Project overview -

Perform proper analysis of datasets and draw conclusions based on your analysis.

Linear Regression -

Linear Regression is a statistical technique which is used to find the linear relationship between dependent and one or more independent variables. This technique is applicable for Supervised learning Regression problems where we try to predict a continuous variable.

Linear Regression can be further classified into two types – Simple and Multiple Linear Regression. In this project, I employ a Simple Linear Regression technique where I have one independent and one dependent variable. It is the simplest form of Linear Regression where we fit a straight line to the data.

Simple Linear Regression (SLR)

Simple Linear Regression (or SLR) is the simplest model in machine learning. It models the linear relationship between the independent and dependent variables.

In this project, there is one independent or input variable which represents the Sales data and is denoted by X. Similarly, there is one dependent or output variable which represents the Advertising data and is denoted by y. We want to build a linear relationship between these variables. This

linear relationship can be modelled by mathematical equation of the form:-

$$Y = \beta 0 + \beta 1 * X \quad \text{----------} \quad (1)$$

In this equation, X and Y are called independent and dependent variables respectively,

β1 is the coefficient for independent variable and

β0 is the constant term.

β0 and β1 are called parameters of the model.

For simplicity, we can compare the above equation with the basic line equation of the form:-

$$y = ax + b \quad \text{----------} \quad (2)$$

We can see that

slope of the line is given by, a = β1, and

intercept of the line by b = β0.

In this Simple Linear Regression model, we want to fit a line which estimates the linear relationship between X and Y. So, the question of fitting reduces to estimating the parameters of the model β0 and β1.

Ordinary Least Square Method

As I have described earlier, the Sales and Advertising data are given by X and y respectively. We can draw a scatter plot between X and y which shows the relationship between them.

Now, our task is to find a line which best fits this scatter plot. This line will help us to predict the value of any Target variable for any given Feature variable. This line is called the Regression line.

We can define an error function for any line. Then, the regression line is the one which minimizes the error function. Such an error function is also called a Cost function.

Cost Function

We want the Regression line to resemble the dataset as closely as possible. In other words, we want the line to be as close to actual data points as possible. It can be achieved by minimizing the vertical distance between the actual data point and fitted line. I calculate the vertical distance between each data point and the line. This distance is called the residual.

So, in a regression model, we try to minimize the residuals by finding the line of best fit. The residuals are represented by the vertical dotted lines from actual data points to the line.

We can try to minimize the sum of the residuals, but then a large positive residual would cancel out a large negative residual. For this reason, we minimize the sum of the squares of the residuals.

Mathematically, we denote actual data points by yi and predicted data points by ŷi. So, the residual for a data point i would be given as

$$d_i = y_i - \hat{y}_i$$

Sum of the squares of the residuals is given as:

$$D = \Sigma\, di^{**}2 \qquad \text{for all data points}$$

This is the Cost function. It denotes the total error present in the model which is the sum of the total errors of each individual data point.

We can estimate the parameters of the model β0 and β1 by minimizing the error in the model by minimizing D. Thus, we can find the regression line given by equation (1).

This method of finding the parameters of the model and thus regression line is called Ordinary Least Square Method.

The problem statement -

Perform proper analysis of datasets and draw conclusions based on your analysis.

Software information -

I did this project using Spyder notebook

The server is running on Python (Python 3.6.5), Anaconda distribution.

Python libraries -

I have an Anaconda Python distribution installed on my system. It comes with most of the standard Python libraries I need for this project. The basic Python libraries used in this project are:-

• Numpy – It provides a fast numerical array structure and operating functions.

• pandas – It provides tools for data storage, manipulation and analysis tasks.

• Scikit-Learn – The required machine learning library in Python.

• Matplotlib – It is the basic plotting library in Python. It provides tools for making plots.

About the dataset -

Data Set has been downloaded from the given link

Exploratory data analysis -

First, I import the dataset into the dataframe with the standard read_csv () function of pandas library and assign it to the df variable. Then, I conducted exploratory data analysis to get a feel for the data.

pandas shape attribute -

The shape attribute of the pandas dataframe gives the dimensions of the dataframe.

Independent and Dependent Variables -

In this project, I refer to the Independent variable as the Feature variable and the Dependent variable as the Target variable. These variables are also recognized by different names as follows: -

Independent variable -

Independent variable is also called Input variable and is denoted by X. In practical applications, an independent variable is also called a Feature variable or Predictor variable. We can denote it as:-

Independent or Input variable (X) = Feature variable = Predictor variable

Dependent variable -

Dependent variable is also called Output variable and is denoted by y.

Dependent variable is also called a Target variable or Response variable. It can be denoted it as follows:-

Dependent or Output variable (y) = Target variable = Response variable

Visual exploratory data analysis -

I visualize the relationship between X and y by plotting a scatterplot between X and y.

Checking dimensions of X and y -

Need to check the dimensions of X and y to make sure they are in right format for the Scikit-Learn API.

It is an important precursor to model building.

Reshaping X and y -

Since we are working with only one feature variable, so we need to reshape using the Numpy reshape() method.

It specifies the first dimension to be -1, which means "unspecified".

Its value is inferred from the length of the array and the remaining dimensions.

Difference in dimensions of X and y after reshaping -

We can see the difference in dimensions of X and y before and after reshaping.

It is essential in this case because getting the feature and target variable right is an important precursor to model building.

Train test split -

I split the dataset into two sets namely - train set and test set.

The model learns the relationships from the training data and predict on test data.

Mechanics of the model -

I split the dataset into two sets – the training set and the test set. Then, I instantiate the regressor lm and fit it on the training set with the fit method.

In this step, the model learned the relationships between the training data (X_train, y_train).

Now the model is ready to make predictions on the test data (X_test). Hence, I predict the test data using the predict method.

In [17]:

Regression metrics for model performance -

Now, it is the time to evaluate model performance.

For regression problems, there are two ways to compute the model performance. They are RMSE (Root Mean Square Error) and R-Squared Value. These are explained below:-

RMSE -

RMSE is the standard deviation of the residuals. So, RMSE gives us the standard deviation of the unexplained variance by the model. It can be calculated by taking the square root of Mean Squared Error. RMSE is an

absolute measure of fit. It gives us how spread the residuals are, given by the standard deviation of the residuals. The more concentrated the data is around the regression line, the lower the residuals and hence lower the standard deviation of residuals. It results in lower values of RMSE. So, lower values of RMSE indicate better fit of data.

R2 Score -

R2 Score is another metric to evaluate performance of a regression model. It is also called the coefficient of determination. It gives us an idea of the goodness of fit for the linear regression models. It indicates the percentage of variance that is explained by the model.

Mathematically,

R2 Score = Explained Variation/Total Variation

In general, the higher the R2 Score value, the better the model fits the data. Usually, its value ranges from 0 to 1. So, we want its value to be as close to 1. Its value can become negative if our model is wrong.

Interpretation and Conclusion -

The RMSE value has been found to be 11.2273. It means the standard deviation for our prediction is 11.2273. So, sometimes we expect the predictions to be off by more than 11.2273 and other times we expect less than 11.2273. So, the model is not a good fit to the data.

In business decisions, the benchmark for the R2 score value is 0.7. It means if R2 score value >= 0.7, then the model is good enough to deploy on unseen data whereas if R2 score value < 0.7, then the model is not good enough to deploy. Our R2 score value has been found to be .5789. It means that this model explains 57.89 % of the variance in our dependent variable. So, the R2 score value confirms that the model is

not good enough to deploy because it does not provide good fit to the data.

In [4]: