Date __/__/____

Saathi Notebooks (1)

Assignment - C3

* **Title :-** Bigmart sales Analysis.

* **Problem statement :-** For data comprising of transaction records of sales store. The data has 8523 rows of 12 variables. Predict the sales of store.

* **Objectives :-**
  - learn regression algorithms.
  - learn to summarize the properties the dataset.
  - learn to split the dataset into training and test datasets.
  - learn to develop a predictive regression model.

* **Outcomes :-** students will be able to develop a predictive model for sales of an item at BigMart.

* **S/W & H/W Requirements :-**
  - OS: 64-bit Ubuntu 18.04.
  - Python 3
  - Jupyter Notebook / Google Colaboratory.
  - Kaggle, Kaggle CI, SKlearn, Pandas, Matplotlib, Pycaret, Graphviz.

* **Theory :-**

Linear Regression :- In statistics, linear Regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one-explanatory variable is called simple linear regression. For more than one,

Page No.

the process is called "multiple linear regression".

If the goal is prediction, forecasting or error reduction, linear regression can be used to fit a predictive model to an observed data set of values of response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted value can be used to make a prediction of the response.

Given a dataset of n statistical units, a linear regression model assumes that the relationship between dependent variable y & the p-vector of regressors x is linear.

Dataset = $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$

Model Equation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i, \quad i = 1 \ldots n.$$

Matrix Notation $\Rightarrow y = X\beta + \varepsilon$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad x = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} 1 & \ldots & x_{11} & \ldots & x_{1p} \\ 1 & \ldots & x_{21} & \ldots & x_{2p} \\ \vdots & & \vdots & & \vdots \\ 1 & \ldots & x_{n1} & \ldots & x_{np} \end{bmatrix}$$

* Dataset Description:

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 find out

sales of each product at particular store.

1) Item-Identifier : Unique product ID.
2) Item-weight : Wt of product.
3) Item-Fat-Content : Whether product is low fat or not.
4) Item-visibility : The % of total display area of all products in a store allocated to a particular product.
5) Item-type : The category to which the product belongs.
6) Outlet-Identifier : Unique store ID.
7) Outlet-Establishment : The year in which store was established.
8) Outlet-Size : Size of the store in terms of ground area covered.
9) Outlet-location-Type : The type of city in which the store is located.
10) Outlet-Type : Whether the outlet is just a grocery store or some sort of supermarket.

* Gaussian Distribution:-

It is a symmetric distribution where most of the observations cluster around central peak and probabilities for values further away from the mean tapes off equally in the both directions.

To deal with missing values in a numerical feature with gaussian distribution, we can calculate the mean of the feature and replace it with the missing values. This is an approximation which add variance to the data set.

* TestCases:-

	Algorithm	Result.
1)	Linear Regression	RMSE : 1239.18. R2score : 0.53.

2) Ridge — RMSE : 1239.17
R2 score : 0.53.

3) Lasso — RMSE : 1239.38.
R2 score : 0.53

4) Elastic Net. — RMSE : 1288.33
R2 score : 0.49.

5) Random Forest Regressor — RMSE : 1119.35
R2 score : 0.61

6) Linear SVR — RMSE : 1291.61
R2 score : 0.49.

* Conclusion:- Successfully developed a predictive model for sales of an item at BigMarket.