Credit Card Fraud Detection Capstone Project

Introduction:

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. With the rise in digital payment channels, the number of fraudulent transactions is also increasing with new and different ways.

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. The objective of this case study is to develop Machine learning models to help these institutions to reduce time-consuming manual reviews, costly chargebacks and fees, and denials of legitimate transactions.

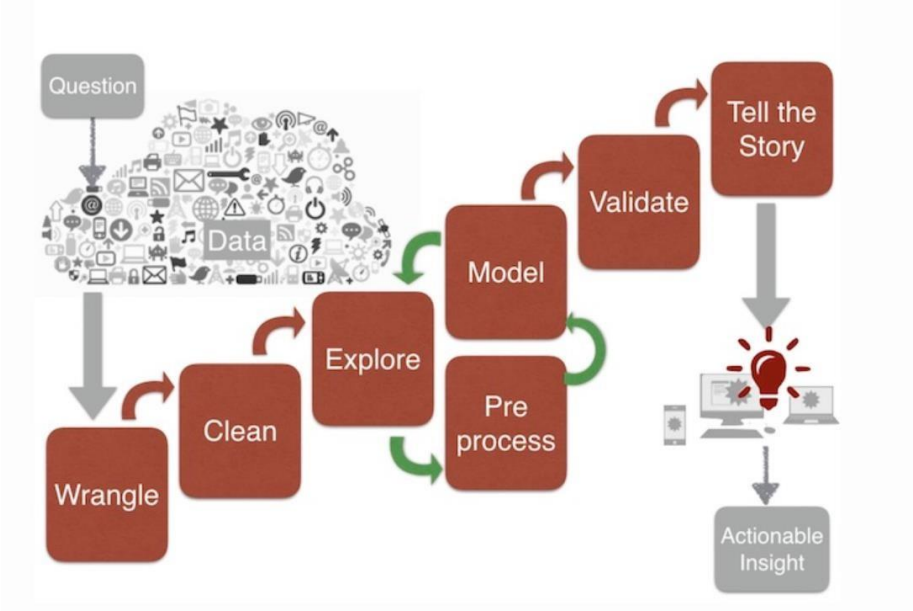
Problem Statement:

The importance of developing a reliable machine learning model to identify fraudulent credit card transactions has long been recognized by the banking industry. However, numerous banks are still plagued by credit card fraudulent transactions, which can be frequently be traced to ineffective identification and treatment of constraints. When a constraint is not properly identified during a suspicious transaction to be able to properly categorize it as an alarming transaction, subsequent frauds are inevitable.

With the increasing number of credit card frauds, manually identifying such transactions can be tedious and leaves a window for human errors. In summary, there is a need for a better understanding of constraints in banking transactions and a structured approach in identifying and modeling constraints to ensure that credit card frauds are curbed.

This case study deals with identifying such fraudulent credit card transactions with the help of various machine learning models.

Problem Approach:



The project pipeline can be briefly summarized in the following four steps:

Data Understanding: We will first load our dataset and try to understand various features present in it.

a. We will check for the shape and the type of data provided.
b. We will check the spread or distribution of our data.
c. We will clean the data by checking the NULL values and remove the discrepancies in the data type if any.

This would help us choose the features that we will need for our final model.

Exploratory data analytics (EDA):

a) In this step, we will perform univariate and bivariate analyses of the data, followed by feature transformations, if necessary.
b) For the current data set, because Gaussian variables are used, we will not perform any scaling.
c) There is skewness in the dataset, and we will handle it using log transformation or Box cox transformation.
d) In this section we will also handle the outliers if any.

Train/Test Split:

a) In this section we will perform the train/test split, in order to check the performance of models with unseen data.
b) For validation, we will use the k-fold cross-validation method. Since there is data imbalance, we would be using stratified K fold method.

Model-Building/Hyper parameter Tuning:

a) This is the final step, where we will create the different models and fine-tune their hyper parameters until we getthe desired level of performance on the given dataset.
b) In this section we will also try various sampling techniques to handle the data imbalance.

Model Evaluation:

In this section, we will evaluate the models using appropriate evaluation metrics. Since there is data imbalance, we would be using metrics like precision, recall, F1-score and AUR-RUC Score to evaluate the performance of the model.