Data Science Bootcamp

Capstone Project

Project 2

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 your 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.

Your 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.

Timeline

We expect you to do your best and submit a solution within 2 weeks.

Deliverables

Please share the following deliverables in a zip file.

- A report (PDF) detailing:
- Description of design choices and Performance evaluation of the model
- Discussion of future work
- The source code used to create the pipeline

Tasks/Activities List

Your code should contain the following activities/Analysis:

- Collect the time series data from the CSV file linked here.
- Exploratory Data Analysis (EDA) - Show the Data quality check, treat the missing values, outliers etc if any.
- Get the correct datatype for date.
- Balancing the data.
- Feature Engineering and feature selection.
- Train/Test Split - Apply a sampling distribution to find the best split.
- Choose the metrics for the model evaluation
- Model Selection, Training, Predicting and Assessment
- Hyperparameter Tuning/Model Improvement
- Model deployment plan.

Success Metrics

Below are the metrics for the successful submission of this case study.

- The accuracy of the model on the test data set should be > 75% (Subjective in nature)

- Add methods for Hyperparameter tuning.
- Perform model validation.

Bonus Points

- You can package your solution in a zip file included with a README that explains the installation and execution of the end-to-end pipeline.
- You can demonstrate your documentation skills by describing how it benefits our company.

Data:

The dataset for this project can be accessed by clicking the link provided below.

creditcard.csv