Sourav Kotkar
TE-3 (N3)
31378

Performance Date - 04/02/2022
Submission Date - 05/02/2022

Page No.:
Date:

YOUVA

DSBDAL

Group - A : Assignment - 4

* **Title :-**
Data Analytics I

* **Problem Statement :-**
Create a Linear Regression Model using Python/R to predict home prices using Boston Housing Dataset. The Boston Housing dataset contains information about various houses in Boston through different parameters. There are 506 samples and 14 feature variables in this dataset.
The objective is to predict the value of the house using the given features.

* **Learning Objectives:-**
To understand:-
1) Techniques for data analytics.
2) Linear Regression and prediction of result.

* **Learning Outcomes :-**
After completion of this assignment, student would be able to create a linear regression model and predict home prices using the Boston Housing dataset.

* **H/W and S/W requirements :-**
8GB RAM laptop, Windows 10, Python 3.9, Jupyter notebook

* Theory :-

⇒ Data Analytics :-

In simple language, data analytics is the science of evaluating raw data to make outcomes from the data. Data analytic techniques help you to carry raw data and find patterns to take out useful ideas from it. Nowadays, data experts use data analytics in their core research. Several companies also use data analytics to make informed decisions.

Data analytics is a wide term that includes numerous assorted sorts of data analysis. Any type of data can be exposed to data analytics strategies to get an understanding that can be used to improve things. For example, gaming corporations use data analytics to set prize timetables for players that keep most of the players dynamic in the game. Similarly, there are other types of corporations that use data analytics according to their needs.

Types of Data Analytics :-

1) Descriptive Analytics
2) Diagnostic Analytics
3) Predictive Analytics
4) Prescriptive Analytics

⇒ Linear Regression :-

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on - the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

⇒ Types of Linear Regression :-

1) Simple Linear Regression :-

In simple linear regression, we try to find the relationship between a single independent variable (input) and a corresponding dependent variable (output). This can be expressed in the form of a straight line.

Equation of line :

$$Y = \beta_0 + \beta_1 X$$

where, Y = output or dependent variable

β_0 & β_1 = two unknown constants that represent the intercept & slope respectively

X = input variable

2) Multiple Linear Regression :-

In multiple linear regression, we try to find the relationship between 2 or more independent variables (inputs) and the corresponding dependent variable (output). The independent variables can be continous or categorical.

Equation of Line :-

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

where,

Y = dependent variable

$\beta_0, \beta_1, \dots, \beta_p$ = coefficients

$x_1, x_2, \dots, x_p$ = independent variables

p = no. of independent variables

⇒ Methods & Functions used :-

1) read_csv() - It is used to read a csv file into a Dataframe.

2) skew() - It returns unbiased skew over requested axis.

3) quantile() - It returns values at the given quantile over requested axis.

4) isnull() - It returns a boolean same-sized object indicating if the values are NA.

5) fillna() - It is used to fill NA/NaN values using the specified method.

6) dtypes - It returns a series with the datatype of each column.

7) describe() - It gives descriptive statistics of the dataset.

8) corr() - It computes pairwise correlation of columns, excluding NA/null values.

9) train_test_split() - It splits arrays or matrices into random train and test subsets.

10) LinearRegression.fit() - Fit linear model from the dataset

11) LinearRegression.predict() - Predict using linear model

12) LinearRegression.score() - Return the coefficient of determination in the prediction

⇒ Packages & Libraries used :-

1) Pandas :-
Pandas is a python library used for working with datasets. It has functions for analyzing, cleaning, exploring and manipulating data.

2) NumPy :-
NumPy is a python library for adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

3) Matplotlib :-
It is a low level graph plotting library in python that serves as a visualization utility.

4) Seaborn :-
It is a library that uses Matplotlib underneath to plot graphs. It is used to visualize random distributions.

5) Scikit-learn :-

It is a free software machine learning library for python. It features various classification, regression and clustering algorithms.

⇒ Plots used :-

1) Displot :-

This function provides access to several approaches for visualizing the univariate or bivariate distribution of data, including subsets of data defined by semantic mapping and faceting across multiple subplots.

2) Pairplot :-

It is used to plot pairwise relationships in a dataset. By default, this function will create a grid of Axes such that each numeric variable in data will be shared across the y-axes across a single row and the x-axes across a single column.

3) Heatmap :-

It is used to plot rectangular data as a color-encoded matrix. This is an Axes-level function & will draw the heatmap into the currently-active Axes if none is provided to the ax argument.

4) Scatterplot :-

A scatterplot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

* Observations :-

- The dataset used is Boston Housing dataset which contains 13 independent variables - CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT and a dependent variable MEDV.
- The MEDV attribute is positively skewed with a skewness of 1.108. The outliers are removed using the Interquartile range method.
- There are null values present in the dataset which are removed by replacing with mean or median values.
- For training the dataset, attributes with a correlation less than -0.4 and greater than 0.4 are taken.
- Accuracy of linear regression model = 81.35 %
  Mean absolute error = 2.11
  Mean squared error = 6.63
  Root mean squared error = 2.57

* Conclusion :-

Through this assignment, we learnt and implemented linear regression and prediction of result using the various Python libraries.