**PROJECT**

**BIGMART SALES PREDICTION**

**Problem Statement:**

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and predict the sales of each product at a particular outlet. Using this model, BigMart will try to understand the properties of products and outlets which play a key role in increasing sales.

Description of the features collected:

|  |  |
| --- | --- |
| Variable | Description |
| Item\_Identifier | Unique product ID |
| Item\_Weight | Weight of product |
| Item\_Fat\_Content | Whether the product is low fat or not |
| Item\_Visibility | The % of total display area of all products in a store allocated to the particular product |
| Item\_Type | The category to which the product belongs |
| Item\_MRP | Maximum Retail Price (list price) of the product |
| Outlet\_Identifier | Unique store ID |
| Outlet\_Establishment\_Year | The year in which store was established |
| Outlet\_Size | The size of the store in terms of ground area covered |
| Outlet\_Location\_Type | The type of city in which the store is located |
| Outlet\_Type | Whether the outlet is just a grocery store or some sort of supermarket |
| Item\_Outlet\_Sales | Sales of the product in the particular store. This is the outcome variable to be predicted. |

**Evaluation Metric:**

The evaluation metric used is **root mean squared error (RMSE)**

**Steps followed to obtain the final model:**

1. Importing Libraries

All the required libraries are imported.

1. Exploratory Data Analysis on both train and test data.

* Checking the correlation of independent variable with the dependent variables.
* Separating categorical, int and float variables.
* Replacing the Item\_Fat\_Content with correct values.
* Checking for duplicate values.
* Missing Value handling of the column Outlet\_Type.
* Checking for Outliers using Boxplot and removing the outliers present in Item\_Outlet\_Sales.
* Dropping unwanted columns which are not required for modelling.
* One Hot Encoding is used for the below columns:

Item\_Fat\_Content,Item\_Type,Outlet\_Identifier,Outlet\_Size,Outlet\_Location\_Type,

Outlet\_Type.

1. Selection of Dependent and Independent variables

The train data will be split into dependent (X-axis) and independent(Y-axis) variables. Her we will be using Item\_Outlet\_Sales as the Independent variable.

1. Split data for train and test

Dividing the train data set for modelling. We choose 25% as the test size.

1. Machine Learning Algorithm selection

Following machine learning supervised algorithms are performed on the dataset and have obtained the RMSE value.

Linear Regression (RMSE : 1022.9000213095405)

Ridge Regressor (RMSE : 1021.5012507643021)

Lasso Regressor (RMSE : 1021.1147863662823)

ElasticNet Regressor (RMSE : 1022.2708139914947)

Random Forest Regressor (RMSE : 981.7494780566408)

XGBoost (RMSE : 991.3272399121147)

RandomForest Regression Model has the least Root Mean Squared Error, so we select **RandomForest Model** as our final model.

Performed hyper parameter tuning using GridSearchCV and the best parameter selected :

Criterion='mse', max\_depth:=5, n\_estimators=60, random\_state=42

**RESULT:**

Final model is Random Forest model which is hyper tuned with GridSearchCV saved as a pickle file for further use.

**RMSE : 981.7494780566408**