DATA SCIENCE PROJECT

Name : RAHUL MALLAAH

Name of Institution : NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

Scholar ID : 1912151

Submitted to : Exposys Data Labs

Mode of Internship : Online

Topic : Profit prediction of a company by using machine learning algorithms

INTERNSHIP CERTIFICATE

Exposys
Data Labs

Certificate of Internship

TO WHOM IT MAY CONCERN

This is to certify that **Mr. RAHUL MALLAH** has completed internship programme on **"Data Science"** from 06.06.2022 to 15.07.2022.

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.

Y Vishnuvardhan
Chief Director

EXPOSYS DATA LABS
BENGALURU

hr@exposysdata.com
www.exposysdata.com

Contents Table

Problem Statement

A company should always set a goal that should be achievable, otherwise, employees will not be able to work to their best potential if they find that the goal set by the company is unachievable. The task of profit prediction for a particular period is the same as setting goals. If you know how much profit you can make with the amount of R&D and marketing you do, then a business can make more than the predicted profit provided the predicted value is achievable. So in this project, I will take you through the task of profit prediction with machine learning using the Python programming language.

Methodology

Linear Regression :

In many cases, the contribution of a single independent variable does not alone suffice to explain the dependent variable Y. If this is so, one can perform a multivariable linear regression to study the effect of multiple variables on the dependent variable. In the multivariable regression model, the dependent variable is described as a linear function of the independent variables Xi, as follows:

$$\mathbf{Y = a + b_1*X_1 + b_2*X_2 + b_3*X_3 + ... + b_n*X_n}$$

The model permits the computation of a regression coefficient bi for each independent variable Xi **Regression line** for a multivariable regression.

where **Y** = dependent variable, Xi=dependent variables a = constant (y-intercept) bi= regression coefficient of the variable. Just as in univariable regression, the coefficient of determination describes the overall relationship between the independent variables Xi (weight, age body-mass index) and the dependent variable Y (blood pressure). It corresponds to the square of the multiple correlation coefficient, which is the correlation between **Y** and $\mathbf{b_1*X_1 + b_2*X_2 + b_3*X_3 + ... + b_n*X_n}$. It is better practice, however, to give the corrected coefficient of determination. Each of the coefficients bi reflects the effect of the corresponding individual independent variable Xi on Y, where the potential influences of the remaining independent variables on Xi have been taken into account, i.e., eliminated by an additional computation. Thus, in a multiple regression analysis with age and sex as independent variables and weight as the dependent variable.

About the Dataset

Dataset link :

https://drive.google.com/uc?export=download&id=1Z7RKmScBO7n9vcDIG3Xeo853Ics4QFaF

This particular dataset holds data from 50 startups in New York, California, and Florida. The features in this dataset are R&D spending, Administration Spending and Marketing Spending, while the target variable is: Profit.

1. **R&D spending:** The amount which startups are spending on Research and development.
2. **Administration spending:** The amount which startups are spending on the Admin panel.
3. **Marketing spending:** The amount which startups are spending on marketing strategies.
4. **Profit:** How much profit that particular startup is making.

Implementation

Libraries Used :

We have implemented the classical machine learning algorithms in Python for prediction. Following is the list of libraries that have assisted us in the code-

```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from xgboost import XGBRegressor
from catboost import CatBoostRegressor
from sklearn.metrics import r2_score,mean_squared_error
import warnings
warnings.filterwarnings('ignore')
```

Pandas : Pandas is used to read a csv and strongly used in data wrangling.
#**Numpy :** Numpy is a mathematical computational library
Scikit Learn : It provides selection of efficient tools for Machine Learning/ Deep learning/ NLP
Matplotlib/ Seaborn : Both are plotting libraries of Python.
Warnings : It is used to reduce warnings that appear while coding.

Correlation between data :

We can clearly see that the profit of a company is more correlated with R&D and Marketing spend as compared to administration spend.

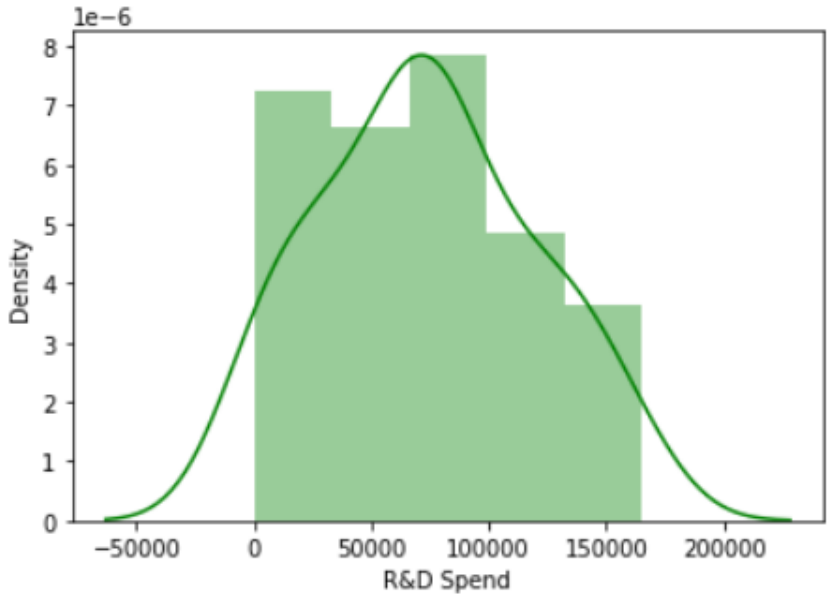
```
sns.heatmap(data.corr(), annot = True, cmap = 'viridis')
plt.show()
```

Univariate Analysis :

It’s clear that companies are spending more or less in R&D

[40]:

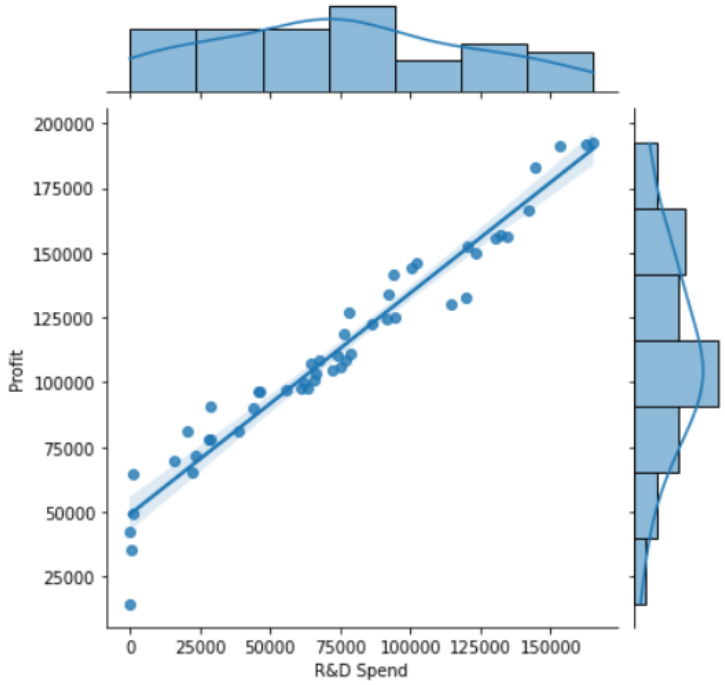
```
sns.distplot(data['R&D Spend'], color = 'green')
plt.show()
```



Bivariate Analysis :

In order to increase the profit of a company we have to spend more on R&D because there is a linear relationship between profit and R&D spend.

[44]:
```
sns.jointplot(x = 'R&D Spend', y = 'Profit', kind = 'reg', data = data)
plt.show()
```



Scaling the data :

Since, all the data are not in the same range. So, we have to scale it to the same range. Here, We are using Min Max Scaler from the Sci-kit learn library.

[19]:
```
scaler = MinMaxScaler(feature_range = (0,1))
X = scaler.fit_transform(X)
X = pd.DataFrame(X,columns = ['R&D Spend', 'Administration', 'Marketing Spend'])
X.head()
```

[19]:

	R&D Spend	Administration	Marketing Spend
0	1.000000	0.651744	1.000000
1	0.983359	0.761972	0.940893
2	0.927985	0.379579	0.864664
3	0.873136	0.512998	0.812235
4	0.859438	0.305328	0.776136

+ Code + Markdown

[20]:

Modeling part :

Here, We have used Linear Regression, Random Forest, Gradient Boosting, Extreme Gradient Boosting and Cat Boost as models from **sci-kit learn** tools.

[52]:
```
lr = LinearRegression(fit_intercept = True)
rfg = RandomForestRegressor(n_estimators = 150, criterion = 'squared_error', random_state = 28)
gbr = GradientBoostingRegressor(loss = 'squared_error',learning_rate = 0.01, n_estimators = 88, criterion = 'squared_error')
xgb = XGBRegressor(n_estimators = 130,learning_rate = 0.05,random_state = 20)
cat = CatBoostRegressor(iterations = 502, learning_rate = 0.01, loss_function = 'RMSE')
```

[53]:
```
reg={

    'Linear Regression':lr,
    'Random Forest':rfg,
    'Gradient Boosting':gbr,
    'Xg Boost':xgb,
    'Cat Boost':cat

    }
```

[24]:
```
def regressor(x,y):

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 101)

    for key in reg.keys():
        reg[key].fit(X_train, y_train)
        y_pred=reg[key].predict(X_test)
        score=r2_score(y_test, y_pred)

        print(key,'------>>>', score)
```

+ Code + Markdown

[25]:
```
%%time
regressor(X, y)
```

```
Linear Regression ------>>> 0.9604774705502481
Random Forest ------>>> 0.9529069852016784
Gradient Boosting ------>>> 0.736896588495445
Xg Boost ------>>> 0.8877151889524597
Cat Boost ------>>> 0.8178753884688377
CPU times: user 1.26 s, sys: 265 ms, total: 1.52 s
Wall time: 2.77 s
```

Results/ Output

Here, We can see the R2-score of all the models. For this dataset Linear Regression and Random Forest Regression are the most efficient one's both have R2-score of 96.04% and 95.29% respectively.

Serial No.	**Model Name**	**R-Squared**
1.	Linear Regression	**96.04%**
2.	Random Forest Regression	**95.29%**
3.	Gradient Boosting Regression	**73.38%**
4.	Extreme Gradient Boosting Regression	**88.71%**
5.	Cat Boost Regression	**81.78%**

CONCLUSION

In this project, we study the system for start-up prediction for business to know the idea of profit to businessmen so that any user wants to know about business profit he/she will use this application for future use. The users have thought about investing and according to that he/she would invest money for business. Our goal is to assist start-up people who are looking for some investment and profit ideas so that they can make a suitable strategy for the future and know how much money their firm will make. To forecast start-up profit, we employ classical machine learning algorithms. After a period of qualitative study includes getting to know the founders and business, start-up investors typically do their own financial examination of possible target companies.

Reference Websites

1. scikit-learn: machine learning in Python — scikit-learn 1.1.3 documentation
2. https://matplotlib.org/
3. https://seaborn.pydata.org/
4. https://ijarcce.com/wp-content/uploads/2022/05/IJARCCE.2022.11561.pdf
5. https://www.analyticsvidhya.com/blog/2021/11/startups-profit-prediction-using-multiple-linear-regression/