Financial Fraud Detection Using Machine Learning

Chitransh Motwani
Spring 2025

Abstract

Financial fraud presents a significant challenge, costing businesses and consumers billions each year. As traditional rule-based fraud detection systems struggle to keep pace with evolving fraudulent behaviors, machine learning (ML) offers a promising alternative. This project aims to harness ML techniques to develop a model capable of accurately identifying fraudulent transactions.

1 Introduction

In today's digital landscape, financial fraud is increasingly prevalent, affecting countless individuals and organizations. Unfortunately, conventional fraud detection methods often fall short against the sophisticated tactics employed by fraudsters. Therefore, this project seeks to create a robust machine learning model that effectively distinguishes between legitimate and fraudulent transactions. To make fraud detection accessible, we will build a user-friendly web interface using Streamlit.
Track: ML Project with Data Exploration Component

2 Problem Statement

The objective of this project is to develop a fraud detection system using a dataset of financial transactions. However, this task presents several challenges:

- Tackling highly imbalanced datasets, as fraudulent transactions are relatively rare.
- Ensuring real-time detection to minimize potential losses.
- Guaranteeing that predictions are interpretable to comply with regulatory standards.

3 Dataset Selection

To kick off the project, I plan to utilize publicly available datasets, including:

- **Kaggle Credit Card Fraud Dataset**: This dataset features anonymized credit card transactions with fraud labels, enabling effective model training.
- **PaySim Synthetic Mobile Transactions**: This dataset simulates mobile payment transactions based on real-world patterns, providing valuable insights for the model.

A crucial part of the data preprocessing phase will involve addressing class imbalance using techniques such as oversampling, undersampling, and the Synthetic Minority Oversampling Technique (SMOTE).

4 Methodology

The project will unfold through several key stages in the fraud detection pipeline:

4.1 Data Preprocessing

Before diving into modeling, I will need to clean and prepare the data:

- Handle missing values and resolve any inconsistencies in the dataset.
- Conduct feature engineering to derive valuable insights (e.g., analyzing correlations between transaction frequency and amounts).
- Address class imbalance through various resampling techniques to ensure the model is well-equipped to identify fraud.

4.2 Model Selection and Training

I will explore a variety of machine learning algorithms to find the best fit for the project's needs:

- **Supervised Learning**: Testing Random Forest, XGBoost, and Logistic Regression for their predictive power.
- **Unsupervised Learning**: Considering techniques like Isolation Forest and Autoencoders for their strengths in anomaly detection.
- **Hybrid Approaches**: Combining multiple methods to enhance the accuracy of fraud detection capabilities.

4.3 Model Evaluation

To ensure optimal performance, I will evaluate the model using metrics effective for imbalanced datasets:

- Metrics such as Precision, Recall, and F1-score will provide insight into the model's effectiveness.
- I will calculate the Area Under Curve - Receiver Operating Characteristic (AUC-ROC) to assess performance.
- A thorough analysis of the confusion matrix will help identify where the model excels and where it may need improvement.

4.4 Explainability and Interpretability

To build trust in the model's predictions, I will incorporate techniques that enhance explainability:

- SHAP values (SHapley Additive exPlanations) will help identify which features are most influential in the decision-making process.
- LIME (Local Interpretable Model-agnostic Explanations) will provide context for individual predictions, making them easier to understand.

4.5 Deployment and Visualization

The fraud detection system will be deployed as an interactive Streamlit web application, allowing users to input transaction details, analyze batch files, and visualize model predictions in real time. I might explore the following technologies:

- An interactive interface for intuitive user input and output.
- Basic visualizations with Matplotlib, with room to consider other libraries as the project evolves.
- Saving the trained model with libraries like joblib or pickle for seamless loading and predictions.
- Depending on progress, additional deployment options or enhanced visualization techniques might be explored.

5 Expected Outcomes

By the end of this project, I aim to deliver:

- A reliable fraud detection model with strong accuracy.
- A comprehensive report detailing the data exploration and strategies employed for fraud detection.
- A functional prototype for detecting fraud, with the potential for future enhancements.
- An analysis of the model's interpretability to ensure transparency in its predictions.
- A fully functional Streamlit-based fraud detection web app.

6 Conclusion

This project aims to create an effective solution for identifying financial fraud using advanced machine learning techniques. By focusing on real-world applications and providing a clear understanding of the model's decision-making process, I hope to make a positive contribution to the field of fraud detection.