DATA SCIENCE PROJECT

Name : RAHUL MALLAAH

Name of Institution : NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR

Submitted to : Exposys data labs

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

ABSTRACT

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 Python.

METHODOLOGY

Multivariable 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 X_i, as follows: $Y = a + b1 \times X1 + b2 \times X_2 + \ldots + b_n \times X_n$. The model permits the computation of a regression coefficient b_i for each independent variable X_i

Regression line for a multivariable regression

$Y = a + b_1 \times X_1 + b_2 \times X_2 + \ldots + b_n \times X_n$, where Y = dependent variable, X_i=dependent variables a = constant (y-intersect) b_i= regression coefficient of the variable.

Just as in univariable regression, the coefficient of determination describes the overall relationship between the independent variables X_i (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 $b_1 \times X_1 + \ldots + b_n \times X_n$.

It is better practice, however, to give the corrected coefficient of determination. Each of the coefficients b_i reflects the effect of the corresponding individual independent variable X_i on Y, where the potential influences of the remaining independent variables on X_i 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.

Two important terms

- **Confounder** (in non-randomized studies): an independent variable that is associated, not only with the dependent variable, but also with other independent variables. The presence of confounders can distort the effect of the other independent variables. Age and sex are frequent confounders.

- **Adjustment:** a statistical technique to eliminate the influence of one or more confounders on the treatment effect. Example: Suppose that age is a confounding variable in a study of the effect of treatment on a certain dependent variable. Adjustment for age involves a computational procedure to mimic a situation in which the men and women in the data set were of the same age. This computation eliminates the influence of age on the treatment effect.

IMPLEMENTATION

Importing the required libraries

```
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')
```

```
#https://drive.google.com/file/d/1Z7RKmScBO7n9vcDIG3Xeo853Ics4QFaF/view
dataset_path="https://drive.google.com/uc?export=download&id=1Z7RKmScBO7n9vcDIG3Xeo853Ics4QFaF"
data=pd.read_csv(dataset_path)
```

```
#Displaying first 5 records of the data
data.head()
```

```
data.shape
```

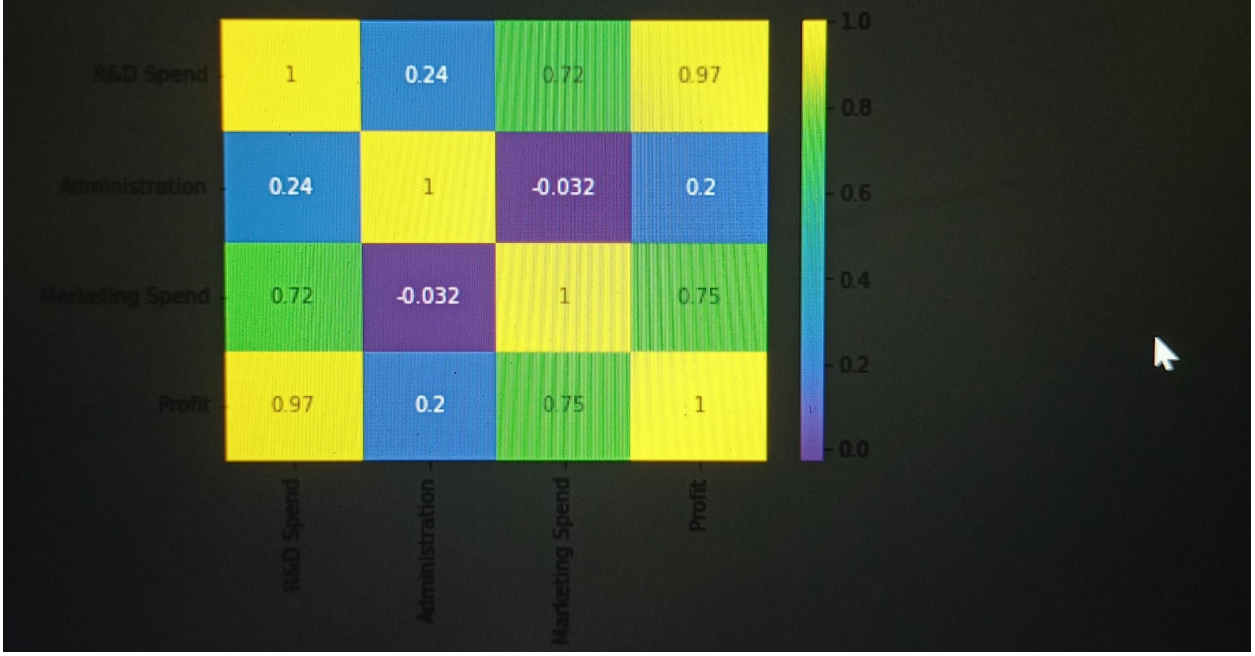
```
(50, 4)
```

```
data.isnull().sum()
```

```
R&D Spend          0
Administration     0
Marketing Spend    0
```

4.Data visualizations

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



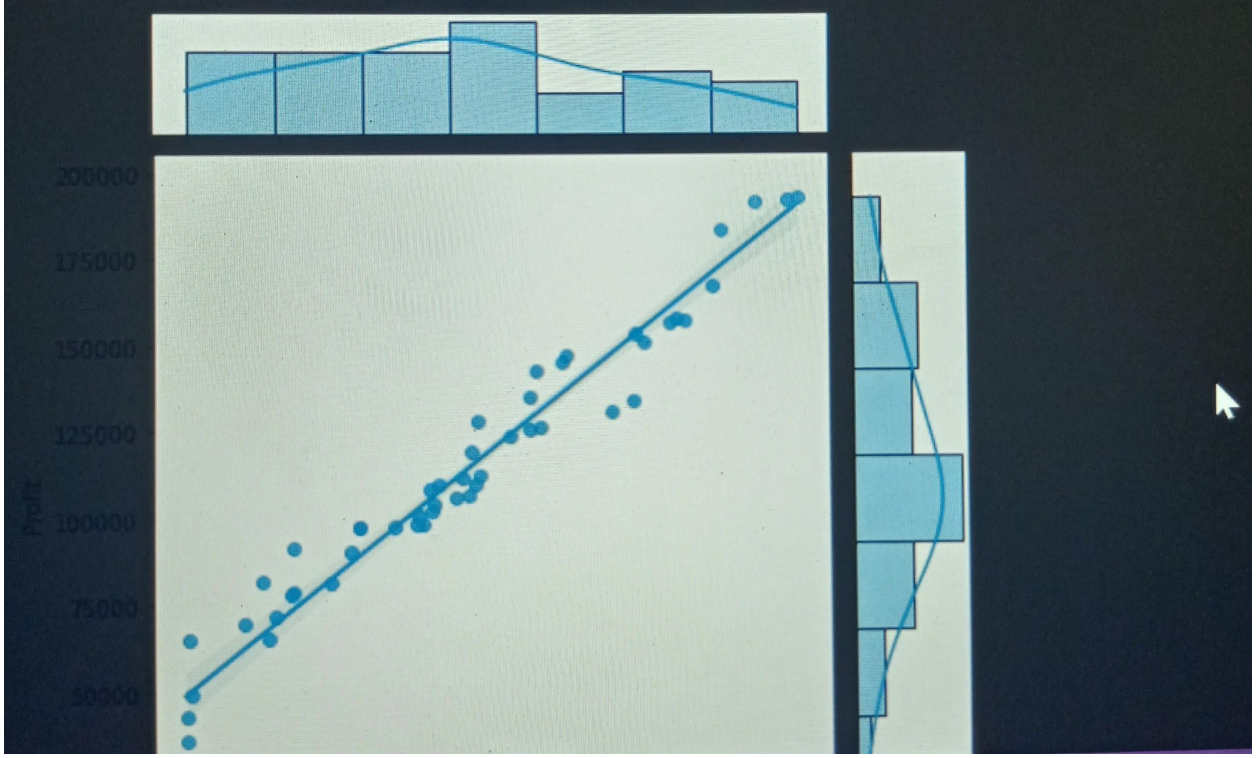
W

5.Univariate analysis

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

6.Bivariate analysis

```
1  sns.jointplot(x='R&D Spend',y='Profit',kind='reg',data=data)
2  plt.show()
```



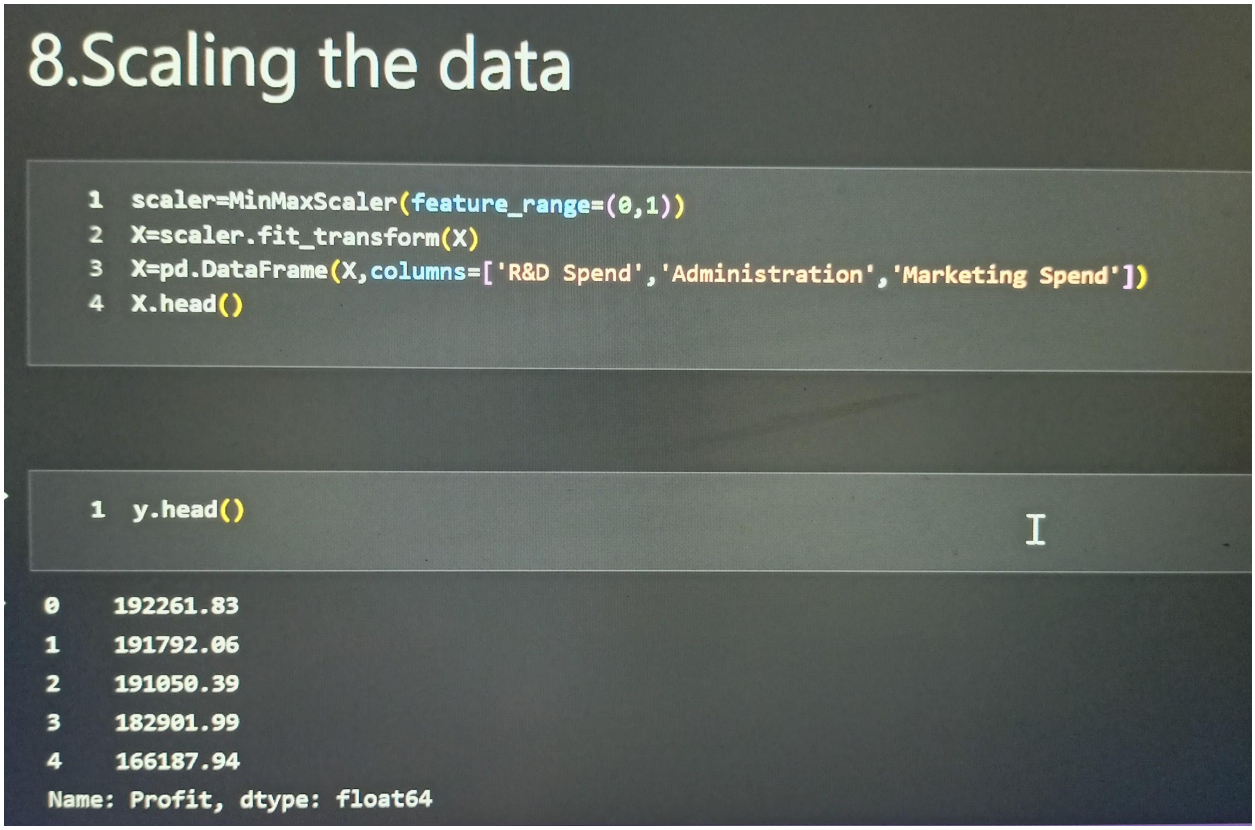
7.Separating the data into X and y

```
1  X=data.drop(columns='Profit',axis=1)
2  y=data['Profit']
```

```
1  X.head()
```

```
1  y.head()
```

```
0    192261.83
1    191792.06
2    191050.39
3    182901.99
4    166187.94
Name: Profit, dtype: float64
```



Using regressor models for profit prediction by using independent variables.

9.Applying the machine learning models

```
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')
```

```
reg={

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

    }
```

```
reg.keys()
```

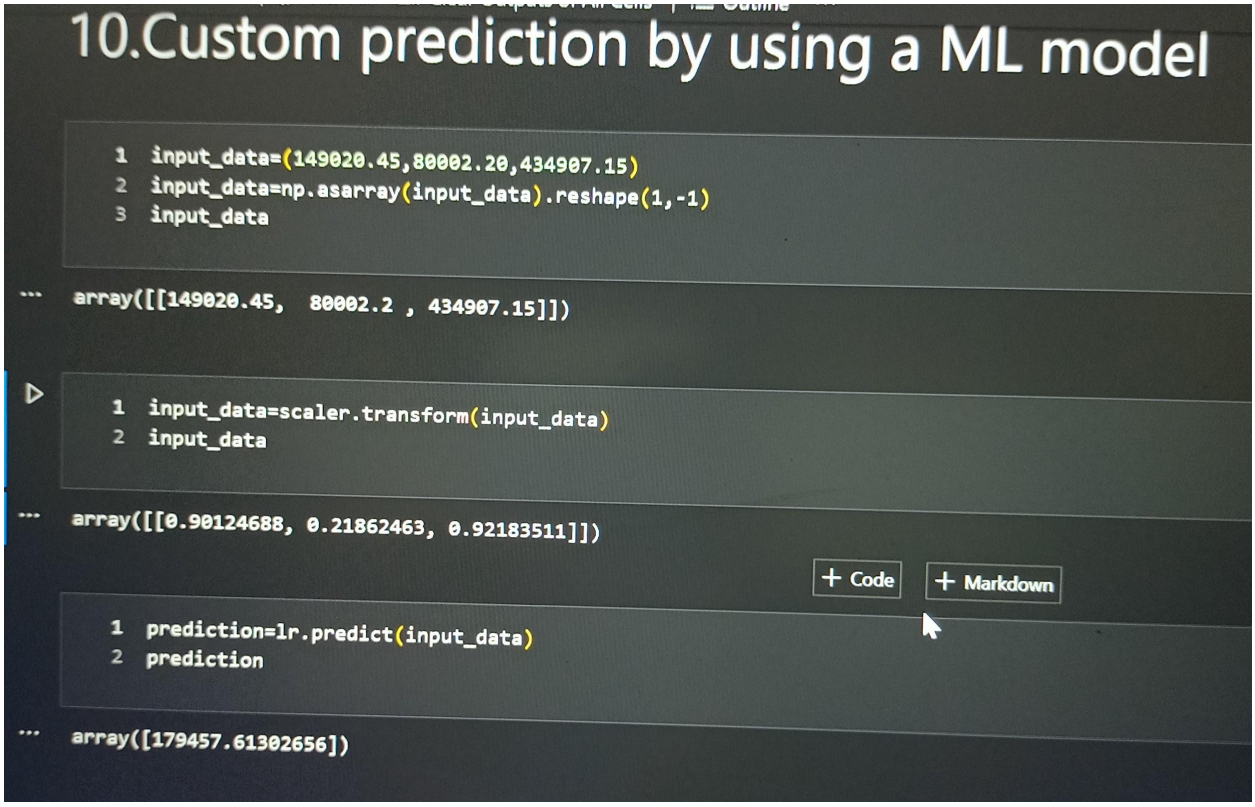
```
dict_keys(['Linear Regression', 'Random Forest', 'Gradient Boosting', 'Xg Boost', 'Cat Boost'])
```

```
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)
```



Results

We had predicted the expected profits of the companies and got R2_score of 96%