Credit Card Fraud Detection - Viva

Explaining the problem statement & your approach for solving this problem.

Business Problem Overview

For many banks, retaining high profitable customers is the number one business goal. Banking fraud, however, poses a significant threat to this goal for different banks. In terms of substantial financial losses, trust and credibility, this is a concerning issue to both banks and customers alike.

In the banking industry, credit card fraud detection using machine learning is not just a trend but a necessity for them to put proactive monitoring and fraud prevention mechanisms in place. Machine learning is helping these institutions to reduce time-consuming manual reviews, costly chargebacks and fees, and denials of legitimate transactions.

Credit card fraud is any dishonest act and behavior to obtain information without the proper authorization from the account holder for financial gain. Among different ways of frauds, Skimming is the most common one, which is the way of duplicating of information located on the magnetic strip of the card. Apart from this, the other ways are:

Data Dictionary

The data set includes credit card transactions made by European cardholders over a period of two days in September 2013. Out of a total of 2,84,807 transactions, 492 were fraudulent. This data set is highly unbalanced, with the positive class (frauds) accounting for 0.172% of the total transactions.

The dataset masked with PCA still provides us with Time, amount and 28 principal components, but their values are far more difficult to interpret. This dataset is clear example totally unbalanced data.

Project Pipeline

Data Understanding: The data set has also been modified with Principal Component Analysis (PCA) to maintain confidentiality. Apart from 'time', 'class' and 'amount', all the other features (V1, V2, V3, up to V28) are the principal components obtained using PCA. The feature 'time' contains the seconds elapsed between the first transaction in the data set and the subsequent transactions. The feature 'amount' is the transaction amount. The feature 'class' represents class labelling, and it takes the value 1 in cases of fraud and 0 in others.

Exploratory data analytics (EDA): we will perform observe the different feature type present in the data and observe the distribution of our classes. Create a bar plot for the number and percentage of fraudulent vs non-fraudulent transactions. Create a scatter plot to observe the distribution of classes with time and Amount. Drop unnecessary columns

Train/Test Split: we will be splitting the data into train & test data in order to check the performance of your models with unseen data use the k-fold cross-validation method.

Model-Building/Hyperparameter Tuning: This is the final step try different models and fine-tune their hyperparameters until the desired level of performance on the given dataset by class imbalanced with Random oversampling, SMOTE, ADASYN and explore other algorithms on balanced dataset by building the model K – nearest neighborhood, SVM, Decision Tree, Random Forest and XGBoost performing as it is structure data and performing hyperparameter by Cross Validation, hyper parameter tuning

Model Evaluation: Evaluate the models using appropriate evaluation metrics since the data is imbalanced it is more important to identify which are fraudulent transactions accurately than the non-fraudulent with Accuracy of 99.82% by selecting the best hyperparameter on the model, oversampling method which shows the best result on a model.