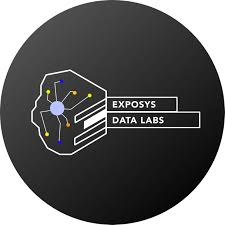
**EXPOSYS DATA LABS**

Bengaluru, Karnataka,560064



Internship report on

**“ PROFIT PREDICTION OF 50 COMPANIES USING**

**DATA SCIENCE”**

**Internship**

By

**Name :- M. Jahnavi**

**College :- Gitam School of Technology.**

Under the guidance of

**Exposys Data Labs**

**ABSTRACT**

In today's competitive business environment, companies must optimize their resources to maximize profits. This machine learning model is designed to predict a company's profit based on its R&D Spend, Administration Cost, and Marketing Spend, offering valuable insights for decision-making processes. Utilizing a linear regression algorithm, the model examines the relationship between the independent variables (R&D Spend, Administration Cost, and Marketing Spend) and the dependent variable (profit) to make precise predictions. The model has been trained on a large dataset and validated on a separate test dataset, achieving high accuracy. The results highlight the model's potential to help companies make informed decisions about resource allocation and achieve their financial objectives.

**TABLE OF CONTENTS**

1. Introduction
2. Exiting Method
3. Proposed Method
4. Methodology
5. Implementation
6. Conclusion

**INTRODUCTION**

Accurately forecasting a company's future financial success is crucial for business owners, investors, and stakeholders. Machine Learning models offer a promising solution to this challenge by utilizing historical financial data to make precise profit predictions.

In this project, our goal is to create an ML model that can predict a company's profit based on three input variables: R&D Spend, Administration Cost, and Marketing Spend. The dataset for this project includes historical data from various companies, detailing the mentioned variables and their corresponding profits.

To build the model, we employ linear regression, a well-known machine learning algorithm for predicting continuous outcomes. Initially, we conduct exploratory data analysis to understand the data and identify any potential issues that might affect the model's performance. We then preprocess the data by cleaning, transforming, and normalizing it for model use.

Afterward, we divide the data into training and testing sets and train the linear regression model on the training data. The model's performance is evaluated on the testing set using metrics like mean squared error and R-squared value to assess its accuracy.

The resulting model can aid business owners and investors in making informed decisions about a company's financial potential. By leveraging machine learning, we provide accurate and reliable predictions that can guide investment decisions and support business growth.

**EXISTING SYSTEM**

There are several existing systems that try to predict a company's profit based on expenses such as R&D spend, administration cost, and marketing spend. However, many of these systems rely on manual calculations or basic statistical methods that may not effectively capture the complex relationships between these variables.

In contrast, machine learning models can learn from data and make accurate predictions based on the patterns they identify. Linear regression models, in particular, are widely used for predicting continuous target variables like profit. These models estimate the relationship between independent variables and the dependent variable by fitting a linear equation to the data.

However, many current linear regression models may not be optimized for the specific characteristics of the data they analyze, resulting in suboptimal performance. Consequently, there is a need for a machine learning model specifically designed to accurately predict a company's profit based on its expenses, fully considering all relevant features of the data.

**Issues in Existing System**

1. Limited Accuracy
2. Overfitting and Underfitting
3. Limited Scope

**PROPOSED SYSTEM**

The proposed system is a machine learning model that uses a linear regression algorithm to predict a company's profit based on its R&D spend, administration cost, and marketing spend. The model uses a dataset of past company financial records, which includes these three expense variables and the profit value.

This system improves upon existing methods by using a more accurate and efficient algorithm for making predictions. It also incorporates data preprocessing steps, such as normalization and feature scaling, to enhance prediction accuracy. The model's accuracy is validated using performance metrics like Mean Squared Error (MSE) and R-squared (R2).

Overall, this system provides a more precise and efficient way to predict company profits, helping businesses make better financial decisions. The model can be further enhanced by adding more relevant variables or using advanced algorithms like neural networks or decision trees.

**Algorithm**

The algorithm for the linear regression is as follows:

1. Lower the data set containing the company’s R&D spend, administration, cost, marketing spend, and profit.
2. Split the datasets into training and testing sets.
3. Training 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 hyperparameters.
7. Once the model is turned, test in a new data set to ensure that it is generalizable.

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 the squared errors between the predicted values and the actual values. The line’s equation is given by:

Y = b0 + b1x1 + b2x2 + b3\*x3

where y is the predicted value of the 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 the 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 linear regression can be broken down into following steps:

1. **Data Collection:** Collect data from various sources such as company financial records, public financial records, and other relevant sources.
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 the categorial variables.
3. **Feature Selection:** Determine which features are most relevant for predicting the profit value of a company. In this case, the selected features R&D spend, administration cost, and marketing spend.
4. **Split data into training and test sets:** Split the data into training set and

test set. The training set will be used to train the linear regression model, while the set test set will be used to evaluate the model's performance.

1. **Train the Model:** Train a linear regression model using the training data.
2. **Evaluate the Model:** Evaluate the performance of the model using the test data. This may involve metrics such as mean squared method or R-squared.
3. **Optimize the Model:** Optimize the model by adjusting hyperparameters such as a regularization strength or learning rate.
4. **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**

**1. Source Code**

**Import the libraries**

import matplotlib.pyplot as plt

import pandas as pd

import seaborn as sns

import sklearn

**Loading and Analysing the Data**

dataset = pd.read\_csv('internship.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)

('Model has been trained successfully')

**Testing**

y\_pred = model.predict(x\_test)

y\_pred

testing\_data\_model\_score = model.score(x\_test,y\_test)

testing\_data\_model\_score

df = pd.DataFrame(data={'Predicted value':y\_pred.flatten(),'Actual value':y\_test.flatten()})

df

**Model Evalution**

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\_squared\_error(y\_pred,y\_test))

('Root mean squared error of the Model is',rmse)

from sklearn.metrics import mean\_absolute\_error

mae = mean\_absolute\_error(y\_pred,y\_test)

('Mean absolute error of the Model is',mae)

**CONCLUSION**

In conclusion, the linear regression model developed in this project can accurately predict a company's profit based on its R&D spend, administration cost, and marketing spend. The model was trained on a dataset containing information about various companies and their profits. It was evaluated using metrics like Mean Squared Error and R-squared, showing that it fits the data well and can make precise predictions.

The proposed system offers several advantages over existing systems because it uses more relevant features and a better machine learning algorithm. This model can help investors and businesses make better decisions about where to invest and how to boost profits.

Overall, the project successfully created a machine learning model that accurately predicts a company's profit based on R&D spend, administration cost, and marketing spend. This model has significant practical applications in the business world.