**PREDICTION OF PROFIT FOR 50 STARTUPS: APPLICATION OF MULTIPLE LINEAR REGRESSION MODEL**

**Prepared by: Tarun Mondal**

**Table of Contents**

[**Objective**](#_wod5tikl95qf) **3**

[**Methodology**](#_hh893m3ny1s9) **3**

[**End-to-end process with solution architecture**](#_kdn45wkqrj3) **3**

[Importing libraries in Python](#_g1iyital8dy6) 3

[Data exploration](#_qra269l97i7c) 3

[Data preprocessing](#_rvhiwoilirvs) 5

[Exploratory data analysis (EDA)](#_u9b7btrzfs1z) 6

[Model development](#_mq1qqg5u4ar3) 10

[**Time Taken**](#_eb5hiccxtul7) **13**

[**Challenges faced**](#_o7jpdcxzc5da) **13**

[**Complexity level**](#_wmlari6ym45d) **13**

# Objective

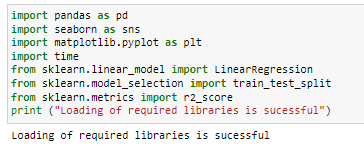
The main objective of this project is to predict the profit of 50 start-ups using a multiple linear regression model.

# Methodology

A machine learning approach (multiple linear regression model) has been incorporated in this study to predict the Profit of the company. Multiple linear regression models (70-30% splitting and 80-20% split) have been performed using Python in Jupyter Notebook.

# End-to-end process with solution architecture

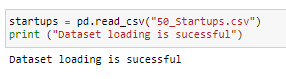
## Importing libraries in Python

****

**Figure 1: Python libraries**

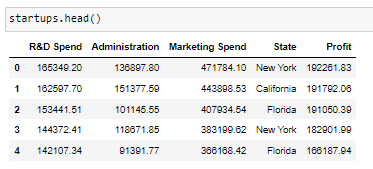
‘Pandas’ library has been imported into Python for data loading and manipulation, whereas Seaborn and Matplotlib libraries have been used for data visualisations. For developing machine learning model (Linear regression), ‘LinearRegression’ module has been imported from scikit-learn framework, whereas the evaluation metric (R2-score) has been imported from ‘sklearn.metrics module (**Refer to Figure 1**). For performing train-test splitting, from sklearn.model\_selection module, train\_test\_split has been imported into Python environment.

## Data exploration

****

**Figure 2: Data loading**

Through the use of ‘read’ function from ‘Pandas’ library, the ‘50\_startups’ dataset has been loaded.

****

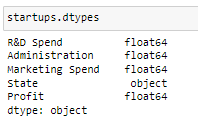
**Figure 3: Data head**

The first few rows of the dataset have been shown using ‘head ()’ function from pandas, from which it can be observed that the dataset contains 5 variables (R&D Spend, Administration, Marketing Spend, State and Profit).

****

**Figure 4: Column of the dataset**

Columns of the dataset have been shown using the ‘columns’ method, whereas data types have been evaluated using ‘dtypes’ method, from which it can be observed that the numerical variables are R&D Spend, Administration, Marketing Spend, and ‘Profit’, whereas ‘State’ is a categorical variable.

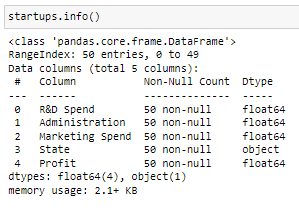
****

**Figure 5: Data types**

****

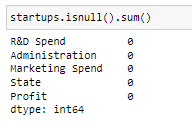
**Figure 6: Shape of the dataset**

Shape of the dataset has been checked using the ‘shape’ method from ‘pandas’ library, from which it has been found that the dataset contains 50 rows and 5 columns (**Refer to Figure 6**).

****

**Figure 7: Information of the dataset**

## Data preprocessing

****

**Figure 8: Checking null values in the dataset**

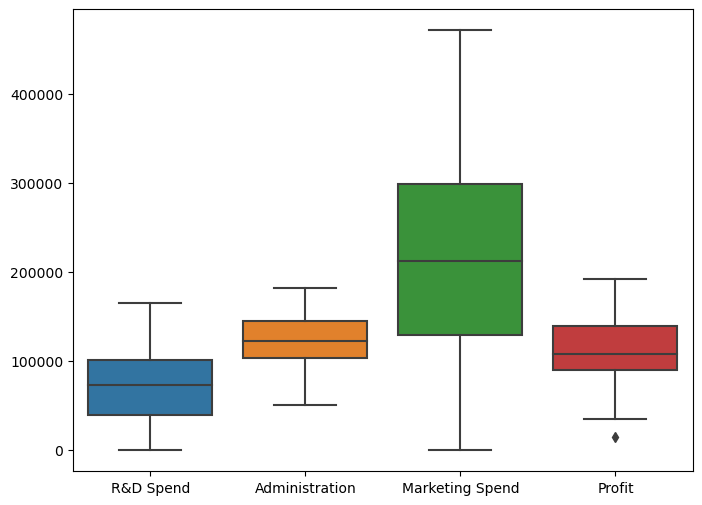
Missing values in the dataset have been checked using ‘isnull ().sum ()’ function, from which it can be observed that the dataset contains 0 missing values for both variables (**Refer to Figure 8**). This indicates that data errors are absent in the dataset, thus, data cleaning steps like dropping null values or filling the null values are not required in this project.

****

**Figure 9: Duplicate values in the dataset**

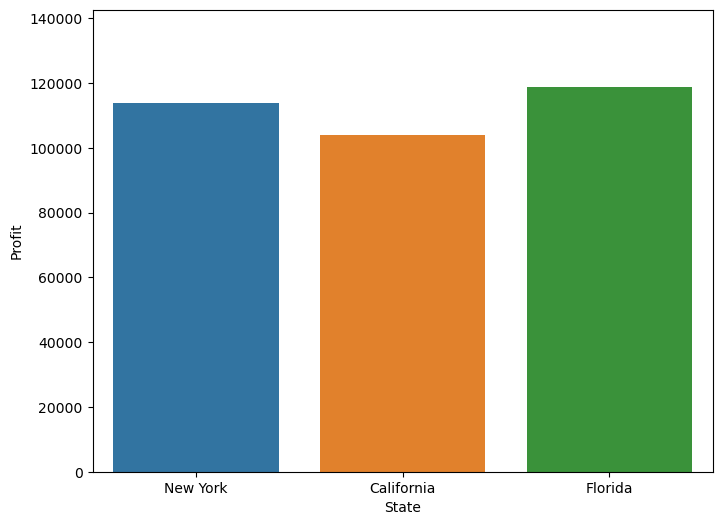
Duplicate values in the dataset have been checked using the ‘duplicated ().sum()’ function, from which total observed duplicate values are 0, indicating non-existence of repetitive observations in the dataset (**Refer to Figure 9**).

## Exploratory data analysis (EDA)

****

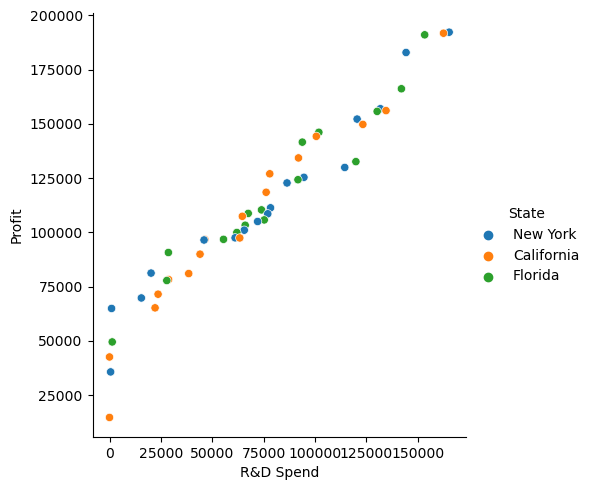
**Figure 10: Box plots**

From the box plot, it can be observed that the spread of marketing spend is considerably higher than R&D spending and administration costs (**Refer to Figure 10**). This leads to the inference that start-ups have employed more investment in marketing to enhance their sales and profitability.

****

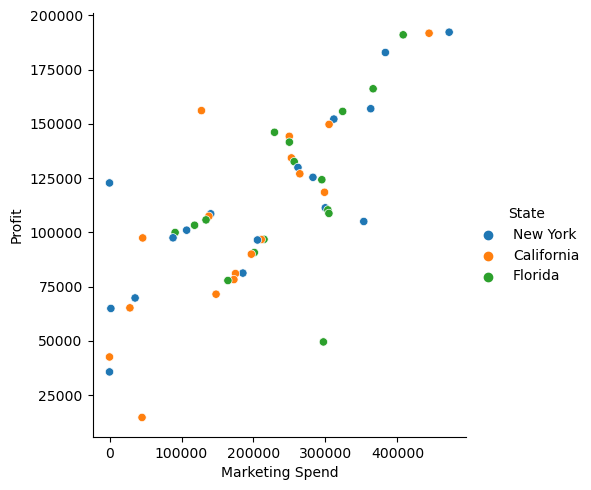
**Figure 11: Bar chart showing the distribution of Profit across the State**

From the above bar chart, it can be observed that Profit of the companies is considerably higher in Florida compared to New York and California. This leads to the potential inference that operating expenses in Florida may be low or the sales in Florida may be considerably higher than in New York and California.

****

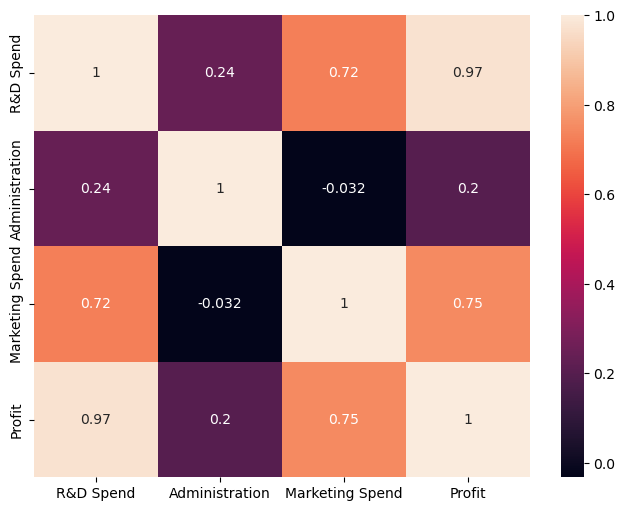
**Figure 12: Association between R&D Spend and Profit**

The scatter plot between R&D Spend and Profit revealed a positive relationship, indicating that with the increase in R&D spend, the Profit of the company has increased across all states (New York, California and Florida) (**Refer to Figure 12**).

****

**Figure 13: Scatter plot between Marketing spend and Profit**

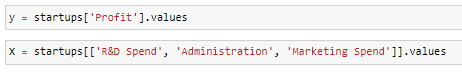
The scatter plot between ‘Marketing Spend’ and Profit revealed a positive relationship, indicating that with the increase in marketing spend, the profitability of the company has increased across all states (New York, California and Florida) (**Refer to Figure 12**). This may be the fact that spending on marketing activities has led to a drive in sales, which has further increased profitability of the company.

****

**Figure 13: Correlation heatmap**

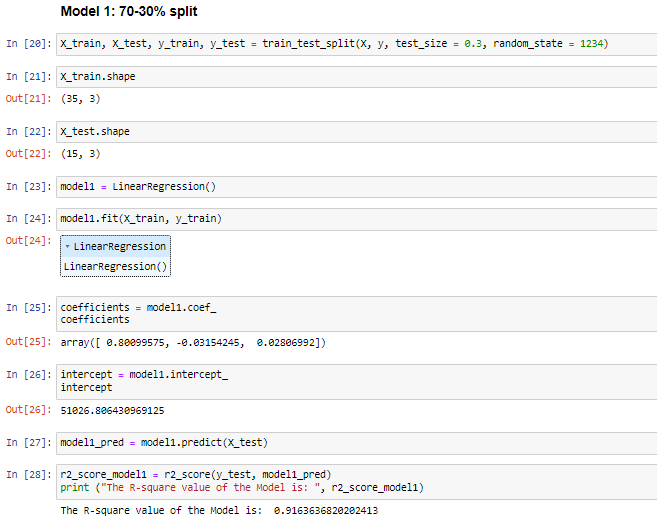
From the correlation heatmap, it can be observed that factors like R&D spend (correlation coefficient = 0.97) and marketing spend (correlation coefficient = 0.75) exhibit a strong positive correlation on Profit. (**Refer to Figure 13**).

## Model development

****

**Figure 14: Features and target variable**

The target variable in this study is ‘Profit’, while the features are ‘R&D Spend’, ‘Administration’ and ‘Marketing Spend’ (**Refer to Figure 14**).

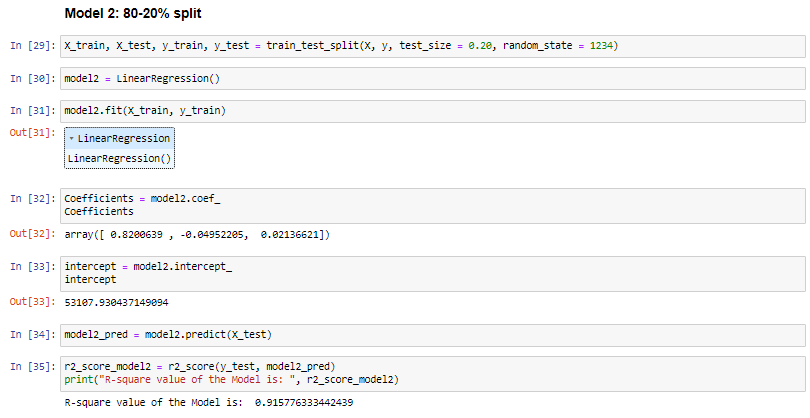
****

**Figure 15: Model architecture and model performance for multiple linear regression with 70-30% split**

From the multiple linear regression model (with 70-30% split), the obtained coefficients are respectively R&D Spend (0.800), Administration (-0.031) and Marketing Spend (0.028). The intercept obtained from the model is 51026.80. Thus, the obtained equation for Profit is:

**Profit = 51026.80 + 0.800 \* R&D Spend + (-0.031) \* Administration + 0.028 \* Marketing Spend …… (Equation 1).**

The obtained R-square value of the multiple linear regression model is 0.9163, indicating the model has the capability to explain 91.63% of variability in Profit (**Refer to Figure 15**).

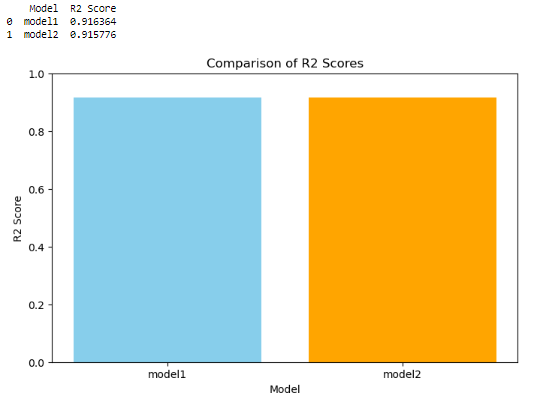


**Figure 16: Model architecture and model performance for multiple linear regression with 80-20% split**

From the multiple linear regression model (with 80-20% split), the obtained coefficients are respectively R&D Spend (0.820), Administration (-0.049) and Marketing Spend (0.0213). The intercept obtained from the model is 53107.93. Thus, the obtained equation for Profit is:

**Profit = 53107.93 + 0.820 \* R&D Spend + (-0.049) \* Administration + 0.0213 \* Marketing Spend …… (Equation 2).**

The obtained R-square value of the multiple linear regression model is 0.9157, indicating the model has the capability to explain 91.57% of variability in Profit (**Refer to Figure 16**).

****

**Figure 17: Model comparison**

The obtained R-square value of model with a 70-30% split is slightly higher than the model with an 80-20% split (**Refer to Figure 17**).

# Time Taken

****

**Figure 18: Execution time**

The total time required for execution of this project is 5.00 seconds, indicating the high speed of the system in estimating delivery time.

# Challenges faced

Challenges have been faced in this project in identifying the appropriate data-splitting ratio.

# Complexity level

The dataset is very small and contains only 5 variables and 50 observations with no data errors (such as missing values, duplicate values and outliers), reflecting the project is very simple. However, due to the robustness of the model and application of a proper Machine Learning flow, this project can also apply to large datasets.