SML PROJECT PROPOSAL

Shobhit Raj (2022482) & Vashu (2022606)

Project Title

Fraud Detection System for Fintech Banking Transactions

Problem Statement

Financial fraud poses a significant threat to the integrity of FinTech banking transactions, leading to substantial financial losses and erosion of trust among customers. As fintech banking and insurance transactions continue to grow in volume and complexity, there is an urgent need for a robust fraud detection system that can effectively detect and prevent fraudulent activities across multiple channels and domains.

Problem Statement: "Detecting and preventing financial fraud in FinTech banking transactions using statistical machine learning techniques."

Motivation

The motivation behind this project stems from the growing adoption of FinTech solutions in the banking sector, which has led to an increase in sophisticated fraudulent activities targeting financial institutions and their customers. By developing effective fraud detection methods, we aim to safeguard financial transactions, protect customer assets, and maintain the trust and confidence of users in FinTech platforms.

Dataset Details

For this project, we will be working with a comprehensive dataset that encompasses transactional data from both fintech banking transactions. The dataset would include a wide range of features such as transaction amount, transaction type, merchant category, user demographics, device informations etc. Additionally, the dataset would contain labels indicating whether each transaction is fraudulent or legitimate.

The dataset should be divided into training and test sets to train and evaluate the performance of your fraud detection models. Cross-validation techniques such as k-fold cross-validation can also be employed to ensure robust model evaluation.

In our analysis, we'll focus on several key parameters to uncover patterns indicative of fraudulent activities. Firstly, we'll scrutinize transaction amounts, identifying outliers and abnormal spending patterns that may signal potential fraud. Secondly, we'll examine transaction types, including online purchases, ATM withdrawals, and fund transfers, to flag irregularities and suspicious behaviors. Additionally, we'll delve into user behavior patterns such as transaction frequency, time of day, and geographic location, aiming to detect deviations from typical activity which will help in detecting fraud activities like money laundering. Through comprehensive analysis of these parameters, we aim to develop a robust fraud detection system capable of identifying fraudulent activities in fintech banking transactions.

Methods

- **Logistic Regression**: Logistic Regression is a linear classification algorithm that models the probability of a binary outcome. It is interpretable, computationally efficient, and well-suited for problems with linear decision boundaries. Through interpretability and computational efficiency, logistic regression will provide valuable insights into the likelihood of fraud occurrence, allowing us to identify significant predictors of fraudulent activities and make informed decisions in real-time fraud detection scenarios.
- **Random Forest**: The Random Forest (RF) method builds a forest of individual decision trees that collectively constitute an ensemble. Each of the trees makes a prediction on the data by taking majority votes. In turn, the class with the majority vote is decided as the final prediction. Thus, individual uncorrelated models come together to perform the best prediction on the data. Through the technique of bagging, Random Forest will maintain minimal correlation between trees, enabling us to achieve high performance in fraud detection while minimizing overfitting. By analyzing

features such as transaction amount, type, and user behavior patterns, Random Forest will identify anomalous patterns indicative of fraudulent activities and generate accurate fraud alerts for further investigation.

- **AdaBoost**: AdaBoost serves the objective of evolving a strong classifier based on a set of weaker classifiers. The common method used with AdaBoost is the decision tree. It develops a strong classifier with the weighted combination of the set of weak classifiers. First, the algorithm tries to fit the training data on a set of classifiers. Then, it picks the one with the least weighted classification error and updates the weights on other data points. This is done by using a normalization factor that ensures that the sum of all the weights of the data points is equal to 1. Thus, after each iteration, the model attempts to minimize the classification error of the classifiers. This process is repeated until the training dataset is classified appropriately or no further pruning can be carried out on the training dataset.
- **XGBoost**: XGBoost is an ensemble ML algorithm based on the concept of decision trees, similar to Random Forest and other Boosting algorithms. XGBoost achieves significant results for classification problems because it applies the principle of boosting a set of weak trees by using a gradient descent approach. Gradient Boosting generally attempts to weed out the less favorable trees with the aim of minimizing errors with a gradient descent algorithm.

Metrics

1. **Accuracy**: Accuracy measures the proportion of correctly classified transactions out of the total number of transactions. It provides an overall assessment of the model's performance but may not be the best metric when dealing with imbalanced datasets.
2. **Precision**: Precision measures the proportion of correctly identified fraudulent transactions out of all transactions predicted as fraudulent. It indicates the model's ability to avoid misclassifying genuine transactions as fraudulent.
3. **F1-score**: The F1-score is the harmonic mean of precision and recall and provides a balanced measure of the model's performance. It is useful when there is an imbalance between the number of genuine and fraudulent transactions in the dataset.
4. **AUC-ROC**: The area under the receiver operating characteristic curve (AUC-ROC) measures the model's ability to discriminate between genuine and fraudulent transactions across different threshold settings. A higher AUC-ROC value indicates better discrimination.