Credit Card Fraud Detection - Capstone Project

Introduction and Problem definition: There are different kind of financial frauds like.

A. Loan given to ineligible person and it never comes back.

B. Real Customers account details used by fake customer for his personal transaction. Transfer money to his account or buy things for himself or for somebody else.

C. Corrupt Bank officers do favor of transferring money to some corrupt person and take commission for that

D. You accidentally enter some details and money get transferred to some incorrect account. It may happen because of customer's mistake or because customer's receiver provided him incorrect details or post phishing some transaction happens.

International Bank Regulator Organization, Central Bank of Country, Individual Banks, has some rule to define what is fraud and what not. For some fraud transactions a bank owns the responsibility and repay to the customer in other case bank fix responsibility on their customer. Irrespective of who pay for loss bank has rules to mark a transaction as fraud.

In many countries in the worlds you can just swipe the credit card and money is deducted from the account, unlike India where OTP or password need to supplied at transaction point. Because of this reason chances of fraud and consequently loss of money, bank litigation are higher compared to India.

Depending upon bank rules and type of fraud, whether bank refund that money or not there is loss of reputation for bank. Loss of reputation, brand is not easy to measure. Secondly arguing court cases with customer and finally losing that to customer is higher loss for bank in terms of legal fee, pentaly, interest, original money refund, reputation loss and loss of energy.

In the context of India - According to the Reserve Bank of India's (RBI) annual report for 2018-19, there is 15% year-on-year increase in fraud cases. In 2017-18 amount involved has raised from ₹41K to ₹72K crore in 2018-19. Frauds related to advances (90%) were predominant while frauds relating to card, internet and deposits constituted only 0.3% of the total value of frauds in 2018-19, amounting to ₹220 crore. Source

In the light of this big issue banks are in need of a solution which can warn or alarm them before a transaction is fully processed. We know thieves are always ahead of the police therefore it is impossible to make these frauds to zero but in rising trends of online transaction if some system can reduce the number of fraudulent transaction or reduce the loss significantly then banks would love to adopt that solution.

Objective: The aim of this project is to make ML based model(s) which can predict fraudulent credit card transactions before it it finally approved by the bank. Transactions are real time and hence response time is very important, unless banks has some rules to approve with some lag. Therefore apart from making a good model we need to ensure that it works seamlessly with existing system.

Dataset: 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 just 0.172% of the total transactions the value of these transactions is .238% of total value of transactions. The given dataset is encrypted and modified with Principal Component Analysis (PCA) to maintain confidentiality. Apart from 'time' and 'amount', all the other features (V1, V2, V3, up to V28) are the principal components obtained using PCA.

Solution Approach

Data Understanding: It inclues following steps

- Loading dataset
- Understanding datatype of each field
- Range/Variance of each field
- Categories in categorical fiels
- Devising strategy how to perform EDA on the dataset.

Exploratory data analytics (EDA): It includes following steps

- Univariate analysis of every field.
- Bivariate analyses of given fields with respect to target and other fields.

Preprocessing:

- Feature transformation (if required)
- Creating new fields (if required)
- Because the given data is Gaussian nature therefore normalization/standardization is not required
- Handling skewness of data. Using power_transform- yeo-johnson (because data has negative values as well) - if required.
- Handling outliers - if required
- The given dataset is hugely imbalance so it need to be balanced using techniques. ADAptive SYNthetic (ADASYN): The aim here is to create synthetic data for minority examples that are harder to learn, rather than the easier ones.
- **Train/Test Split**: The train/test split will be done in order to check the performance of the model with unseen data. Here, for validation, the k-fold cross-validation method will be deployed. The appropriate value of k value will be chosen so that the minority class is correctly represented in the test folds.

Model-Building: Following algorithms will be tried

- Logistic Regression

- KNN,
- SVM,
- Decision Tree,
- Random Fores,
- XGBoost,
- Neural Network

Metrics Selection: Because this is an imbalance dataset therefore choosing right metrics is important. Recall, Precision, AUC & F1 score are good metrics for this project. Depending upon computing resources one or more of these metrics will be used.

Hyperparameter Tuning: Every algorithm will be tried with various hyper parameters using cross validation. The best hyper parameter will be used to know the best performance of a given model. Depending upon computing resources broad range of hyperparameter will be expertimented.

Final Model Selection

Benefit Measurement (ROI Analysis)