A Project report on

PROFIT PREDICTION

By

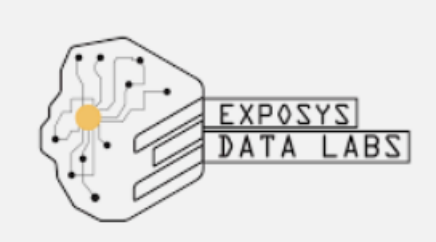
SHAMEENA M K

Rajiv Gandhi Institute of Technology, Kottayam

Submitted to

Exposys Data Labs

Bangaluru, Karnataka, 560064



ABSTRACT

When there is no computerized system there is always a difficulty in predicting profit as there are many considerations to be done. Machine learning models used for predicting profit via R&D costs, administration cost and marketing spend in a business. Models focusses on three machine learning algorithms, LinearRegression, RandomForestRegressor, KNeighborsRegressor.

TABLE OF CONTENTS

1.INTRODUCTION

1.1 Data Science

Data science is the domain of study that deals with vast volumns of data using modern tools and techniques to find unseen patterns, drive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.The data used for analysis can come from many different sources and presented in various formats.

The Data Science Lifecycle

Data Science's lifecycle consists of five distinct stages, each with its own tasks:

1**.Capture**: Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.

2.**Maintain**: Data Warehousing, Data Cleaning, Data Processing, Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.

3.**Process**: Data Mining, Clustering/Classification, Data Modeling, Data Summerization, Data scientists take the prepared data and examine its patterns,ranges and biases to determine how useful it will be in predictive analysis.

4**.Analyze**: Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. This stage involves performing the various analysis on the data.

5.**Communicate**: Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the nalyses in easily readable forms such as charts, graphs, and reports.

1.2 Machine Learning

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition**,** speech recognition**,** email filtering**,** Facebook auto-tagging**,** recommender system, and Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions many more.

This machine learning tutorial gives you an introduction to machine learning along with the wide range of machine learning techniques such as Supervised**,** Unsupervised, and Reinforcement learning. You will learn about regression and classification models, clustering methods, hidden Markov models, and various sequential models.

2. EXISTING SYSTEM

By using a single independent variable such as the investment cost of a company's project, the value of the dependent variable i.e., the profit of the company by the means of that project is approximately predicted. Linear regression makes use of a single independent variable to predict the value of a dependent variable by developing a regression line along the given data and thereby predicting dependent variable using that regression line. There are some other techniques viz., the Classification tree and Random Forest that makes use of a lot of dependent variable to predict the value of the dependent variable and these techniques works best for some of the given values but not for all.

2.1 Disadvantages of existing system

- Linear regression makes use of only one independent variable and so results are less accurate
- Data are not completely consumed by a linear regression model.

3.PROPOSED SYSTEM

The main intention is to predict the value of the dependent variable i.e., the value of the profit of the company based on the data of the company over the previous years. So, from all the techniques used before for the prediction of profit an average from all those predicted values of the dependent variable is computed and made as the predicted dependent variable.

3.1 Advantages of Proposed System

- It makes use of all the data given to it to predict the value of independent variable.
- Theoretically it is better than all the other existing algorithms

4.REGRESSORS

4.1 Linear Regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

4.2 Random Forest Regression

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning**,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

4.3 KNeighbours Regressor

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics we might have learned in our childhood— calculating the distance between points on a graph. There are other ways of calculating distance, and one way might be preferable depending on the problem we are solving. However, the straight-line distance (also called the Euclidean distance) is a popular and familiar choice.

5. MODEL

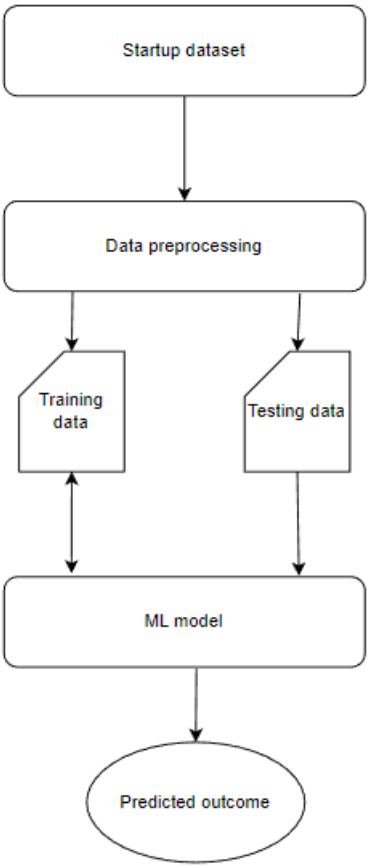


Fig: working flow of a model

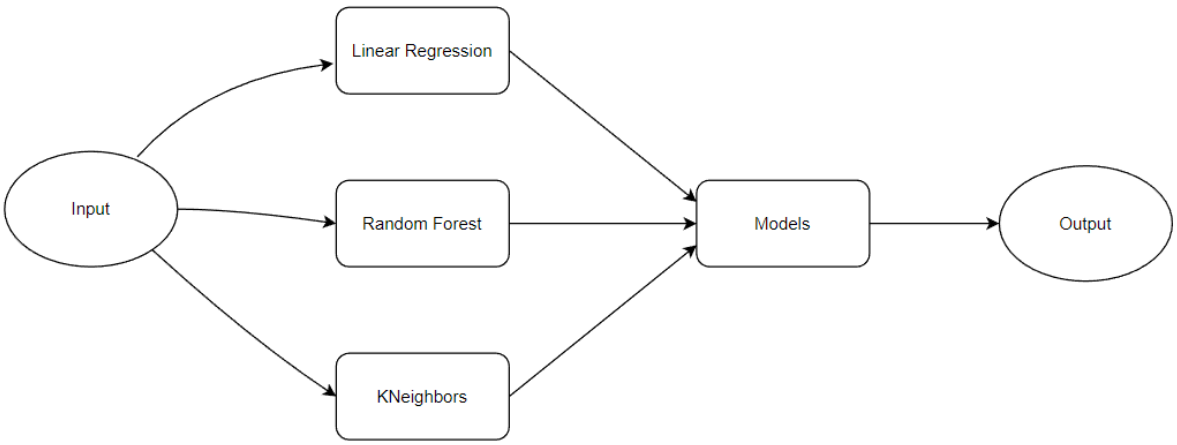


Fig:working of models

6.IMPLEMENTATION

6.1 Source code

Import the libraries

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
```

Load the dataset

```
data=pd.read_csv("50_Startups.csv")
data.head()
print(data.shape)
print(data.size)
data.describe()
data.isnull().sum()
data.info()
```

Visualising the data

```
data.plot(kind='bar', stacked='true',figsize=(10,6))
plt.figure(figsize=(8,6))
sns.heatmap(data.corr(),annot=True,cmap='Blues');
sns.pairplot(data)
plt.show()
data.plot(kind='box',figsize=(10,6))
# Correlation Matrix
```

```
companies_correlation = data.corr()
companies_correlation['Profit'].sort_values(ascending=False)
```

Training and testing the data

```
X = data.drop(["Profit"],axis=1)
y = data['Profit']
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.3,rand
om_state=21)
```

Using Linear Regression

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import accuracy_score,r2_score
model_LinearRegression = LinearRegression()
model_LinearRegression.fit(X_train, y_train)
y_pred = model_LinearRegression.predict(X_test)
r2 = r2_score(y_test, y_pred).round(4)
mse = mean_squared_error(y_test, y_pred).round(4)
rmse = np.sqrt(mean_squared_error(y_test, y_pred)).round(4)
mae = mean_absolute_error(y_test,y_pred).round(4)
print('R2 Score : ', r2)
print('MSE      : ', mse)
print('RMSE     : ', rmse)
print('MAE      : ', mae)
```

using Random Forest

```
from sklearn.ensemble import RandomForestRegressor

model_RandomForestRegressor = RandomForestRegressor()
model_RandomForestRegressor.fit(X_train, y_train)
```

```
pred_R= model_RandomForestRegressor.predict(X_test)
r2_R = r2_score(y_test, pred_R).round(4)
mse_R= mean_squared_error(y_test, pred_R).round(4)
rmse_R= np.sqrt(mean_squared_error(y_test, pred_R)).round(4)
mae_R= mean_absolute_error(y_test, pred_R).round(4)

print('R2 Score  : ', r2_R)
print('MSE      : ', mse_R)
print('RMSE     : ', rmse_R)
print('MAE      : ', mae_R)
```

using Kneighbors

```
from sklearn.neighbors import KNeighborsRegressor

model_KNeighborsRegressor = KNeighborsRegressor()
model_KNeighborsRegressor.fit(X_train, y_train)
pred_K = model_KNeighborsRegressor.predict(X_test)
r2_K = r2_score(y_test, pred_K).round(4)
mse_K = mean_squared_error(y_test, pred_K).round(4)
rmse_K= np.sqrt(mean_squared_error(y_test, pred_K)).round(4)
mae_K = mean_absolute_error(y_test, pred_K).round(4)

print('R2 Score  : ', r2_K)
print('MSE      : ', mse_K)
print('RMSE     : ', rmse_K)
print('MAE      : ', mae_K)
```

Compare the models

```
models = pd.DataFrame({


   'Model': [
      'LinearRegression',
'RandomForestRegressor','KNeighborsRegressor'
   ],
    'R2 Score': [
      r2,r2_R,r2_K
   ],
    'MSE': [
      mse,mse_R,mse_K
   ],
     'RMSE': [
      rmse,rmse_R,rmse_K
   ],
    'MAE': [
      mae, mae_R,mae_K
   ]
})
Models
p = plt.figure(figsize=(18,16))
p = sns.set_theme(style="white")
p= models=models.sort_values(by='R2 Score',ascending=False)[:20]
p = sns.barplot(y= 'Model', x= 'R2 Score', data= models)
for container in p.containers:
   p.bar_label(container,label_type = 'center',padding = 2,size =
15,color = "Red",rotation = 0,
```

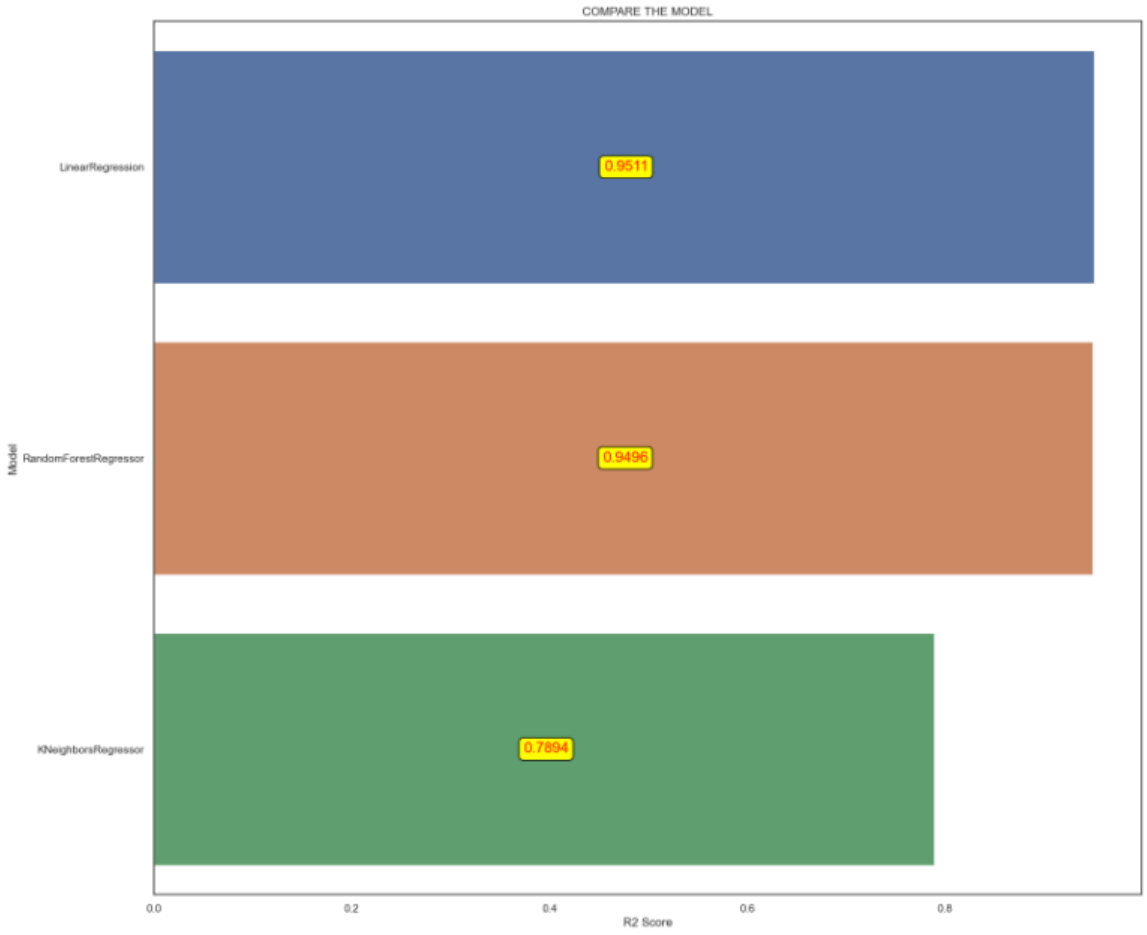
```
    bbox={"boxstyle": "round", "pad": 0.3, "facecolor": "yellow",
"edgecolor": "black", "alpha": 1})

plt.title('COMPARE THE MODEL')

plt.xlabel('R2 Score')

plt.ylabel('Model');
```

7.COMPARE THE MODELS



8.CONCLUSION

In the given dataset, R&D Spend, Administration Cost and Marketing Spend of 50 Companies are given along with the profit earned. The target is to prepare an ML model which can predict the profit value of a company if the value of its R&D Spend, Administration Cost and Marketing Spend are given. The models used are Linear Regression , Random Forest Regressor and KNeighbors Regressor.By comparing these models, I concluded that the best model is Linear Regression and the worst model is KNeighbors Regressor.

REFERENCE

[1] https://www.linkedin.com/learning/using-python-with-excel

[2] https://www.linkedin.com/learning/machine-learning-with-python-foundations

[3] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[4] www.kaggle.com

[5] https://www.javatpoint.com/machine-learning-random-forest-algorithm