Exploratory Data Analysis

Commonly called as EDA is referred to the Data Understanding and Data Analysis which mainly comprises of the relationship identifications between the various variables in the form of visualizations and numbers. It can be a Qualitative Analysis or Quantitative Analysis. For exaample, Pie Charts, Bar Charts, Histograms, Distributions, Skewness, Proportions, Inter-quartile Ranges, Dispersion, Central Tendency of Data etc. Most of the time a good data exploration can provide the useful insights within the data as well as solve almost 70% of the problem in the EDA stage only.

Data Pre-Processing & Data Cleaning

Data Pre-Processing & Data Cleaning are more of the processes where an Engineer will make the data ready for the consumption of the Machine Learning Model. The most commonly used techniques are as follows:

1. Missing Value Checks & Missing Value Imputations
2. Removal of the unwanted data
3. Data Optimization on the basis of Domain or Business recommendations.
4. Outlier Detection & Removal
5. Dimension Reduction
6. Duplicate records removal

Feature Engineering & Feature Selection

Feature Engineering or Feature Selection is a technique to identify the most important features within a dataset. Features can be derived also from the existing feature space as well as can be reduced.

For my training and validation data, I use Sklearn's Train_test_split function and use 80-20 portion. The dataset itself is large enough so I am not warried about splitting too much to validation data. 20% of validation data seems enough for me. 80% of the data for training ensures that the model has access to a large enough dataset to learn the underlying patterns. This is important because the performance of machine learning models generally improves with more data, up to a certain point. A larger training set provides a rich variety of examples from which the model can learn, leading to a more accurate and robust model.

Machine Learning Model Selection

Machine Learning Model selection is based on the type of business problem we are handling or more than that depends on the application and end results. Few of the most common problems available in the Machine Learning area are Classification, Regression, Clustering etc. As far as a pure Machine Learning project is concerned the below mentioned Algorithms are highly used ones industry wide:

1 . Decision Tree

2 . Random Forest

3 . Regression

After comparesion, A random forest model is selected for this project. Random forest is the only model that can reach about 99.6% validation accuracy so a random forest model is choose to be the winner of all 4 models and will be used for this project later. Other models may also be able to reach this score or be even better than Random Forest, but they will require more tuning and more hyperparameters needs to be tested.

The hyperparameters for random forest will be n_estimators = 1000 and max depth of 10. This is tested to be the best for random forest on this problem.