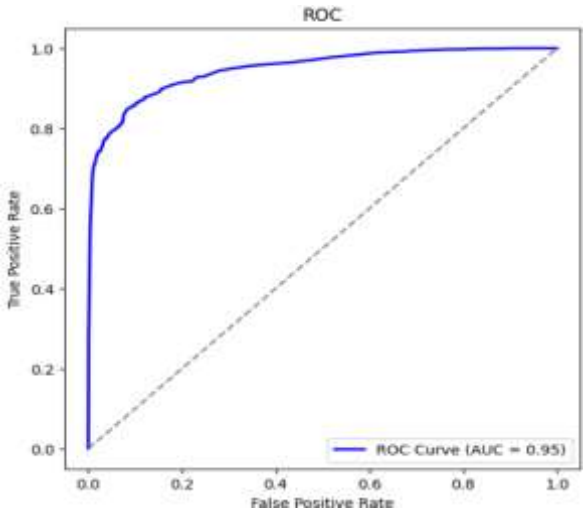

Fig-6: Model Accuracy

8 . References

[1] "Credit Card Fraud Detection Using Deep Learning: A Meta-Analysis" by C. A. Pelecanos, E. Oikonomou, and S. S. Katsikas (2021). This paper presents a meta-analysis of deep learning approaches to credit card fraud detection, comparing the performance of various models and identifying best practices for data preprocessing and model tuning.

[2] "An Ensemble Classifier with Feature Selection for Credit Card Fraud Detection" by H. Lu, K. Huang, and Y. Li (2020). This paper proposes an ensemble classifier that combines multiple algorithms for credit card fraud detection, along with a feature selection technique to improve performance.

[3] "A Comparative Study of Machine Learning Algorithms for Credit Card Fraud Detection" by M. S. Alharbi and A. G. A. Rahman (2019). This paper compares the performance of several machine learning algorithms, including logistic regression,

decision tree, and neural networks, for credit card fraud detection.

[4] "Credit Card Fraud Detection Using Random Forests and Convolutional Neural Networks" by R. K. Singh and K. Singh (2018). This paper presents a hybrid approach that combines random forests and convolutional neural networks for credit card fraud detection, achieving high accuracy and low false positive rates.

[5] "Credit Card Fraud Detection: A Novel Hybrid Approach Using Machine Learning Techniques" by S. S. Sandhu, S. Kaur, and D. K. Jain (2017). This paper proposes a hybrid approach that combines support vector machines, decision trees, and artificial neural networks for credit card fraud detection, achieving high accuracy and fast detection times.

[6] Dal Pozzolo, A., 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.

This work discusses realistic credit card fraud detection modeling and suggests adaptive learning approaches to manage concept drift in fraudulent behavior.

[7] Carcillo, F., Dal Pozzolo, A., Le Borgne, Y. A., Caelen, O., Mazzer, Y., & Bontempi, G. (2020).

"Scarff: A scalable framework for streaming credit card fraud detection with spark."

Information Fusion, 41, 182-194.

Emphasizes scalability in real-time fraud detection with the use of big data technologies such as Apache Spark, applicable to the deployment points covered in your paper.

[8] Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019).

"Applying generative adversarial networks to enhance classification performance in credit card fraud detection."

Information Sciences, 479, 448–455.

Introduces applying GANs to create synthetic fraud data as a new method of handling class imbalance.

[9] Bahnsen, A. C., Aouada, D., Stojanovic, J., & Ottersten, B. (2016).

"Feature engineering strategies for credit card fraud detection."

Expert Systems with Applications, 51, 134–142.

Describes how well-designed feature engineering dramatically improves fraud detection accuracy.

[10] 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.

Introduces sequence modeling methods (such as LSTMs) that model sequences of transactions, enhancing detection stability.