**ABSTRACT**

The Aim to build this model using Multiple Linear Regression is that we deal with a dataset which contains the details of 50 start-up’s and predicts the profit of a new Start-up based on certain features. To Venture Capitalists this could be a boon as to whether they should invest in a particular Start-up or not.

So, let’s say I work as a Venture Capitalist and I be hired as a Data Scientist to derive insights into the data, and to predict whether a particular start-up would be safe to invest in or not.

We can also derive useful insights into the data by actually seeing as to what difference does it make, if a start-up is launched and which start-up’s end up performing better by seeing that if they spent more money on marketing or was it their stellar R&D department which led them to this huge profit and in turn huge fame and success.

After EDA (Exploratory Data Analysis) of dataset we will Train and Test using train\_test\_split function. Now Inserting Linear regression () function to finally predict our profit. Before that one must check model score. So, we can define the accuracy of model. Here, we reached on conclusion that our dataset is small, so we are not accurate. If we had large amount of data for same process we had more accuracy of prediction the Profit.

**TABLE OF CONTENT**

1. Introduction
   1. Introduction to Regression Algorithms.
   2. Introduction to Dataset.
2. File Pre-processing.
   1. Importing Libraries
   2. Importing Dataset
3. Train and Test.
4. Fitting Regression models.
5. Predicting Test Set result.
6. **Introduction**

**1.1 Introduction** of Multiple linear regressions.

In simple words, Linear regression is approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables.)

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Chart, scatter chart

Description automatically generated

* 1. **Introduction to Dataset.**

Before starting analysis of data, one must know what is in the data, what to do, how to do things with that data.

In our project we have a .CSV file that has 50 rows and 4 columns in total. 50 rows contain information about how they spend their funds to grow their start-ups i.e. there Research and development spendings, administration spendings and marketing spendings.

Fourth column of our file represents the profit made by them in a year

So, we have R&D spend, Administration spends, and Marketing spend as our independent variables and Profit column will be our dependent variable that depends on independent variables.

DATASET -

Table

Description automatically generated

1. **FILE PRE-PROCESSING**

Our first task of creating a prediction model or say making any machine learning model we need to import some useful libraries in Jupyter Notebook, or we can also use Google Colab for the same task.

**2.1 Importing Libraries.**

In this Model we will import libraries like Numpy, Pandas, Seaborn, Matplotlib, Sklearn, Math, Statsmodels. Text

Description automatically generated

* 1. **Importing Dataset.**

our next task is to import our dataset using pandas (the dataset introduced previously) we have different-different syntax for different tasks.

Before EDA (Exploratory Data Analysis) we must check no. of columns and rows in our dataset and a complete description of data is to be known to analyst. Then comes our EDA part of the dataset, where we must create some graphs to visualise the dataset and handling outliers.

Correlation matrix of the dataset.

Chart

Description automatically generated with medium confidence

Reviewing Outliers-

Chart, box and whisker chart

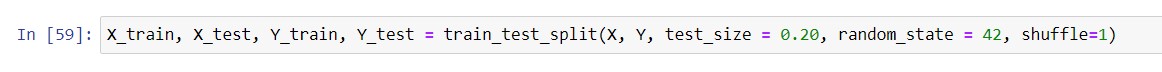
Description automatically generated

As we have a small datasets we can ignore handling outliers here.

1. **TRAIN AND TEST.**

Using appropriate syntax, training and testing the model is a very important and basic part.

Following steps are required to do train and test.

Step-1Table

Description automatically generatedTable

Description automatically generatedTable

Description automatically generated

Comparing the predicted value and actual values.

Graphical user interface, table

Description automatically generated

1. **Fitting Regression models.**

we can see that the predicted value is close to the actual values i.e., the one present in the testing set, hence we can use this model for prediction. But first, we need to calculate how much is the error generated.

Model evaluation

1. R2 score: R2 score – R squared score. It is one of the statistical approaches by which we can find the variance or the spread of the target and feature data.

Graphical user interface, application

Description automatically generated

1. **MAE:** MAE – **Mean Absolute Error**. By using this approach we can find the difference between the actual values and predicted values but that difference is absolute i.e. the difference is positive.

Graphical user interface, text, application

Description automatically generated

1. MSE: MSE – Mean Squared Error. By using this approach we can find that how much the regression best fit line is close to all the residual.

Graphical user interface, text, application

Description automatically generated

1. RMSE: RMSE – Root Mean Squared Error. This is similar to the Mean squared error (MSE) approach, the only difference is that here we find the root of the mean squared error i.e. root of the Mean squared error is equal to Root Mean Squared Error. The reason behind finding the root is to find the closer residual to the values found by mean squared error.

Graphical user interface, text, application

Description automatically generated

1. **Predicting Test Set result.**

Finding the profit of a start-up that invested in following ratio:

> R&D = 165444

> Administration = 90000

> Marketing = 300000

Graphical user interface, text, application, email

Description automatically generated