VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANASANGAMA, BELAGAVI – 590018

An Internship Report
on

Startup Profit Prediction Using Machine Learning

Submitted in partial fulfillment for the award of degree of

Bachelor of Engineering
In

Computer Science and Engineering

Submitted by

PRANITH RAO

4SO18CS088

Internship Carried Out
at

EXPOSYS DATA LABS DODDABALLAPUR
MAIN ROAD SINGANAYAKANAHALLI,
BENGALURU, KARNATAKA - 560064

Internal Guide
Ms GAYANA M N
Assistant professor
St Joseph Engineering College

External Guide
Mr. Vishnuvardhan Y
Company Head
Exposys Data Labs

Department of Computer Science and Engineering
St Joseph Engineering College
Mangaluru - 575028
2021-2022

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANASANGAMA, BELAGAVI – 590018

An Internship Report
on

Startup Profit Prediction Using Machine Learning

Submitted in partial fulfillment of the requirements for the degree

Bachelor of Engineering
in

Computer Science and Engineering

Submitted by

PRANITH RAO **4SO18CS088**

Department of Computer Science and Engineering
St Joseph Engineering College
Mangaluru - 575028
2021-2022

St Joseph Engineering College
Mangaluru – 575 028
Department of Computer Science and Engineering

CERTIFICATE

Certified that the Internship Work titled **“Startup Profit Prediction using Machine Learning”** was carried out by **Mr. PRANITH RAO**, bearing USN **4SO18CS088**, a bonafide student of final year B.E. in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi, during the year 2021-2022. Further, it is certified that all corrections/suggestions indicated during Internal Evaluation have been incorporated in this report.

------------------------	--------------------------------	-------------------------------
Ms Gayana M N **Internal Guide**	**Dr Sridevi Saralaya** **Head of the Department**	**Dr Rio D’Souza** **Principal**

External Viva Voce Examination

Name of the Examiners **Signature with Date**

1. .... ------------------------------------

2. ....

Exposys
Data Labs

EXPOSYS
DATA LABS

Certificate of Internship

TO WHOM IT MAY CONCERN

This is to certify that **Mr. PRANITH RAO** has completed internship programme on **“Data Science”** from 06.04.2022 to 05.05.2022.

He took keen interest in the work assigned and successfully completed it. During the period of internship we found him to be punctual, hardworking and inquisitive.

We wish him luck and success in all his future endeavours.

EXPOSYS DATA LABS
BENGALURU

Y Vishnuvardhan

Chief Director

hr@exposysdata.com
www.exposysdata.com

DECLARATION

I, **Pranith Rao,** bearing USN **4SO18CS088**, student of final year B.E. in Computer Science and Engineering, St Joseph Engineering College, Mangaluru, hereby declare that the Internship Work titled **“Startup Profit Prediction Using Machine Learning”** has been duly executed by me from 6th April 2022 to 5th May 2022, at Exposys Data Labs, Bangalore. Further, the “Task Performed” of this report represents the work done solely by me and does not contain any statements falsely claiming work done by others, as my own.

Date: 07/05/2022

Place: Mangaluru **Pranith Rao**

ACKNOWLEDGMENT

I dedicate this page to acknowledge and thank those responsible for shaping this project. Without their guidance and help, this experience would not have been so smooth and efficient.

I would like to extend my sincere gratitude to **Mr. Vishnuvardhan Y, Company Head, Exposys Data Labs** for giving me the opportunity to complete my internship at his start-up, his guidance and encouragement helped me throughout the internship.

I sincerely thank **Ms Gayana M N, Assistant Professor**, Department of Computer Science and Engineering for her guidance which helped us fulfil the requirements prescribed by the university and her valuable suggestions which brought this internship to fruition.

I am indebted to **Dr Sridevi Saralaya, Head of the Department** of Computer Science and Engineering, whose kind consent and guidance helped us complete this internship successfully.

I am extremely thankful to our Principal **Dr Rio D'Souza**, our Director **Rev.Fr. Wilfred Prakash D'Souza** and our Assistant Director **Rev.Fr. Alwyn Richard D'Souza** for their support and encouragement.

I would like to thank all our Computer Science and Engineering staff members who have always been with us extending their support, precious suggestions, guidance and encouragement through the project.

I also like to extend thanks to my friends and family members for their continuous support

Executive Summary

I carried out my internship in Machine Learning and Artificial Intelligence at Exposys Data Labs, Bangalore from 6th April 2022 to 5th May 2022.

Exposys Data Labs is a Bengaluru based company which aims to solve real world business problems like Automation, Big Data and data Science. Exposys Data labs also helps businesses to identify issues, opportunities and prototype solutions using trending technologies like AI, ML, Deep Learning and Data Science.

The Objective of the internship was to build Startup Profit Prediction Using Machine Learning. The languages and tools used were python, data pre-processing, Data Visualization Using Matplotlib, machine learning algorithms such as Random Forest and Multi-linear Regression and Jupyter Notebook.

This internship has guided me to gain knowledge on Machine learning and algorithms and also has helped me to become familiar with tools such as Jupyter notebook and Python. This internship has also given me an opportunity to enhance my work ethic, professional skills and gain industry knowledge. It also encouraged me to learn new technologies and tools like machine learning and its algorithms, Python and Jupyter notebook. Further it has helped me in decision making, communication and time management skills.

PRANITH RAO

CONTENTS

Figure Index

CHAPTER 1
ABOUT THE COMPANY

1.1 Brief History

Exposys Data Labs is a Bengaluru based company which aims to Solve real world business problems like Automation, Big Data and data Science. Exposys Data labs also helps businesses to identify issues, opportunities and prototype solutions using trending technologies like AI, ML, Deep Learning and Data Science.

1.2 Services Offered by the Company

Software Development: They develop and test the product idea and deliver a software that will satisfy their clients, sustain competition and achieve highest returns.

Web Application Development: They provide software and hardware; design and development services range from business-critical applications to powerful Augmented and Virtual Reality applications.

Internships: Exposys has partnered with several companies ranging from startups to well-established MNCs providing a unique interning experience which equips students.

1.3 Contact Details

Address:
P M R. Residency, Ground Floor, No-5/3 Sy. No.10/6-1
Opp Nithyotsava Wedding Hall, Doddaballapur Main Road
Singanayakanahalli, Yelahanka. Bengaluru, Karnataka
560064

Phone: +91 7795207065

E-mail: hr@exposysdata.com

CHAPTER 2
ABOUT THE DEPARTMENT

2.1 Introduction

This was project was carried out under the guidance of Exposys Data Labs. The trainers are specialists in areas like Machine Learning and Artificial Intelligence, Cybersecurity, Cross Platform Application Development etc.

The mission of the department: Create a learning and development ecosystem using modern computing technologies to solve real-time problems.
The vision of the department: Learning and Training in IT education by exploring modern computing technologies and trends.

2.2 Roles and responsibilities

I was assigned to work on a project, by the external guide. The project assigned was to create a Machine learning model to predict the profit earned by the start-ups. The model that we have implemented in our project for the given data set were using different Machine Learning Algorithms like multi-linear regression, Random Forest Regressor, etc.

To build the required model, following steps were performed:

- Importing Dataset: The 50 startups Dataset was shared to me via email by the company. Additionally, I downloaded the dataset that was available on Kaggle for better training of the model.
- Pre-processing Dataset: Each review undergoes through a pre-processing step, where all the vague information is removed.
- Analysis Conclusion: In this study, on implementation, the prediction results show the correlation among different attributes considered. Multiple instances, parameters and various factors can be used to make classification more innovative and successful.

CHAPTER 3

TASKS PERFORMED

3.1 Daily Work Schedule

Day 1: Joining date

Date	06th April 2022
Task Assigned	Joining formalities
Task Objective	Communication with Mentor
Task Outcome	Establish means of communication with the company and understand the work assigned.
Brief Description of the Work	
Mentor contacted me via WhatsApp and explained me the work assigned and in what ways I can approach the problem statement.	

Day 2-4: Understanding Python

Date	07th April 2022 - 09th April 2022
Task Assigned	Basics of Python
Task Objective	Understand python basics
Task Outcome	Understand the logic and implementation of the python basics
Brief Description of the Work	
I had to set up the Environment Variables and install Python software with Jupyter Notebook initially. Following that, I had to look over the documentation for various keywords, identifiers, statements, datatypes, and understand the usage of variables, functions, and arguments.	

Day 5-7: Python and its application in Machine Learning

Date	10th April 2022 – 12th April 2022
Task Assigned	Python in Machine Learning

Task Objective	Understand python in ML
Task Outcome	Understand Simple Machine Learning model in Python
Brief Description of the Work	
I had to train a simple Linear Regression Model and compare its outcomes to the expected outcomes.	

Day 8-10: Supervised and Unsupervised Learning

Date	13th April 2022 – 15th April 2022
Task Assigned	Simple Linear Regression, Multiple Linear Regression, Clustering.
Task Objective	To understand the concept of Simple Linear Regression, Multiple Linear Regression.
Task Outcome	To be able to comprehend and apply the Pre-processing, Simple Linear Regression, Multiple Linear Regression, Concepts.
Brief Description of the Work	
Learning about pre-processing was necessary as format of the data has to be in a proper manner for achieving better results from the applied model in Machine Learning projects. And as multiple variables were responsible in deciding the value of the target variable, I went through some models which were helpful in predicting the same.	

Day 11-12: Knowledge about different techniques

Date	16th April 2022 – 17th April 2022
Task Assigned	Homoscedasticity and Heteroscedasticity error terms
Task Objective	To understand the heteroscedasticity of error terms.
Task Outcome	To be able comprehend metrics of regression techniques.
Brief Description of the Work	

I learned about various strategies as well as why and how the heteroscedasticity variance of mistakes does not remain constant across data.

Day 13-14: Clustering and Sentiment Analysis

Date	18th April 2022 – 19th April 2022
Task Assigned	KMeans Clustering, Random Forest approach, Lasso Regression
Task Objective	To understand about the types of clustering and recommendation engine. Also, about interaction between machine such as computers and natural languages used by human beings.
Task Outcome	To be able to comprehend and apply the learned concepts in project work.
Brief Description of the Work	
Since I had to use at least 2 models to predict the target variable and chose the best among it I went through different training models and used the best among them for my project.	

Day 15: Project work

Date	20th April 2022
Task Assigned	The project assigned was to create a ML model that can be used in ‘Start-up profit prediction’
Task Objective	To be able to understand the concepts which was learnt during this internship, and apply it to the project work.
Task Outcome	To be able to apply the learnt concepts in the project works.
Brief Description of the Work	
I started to work on the project assigned to me by the external guide. The project assigned was to create ML models that can be used in ‘Start-up profit prediction’ and to work on its accuracy score.	

Day 16-17: Collection and preparation

Date	21st April 2022 – 22nd April 2022
Task Assigned	Collecting data for training the ML model
Task Objective	The data set is collected from Kaggle
Task Outcome	List of attributes used for prediction were collected
Brief Description of the Work	
Collecting data for training the ML model is the basic step in the machine learning pipeline. The data set and the attributes were shared to me via email and additional dataset were collected from Kaggle for better training of the model.	

Day 18-19: Data pre - processing

Date	23rd April 2022 – 24th April 2022
Task Assigned	Cleaning the data and making and it suitable for a machine learning model. Finding missing values and eliminating inconsistencies.
Task Objective	Replace null and eliminate duplicate values.
Task Outcome	Improve the accuracy and efficiency of a machine learning model.
Brief Description of the Work	
Data pre-processing operations are required for cleaning the data and making it acceptable for a machine learning model, which improves the model's accuracy and efficiency. I checked for null and duplicate values but none were present so there was no need of any data replacement and elimination.	

Day 20-22: Technique selection

Date	25th April 2022 – 27th April 2022
Task Assigned	Selection of techniques relevant to the given data.

Task Objective	Techniques that train the model better to determine the target attribute were to be selected.
Task Outcome	Techniques like Multi-linear regression and random forest regressor were selected.
Brief Description of the Work	
I chose multi-linear regression and random forest regressor techniques to train my model to determine the profit of start-ups as multiple variables were responsible for determining the value of the target variable.	

Day 23-24: Splitting data into train data and test data

Date	28th April 2022 – 29th April 2022
Task Assigned	Splitting data into two subsets: training data and testing data.
Task Objective	To make predictions on the test data using Linear Regression and Random Forest Regressor.
Task Outcome	The split percentage is 70 and 30 respectively for train and test data
Brief Description of the Work	
I divided the data into two categories: training and testing. For train and test data, the split percentage is 70% and 30%, respectively. I used Linear Regression and Random Forest Regressor to fit the model to the train data in order to generate predictions on the test data.	

Day 25-27: Performance evaluation

Date	30th April 2022 – 2nd May 2022
Task Assigned	Improve model performance
Task Objective	Accuracy Prediction

Task Outcome	The training score was found for both the models and the best model was chosen to predict the profit of start-ups based on different expenditures.
Brief Description of the Work	
A predictive system was built to estimate start-up profit based on several spending metrics that will be provided as inputs to the machine learning model.	

Day 28-30: Project Submission

Date	3rd May 2022 – 5th May 2022
Task Assigned	To prepare the final internship report
Task Objective	To prepare the report and submit the same as a sign of successful completion of the internship.
Task Outcome	To document the entire internship process.
Brief Description of the Work	
I submitted a video demo explaining the entire project, project report and a PPT presentation all zipped in a file via email marking the completion of my internship.	

3.2 Project Implementation

The 50 start-ups dataset which was shared to me via email is used in this project. The Machine Learning algorithms performed on this data set are Linear Regression, Random Forest Regressor, Lasso and Decision Tree Regressor. The main features in the dataset are R&D Spend, Administration and Marketing Spend. These features together will help in predicting the profit that could be earned by the respective startup.

Collection and Preparation:

Collecting data for training the ML model is the basic step in the machine learning pipeline. The predictions made by ML systems can only be as good as the data on which they have been trained. The description of the dataset collected is as follows:

List of attributes used for prediction:

- R&D Spend
- Administration
- Marketing Spend

Target Attribute:

- Profit

Data pre-processing:

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing operations are required for cleaning the data and making it acceptable for a machine learning model, which improves the model's accuracy and efficiency. This includes following steps:

- Finding missing values: No missing values were found in our dataset.
- Finding duplicate values: No duplicate values were found in our dataset.
- Exploratory data analysis: This is necessary in order to discover patterns, detect anomalies outliers and to find interesting relations among attributes.

Splitting data into train and test:

We split the data into two subsets: training data and testing data. The split percentage is 70 and 30 respectively for train and test data. We fit our model on the train data, in order to make predictions on the test data.

Machine Learning Algorithms:

1. **Multiple Linear Regression:** Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

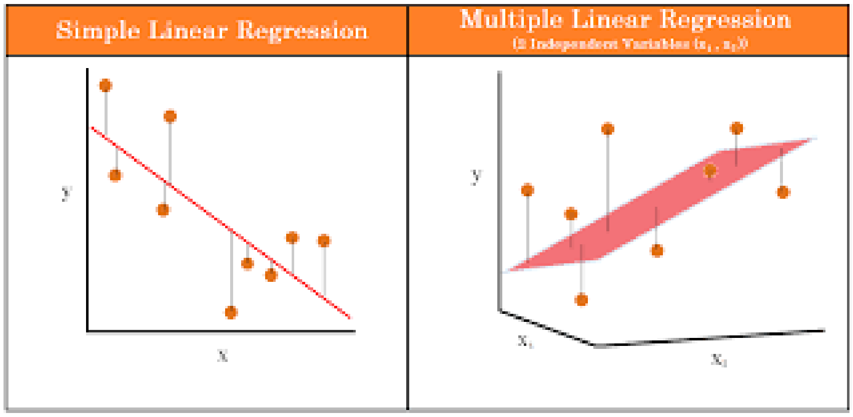


Fig 3.2.1 Multiple Linear Regression

2. **Random Forest Regressor:** A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter and if bootstrap=True (default), otherwise the whole dataset is used to build each tree. It combines ensemble learning methods with the decision tree framework to create multiple randomly drawn decision trees from the data, averaging the results to output a new result that often leads to strong predictions/classifications.

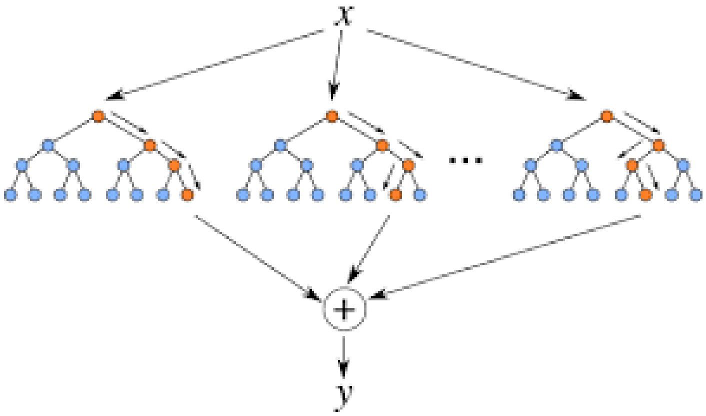


Fig 3.2.2 Random Forest Regressor

3. **Lasso:** The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.

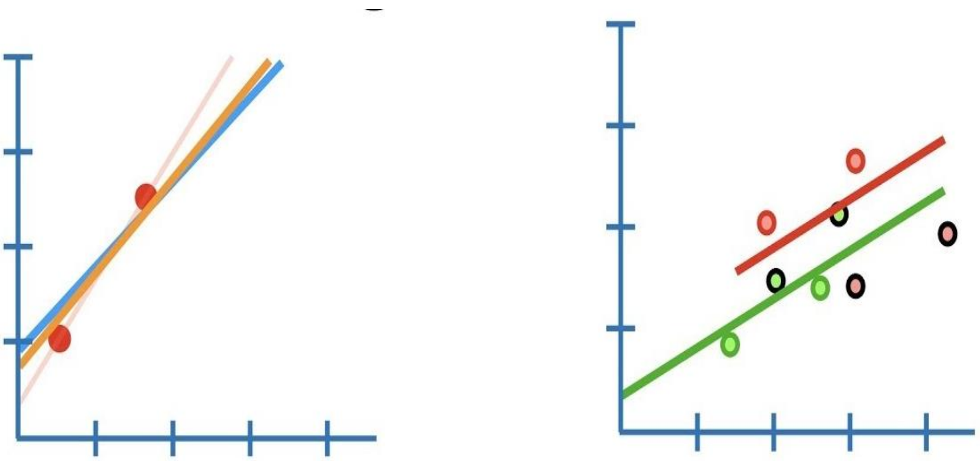


Fig 3.2.3 Lasso Regression

4. **Decision Tree Regressor:** Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. Leaf node represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node.

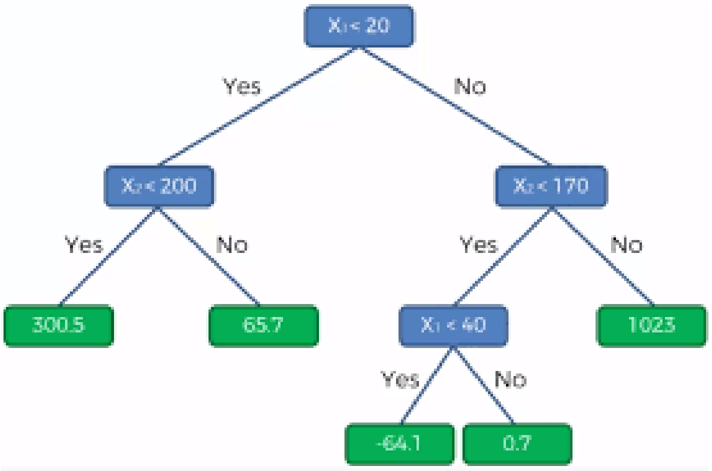


Fig 3.2.4 Decision Tree Regressor

Workflow:

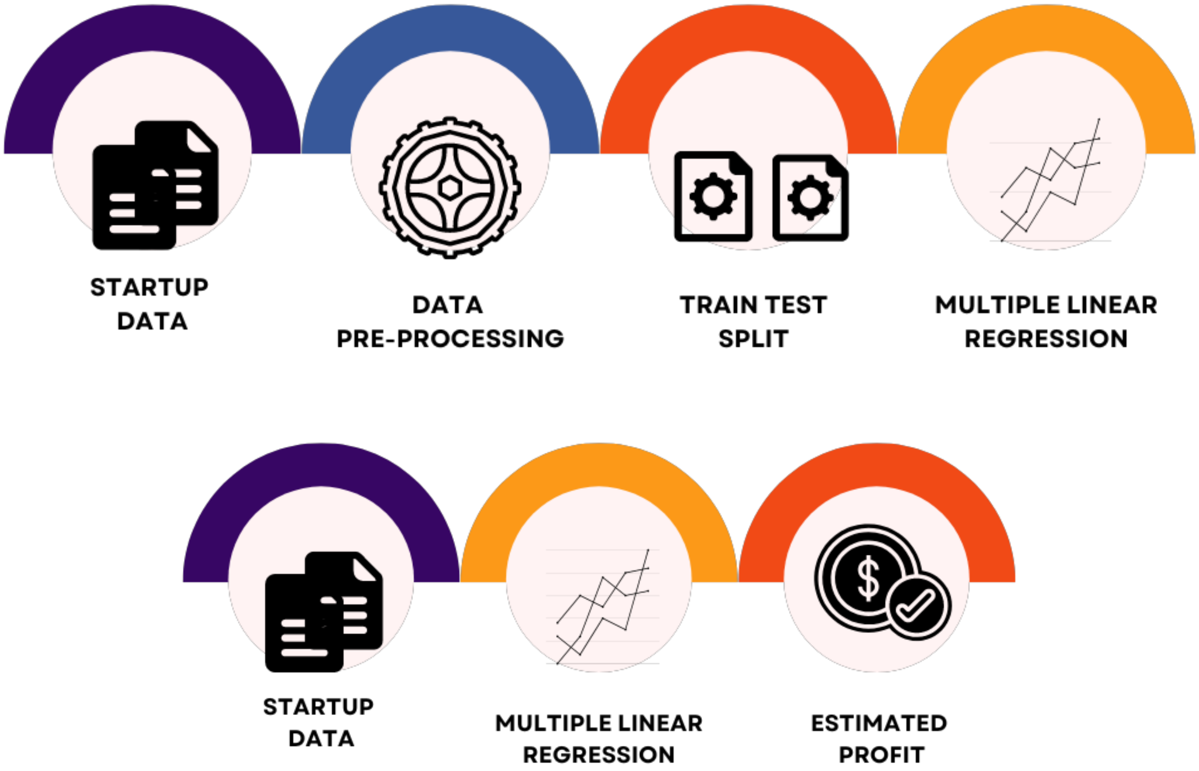
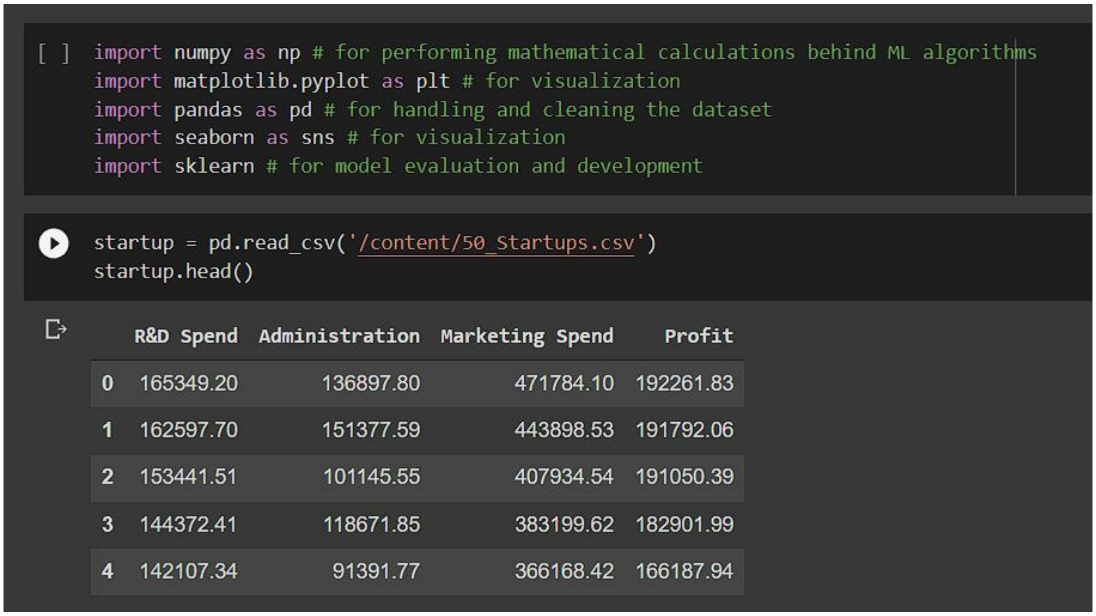


Fig 3.2.5 Workflow of the model

3.3 Snapshots

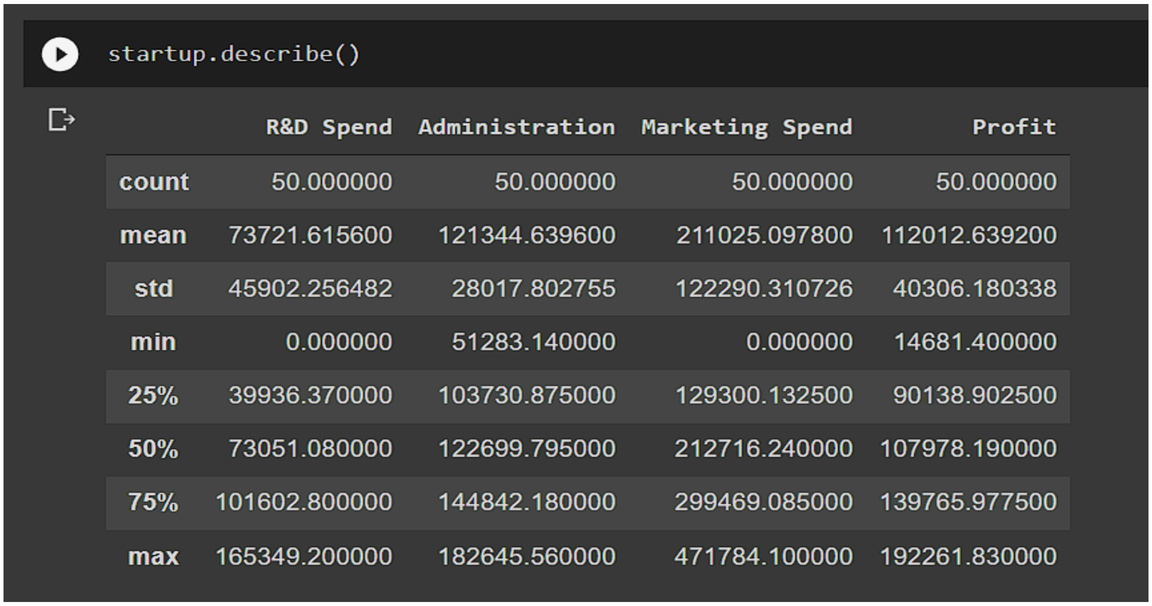


```
import numpy as np # for performing mathematical calculations behind ML algorithms
import matplotlib.pyplot as plt # for visualization
import pandas as pd # for handling and cleaning the dataset
import seaborn as sns # for visualization
import sklearn # for model evaluation and development
```

```
startup = pd.read_csv('/content/50_Startups.csv')
startup.head()
```

```
   R&D Spend  Administration  Marketing Spend     Profit
0  165349.20       136897.80        471784.10  192261.83
1  162597.70       151377.59        443898.53  191792.06
2  153441.51       101145.55        407934.54  191050.39
3  144372.41       118671.85        383199.62  182901.99
4  142107.34        91391.77        366168.42  166187.94
```

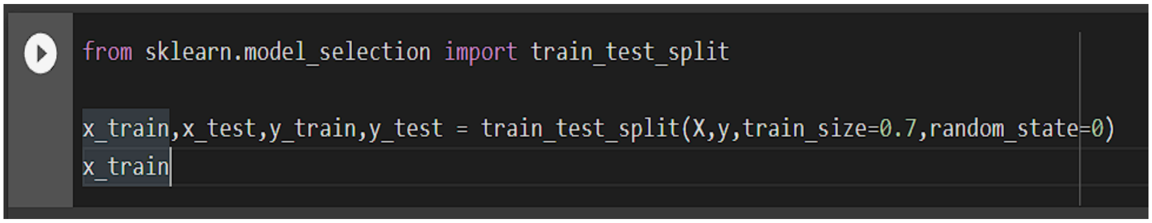
Fig 3.3.1 Preprocessing the Dataset



```
startup.describe()
```

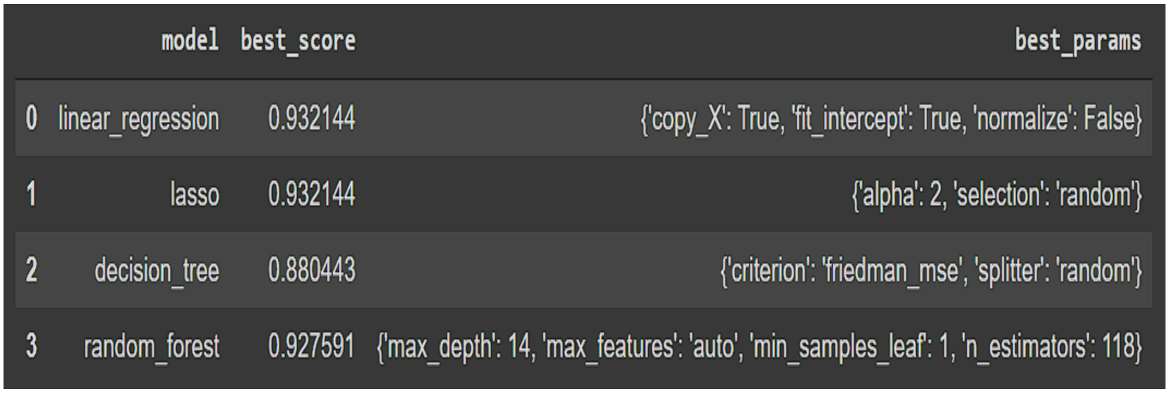
	R&D Spend	Administration	Marketing Spend	Profit
count	50.000000	50.000000	50.000000	50.000000
mean	73721.615600	121344.639600	211025.097800	112012.639200
std	45902.256482	28017.802755	122290.310726	40306.180338
min	0.000000	51283.140000	0.000000	14681.400000
25%	39936.370000	103730.875000	129300.132500	90138.902500
50%	73051.080000	122699.795000	212716.240000	107978.190000
75%	101602.800000	144842.180000	299469.085000	139765.977500
max	165349.200000	182645.560000	471784.100000	192261.830000

Fig 3.3.2 Data set parameters



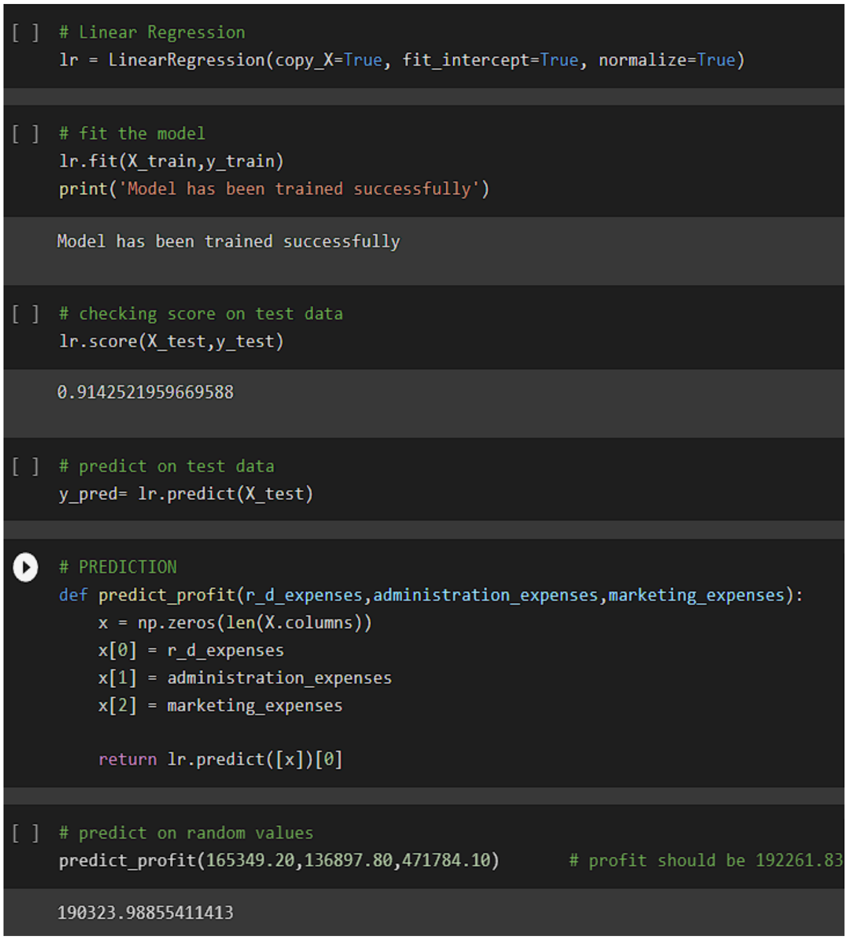
```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test = train_test_split(X,y,train_size=0.7,random_state=0)
x_train
```

Fig 3.3.3 Splitting the data into training and testing data



	model	best_score	best_params
0	linear_regression	0.932144	{'copy_X': True, 'fit_intercept': True, 'normalize': False}
1	lasso	0.932144	{'alpha': 2, 'selection': 'random'}
2	decision_tree	0.880443	{'criterion': 'friedman_mse', 'splitter': 'random'}
3	random_forest	0.927591	{'max_depth': 14, 'max_features': 'auto', 'min_samples_leaf': 1, 'n_estimators': 118}

Fig 3.3.4 Calculation of best score using different algorithms



```
[ ]  # Linear Regression
     lr = LinearRegression(copy_X=True, fit_intercept=True, normalize=True)
```

```
[ ]  # fit the model
     lr.fit(X_train,y_train)
     print('Model has been trained successfully')
```

```
     Model has been trained successfully
```

```
[ ]  # checking score on test data
     lr.score(X_test,y_test)
```

```
     0.9142521959669588
```

```
[ ]  # predict on test data
     y_pred= lr.predict(X_test)
```

```
     # PREDICTION
     def predict_profit(r_d_expenses,administration_expenses,marketing_expenses):
         x = np.zeros(len(X.columns))
         x[0] = r_d_expenses
         x[1] = administration_expenses
         x[2] = marketing_expenses

         return lr.predict([x])[0]
```

```
[ ]  # predict on random values
     predict_profit(165349.20,136897.80,471784.10)      # profit should be 192261.83
```

```
     190323.98855411413
```

Fig 3.3.5 Using the best model to predict

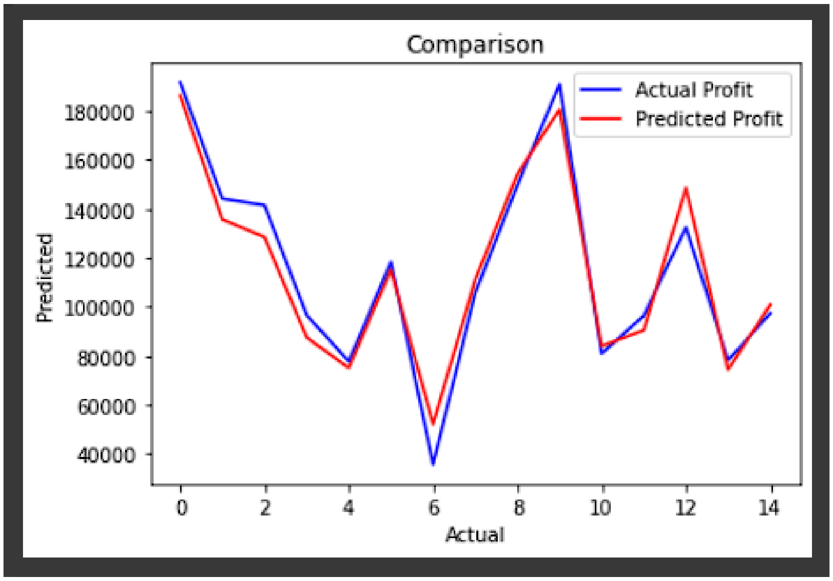
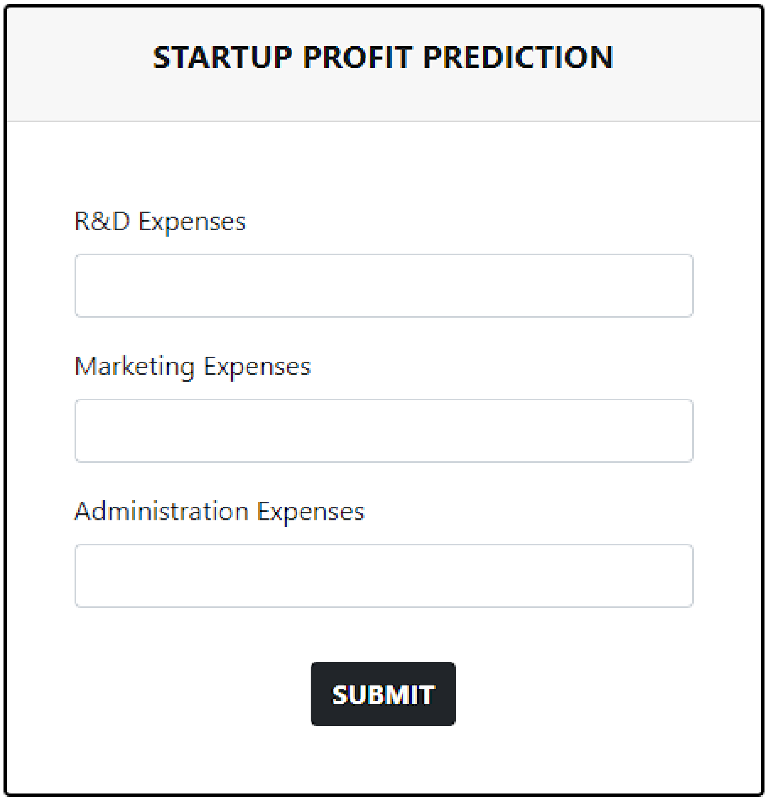


Fig 3.3.6 Comparison between Actual & Predicted profit



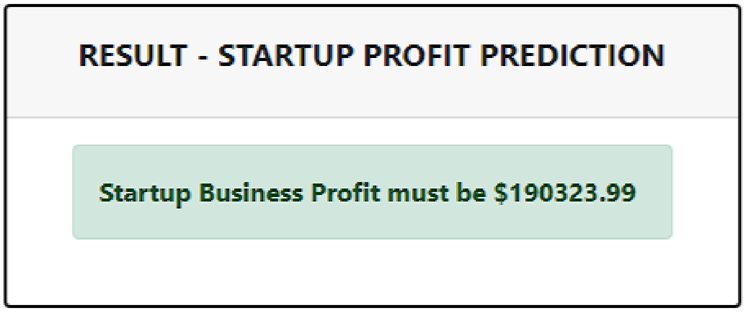
STARTUP PROFIT PREDICTION

R&D Expenses

Marketing Expenses

Administration Expenses

Fig 3.3.7 Flask Application



RESULT - STARTUP PROFIT PREDICTION

Startup Business Profit must be $190323.99

Fig 3.3.8 Result Page

CHAPTER 4
REFLECTION NOTES

4.1 Experience

My time at Exposys Data Labs was extremely beneficial because it allowed me to develop my technical abilities and learn more about Machine Learning. Mr. Abhishek Kumar, my guide, was really encouraging and accommodating. Working for the organization was a fantastic experience. They assisted me in improving my communication skills and teamwork. I was able to learn essential information about Machine Learning and different new technologies and tools. It enabled me to get professional expertise and experience working in a real-world production setting. It improved my work ethic, as I had to meet deadlines for various tasks. Overall, I had a very fruitful experience during my internship at Exposys.

4.2 Technical Outcomes

4.2.1 Learned new concepts and Technologies: During the internship, I had to constantly learn new concepts and technologies such as Data Pre-processing and preparing it for machine learning algorithms. After that, I learnt about Classification and Regression algorithms with the basics of Natural Language Processing.

4.2.2 Understood the importance of accuracy in Machine Learning: Machine learning models are used by businesses to make practical business decisions, and more accurate model outcomes lead to better conclusions. Errors have a high cost, but improving model accuracy lowers that cost. Of course, there is a point at which the benefit of constructing a more accurate model does not result in a matching rise in profit, but it is often advantageous across the board. For example, a false positive cancer diagnosis costs both the hospital and the patient. The advantages of enhancing model accuracy include saving time, money, and worry.

4.2.3 **Understood the importance of Machine Learning and its different types:** Machine learning is significant because it allows businesses to see trends in customer behavior and business operating patterns while also assisting in the development of new goods. Machine learning is at the heart of many of today's most successful businesses, like Facebook, Google, and Uber. For many businesses, machine learning has become a crucial competitive differentiation. There are four main strategies: Reinforcement Learning, Supervised Learning, Unsupervised Learning, Semi-supervised Learning.

4.3 Non-Technical Outcomes

4.3.1 Time Management:

Because frequent assignments were assigned with tight deadlines during the internship, time management was essential. The deadlines were met thanks to careful prioritization and hard labor.

4.3.2 Teamwork:

The internship provided me with the opportunity to work alongside Machine Learning experts. Because the majority of the tasks assigned were team activities, it also strengthened my capacity to collaborate with my peers.

4.3.3 Communication:

During my internship, I was able to improve my communication skills and convey ideas in a better way.

4.3.4 Adaptability Skills:

During this internship, we had a lot of tasks which had to be completed within the week itself. So, managing both the assignments and session, I finally adapted to the work environment.

CONCLUSION

Overall, this internship was a fantastic opportunity. I have learned new information and abilities, as well as met some of my learning objectives in this sector. Got knowledge of practically every element of working in an organization. I learned the value of punctuality, utmost dedication, and team spirit while working in a corporation. The internship also assisted me in better understanding my own strengths and shortcomings. It also helped me recognize what abilities and information I need to have in order to flourish in my career. During the first two weeks, I learnt the fundamentals of Python, how to use Google Colab, and discussed several machine learning techniques for developing a model. I was able to finish the project utilizing machine learning methods at the conclusion of the internship. Throughout the internship, I learned the importance of critical and analytical thinking, time management, goal management, and colleague relationships while working as part of a team for a company.

REFERENCES

[1] Exposys Data Labs: http://exposysdata.com/

[2] Data Science: https://towardsdatascience.com/artificial-intelligence/home

[3] Kaggle: https://www.kaggle.com/datasets/root64shivansh/profit-in-startup-of-a-company

[4] ML Algorithms: https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/