**REPORT ON PROJECT OF DATA SCIENCE**

**-BY SHREYA MATHUR**

**ABSTRACT**

This project is specifically on the Topic given in Data Science domain. In that we have to make evaluation on profit and ML model of the 50 companies on the basis of the data set given in which data of its R&D spend, Administration, Marketing Spend was given.

As In today’s world, people are more attracted towards starting their own new startup but there is cut throat competition in the market in terms of product, profit, etc. To survive in the market and earn something high, different strategies are followed and several opinions and points are considered. The model aims at predicting the profit of any start up based on the values of certain parameters such as R&D Spend, Marketing Spend, Administration. Our model will be helpful in providing the profit amount to the users beforehand. Prediction of profit in advance, will be helpful for users to adopt and examine different strategies to enhance it.

To solve the problem we have taken the procedure of Linear Regression model as it is one of the most-used regression algorithms in Machine Learning. A significant variable from the data set is chosen to predict the output variables (future values). Ordinary least squares (OLS) and linear least squares, is the real workhorse of the regression world. Use linear regression to understand the mean change in a dependent variable given a one-unit change in each independent variable.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **SL.NO.** | **TOPICS** | **PAGE NUMBERS** |
| 1 | Abstract | 2 |
| 2 | Table of Contents | 3 |
| 3 | Introduction | 4,5 |
| 4 | Existing Method | 6,7 |
| 5 | Proposed method with Architecture | 8 |
| 6 | Methodology | 9 |
| 7 | Implementation | 10,11,12 |
| 8 | Conclusion | 13 |

**INTRODUCTION**

Start-up Companies plays an indispensable role in uplifting the economy by providing employment to experienced and young professionals. Start-up companies are generally set up with high costs and limited revenue and require capital from a variety of sources such as venture capitalists. Therefore, these reasons make start-up companies an important target of analysis. It becomes very difficult for the start-up companies to operate and tackle the problems in a very highly competitive environment. Profit earned by the start-up company is an important factor in determining its ability to survive and enhance its business. Profit of any start-up company largely depends on the structure of money distribution for different causes. Evaluation of profit will not only provide an overview of the gain but also helpful in developing future strategies and to reorder the money distribution for different parameters.

Since Machine Learning possess the ability to solve complex problems, therefore machine learning is implemented in our model to make prediction of profit. Using appropriate dataset and machine learning algorithms, it is possible to predict the profit. The dataset and algorithm used in the model plays an indispensable role to determine the accuracy of result obtained. This Profit Prediction Model aims at predicting the profit of the start-up based on the values of Administration Spend, R&D Spend, Marketing Spend using different machine learning algorithms. The dataset consists of the data of 50 start-ups. The dataset consists of the values of four parameters which are R&D Spend, Marketing Spend, Administration Spend and State in which the start-up is located. The selected dataset is then fed for further processing. Data Pre-processing including data cleaning and outlier removal is done. The pre-processed data is then trained and tested. Out of the complete dataset, 20% of the dataset is used for testing purpose and the remaining 80% precent of the dataset is used for training.

Linear Regression algorithm Implemented on the dataset after training.

As it is one of the most-used regression algorithms in Machine Learning. A significant variable from the data set is chosen to predict the output variables (future values). Ordinary least squares (OLS) and linear least squares, is the real workhorse of the regression world. Use linear regression to understand the mean change in a dependent variable given a one-unit change in each independent variable.

**EXISTING METHOD**

There are various regression methods for making Ml model. Some of the existing method are : -

**1. Linear Regression**

Linear regression is one of the most basic types of regression in machine learning. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other. In case the data involves more than one independent variable, then linear regression is called multiple linear regression models

### ****2. Logistic Regression****

Logistic regression is one of the types of regression analysis technique, which gets used when the dependent variable is discrete. Example: 0 or 1, true or false, etc. This means the target variable can have only two values, and a sigmoid curve denotes the relation between the target variable and the independent variable.

### ****3. Ridge Regression****

This is another one of the types of regression in machine learning which is usually used when there is a high correlation between the independent variables. This is because, in the case of multi collinear data, the least square estimates give unbiased values. But, in case the collinearity is very high, there can be some bias value. Therefore, a bias matrix is introduced in the equation of Ridge Regression. This is a powerful regression method where the model is less susceptible to overfitting.

### ****4. Lasso Regression****

Lasso Regression is one of the types of regression in machine learning that performs regularization along with feature selection. It prohibits the absolute size of the regression coefficient. As a result, the coefficient value gets nearer to zero, which does not happen in the case of Ridge Regression.

**PROPOSED METHOD WITH ARCHITECTURE**

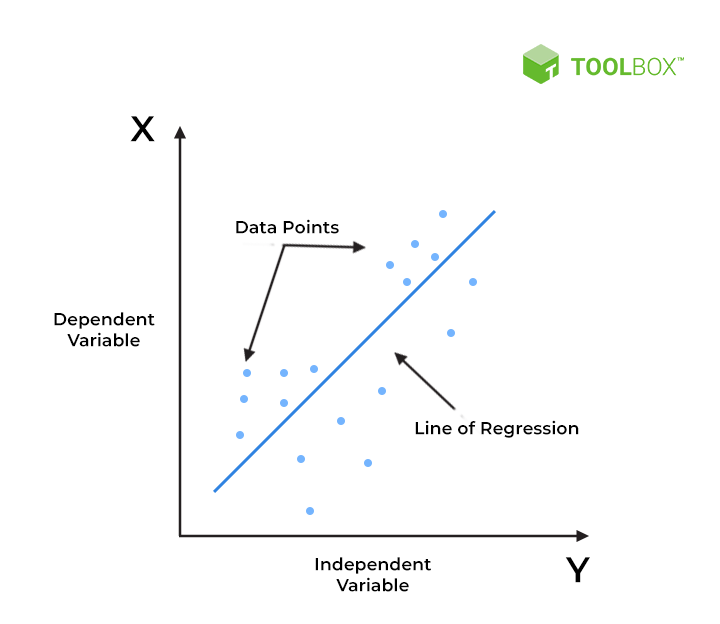
We have taken the method of Linear Regression

**Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.**

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.



**METHODOLOGY**

Proposed Methodology consists of the following steps: -

1.Data Analysing: - This is the first step involved in the process of making profit prediction. It consists of the data values of 50 start-up companies.

2.Data Pre-processing: - Data Pre-processing is done after loading the dataset. The entire dataset is split in the ratio of 2:8 for testing and training the model respectively.

3.Selecting a Model: - Multiple Linear Regression is the machine learning algorithms implemented in this model to predict the profit.

Multiple Linear Regression is usually used when there is a continuous dependent variable and two or greater than two independent variables. Similarly, for this model we have profit as a dependent variable and rest other parameters Marketing Spend, Administration Spend and R&D Spend as the independent variables .

4.Prediction of Profit: - One of the above-mentioned models can be used to make prediction of profit by providing the values of the required variables. These parameters include money spent for different causes such as R&D, Marketing and Administration purpose.

**IMPLEMENTATION**

1. **Importing libraries and datasets**

We have imported libraries such as numpy, pandas, matplotlib.pyplot and seaborn and loaded the dataset in the given excel csv file.

1. **Numerical or Statistical Analysis of dataset**

We have described the dataset in count , mean, std and others parameters.

1. **Schema of dataset**

In this with the help of info parameter we get the data type of each and every variable in the table.

1. **Correlation**

Here we can get that all our three columns (R&D spend, Administration, Marketing Spend) have a direct relationship with the profit, which is the target variable.

1. **Correlation Matrix**

It provides the direct correlation with the profit by giving the heatmap of the correlation plot.

1. **OLS Model**

In Ordinary least squared model, since r squared value determine that how the regression predictions approximate to the real data points. Since the respective value is 0.951(close to 1) which means it is close to perfectly fit data.

1. **Outliers Detection in the Target Variable**

While looking at the **boxplot** we can see the **outliers in the profit (target variable)**, but the amount of data is not much (just 50 entries) so it won’t create much negative impact.

1. **Histogram on profit**

**The average profit (which is 100k)** is the most frequent i.e. this should be in the category of distribution plot.

1. **Pair Plot**

As we can see in the pair pot, Research and development are directly proportional to the investment that we can do. The marketing spend seems to be directly proportional (though a little bit outliers are there) with the profit. There is no relationship between the second column and profit i.e. our target column.

1. **Calculation of Regression Metrices**

Here the intercept values, coeff values, and coefficient of determination is calculated.

1. **Data is splitted in training data and testing data**

Data is divided in two parts where 80 percent data go for training and 20 percent as testing data.

1. **Testing the model using the predict function**

**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**.

1. **Error Calculation**

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

**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 more close residual to the values found by mean squared error.

**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.

**CONCLUSION**

So, the mean absolute error is **6979.152252370402**. Therefore our predicted value can be **6979.152252370402 units more or less than the actual value.**

The best model is LINEAR REGRESSION, also known as ordinary least squares (OLS) and linear least squares, is the real workhorse of the regression world. We can use linear regression to understand the mean change in a dependent variable given a one-unit change in each independent variable. Since the R squared value of our model is **0.951** S is close to 1 and hence the data fit well in this regression model .As high r squared is good for the regression model. Here we also see that

R&D spend with profit as target variable shows the highest correlation.