PRODUCT DEMAND FORECASTING

Evaluation Metrices

R-Squared:

- Ranging from 0 to 1 with 1 being the perfect prediction.

Root Mean Square Error:

- It describes the closeness of the observed data points and model's predicted values.
- lower values of RMSE indicates better model.

Baseline Model

Linear Regression:

- Linear Regression model is used as the baseline initially to produce the results.
- Since the final model helps in identifying the item's demand at various stores, initially we check the item's sales at the stores by applying this baseline method as this helps in giving the general ideology about the demand of the considered item.

Proposed Methodology:

- The dataset which is being considered has 14,204 records. From which, I would be considering 8523 records for training set and 5681 records for testing set.
- The considered test set is Identically Independent Dataset as the features that are considered do not have any dependency among them and are independent. Also, all the features are generated considering the same underlying idea.

Various Machine Learning techniques that I would use are:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Support Vector Machine
- Decision Tree

- The hyperparameters are selected such that they help define model in a better way to predict the item sales at an outlet than that of the train model. The hyperparameters that will be used in Decision Tree include min_samples_split, max_depth and max_features. For the SVM, gamma and slack variable are considered to further tune the model.

Feature selection:

- The dataset has 12 features which are related to Item such as: Item weight, item fat content, item visibility, item type, item mrp and related to Outlet such as: outlet identifier, outlet establishment year, outlet size, outlet location type, outlet type and our target to predict the item outlet sales.

- The features are selected carefully such that these features support store level and product level hypothesis.

Feature Scaling:

- There are null values present in the ‘item_weight’ and ‘outlet_size’ features, hence we will fill those null values.
- Replacing the repetitive values (i.e., repetitive in terms of abbreviations) in ‘Item_Fat_content’ , ‘Item_Identifier’ into singular value.

Feature Engineering:

- Diving feature ‘Item_Fat_Content’ into different levels based on the fat content.
- 'Item_Identifier', 'Item_Type', 'Outlet_Identifier' are non-numerical in nature so applying label encoding.
- 'Item_Fat_Content', 'Outlet_Location_Type', 'Outlet_Size','Item_Category', 'Outlet_Type' are having ranking levels in them so applying one-hot encoding.