Credit Card Fraud Detection Capstone Project

FindDefault (Prediction of Credit Card fraud)

Problem Statement:

A credit card is one of the most used financial products to make online purchases and payments. Though the Credit cards can be a convenient way to manage the finances, they can also be risky. Credit card fraud is the unauthorized use of someone else's credit card or credit card information to make purchases or withdraw cash.

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

The dataset contains transactions made by credit cards in *September 2013* by European cardholders. This dataset presents transactions that occurred in two days, where we have ***492 frauds out of 284,807 transactions***. The dataset is highly unbalanced, the positive class (frauds) account for *0.172%* of all transactions.

We have to build a classification model to predict whether a transaction is fraudulent or not.

Our focus in this project should be on the following:

The following is recommendation of the steps that should be employed towards attempting to solve this problem statement:

- **Exploratory Data Analysis:** Analyze and understand the data to identify patterns, relationships, and trends in the data by using Descriptive Statistics and Visualizations.
- **Data Cleaning:** This might include standardization, handling the missing values and outliers in the data.
- **Dealing with Imbalanced data:** This data set is highly imbalanced. The data should be balanced using the appropriate methods before moving onto model building.
- **Feature Engineering:** Create new features or transform the existing features for better performance of the ML Models.
- **Model Selection:** Choose the most appropriate model that can be used for this project.
- **Model Training:** Split the data into train & test sets and use the train set to estimate the best model parameters.
- **Model Validation:** Evaluate the performance of the model on data that was not used during the training process. The goal is to estimate the model's ability to generalize to new, unseen data and to identify any issues with the model, such as overfitting.
- **Model Deployment:** Model deployment is the process of making a trained machine learning model available for use in a production environment.

Data Understanding and EDA:

This dataset presents transactions that occurred in two days, where we have ***492 frauds out of 284,807 transactions***. The dataset is highly unbalanced, the positive class (frauds) account for *0.172%* of all transactions.

Features V1, V2, …, V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

We will use **Heatmap, distplots, histplots** and **piechart** for basic EDA.

Model Selection, Building and Evaluation:

We will start building the model with train-test split. Then we need to find which ML model works well with the imbalanced data and have better results on the test data.

- **Linear Regression** works best when the data is linearly separable and needs to be interpretable.

- Though **KNN** is highly interpretable, but it consumes a lot of computation when we have a huge amount of data.

- **XGBoosting** is an extended version of gradient boosting, with additional features like parallel tree learning algorirthm and regularization for finding the best split.

We will use **PowerTransformer** to scale the data. It's a method used for transforming features by raising them to a certain power. It's commonly used for transforming skewed data distributions to more closely resemble a normal distribution, which can be beneficial for certain machine learning algorithms that assume normally distributed data.

We will use **ROC curve, AUC score** and **confusion matrix** for this task as the metrics. ROC curve is used to understand the strength of the model by evaluating the performance of the model. The AUC score represents the area under the ROC curve. From the confusion matrix, we can calculate accuracy, precision, recall, specificity, and F1-score of the models.

Hyperparameter Tuning:

We will use **K-Fold cross-validation** using **StratifiedKFold** to get a better performance of the model.

- **Cross-validation** is a technique used in machine learning and statistics to assess the performance of a predictive model. It's particularly useful when you have a limited amount of data and want to estimate how well your model will generalize to new, unseen data.

- **StratifiedKFold** is a cross-validation technique used in machine learning for evaluating the performance of a model. It is an extension of k-fold cross-validation, which splits the dataset into k consecutive folds (or subsets).

  However, unlike regular k-fold cross-validation, StratifiedKFold ensures that each fold maintains the same proportion of classes as the original dataset. This means that for classification problems with imbalanced class distributions, each fold will have a representative distribution of classes, which helps in producing more reliable and unbiased estimates of the model's performance.

Benefits:

Depending on the use case, we have to account for what we need: high precision or high recall.

For banks with smaller average transaction value, we would want high precision because we only want to label relevant transactions as fraudulent. For every fraudulent transaction, you can add the human element to verify whether the transaction was done by calling the customer. However, when precision is low, such tasks are a burden because the human element has to be increased.

So here, to save banks from high-value fraudulent transactions, we have to focus on a high recall in order to detect actual fraudulent transactions. We need to determine how much profit we are saving with our best selected model.