ISyE 6740 – Fall 2024
Project Proposal

Team Member Names: Akhil Bhargava, Ankit Kumar, Ankur Asthana

Project Title:

Fraud Detection in Financial Transactions Using Machine Learning

1. Introduction

Fraudulent activities in financial transactions have been a significant challenge for businesses, banks, and other financial institutions. The rapid growth of digital transactions has increased the exposure to potential fraud, making it crucial to develop effective and automated fraud detection systems. This project will utilize a publicly available dataset from Kaggle to build a machine learning model capable of detecting fraud in financial transactions with high accuracy.

The dataset available at Kaggle contains transaction records with a mix of numerical and categorical features, including the target variable that indicates whether a transaction is fraudulent or not.

2. Problem Statement

Fraudulent transactions cause significant financial losses and damage reputations. Early detection of such transactions can prevent fraud from occurring, thus saving financial institutions considerable losses. The challenge is to accurately classify whether a financial transaction is fraudulent based on transaction features while minimizing false positives. However, the data is often highly imbalanced, with fraudulent transactions making up only a small percentage of the total, adding complexity to the classification task.

3. Objectives

The objectives of this project are:
1. To analyze and preprocess the financial transaction data for detecting patterns and trends.
2. To build and compare several machine learning models to classify fraudulent and non-fraudulent transactions.
3. To implement a highly accurate and efficient fraud detection model that can be integrated into a real-time transaction processing system.
4. To evaluate the performance of these models based on accuracy, precision, recall, F1-score, and Area Under the Curve (AUC).

4. Dataset Overview

The Fraud Detection dataset is a simulated credit card transaction dataset available on Kaggle, containing records of legitimate and fraudulent transactions from 1st January 2019 to 31st December 2020. The dataset includes credit card transactions from 1,000 customers across 800 merchants, with a mix of features capturing various

aspects of the transactions.

The dataset was generated using the Sparkov Data Generation tool created by Brandon Harris, and the files were combined and converted into a standard format for ease of use. It consists of two synthetic datasets, namely fraudTrain.csv and fraudTest.csv.
Sample Structure:

Column Name	Description
trans_date_trans_time	Date and time of the transaction
cc_num	Credit card number
Merchant	Merchant where the transaction occurred
Category	Purchase category (e.g., travel, personal care)
Amt	Transaction amount
First	Cardholder's first name
Last	Cardholder's last name
Gender	Gender of the cardholder
street, city, state, zip	Address details of the cardholder
lat, long	Latitude and longitude of the cardholder's location
city_pop	Population of the cardholder's city
Job	Cardholder's job
Dob	Date of birth of the cardholder
trans_num	Unique transaction identifier
unix_time	Unix timestamp of the transaction
merch_lat, merch_long	Latitude and longitude of the merchant's location
is_fraud	Indicator if the transaction is fraudulent (1 for fraud, 0 for legitimate)

5. Methodology

The project will follow these key steps:

5.1 Data Exploration and Preprocessing:

- Exploratory Data Analysis (EDA) will be conducted to understand the distribution of data, feature correlations, missing values, and patterns in fraudulent versus legitimate transactions.
- Data Cleaning: Handling missing values, normalizing the 'Amount' and 'Time' features.
- Feature Selection and Engineering: Find significant features from the dataset and potentially generate new features based on the existing data.
- Creating attributes using PCA to test with different classifiers
- Imbalanced Data Handling: Since fraud detection datasets are usually imbalanced, techniques such as resampling (SMOTE, undersampling), cost-sensitive learning, or using class weights will be explored.

5.2 Model Selection:

We will evaluate several machine learning algorithms, including:
1. Logistic Regression
2. Naive Bayes Classifier
3. Random Forest Classifier
4. Gradient Boosting Machines (XGBoost)
5. Support Vector Machines (SVM)
6. Neural Networks
7. Anomaly detection

5.3 Model Evaluation:
Performance metrics: Accuracy, Precision, Recall, F1-Score, AUC-ROC.

6. Tools and Technologies

The project will use the following tools and technologies:
- Programming Language: Python
- Data Processing Libraries: Pandas, NumPy
- Visualization Libraries: Matplotlib, Seaborn
- Machine Learning Libraries: Scikit-learn, XGBoost, TensorFlow/Keras
- Model Evaluation Tools: ROC curves, Cross-Validation

7. Expected Outcomes

By the end of this project, we expect the following outcomes:
1. A comprehensive report on data analysis, preprocessing, and model-building steps.
2. A machine learning model with high fraud detection accuracy and low false positives.

8. Conclusion

By leveraging machine learning techniques and the available dataset, this project aims to deliver a robust solution for fraud detection in financial transactions. The final solution will improve the accuracy of fraud detection and help the banks as well as customers from the heavy fraud losses.