

## EXPOSYS INTERNSHIP REPORT

Title: Data Science Company profit prediction

Name: Anirudh Uday Parvatikar

Class: MCA 2<sup>nd</sup> Semester

USN: 1BY22MC007

College: BMSIT Bangalore

## **ABSTRACT**

The internship is mainly aimed to test our data analytical skills which has to be applied in solving any data science problem as given here “Company profit prediction”, This project aims to predict profit which can be gained by any company based on how much funds have been invested/used for Research & development(R&D), money spent on administration, fund invested for marketing purposes. These 3 components can be used as input to determine the profit which we can have by launching product into the market.

This internship aims to contribute into the field of financial sector by presenting innovative, data-driven approach for predicting profits for any given company. The model which has been developed in this project can be used by the company in financial analytics, investors and corporate strategies to enhance their decision-making processes.

## Table of Contents

1. Introduction	01
2. Existing Method	02
3. Proposed method	03
4. Methodology	05
5. Implementation	07
6. Conclusion	10

## INTRODUCTION

In today's market it is important for any company to estimate their profits for whatever they are investing for their product, this helps them in efficient usage of their capital and decide if its optimal working on any given product. In recent times as technology is evolving at much faster pace companies can make use of large data which is produced by the company and historical data to make predictions using data science algorithms like Linear Regression, Decision Tree regressors and other ensemble functions.

The dataset that is used in this project comprises of historical financial data, market trends and other economic indicators. These variables collectively shape a company's financial performance and are very important for accurate profit projection. To ensure a thorough representation of a company's operational landscape, we utilize a diverse dataset encompassing various industries, coupled with historical financial records.

This project also explores various machine learning algorithms, including some of the complex functions like ensemble methods and deep learning models. Each algorithm is rigorously evaluated for accuracy and compared with outcome of every other algorithm used to determine its effectiveness on the given data. The models are cross-validated to estimate their robustness and ability to generalize beyond trained data.

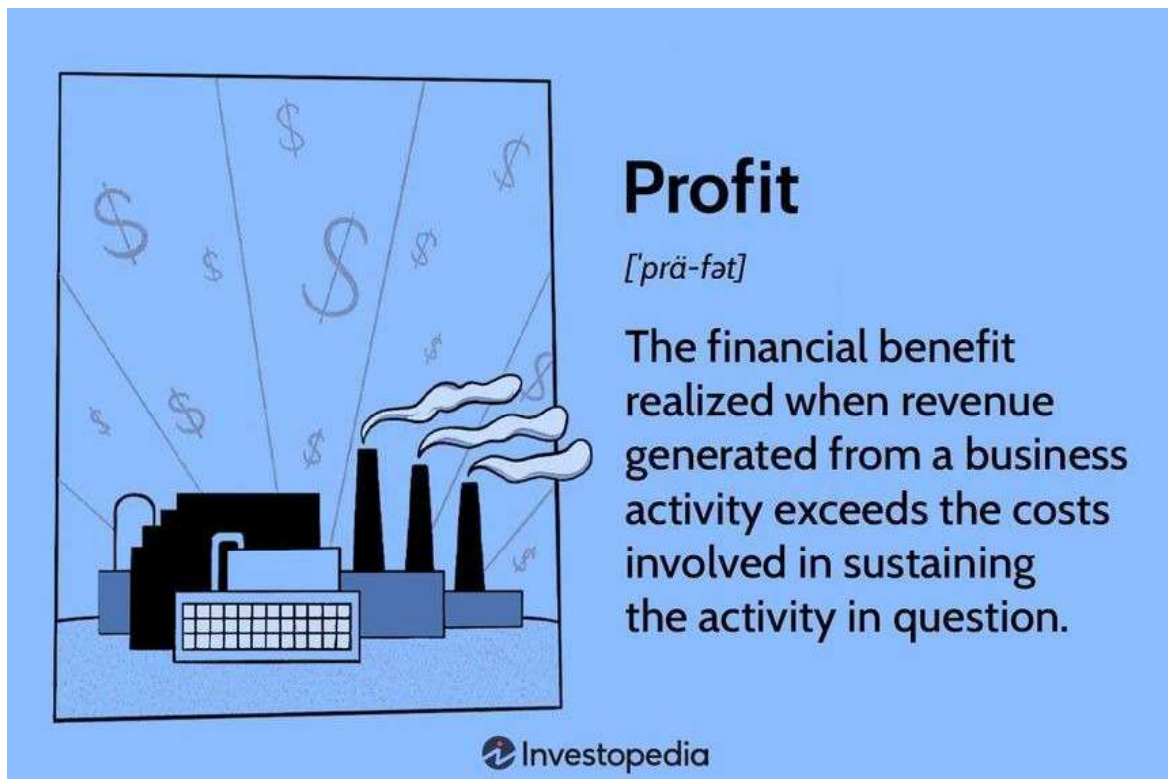
The culmination of this research aims to present a machine learning-based profit prediction model that surpasses the capabilities of conventional forecasting methods. By employing this model, businesses and investors can make more informed decisions, bolstering their competitive edge in today's dynamic markets. Moreover, the methodology and insights derived from this study serve as a foundational framework for further advancements in predictive analytics for financial performance.

## EXISTING SYSTEM

The existing system typically relies on traditional methods for forecasting purposes. As these are very old methods in some scenarios it cannot be applied and even if applied the accuracy can be an issue. Some of traditional methods are: Financial ratio analysis, trend analysis & time series forecasting techniques. Financial ratio analysis involves testing various financial ratios such as liquidity, solvency, and profitability ratios to assess a company's financial health. Trend analysis involves studying historical financial data to identify patterns and extrapolate future performance.

While these traditional methods have been a go to method for many years for companies to estimate their profit but they may fall short in capturing the complexity of today's rapidly growing/evolving business environment. They may have problem to account for non-linear relationships, sudden market shifts and impact of external factors such as regulatory changes in the market.

Furthermore these traditional methods may not fully utilize the vast data that is being generated in the recent times therefore we might not be able to predict accurate profits for company. This can lead to less optimal decision which can be taken by board members of the company and in turn can have ill effect on the product they might be launching.



## **PROPOSED METHOD**

Our approach for solving this prediction problem involves using advanced prediction algorithm based on machine learning which significantly differs from traditional methods. Instead of relying solely on historical data for prediction we also make use of other external data like market trends, customer opinion on brand value and other factors which might affect performance of the product in the market.

Steps in performing prediction are:

1. Feature Selection and Engineering:

We begin by carefully selecting input features. These consists a diverse set of financial metrics, operational indicators, market trends, and other relevant data points. Such as cost incurred for research and development of the product, Administration cost during the development of the product, amount spent on marketing of the product. This step ensures that the model is equipped with a rich set of information, enabling it to recognize intricate patterns and relationships.

2. Data Pre-processing:

Raw data can be in any format, therefore it requires preparation before it can be fed into the machine learning model. This involves tasks such as handling missing values, scaling numerical features, and encoding categorical variables. By doing so, we ensure that the data is in a suitable form for the model to learn from.

3. Model Selection and Training:

We explore a variety of machine learning algorithms, including algorithms like linear regression, knn mapping, ensemble methods like Random Forest regressor & as well as advanced deep learning models. These algorithms are trained on the pre-processed data to learn the underlying patterns that drive a company's profitability.

4. Cross-Validation:

To ensure the model's performance is robust and not overly specialized for only the training data, we employ cross-validation. This technique involves partitioning the data into multiple subsets, training the model on a portion, and validating it on another. This process is repeated to provide a comprehensive assessment of the model's predictive capability.

#### 5. Model Evaluation and Hyperparameter Tuning:

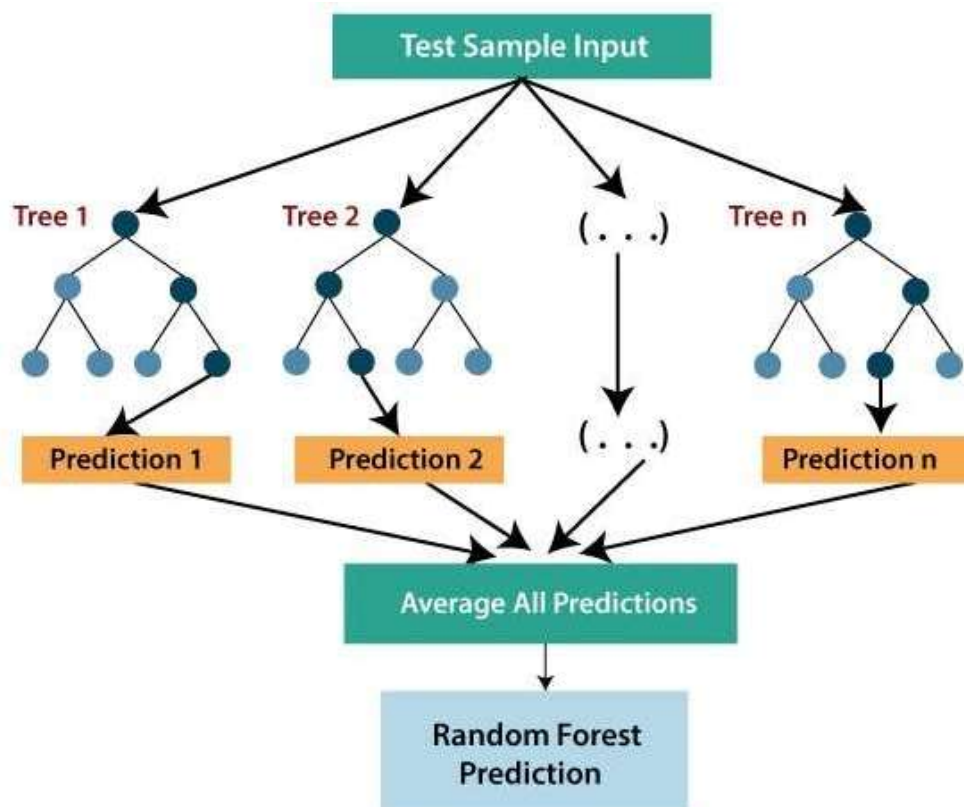
We rigorously assess the performance of each model based on metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. Additionally, we fine-tune the model's hyperparameters to optimize its performance.

#### 6. Feature Importance Analysis:

Understanding which features have the most influence on profit prediction is crucial for interpreting the model's decisions. We conduct a thorough analysis to identify the key drivers of profitability. This information is invaluable for decision-makers seeking actionable insights.

#### 7. Back-Testing and Validation:

To validate the model's effectiveness, we conduct extensive back-testing using historical data. This process involves applying the model to past periods and comparing its predictions to actual profit figures.



## METHODOLOGY

As this project required us to predict profit, profit for any company will always be continuous value and for prediction of continuous values we have to employ regression models as classification models can only handle discrete values and not continuous values.

There are many regression algorithm that can be used for this scenario, it depends on user which algorithm he/she wants to use and also accuracy of any given algorithm, for this project we used mainly 2 different algorithm and calculated their accuracy of the model produced by both the model and based on model having higher accuracy we decided to employ the same.

The algorithm we used were:

1. Linear Regression
2. Random forest Regressor.

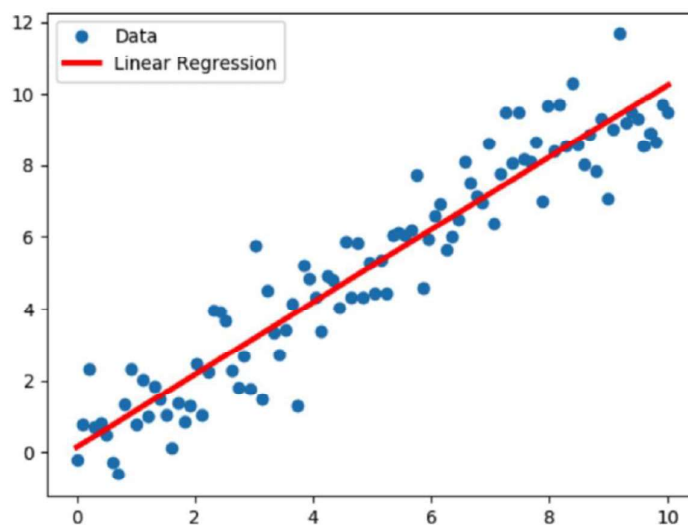
### 1. Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Linear regression is a baseline modelling technique to model a continuous target/dependent variable, which assumes a linear connection between a continuous dependent variable (Y) and an independent variable (X).

Linear regression fits a straight line that minimizes the discrepancies between predicted and actual output values.

The linear regression model can be simple (with only one dependent and one independent variable) or complex (with numerous dependent and independent variables) (with one dependent variable and more than one independent variable).





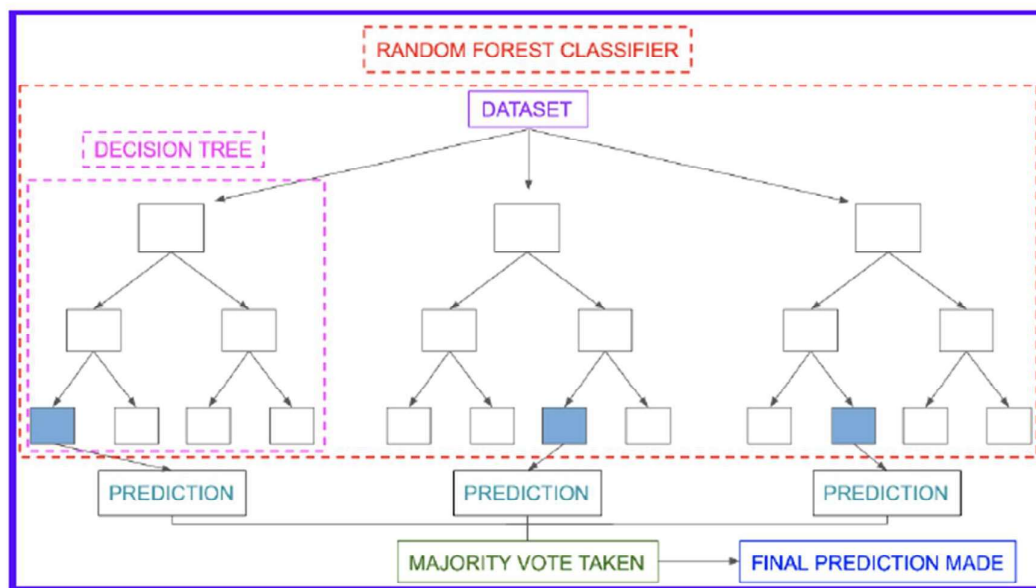
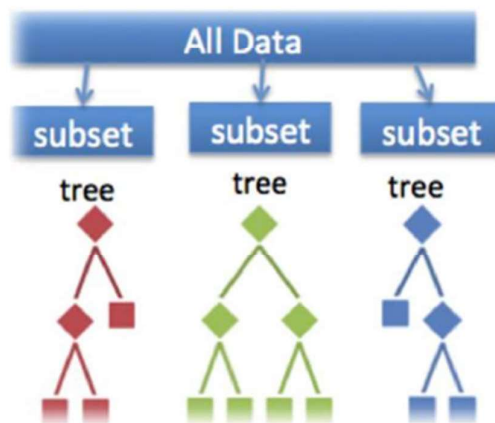
## 2. Random forest trees:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest splits out a class prediction and the class with most votes becomes our model's prediction.

Random forests are based on two key concepts namely bagging and feature selection. The key principle underlying the random forest approach comprises the construction of many "simple" decision trees in the training stage and the majority vote across them in the classification stage.

In the training stage, random forests apply the general technique known as bagging to individual trees in the ensemble. Random forest models decide where to split based on a random selection of features.

Rather than splitting at similar features at each node throughout, Random Forest models implement a level of differentiation because each tree will split based on different features. This level of differentiation provides a greater ensemble to aggregate over, producing a more accurate predictor. At the end, the final prediction is given based on the concept of voting.



## IMPLEMENTATION

In this section we discuss about how we implemented the algorithm and created different models based on the data and how we evaluated the models and which were the best models produced based on the data and which were selected for the final project as outcome/model for prediction.

To start with the data analytics, in market there are couple of programming language like: Python, R programming, it is solely user preference on choosing which language he/she wants to use. For this project we went ahead and used python as primary language for prediction.

System Requirements:

1. CPU- Intel i3 3<sup>rd</sup> gen and above, AMD Ryzen 3<sup>rd</sup> generation and above.
2. RAM- 8GB & above.
3. Storage- Minimum 100 GB to be available.
4. IDE- Jupyter notebook, Visual Studio Code.
5. Programming language- Python 3.8+

Getting Started:

Before starting the analytics process, we need to make sure that we have all the necessary libraries installed in our development computer.

The zip folder of this project consists of file “requirements.txt” which consists of all the packages required to be installed in order to run the program.

To install packages from the file, open command prompt and navigate to the directory where files of extracted zip file is present, execute this command `pip install -r requirements.txt` this command will install all the packages present in the text file automatically, there is no need of human intervention in this process.

Once the packages are installed now it's the time to open the jupyter notebook and start doing the analytical process.

To start jupyter notebook, open command prompt and type `jupyter notebook`, this will open browser with jupyter notebook, from here we can start doing the prediction process.

Start the project:

Import all the packages required for the program

```
In [2]: import pandas as pd
import numpy as np
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
data=pd.read_csv('datasets/50_Startups.csv')
print(data.head())
```

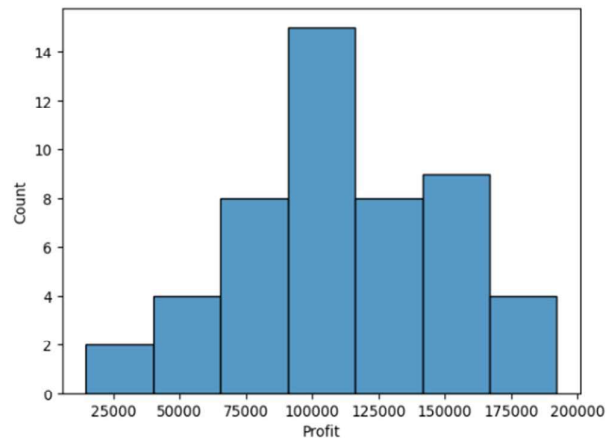
	R&D Spend	Administration	Marketing Spend	Profit
0	165349.20	136897.80	471784.10	192261.83
1	162597.70	151377.59	443898.53	191792.06
2	153441.51	101145.55	407934.54	191050.39
3	144372.41	118671.85	383199.62	182901.99
4	142107.34	91391.77	366168.42	166187.94

```
In [73]: data.isnull().sum()
Out[73]: R&D Spend      0
Administration    0
Marketing Spend    0
Profit             0
dtype: int64
```

Perform the analytics process by employing graphs and scatter plots and analyze the relation between input and output fields.

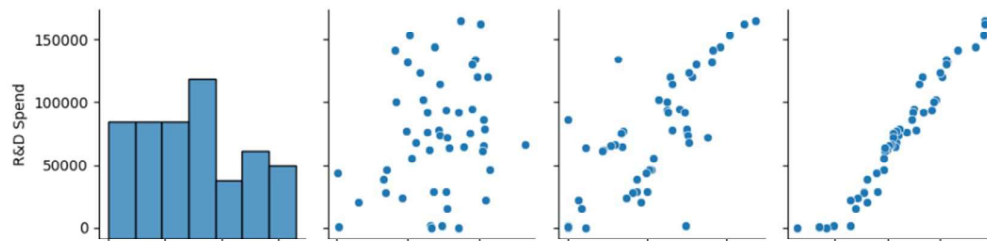
```
In [115]: # Find relation between profit through histogram
sns.histplot(data.Profit)
```

```
Out[115]: <Axes: xlabel='Profit', ylabel='Count'>
```



```
In [117]: sns.pairplot(data)
```

```
Out[117]: <seaborn.axisgrid.PairGrid at 0x23cb587ea40>
```



Split the data into input and output fields and initialise linear regression model and fit the trained data field into the model.

### Preparing test and train data

```
In [75]: from sklearn.model_selection import train_test_split
x1,x2,y1,y2=train_test_split(X,Y,test_size=0.2,random_state=42)
```

```
In [76]: from sklearn.linear_model import LinearRegression
model_linear=LinearRegression()
model_linear.fit(x1,y1)
```

```
Out[76]: LinearRegression()
```

Now that we have linear regression model ready, do the same for random forest regressor so that we can compare both regression model and evaluate which of the model is better in provide accurate results.

### Train the Random Forest Regressor

```
In [92]: from sklearn.ensemble import RandomForestRegressor
         forest=RandomForestRegressor()
         forest.fit(X_train,y_train)
```

```
Out[92]: RandomForestRegressor()
```

Now that we have both the models ready, find the accuracy of them to decide which one is better and save model which suits the user requirement.

### Calculate Accuracy of improved random forest

```
In [97]: best_forest.score(X_test,y_test)
```

```
Out[97]: 0.934648260560111
```

```
In [111]: linear_accuracy=model_linear.score(X_test,y_test)*100
          forest_accuracy=forest.score(X_test,y_test)*100
          best_forest_accuracy=best_forest.score(X_test,y_test)*100

          print("Linear Regression Accuracy is: {:.2f}%".format(linear_accuracy))
          print("Simple random forest Accuracy is: {:.2f}%".format(forest_accuracy))
          print("Optimized random forest Accuracy is: {:.2f}%".format(best_forest_accuracy))
```

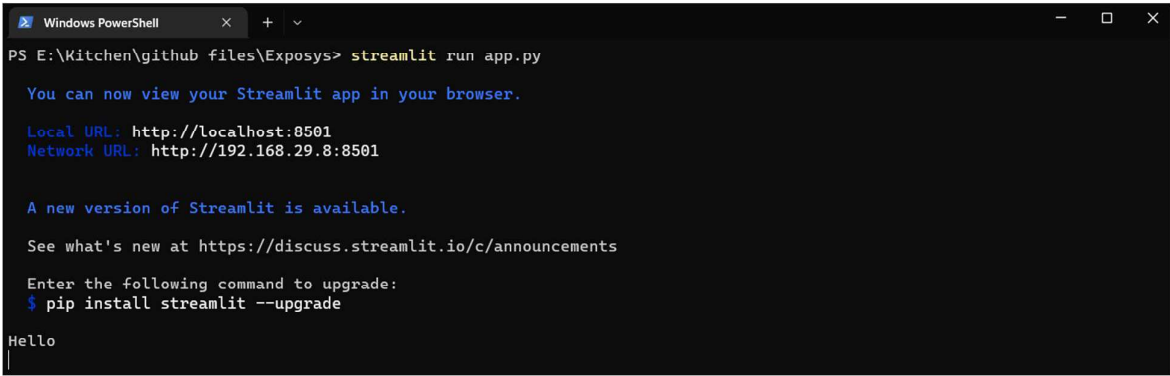
```
Linear Regression Accuracy is:96.82%
Simple random forest Accuracy is:92.75%
Optimized random forest Accuracy is:93.46%
```

As we can see linear regression gives out better accuracy than other 2 models, hence we can go ahead and save and use linear regression model for our prediction purpose.

On implementing this model in real life, we can use streamlit library created for python programming which helps in creating intuitive web interface for loading of the prediction model and accepts user inputs and gives out the prediction results.

To use the streamlit implementation of program follow these steps:

1. Open command prompt in project folder.
2. Make sure you have installed streamlit dependency.
3. `pip install streamlit`
4. Type the command `streamlit run app.py`.
5. New browser windows will open with the web implementation of the project.



```
Windows PowerShell
PS E:\Kitchen\github files\Exposys> streamlit run app.py

You can now view your Streamlit app in your browser.

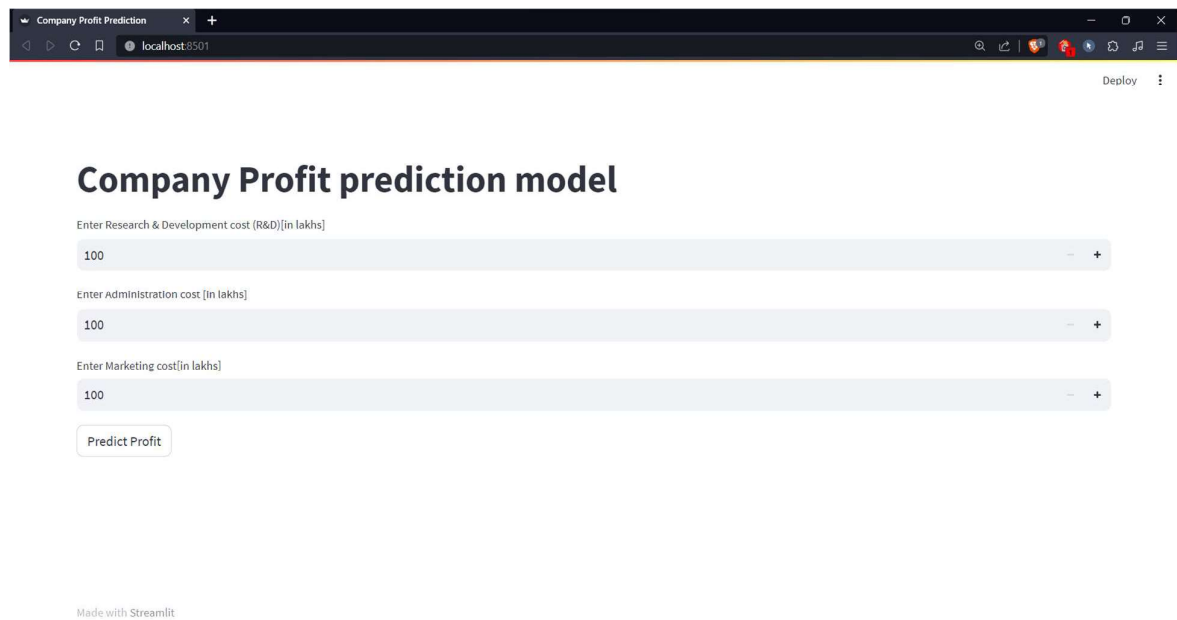
Local URL: http://localhost:8501
Network URL: http://192.168.29.8:8501

A new version of Streamlit is available.

See what's new at https://discuss.streamlit.io/c/announcements

Enter the following command to upgrade:
$ pip install streamlit --upgrade

Hello
```



(Web implementation of the project)

We can write the python script by referring to documentation of streamlit and we can host the same on any of cloud services like AWS, Google Cloud Platform, Microsoft Azure, etc...

## CONCLUSION

This project displays significant advancement in the domain of company profit prediction by employing advanced machine learning techniques. Through a systematic and thorough methodology, we have successfully developed a robust model capable of accurately forecasting company profits.

The incorporation of a diverse set of features, including financial metrics, and market trends, allowed for a comprehensive representation of a company's operational environment. This enhanced dataset provided the model with valuable insights, enabling it to discern intricate patterns and relationships that traditional methods may have overlooked.

The evaluation of various machine learning algorithms, including ensemble methods and deep learning models, demonstrated the advancement of our approach over traditional forecasting methods. The model's performance was rigorously assessed through cross-validation, ensuring its robustness and generalization capabilities.

Additionally, the feature importance analysis helped in understanding the key drivers influencing company profitability. This critical insight not only enhances our understanding of profit margins but also provides actionable information for strategic decision-making.

In conclusion, this project presents a powerful tool for businesses and investors seeking to enhance their decision-making processes. The machine learning-based profit prediction model offers a substantial improvement over traditional methods, particularly in navigating today's dynamic and evolving business landscape. The methodology and insights derived from this study provide a solid foundation for further advancements in predictive analytics for financial performance analysis sector. This project not only contributes to the academic field but also holds practical implications for industry professionals, enabling them to make more informed and strategic financial decisions.