Approach for the Capstone Project – Credit Card Fraud Detection

Problem Statement

The Credit Card Fraud Detection problem is a huge problem of misuse of sensitive customer credit card data to perform fraudulent transactions. Since the financial institutions are the ones who are responsible for the security of these transactions, and thus cause heavy losses in the cases where these occur, it becomes extremely important to deal with such problem with utmost sincerity. It is also very important to work with the data very carefully keeping in mind the extremely sensitive nature of the data. We also need to keep in mind to flag the fraudulent transactions with very high accuracy since the frauds are generally of huge amounts and even one of these transactions make a huge difference.

Performing Exploratory Data Analysis

Beginning the EDA process with basic dataset exploration, shape, data types and null values. Null values are dropped/imputed. After being satisfied with the above EDA, we move on to the visualization chunk. Firstly, plotting the columns on a heatmap for spotting multicollinearity and then plotting histograms of the columns to check the skewness of the data, if the columns are not normally distributed then we move to scaling the data otherwise it's good to pass on to the model building stage.

Handling Class Imbalance

Since the pre-requisites tell us that the data is highly unbalanced amongst its classes with a very poor class distribution ratio of 1 minority class for every 578 majority classes, we begin with balancing the data otherwise the model training for the minority class will be extremely poor. We can use multiple techniques such as Random Under-sampling, Random Over-sampling, SMOTE (Synthetic Minority Over-sampling Technique) or ADASYN (Adaptive Synthetic Sampling). It is always advised to use ADASYN due to its adaptive approach where it tries to populate the less dense minority class for better data understanding.

Model Selection and Building

Since the problem is quite evidently a classification problem we have a bunch of models to choose from. We will try to train various models on the data and then test their accuracy to finalize one of them. We can choose form a logistic Regression Model, Decision Tree, Random Forest or XGBoost Classifier. We will not use the KNN classifier because of its poor computational performance when modeling on huge datasets like this one. We will build each model in combination with different class imbalance handling techniques to check which combination works the best.

Hyperparameter Tuning

For better results we will optimize each of the model to perform at its absolute maximum and will try to ascertain the best possible performance. Hyperparameters are user entered parameters that mold the model to work with the given problem statement and dataset for greater accuracy and efficiency.

Model Evaluation

Since accuracy is not the only parameter of the performance of the model, we will utilize other powerful metrics to understand and improve the strength of the model, ROC Curve and the Area Under the Curve (AUC). We will also tune the threshold of the model to optimize the performance. The f1 score that depends on the Recall and Precision of the model, depending on the amount of the transactions, since high value fraudulent transactions shouldn't pass undetected at any cost.