T.Raghavendra Prasad, Technical Officer ‘B’

Combat Aircraft Systems Development and Integration Centre

Defence Research and Development Organisation

Using Multiple Linear Regression to Predict Profits for Startups

Introduction

As a part of IETE Project work, Machine Learning Model is used for predicting the profit from the startup's dataset with the given features. We will use the 50-startups dataset for this problem statement and we will be using the concept of Multiple linear regression to predict the profit of startups companies.

How this project helps the startup?

Startups are typically funded through various sources and methods, such as community-based funding, crowdfunding, angel investors, venture capitalists, and government grants/programs, among others. The funding sources and methods used by startups are similar to those used by other multinational corporations (MNCs), but the main difference is that startups work on a small scale to create innovative products that benefit customers, while established companies work on a larger scale by improving existing products.

How Startups are funded?

Startups are often funded through seed funding, which involves early-stage investors providing capital to help turn an idea into a product. This type of funding is typically provided by angel investors or venture capitalists who are willing to take on the risk of investing in a company that does not yet have a proven track record or market value.

To analyze investments in startups and determine profitability, there are several methods that can be used, including discounted cash flow analysis, net present value analysis, and internal rate of return analysis. These methods involve projecting future

cash flows and discounting them to determine their present value, which can help investors assess the potential return on their investment.

How the Linear Regression Model helps?

This particular machine learning model can be beneficial in situations where there is a need to determine profitability based on market expenditures. Essentially, the model can assist in identifying profits based on the amount of money spent from the 50 startup dataset.

About the Startup Dataset:

This dataset contains information on 50 startups located in New York, California, and Florida. The dataset includes data on R&D spending, administration spending, marketing spending, and location features, with the target variable being profit. The dataset provides information on how much each startup spends on research and development, administration, and marketing, as well as the state in which each startup is located, and the amount of profit each startup generates.

About the Source of Dataset:

Data Source: https://www.kaggle.com/datasets/farhanmd29/50-startups

Difference between Linear regression and Multiple linear regression

The primary distinction between linear regression and multiple linear regression is that linear regression involves a single independent variable, whereas multiple linear regression involves more than one independent variable.

To illustrate, consider the following examples:

Linear regression: Predicting an individual's height based on their weight.

Multiple linear regression: Modifying the above scenario slightly by including additional features such as age, gender, and height to predict an individual's weight, which would require the use of multiple linear regression.

For Numerical/ Statistical analysis of the dataset, we use the describe method in pandas.

```
dataset.describe()
```

Using the describe method, we get the count, min, max and standard deviation values.

To find out the number of rows and columns in the dataset we using the below function.

```
# Dimensions of dataset
print('There are ',dataset.shape[0],'rows and ',dataset.shape[1],'columns in the dataset.')
```

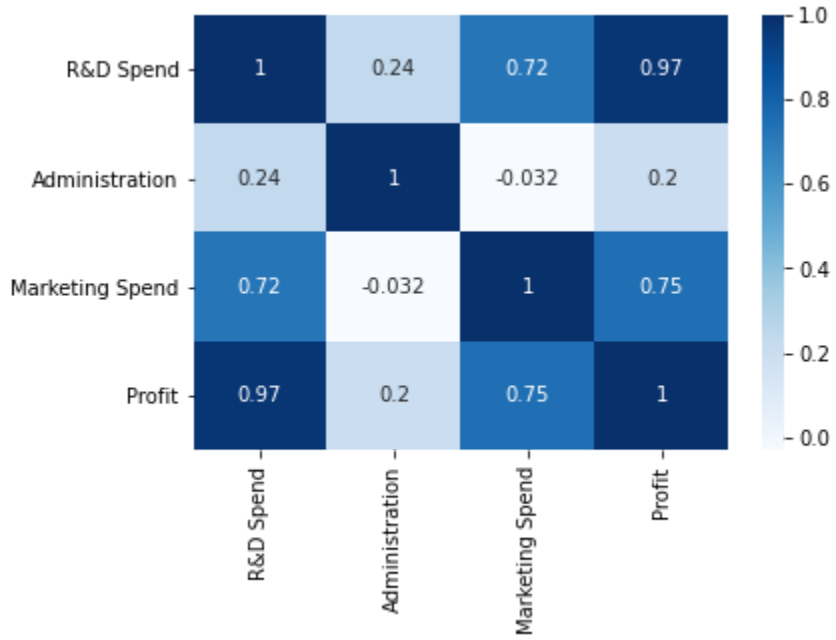
Data Cleaning

To check for null values in the dataset, we use isnull() function as below:

```
# Dimensions of dataset
print('There are ',dataset.shape[0],'rows and ',dataset.shape[1],'columns in the dataset.')
```

Correlation between the columns

To find out the correlation between the independent variables R&D Spend, Administration and Marketing Spend against the dependent variable profit we use the corr() function in pandas.



The heatmap of correlation also demonstrates the. **direct relationship of R&D Spend and Marketing Spend with profit.**

Development of the Linear Regression Model:

- Spliting Dataset in Dependent & Independent Variables
    - X = dataset.iloc[:, :-1].values
    - y = dataset.iloc[:, 4].values
- Encode labels with value between 0 and n_classes-1
    - labelencoder = LabelEncoder()
    - X[:, 3] = labelencoder.fit_transform(X[:, 3])
- split the data into training and testing data
    - x_train,x_test,y_train,y_test = train_test_split(X,y,train_size=0.7,random_state=0)
    - x_train

- Train the Machine Learning model with the dataset
- Predict the Profits
- We check the accuracy of the model with test set and training set
    - testing_data_model_score = model.score(x_test, y_test)
    - print("Model Score/Performance on Testing data",testing_data_model_score)

    - training_data_model_score = model.score(x_train, y_train)
    - print("Model Score/Performance on Training data",training_data_model_score)
- Model Score/Performance on Testing data 0.9355139722149948
- Model Score/Performance on Training data 0.9515496105627431

Predictedl Values	Actual Values
104055.184238	103282.38
132557.602897	144259.40
133633.012845	146121.95

As we can see that the **predicted value is somewhat near to the actual values**, therefore, we can use this model for prediction.

Model Evaluation

R squared score: It is one of those statistical approach by which we can find the variance or the spread of the dependent variables with the help of independent data.

R2 score of model is : 93.39448007716636

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

Root Mean Squared Error: This is the similar to Mean squared error(MSE) approach. The reason behind finding the root is to find the more close residual error as compared 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

Model Evaluation Parameters	**Value**
R Squared Score	93.39448007716636
Mean Squared Error	6224496238.94644
Root Mean Squared Error:	788954.7666974603
Mean Absolute Error:	6503.577323580025

For the above Model Evaluation parameters, we can conclude that since R Squared score is 93.94%, the Linear Regression Line is best fit for the give data set and Mean Absolute Error of 6503.577. Therefore our predicted value can be 6503.577323580025 units greater than or less than the actual value.

References

- http://www.ltcconline.net/greenl/courses/201/regression/scatter.htm
- https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/
- https://www.statology.org/scatterplot-with-regression-line-python/
- https://www.kaggle.com/datasets/farhanmd29/50-startups