Assignment No-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 feature variables in this dataset.

Learning Objective :-

1] To understand the concept of Linear Regression
2] To predict the value of prices of the house using given features.

Learning Outcomes :-

- After performing the assignment one should be able to
- Apply Linear Regression to fit model & predict values.
- Understand various metrics using Sklearn Library.

Software Requirements :-
- Anaconda Navigator
- Jupyter Notebook
- Python 3.8

Hardware Requirements :-
- Windows 10
- 8 GB RAM
- Intel i5 processor, 64 bit OS

Theory :-

* Data Analysis / Analytics :-

Data analytics is the science of analyzing raw data to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.

Data Analysis Steps :-

1. The first step is to determine the data requirements or how the data is grouped. Data may be separated by age, demographic, income or gender. Data values may be numerical or be divided by category.

2. The second step in data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources cameras or through personal.

3. Once the data is collected, it must be organised so it can be analyzed. This may take place on a spreadsheet or other form of software that can take statistical data.

4. The data is then cleaned up before analysis. This means it is scrubbed and checked to ensure there is no duplication or errors, and that it is not incomplete. This step helps correct any errors before it goes or to a data analyst to be analyzed.

Types of Data Analytics :-

1] Descriptive Analytics
11] Diagnostic Analytics
111] Predictive Analytics
1v] Perspective Analytics.

Linear Regression:-

Linear Regression is used for finding linear relationship between target and one or more predictors. It is a linear model e.g. a model that assumes a linear relationship betn the input variables (n) & the single output variable (y).

Types of Linear Regression:-

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

Eqn of line:-

$$y = B_0 + B_1 x$$

where y = output or dependent variable

B_0 & B_1 = two unknown constants that represent the intercept & slope

x = input variable.

Multiple Linear Regression:-

In Multiple Linear Regression, we try t

find the relationship between 2 or more independent variables and the corresponding dependent variable. The independent variables can be continous or categorical.

Eqn of Line :- $y = B_0 + B_1x_1 + B_2x_2 + \dots + B_nx_n$

where,

N = dependent variable

$B_0, B_1, \dots B_n$ = coefficients

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

Various Plots Used :-

1] Displot :-

The displot figure displays a combination of statistical representations of numerical data, such as histogram, kernel density estimation or normal curve & rug plot.

2] Scatter Plot :-

A scatter plot is a diagram where each value in the dataset is represented by dot. The Matplotlib

module has a method for drawing scatter plots, it needs two arrays of same length, one for X-axis & other for Y-axis.

3] Pair Plot :-

It is used to plot pairwise relationships in a dataset. By default this func 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.

Methods & Func's used :-

1] Pandas.read_csv() :-

Read a comma seperated values file into a Dataframe.

Also supports optionally iterating or breaking of the file into chunks.

iii) data-frame.head(limit) :-

Returns the first 5 rows of data frame. To override the default, you may insert a value between the parenthesis to change the number of rows returned

iii) data-frame.shape

Returns a tuple representing the dimensions.

iv) data-frame.describe

Provides descriptive statistics that summarizes the central tendency, dispersion and shape.

v) .isnull()

Returns data frame with value 'True' where it finds Null values and 'False' where it encounters any type of value.

vi) data-frame.dtypes

This attributes return the dtypes in the Dataframe. It returns a series with the data type of each column.

Packages / Module / Libraries :-

1) Pandas :

Pandas is a software library written for the python programming language for data manipulation and analysis. In particular it offers data structures and operations for manipulating numerical tables and time series. It is a free software release under the three clause BSD License. The name is derived from the term "panel data", an econometrics terms for data sets that include observations over multiple time periods for the same individuals.

It allows importing data from various file formats such as CSV, JSON or queries

2) Numpy :-

Numpy is a python library used for working with arrays. It also has many functions for working in domains of linear algebra, matrices etc.
It is an open source project and you can use it freely. Numpy stands for Numerical Python.

3) Matplotlib :-

Matplotlib is a plotting library for the python programming language and its numerical mathematics extension Numpy. It is a library used for plotting simple graphs and to be used for data visualization.

4) Scipy :-

Scipy is a free and open-source python library used for scientific computing and technical computing. Scipy contains modules for linear algebra, integration, image processing interpolation etc.

Pylab :-

Pylab is a module that provides a Matlab like namespace by importing functions from the modules Numpy and Matplotlib.

Syntax :- import pylab

Analysis & Observations :-

1] For this assignment Boston Housing Dataset is used which contains 13 independent variables.

2] First the dataset is preprocessed the NULL values are replaced with mean value the outliers for targed variable MEDV is detect & remove using Inter-Quartile Range.

3] For training the dataset the attributes which have correlation aboves 0.5 are considered.

Conclusion :-

From this assignment we learnt the concept of Linear Regression and implemented it successfully.