Page :
Date : / /

Title : Bigmark sales analysis.

Problem Statement:

Bigmart sales analysis. For data comprising of transaction records of a sales store. The data has 8523 rows and 12 columns.
Predict the sales of a store.

Objective : To apply different regression techniques to find / predict the sales of a store.

Outcome : The outcomes are :

- To learn to preprocess tabular data.
- To apply different regression techniques.

Theory :

The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 different stores in different cities.

Data :

- 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 of product.

- Outlet_Identifier: unique store id.
- Outlet_Size: The size of the stores 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.

- The different steps involved are:
  - Data exploration: Looking at the categorical and continuous feature summaries and making inferences about the data.
  - Data cleaning: Inputing missing values in the data and checking for outliers.
  - Feature engineering: Modifying existing variables and creating new ones for analysis.
  - Model Building: Making predictive models on the data.

- Since we are dealing with continuous values as our target value (Item_Outlet_Sales) this would come under regression problem.

P.T.O

Date / /

Algorithms:

- **Linear Regression:**

It is a linear approach to modeling the relationships between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

- **Random Forest:**

Random forests or random forest trees are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

Different libraries used:
- numpy
- pandas
- scikit learn.

- We split the train data into training and validation at 70:30 ratio.

P.T.O

Page :
Date : / /

Analysis :

Algorithm	Validation Score	Test Score
1. Linear Regression	1148.49	1277.805
2. Random Forest Regressor	1138.185	1226.34

Evaluation Metric = Root Mean Squared Error.

CONCLUSION: We have thus build a machine learning model to predict Outlet Sales using Big Mart Dataset.