Credit Card Fraud Detection Using Machine Learning

Eershan Reza Khan
Electrical & Computer Engineering
North South University
Dhaka, Bangladesh
eershan.khan@northsouth.edu

Mohammed Nafees Imtiaz
Electrical & Computer Engineering
North South University
Dhaka, Bangladesh
nafees.imtiaz@northsouth.edu

Most Yeanur Akhter
Electrical & Computer Engineering
North South University
Dhaka, Bangladesh
yeanur.akhter@northsouth.edu

Most Aysha Siddika Sumona
Electrical & Computer Engineering
North South University
Dhaka, Bangladesh
most.sumona@northsouth.edu

***Abstract*— Credit card fraud causes billions in annual losses worldwide. Traditional systems lack adaptability to evolving fraud patterns. This paper proposes a supervised machine learning-based fraud detection model utilizing the Kaggle credit card dataset. Techniques such as data preprocessing, feature engineering, and imbalanced learning (SMOTE, undersampling) are employed to improve model accuracy and recall. The system is built using Python-based tools and achieves ~99.7% accuracy with ~89% recall for fraudulent transactions.**

I. Introduction

Credit card fraud has become a significant global issue, resulting in billions of dollars in losses annually. According to a Nilson Report, global card fraud losses reached $32.3 billion in 2021 and are projected to exceed $40 billion by 2030. This rapid rise underscores the urgent need for intelligent, real-time detection mechanisms. While traditional rule-based systems are still in use, they lack adaptability to novel fraud patterns and often generate high false positives. Machine learning introduces dynamic capabilities, offering robust, data-driven approaches. This report aims to design a fraud detection model that leverages supervised learning algorithms to address the challenges of class imbalance, large-scale data processing, and predictive accuracy.

II. Literature review

Numerous studies have addressed credit card fraud detection through machine learning. For instance, Dal Pozzolo et al. (2017) proposed a realistic modeling framework that accounts for concept drift. In contrast, Carcillo et al. (2018) utilized active learning to dynamically query models. Traditional models like Logistic Regression and Decision Trees are simple and interpretable but perform poorly on highly imbalanced data. On the other hand, Random Forests and XGBoost offer higher accuracy and are less prone to overfitting. Recent advancements have focused on ensemble learning and deep learning methods, such as autoencoders and LSTMs. Despite these innovations, many models still suffer from low recall, particularly in detecting rare fraud events. The literature emphasizes the necessity of resampling techniques, cost-sensitive learning, and hybrid models to combat this issue.

III. Methodology

A. Software Requirement

The implementation was conducted using Python in Jupyter Notebook and Google Colab. Key libraries used include Pandas, NumPy for data processing; Scikit-learn for modeling; Matplotlib and Seaborn for visualization.

B. Dataset Description

The dataset used is from Kaggle's Credit Card Fraud Detection collection, containing 30 anonymized features including transaction amount and time. The target label is binary: 1 for fraud, 0 for normal transactions. Class imbalance was addressed using SMOTE and random undersampling.

C. Data Preprocessing

Preprocessing steps included scaling the features, handling missing values, balancing class distributions, and splitting data into training and testing sets. TF-IDF was not required as the features were numeric.

D. Classification Models

Three models were implemented: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC. Performance was analyzed with and without class balancing.

IV. Result Analysis

Our Random Forest model achieved the best performance with 99.7% accuracy and approximately 89% recall, which is crucial in minimizing false negatives. Logistic Regression and SVM also demonstrated competitive results with optimized hyperparameters. Class balancing using SMOTE significantly improved the F1-score.

V. conclusion

This study demonstrates that supervised machine learning, combined with proper data preprocessing and balancing, can be highly effective for credit card fraud detection. With continued refinement, these models can be deployed in real-time systems to combat financial fraud proactively.

VI. FUTURE WORK

Planned extensions include deploying the model via an API for real-time inference, integrating deep learning techniques (e.g., LSTM), and implementing regular re-training strategies using incoming transactional data.

.

REFERENCES

[1] A. Kumar Das et al., "Bangla hate speech detection on social media using attention-based recurrent neural network," Journal of Intelligent Systems, vol. 30, no. 1, pp. 578–591, 2021.

[2] M. M. Nabi et al., "Detecting sentiment from Bangla text using machine learning technique and feature analysis," International Journal of Computer Applications, vol. 153, no. 11, pp. 28–34, 2016

.

[3] B. Mathew et al., "Analyzing the hate and counter speech accounts on Twitter," arXiv preprint arXiv:1812.02712, 2018.

[4] Scikit-learn Documentation. Available: https://scikit-learn.org

[5] Kaggle: Credit Card Fraud Detection Dataset. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud