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.

Focus of 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.

Exploratory Data Analysis:

This dataset captures transactions spanning a period of two days, totaling 284,807 entries, among which 492 are flagged as fraudulent, representing a mere 0.172% of the overall transactions. The dataset is notably imbalanced, with the positive class (frauds) being significantly underrepresented.

By looking at the data we can say that the features V1 through V28 are principal components derived from PCA, except for 'Time' and 'Amount'. 'Time' indicates the elapsed seconds between each transaction and the first one recorded in the dataset, while 'Amount' represents the transaction amount. Notably, 'Amount' can be leveraged, for example-dependent cost-sensitive learning.

The response variable, denoted by 'Class', takes a value of 1 in instances of fraud and 0 for normal transaction.

For our exploratory data analysis, we plan to utilize various visualizations, including heatmaps, countplots, histogram plots, barplots, boxplots, and pie charts, to gain insights into the dataset's structure and uncover any underlying patterns or trends.

Model Selection, Building and Evaluation:

Our initial step involves constructing the model through a train-test split. Subsequently, our aim is to identify the machine learning model that effectively handles imbalanced data and yields superior results on the test dataset.

- Linear Regression performs optimally in scenarios where the data exhibits linear separability and requires interpretability. It is particularly effective in situations where the relationship between the independent and dependent variables can be represented linearly.
- KNN (K-Nearest Neighbors), while offering high interpretability, can be computationally intensive, especially with large datasets. This algorithm classifies data points based on the majority class of their nearest neighbors, making it intuitive to understand and implement. However, its computational complexity grows significantly as the dataset size increases due to the need to calculate distances between data points.
- XGBoost (Extreme Gradient Boosting) is an advanced version of gradient boosting, featuring enhancements such as parallel tree learning algorithms and regularization techniques. It is renowned for its superior performance in predictive modeling tasks, particularly in structured datasets with many features. XGBoost constructs a series of decision trees iteratively, with each subsequent tree correcting the errors of the previous ones, ultimately producing a robust ensemble model capable of handling complex relationships within the data.

In evaluating our models, we will utilize several key metrics: ROC curve, AUC score, and confusion matrix.

- The ROC curve provides insight into the model's performance by illustrating the trade-off between its true positive rate (sensitivity) and false positive rate (1 - specificity) across different classification thresholds. A steeper ROC curve indicates better model performance.
- The AUC score, or Area Under the ROC Curve, quantifies the overall performance of the model. It represents the probability that the model will correctly classify a randomly chosen

positive instance higher than a randomly chosen negative instance. A higher AUC score (closer to 1) indicates better discrimination ability of the model.

- The confusion matrix is a table that summarizes the model's performance by presenting the counts of true positive, true negative, false positive, and false negative predictions. From the confusion matrix, we can calculate various metrics including:
    - Accuracy: the proportion of correct predictions among all predictions.
    - Precision: the proportion of true positive predictions among all positive predictions, indicating the model's ability to correctly identify positive instances.
    - Recall (Sensitivity): the proportion of true positive predictions among all actual positive instances, indicating the model's ability to capture all positive instances.
    - Specificity: the proportion of true negative predictions among all actual negative instances, indicating the model's ability to correctly identify negative instances.
    - F1-score: the harmonic mean of precision and recall, providing a balanced measure of a model's accuracy.

Hyperparameter Tuning:

To enhance our model's performance, we will employ K-Fold cross-validation using StratifiedKFold.

- Cross-validation is a pivotal technique in machine learning and statistics utilized to evaluate predictive models. It's especially valuable when dealing with limited data, providing insights into how well the model will generalize to unseen data. By partitioning the dataset into multiple subsets, cross-validation enables robust estimation of the model's performance across various data samples.
- StratifiedKFold is an extension of traditional k-fold cross-validation, tailored for evaluating model performance in classification tasks. Unlike standard k-fold cross-validation, StratifiedKFold ensures that each fold maintains the same class distribution as the original dataset. This approach is particularly advantageous in scenarios with imbalanced class distributions, as it guarantees that each fold retains a representative mix of classes. Consequently, StratifiedKFold aids in generating more accurate and unbiased assessments of the model's performance.

Benefits:

In different scenarios, it's crucial to emphasize either high precision or high recall, depending on the specific use case.

In scenarios where banks typically handle smaller average transaction values, emphasizing high precision is crucial. This ensures that only relevant transactions are labeled as fraudulent, minimizing the burden of additional verification tasks such as contacting customers. However, when precision is low, the increased reliance on human intervention becomes cumbersome and resource intensive.

Conversely, in situations where the primary concern is to mitigate the risk posed by high-value fraudulent transactions, focusing on high recall becomes imperative. Detecting as many actual fraudulent transactions as possible is essential for safeguarding against potential financial losses.

To determine the effectiveness of our selected model, it's essential to quantify the amount of profit saved through fraud detection. This evaluation provides valuable insights into the model's real-world impact and its ability to address the specific needs and challenges faced by financial institutions.