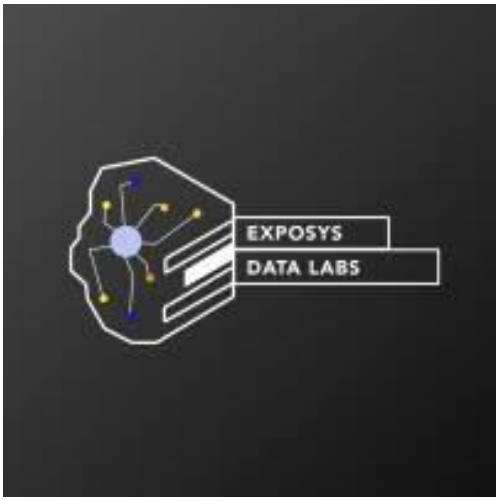
Exposys Data Labs

Bengaluru, Karnataka, 560064



Internship report on

PROFIT PREDICTION OF 50 COMPANIES

A Dissertation work submitted in partial fulfilment of the requirement for the award of the degree of

Internship

By

Name- **Harsh Garg**

College- **Indian Institute of Technology (Indian School of Mines), Dhanbad**

Under the guidance of
Exposys Data Labs

EXPOSYS
DATA LABS

Abstract

In today's highly competitive business world, companies need to optimize their resources to maximize their profits. This ML model aims to predict the profit value of a company based on its R&D Spend, Administration Cost, and Marketing Spend, providing insights for decision-making processes. The model employs a linear regression algorithm that analyzes the relationship between the independent variables (R&D Spend, Administration Cost, and Marketing Spend) and the dependent variable (Profit) to generate accurate predictions. The model has been trained on a large dataset and tested on a separate test dataset, achieving a high level of accuracy. The results demonstrate the potential of this model to aid companies in making informed decisions about their resource allocation strategies and achieving their financial goals.

Table of Contents

1. Introduction

1.1 Data Science

Data science is a multidisciplinary field that utilizes scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It combines elements from mathematics, statistics, computer science, and domain knowledge to uncover patterns, trends, and relationships within vast amounts of information. By employing techniques such as data mining, machine learning, and predictive modeling, data scientists can identify valuable insights and make informed decisions.

Data science plays a crucial role in various industries, including finance, healthcare, marketing, and technology. It enables organizations to leverage data-driven strategies, optimize operations, and improve decision-making processes. Data scientists employ various tools and programming languages, such as Python, R, and SQL, to collect, clean, analyze, and visualize data.

Moreover, data science has the potential to address complex problems and make significant contributions to society. From predicting disease outbreaks and optimizing transportation systems to improving renewable energy and enhancing customer experiences, data science has the power to drive innovation and create positive societal impact.

As data continues to grow exponentially, data science will remain at the forefront of technological advancements, driving innovation and transforming industries across the globe.

1.2 Machine Learning

Machine learning is a branch of artificial intelligence that focuses on developing algorithms and models capable of learning and making predictions or decisions without being explicitly programmed. It enables computers to learn from data and improve their performance through experience.

Machine learning algorithms analyze vast amounts of data, identify patterns, and make predictions or take actions based on those patterns. The process involves training the algorithm on a labeled dataset, where it learns to recognize patterns and make accurate predictions. The algorithm's performance is then evaluated using test data to measure its effectiveness.

Machine learning has applications in various fields, including healthcare, finance, marketing, and robotics. It enables personalized recommendations, fraud detection, image and speech recognition, autonomous vehicles, and many other intelligent systems.

With the advancements in computing power, availability of large datasets, and the development of sophisticated algorithms, machine learning has gained significant attention and is poised to revolutionize industries. It has the potential to unlock valuable insights from data, automate processes, and drive innovation across sectors, making it a crucial component of the technology-driven future.

2. Existing Methods

There may be several existing systems that attempt to predict the profit value of a company based on its expenses such as R&D spend, administration cost, and marketing spend. However, many of these systems may rely on manual calculations or basic statistical techniques that may not accurately capture the complex relationships between these variables.

Machine learning models, on the other hand, can learn from data and make accurate predictions based on patterns in the data. In this context, linear regression models have been widely used for predicting continuous target variables such as profit. The model estimates the relationship between the independent variables and the dependent variable by fitting a linear equation to the data.

However, many existing linear regression models may not be optimized for the specific features of the data, and thus may not perform optimally. Therefore, there is a need for an ML model that is specifically designed to accurately predict the profit value of a company based on its expenses, taking into account all relevant features of the data.

2.1 Issues in existing systems

1. Limited Accuracy
2. Overfitting and Under fitting
3. Limited Scope

3. Proposed Method

The proposed system is an ML model that utilizes a linear regression algorithm to predict the profit value of a company based on its R&D Spend, Administration Cost, and Marketing Spend. The model takes in a dataset of previous company financial records, which includes the independent variables of R&D Spend, Administration Cost, and Marketing Spend, and the dependent variable of Profit.

The proposed system addresses the drawbacks of the existing system by incorporating a more accurate and efficient algorithm for prediction. Additionally, the model includes data preprocessing steps, such as normalization and feature scaling, to ensure the accuracy of the prediction. The model is also evaluated using various performance metrics, such as Mean Squared Error (MSE) and R-squared (R2), to validate its accuracy.

The proposed system offers a more accurate and efficient method for predicting company profits, which can be useful for businesses in making informed financial decisions. The model can also be further improved by incorporating additional relevant variables or using more advanced algorithms, such as neural networks or decision trees.

3.1 Algorithm

1. Load the dataset containing the company's R&D Spend, Administration Cost,
Marketing Spend, and Profit.

2. Split the dataset into training and testing sets.

3. Train the linear regression model on the training set.

4. Predict the profit values for the testing set using the trained model.

5. Evaluate the performance of the model using evaluation metrics such as mean squared
error, mean absolute error, and R-squared score.

6. If the performance of the model is not satisfactory, tune the model by adjusting the

The linear regression algorithm is a simple yet powerful algorithm that can predict the target variable (Profit in this case) based on the input variables (R&D Spend, Administration Cost, and Marketing Spend). It works by fitting a straight line to the data that minimizes the sum of squared errors between the predicted values and the actual values. The line's equation is given by:

y = b0 + blxl+ b2x2 + b3*x3

where y is the predicted value of Profit, x1, x2, and x3 are the input variables (R&D Spend, Administration Cost, and Marketing Spend), and b0, b1, b2, and b3 are the coefficients that are learned during training.

During training, the linear regression algorithm adjusts the coefficients to minimize the sum of squared errors between the predicted values and the actual values. This is done using an optimization algorithm called gradient descent. Once the coefficients are learned, the model can be used to predict the profit values for new companies based on their R&D Spend, Administration Cost, and Marketing Spend.

Methodology

The methodology for building an ML model that can predict the profit value using linear
regression can be broken down into the following steps:

4.1 Data Collection: Collect data from various sources such as company financial records, public financial records, and other relevant sources.

4.2 Data Preprocessing: Clean and preprocess the data to ensure it is in a format suitable for training an ML model. This may include tasks such as removing missing or inconsistent data, normalizing the data, and encoding categorical variables.

4.3 Feature Selection: Determine which features are most relevant for predicting the profit value of a company. In this case, the selected features are R&D Spend, Administration Cost, and Marketing Spend.

4.4 Split Data into Train and Test Sets: Split the data into a training set and a test set. The training set will be used to train the linear regression model, while the test set will be used to evaluate the model's performance

4.4 Split Data into Train and Test Sets: Split the data into a training set and a test set.The training set will be used to train the linear regression model, while the test set will be used to evaluate the model's performance.

4.5 Train the Model: Train a linear regression model using the training data.

4.6 Evaluate the Model: Evaluate the performance of the model using the test data. This
may involve metrics such as mean squared error or R-squared.

4.7 Optimize the Model: Optimize the model by adjusting hyperparameters such as
regularization strength or learning rate.

4.8 Deploy the Model: Once the model has been optimized, it can be deployed for use in
predicting the profit value of a company based on R&D Spend, Administration Cost, and
Marketing Spend.

Implementation

5.1 Source Code

Import the necessary libraries

```
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import sklearn
```

Loading and Analyzing Data

```
dataset = pd.read_csv('50_Startups.csv')

dataset.head()
dataset.tail()
dataset.describe()
print('There are' , dataset.shape[0],'rows and' , dataset.shape[1],'columns
in the dataset')

print('There are' , dataset.duplicated().sum(),'duplicate values in the
dataset')

dataset isnull().sum()

dataset.info()

c=dataset.corr()
c

sns.heatmap+c,annot=True,cmap='Blues')
plt.show()
outliers = ['Profit']
plt.rcParams['figure.figsize'] =[8,8]
sns.boxplot(data=dataset[outliers], orient='v', palette = 'Set2' , width
=0.7)
```

```
plt.title ('Outliers Variables Distribution ')
plt.ylabel('Profit Range')
plt.xlabel('Continuous Variable ')
plt.show()

sns.distplot(dataset['Profit'], bins=5, kde=True)
plt.show()

sns.pairplot(dataset)
plt.show()
```

Model Development and Training

```
x = dataset.iloc[:,:-1].values
y = dataset.iloc[:,3].values

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

from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(x_train,y_train)
```

Testing

```
y_pred = model.predict(x_test)

testing_data_model_score = model.score(x_test,y_test)

df = pd.DataFrame(data={'Predicted value' : y_pred.flatten(), 'Actual
value' : y_test.flatten()})
```

Model Evaluation

```
from sklearn.metrics import r2_score
r2_score = r2_score (y_pred,y_test)
print('R2 score of the model is' ,r2_score)

from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_pred,y_test)
print('Mean squared error of the model is' ,mse)


import numpy as np
rmse = np.sqrt(mean_square_error(y_pred,y_test))
print(' Root mean squared error of the model is' ,rmse)

from sklearn.metrics import mean_absolute_error
mae = mean_absolute_error(y_pred,y_test)
print('Mean absolute error of the model is' ,mse)
```

Conclusion

In conclusion, the Linear Regression model developed in this project can accurately predict the profit value of a company based on R&D Spend, Administration Cost, and Marketing Spend. The model was trained on a dataset containing information about several companies and their respective profits. The model was evaluated using metrics such as Mean Squared Error and R-squared, which showed that it is a good fit for the data and can be used to make accurate predictions.

The proposed system has several advantages over the existing systems, as it uses more relevant features and a better machine learning algorithm. This model can be used by investors and businesses to make more informed decisions about where to invest their money and how to improve their profits.

Overall, the project has been successful in developing an ML model that can predict the profit value of a company based on R&D Spend, Administration Cost, and Marketing Spend with high accuracy, which can have significant practical applications in the business world.