Credit Card Fraud Detection

Project Proposal for
CS F415 Data Mining Project

BITS Pilani Hyderabad Campus

By

Dhairya Luthra	**2022A7PS1377H**
Animish Agrahari	**2022A7PS1367H**
B. Vaishnavi	**2022A7PS1357H**

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI,HYDERABAD CAMPUS

February,2024

TABLE OF CONTENTS

Team Members and Emails

DHAIRYA LUTHRA f2022177@hyderabad.bits-pilani.ac.in

ANIMISH AGRAHARI f20221367@hyderabad.bits-pilani.ac.in

B. VAISHNAVI f20221357@hyderabad.bits-pilani.ac.in

Project Proposal

1)Who is in your group ?

Dhairya Luthra - 2022A7PS1377H

Animish Agrahari - 2022A7PS1367H

B. Vaishnavi - 2022A7PS1357H

2) What data do you plan to use?

The dataset contains transactions made by credit cards in **September 2013** by **European cardholders**

.

This dataset presents transactions that occurred in two days, where we have **492 frauds out of 284,807 transactions.** The dataset is highly unbalanced, the positive class (frauds) account for **0.172%** of all transactions.

3) What is the problem you are trying to solve?

In our pursuit of addressing the critical issue of fraud detection, our primary focus is to **leverage advanced data mining techniques** for a comprehensive analysis aimed at identifying and preventing fraudulent activities. By harnessing the power of data mining, we aim to provide credit card companies with sophisticated tools and methodologies to scrutinize transactions effectively. The overarching goal is to enable these financial institutions to swiftly **recognize and mitigate the impact of fraudulent credit card transactions**. Our initiative is driven by the commitment to safeguard the financial interests of customers, ensuring they are not unjustly charged for items they did not authorize or purchase.

4) Why is this problem interesting?

Fraud detection is a compelling challenge in the ever-evolving financial landscape, with increasing frequency and sophistication of fraudulent activities. Innovative solutions are crucial due to the dynamic nature of fraudulent tactics.

Data mining, a key player in addressing this challenge, allows for the analysis of vast datasets to **uncover patterns and anomalies** indicative of fraudulent behavior.

Classification, a fundamental data mining technique, helps develop models trained on historical data to **distinguish between legitimate and fraudulent transactions.**

Outlier analysis is integral to our approach, as fraudulent transactions often deviate significantly from the norm. **Employing outlier detection techniques systematically identifies irregularities**, enhancing the accuracy of fraud detection systems.

The significance of this problem lies in the potential financial losses and reputational damage for credit card companies and the erosion of consumer trust in the financial system. Effectively combating fraud through advanced data mining safeguards individual financial interests and contributes to the overall integrity of the financial ecosystem. Our goal is to leverage data mining to stay ahead of evolving fraudulent tactics, creating a safer and more secure financial environment for all stakeholders

5) What is new, you will implement/learn?

We will explore and implement cutting-edge methodologies in classification and outlier analysis to enhance the accuracy and effectiveness of our fraud detection system. Here are some potential approaches:

Classification Models:

Logistic Regression:

- A simple yet effective model for binary classification, Logistic Regression can serve as a baseline algorithm. It's interpretable and well-suited for understanding the contribution of features.

Random Forest Classifier:

- Random Forests are ensemble models that can handle imbalanced datasets well. By aggregating multiple decision trees, they provide robustness and can capture complex relationships within the data.

Support Vector Machines (SVM):

- SVMs are effective in high-dimensional spaces, making them suitable for credit card fraud detection where the feature space might be extensive. Kernel tricks can help capture nonlinear relationships.

XGBoost Classifier:

- XGBoost is an advanced boosting algorithm that can handle imbalanced datasets and is known for its speed and performance. It can capture intricate patterns in the data and has a built-in regularization mechanism.

Outlier Analysis Models:

Isolation Forest:

- Isolation Forest is well-suited for detecting anomalies in high-dimensional datasets. It isolates instances by randomly selecting features, making it effective for identifying unusual patterns indicative of fraud.

Local Outlier Factor (LOF):

- LOF is a density-based algorithm that calculates the local density of instances. It is effective in identifying outliers by flagging instances with significantly lower density, making it adaptable to the unbalanced nature of the dataset.

One-Class SVM:

- Designed for novelty detection, One-Class SVM can identify outliers by constructing a hyperplane that encapsulates the majority of instances. It is particularly useful when dealing with imbalanced datasets.