Problem Statement

Banking fraud is defined as any dishonest act or behavior to obtain privileged banking information without authorization from the user for monetary gain.

Retaining profitable customers takes the top priority for most banks and credit card fraud is a significant threat to this goal. The rise in digital payment platforms is leading to an exorbitant rise in fraudulent transactions that threaten the sanctity of these banks and the trust of customers. Thereby, credit card fraud detection using machine learning is not just a trend but a necessity to proactively monitor and deter its occurrence.

Approach

The dataset used in this study contains credit card transactions for September 2013 by several European cardholders. The dataset, however, is heavily unbalanced; only 0.172% of all transactions are classified as fraud and its features had been transformed using PCA to protect classified information, leaving only amount and time fields as it is.

The objective of the study is to create a model using the given dataset that best identifies fraudulent transactions.

Steps:

1. Data pre-processing and Exploratory Data Analysis
2. Model Building
3. Model Evaluation

1. Data pre-processing

The first step is to perform the preliminary checks such as the presence of null values and distribution of the class variable.

Since PCA transformed variables are already supposed to be Gaussian distributed, I will checking for skewness and outliers on these 28 variables as the models used in this project can be very sensitive to feature distribution. Histograms can be used to check for skewness and they can be treated using PowerTransformer (Box-cox or yeo-Johnson transformation)

Amount and Time fields are not PCA transformed and hence we need to check their distribution and evaluate if they need to be scaled or not.

The next step is to remove any unnecessary fields, I am planning to use box plots and check for correlation with the class variable to make this decision.

2. Model Building

Once the dataset is cleaned and structured, split it into the train and the test data; stratified based on the class variable (Y variable). Stratify will make sure that the test and train data will have the same proportion of fraudulent and genuine transactions. But this won't solve our problem of class imbalance, which is very significant that merely classifying all the transactions as genuine will yield a model with very high accuracy.

Therefore we will be creating models based on these 4 criteria's:

a. Unbalanced dataset
b. Balanced dataset with Random Oversampling
c. Balanced dataset with SMOTE (Synthetic Minority Oversampling Technique)
d. Balanced dataset with ADASYN (Adaptive Synthetic Sampling)

Models used in the study:

- Logistic Regression
- KNN
- SVM
- Decision Tree
- Random Forrest
- XGBoost

The best model among unbalanced dataset models will be considered as the baseline model and the models created using the balanced dataset are expected to outperform the baseline model.

All 6 models will be used with each criterion to find the best model. GridSearchCV or RandomizedSeachCV will be used to tune the hyperparameters of these models with Stratified KFold cross-validation.

3. Model Evaluation

Model evaluation is not as simple as to check for accuracy due to the heavy class imbalance. Instead, we will be adjusting the models to maximize the AUC score. The ROC curve shows the tradeoff between the true positive rate and the false positive rate at various threshold values.

Once the best model is identified using cross-validation, it is used to predict on the test data to evaluate the final model performance. (Precision, Recall, F1-Score)

A cost-benefit analysis will be done on the final model metrics to understand the value of implementing the final model at the bank.