Credit Card Fraud Detection Summary

Problem Statement:

The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the knowledge of the ones that turned out to be fraud. This model is then used to identify whether a new transaction is fraudulent or not. Our aim here is to detect 100% of the fraudulent transactions while minimizing the incorrect fraud classifications.

About Data and Source:

The data set includes credit card transactions made by European cardholders over a period of two days in September 2013.

The dataset is from Kaggle: https://www.kaggle.com/mlg-ulb/creditcardfraud

Project Pipeline:

Data Understanding: Loading the data and understanding the features present in it. This would help us choose the features that are needed for the final model.

Exploratory data analytics (EDA): Performing univariate and bivariate analyses of the data, followed by feature transformations, if necessary. Since Gaussian variables are used, we do not need to perform Z-scaling. However, we need to check if there is any skewness in the data and try to mitigate it, as it might cause problems during the model-building phase.

Train/Test Split: Performing the train/test splitin order to check the performance of our models with unseen data. Here, for validation, weare using the k-fold cross-validation method. We also need to choose an appropriate k value so that the minority class is correctly represented in the test folds.

Model-Building/Hyperparameter Tuning: Choosing different models and fine-tune their hyperparameters until we get the desired level of performance on the given dataset. We should try and see if we get a better model by the various sampling techniques.

Model Evaluation: Evaluate the models using appropriate evaluation metrics.

Observations:

- The data set is highly skewed, consisting of 492 frauds in a total of 284,807 observations. This resulted in only 0.172% fraud cases. This skewed set is justified by the low number of fraudulent transactions.
- The dataset consists of numerical values from the 28 ‘Principal Component Analysis (PCA)’ transformed features, namely V1 to V28. Furthermore, there is no metadata about the original features provided, so pre-analysis or feature study could not be done.
- The ‘Time’ and ‘Amount’ features are not transformed data.
- There is no missing value in the dataset.

Inferences :

- Owing to such imbalance in data, an algorithm that does not do any feature analysis and predicts all the transactions as non-frauds will also achieve an accuracy of 99.82%. Therefore, accuracy is not a correct measure of efficiency in our case. We need some other standard of correctness while classifying transactions as fraud or non-fraud.

- The ‘Time’ feature does not indicate the actual time of the_transaction and is more of a list of the data in chronological order. So we assume that the ‘Time’ feature has little or no significance in classifying a fraud transaction. Therefore, we eliminate this column from further analysis.

Outcome:

Building machine learning model that helps us find fraudulent customers which will help the Organization saving millions of dollars