Payment Fraud Detection — Project Report

# Abstract

This report details the development of a machine learning model for detecting fraudulent payment transactions. The project utilized a large transactional dataset and explored three different classification models: Logistic Regression, Decision Tree, and XGBoost. The models were trained and evaluated on a dataset preprocessed with one-hot encoding and balanced using the Synthetic Minority Over-sampling Technique (SMOTE). The XGBoost model demonstrated superior performance, achieving a near-perfect accuracy of 99.6. The final model was deployed in a user-friendly web application using Streamlit, enabling users to upload their data and receive fraud predictions.

# 1. Introduction

Payment fraud detection is a critical task for financial institutions and payment service providers. This project implements a lightweight Streamlit-based dashboard for predicting fraud in transaction data and demonstrates preprocessing, model loading, and prediction steps.

# 2. Dataset

The dataset used in this project is a large transactional dataset containing 6,362,620 rows and 10 columns. The features in the dataset include:

* step: Represents a unit of time in the real world.
* type: The type of transaction (e.g., PAYMENT, TRANSFER, CASH\_OUT).
* amount: The amount of the transaction.
* nameOrig: The customer who started the transaction.
* oldbalanceOrg: The initial balance of the sender's account.
* newbalanceOrig: The new balance of the sender's account after the transaction.
* nameDest: The customer who is the recipient of the transaction.
* oldbalanceDest: The initial balance of the recipient's account.
* newbalanceDest: The new balance of the recipient's account after the transaction.
* isFraud: A binary variable indicating whether the transaction is fraudulent (1) or not (0).

# 3. Data Preprocessing

The 'type' column, which is categorical, was converted into a numerical format using one-hot encoding. This is a necessary step for most machine learning algorithms to process the data correctly.

The dataset was highly imbalanced, with a very small percentage of fraudulent transactions. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training data. SMOTE creates synthetic samples of the minority class (fraudulent transactions) to balance the class distribution, which helps in preventing the model from being biased towards the majority class.

# 4. Models

Three different machine learning models were trained and evaluated:

1. Logistic Regression: A linear model that is commonly used for binary classification tasks. It is simple, interpretable, and computationally efficient.
2. Decision Tree Classifier: A non-linear model that creates a tree-like structure of decisions to classify the data. It is easy to visualize and understand.
3. XGBoost Classifier: An advanced and powerful gradient boosting algorithm that is known for its high performance and accuracy. It builds an ensemble of decision trees and combines their predictions to make the final classification.

# 5. Results and Discussion

The models were evaluated based on their accuracy, precision, recall, and F1-score. The performance of the models on the test set is summarized below:

|  |  |
| --- | --- |
| Model | Accuracy |
| Logistic Regression | 0.942 |
| Decision Tree | 0.995 |
| XGBoost | 0.996 |

The XGBoost model significantly outperformed the other two models, achieving a near-perfect accuracy of 0.99. The detailed classification report for the XGBoost model is as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-score | Support |
| 0 (Not Fraud) | 1.00 | 1.00 | 1.00 | 1,906,351 |
| 1 (Fraud) | 0.99 | 1.00 | 0.99 | 762,540 |
| Accuracy |  |  | 1.00 | 2,668,891 |
| Macro Avg | 0.99 | 1.00 | 1.00 | 2,668,891 |
| Weighted Avg | 1.00 | 1.00 | 1.00 | 2,668,891 |

The results show that the XGBoost model can accurately identify both fraudulent and non-fraudulent transactions with high precision and recall. The high F1-scores for both classes indicate that the model is robust and reliable. The superior performance of XGBoost can be attributed to its ability to handle complex non-linear relationships in the data and its regularization features that prevent overfitting.

# 6. Conclusion and Future Work

This project successfully developed a highly accurate machine learning model for payment fraud detection. The XGBoost model, trained on a SMOTE-balanced dataset, demonstrated excellent performance and can be a valuable tool for financial institutions to combat fraud.

Future work could involve:

* Exploring other advanced models like deep learning and neural networks.
* Incorporating more features and performing advanced feature engineering to further improve the model's performance.
* Deploying the model in a real-time environment to detect fraud as it happens.