**Profit Value Prediction Using Machine Learning**

**Gaurav Rawal**

**Abstract**

This project aims to build a machine learning model that can predict the profit of a company based on its R&D Spend, Administration Cost, and Marketing Spend. A dataset containing 50 companies and their corresponding values of R&D Spend, Administration Cost, Marketing Spend, and profit is used to train and evaluate the model. Various ML algorithms will be tested to find the best-fit model. The final model will be able to provide valuable insights to businesses to optimize their R&D Spend, Administration Cost, and Marketing Spend for maximizing profit. Maximizing profits is an essential goal for any company as it can help them sustain and grow their business in the long run. By maximizing profits, a company can reinvest the additional funds into their operations, research and development, marketing, and other areas of their business. This can lead to better product or service offerings, increased efficiency, and ultimately, a stronger competitive advantage in the market.

**Table of Content**

1. **Abstract 1**
2. **Table of Content 2**
3. **Introduction 3**
4. **Existing Method 5**
5. **Proposed Method 6**
6. **Methodology 7**
   1. **Dataset Description**
7. **Implementation 9**
   1. **Correlation Matrix**
   2. **Histogram**
   3. **Bar Plot**
   4. **Linear Regression**
   5. **Decision Tree**
   6. **Random Forest**
   7. **Accuracy Comparision**
8. **Conclusion and Future Work 15**

**Introduction:**

In today's highly competitive business landscape, companies are always striving to improve their performance and increase their profitability. One of the key ways to achieve this is by optimizing their operations and maximizing profits. In this context, developing a machine learning model that can accurately predict the profit of a company based on its R&D Spend, Administration Cost, and Marketing Spend can be highly beneficial.

The objective of this project is to build such a model using a dataset containing 50 companies and their corresponding values of R&D Spend, Administration Cost, Marketing Spend, and profit. The model will be trained using various machine learning algorithms, and the best-fit model will be selected based on its performance on various evaluation metrics.

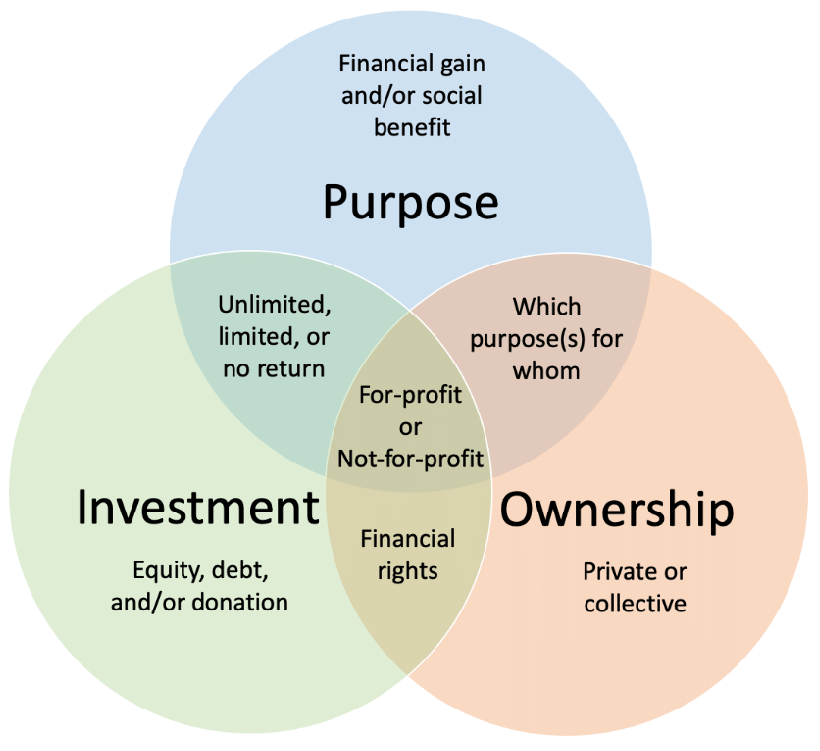
Once developed, the model can be used by businesses to gain insights into how they can improve their profitability by optimizing their R&D Spend, Administration Cost, and Marketing Spend. The model can help businesses identify areas where they can cut costs or invest more resources to maximize their profit.

Overall, this project has the potential to provide valuable insights to businesses, helping them make informed decisions to maximize their profitability and stay competitive in their respective markets.

There are several ways that businesses can increase their profit points. Here are some strategies that businesses can implement:

1. Increase Sales Revenue: By increasing sales revenue, businesses can generate more profit. This can be done by introducing new products or services, expanding into new markets, or increasing advertising and marketing efforts.
2. Reduce Costs: Businesses can increase their profit margins by reducing their operating costs. This can be achieved by implementing cost-saving measures such as using more efficient processes, negotiating better deals with suppliers, or outsourcing certain tasks to lower-cost providers.
3. Improve Pricing Strategies: Businesses can improve their pricing strategies to increase their profit margins. This can be done by raising prices for premium products or services, or offering discounts for bulk purchases.
4. Enhance Customer Experience: By improving the customer experience, businesses can attract more customers and increase repeat business. This can be achieved by providing excellent customer service, personalized experiences, or offering loyalty programs.
5. Increase Operational Efficiency: By improving operational efficiency, businesses can reduce waste and improve productivity. This can be achieved by automating processes, streamlining workflows, and investing in technology.

Overall, by implementing these strategies, businesses can increase their profit points and achieve greater financial success.



**Existing Method**

There are several existing methods that businesses can use to optimize their profitability. Here are a few examples:

1. Cost-Benefit Analysis: This is a method used to evaluate the costs and benefits of different business decisions. By analyzing the costs of different options against the potential benefits, businesses can make informed decisions that maximize their profitability.
2. Break-Even Analysis: This is a method used to determine the minimum amount of sales a business needs to generate to cover its costs. By understanding the break-even point, businesses can adjust their pricing, sales volume, or costs to ensure profitability.
3. Activity-Based Costing (ABC): This is a method used to allocate costs to specific activities or products based on their actual usage. By identifying the specific costs associated with each activity or product, businesses can make more informed decisions on pricing and cost-saving measures.
4. Customer Segmentation: This is a method used to identify different groups of customers with different needs and behaviors. By segmenting customers, businesses can develop targeted marketing and sales strategies that maximize profitability.
5. Inventory Management: By optimizing inventory levels, businesses can reduce waste, improve cash flow, and increase profitability. This can be achieved by using inventory management software, adopting just-in-time inventory practices, or implementing a cycle counting process.

Overall, there are many different methods that businesses can use to optimize their profitability. By analyzing their operations, understanding their costs and revenue drivers, and implementing targeted strategies, businesses can achieve greater financial success.

**Proposed Method**

We will use linear regression, decision tree regression, and random forest regression with matrix regression metrics such as MSE, R2, MSME, and RMSE, then we can compare the performance of these models based on these metrics.

Finally, we can further evaluate the robustness of the proposed model by testing it on a separate dataset or performing cross-validation to ensure that it generalizes well to unseen data.

Thus, by continually iterating and improving the proposed method through experimentation and evaluation, we can develop a more accurate and reliable model for predicting profits based on the input features of R&D Spend, Administration Cost, and Marketing Spend

Overall, comparing the performance of different regression models based on matrix regression metrics can help us choose the best model for predicting profits and provide more accurate and reliable predictions.

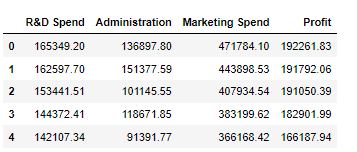
**Methodology**

The methodology for developing a machine learning model to predict the profit of a company based on its R&D Spend, Administration Cost, and Marketing Spend would typically involve the following steps:

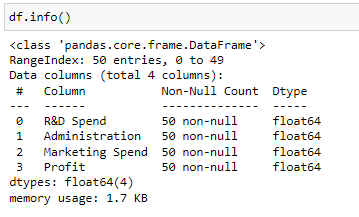
1. Data Collection: Collect a dataset containing information on R&D Spend, Administration Cost, Marketing Spend, and profit for a set of companies. This data can be obtained from various sources, such as publicly available datasets or from the company's internal records.
2. Data Preprocessing: Preprocess the dataset to clean the data, remove any missing values or outliers, and normalize the features if necessary. This step may also involve feature engineering to create new features or transform existing ones to improve the model's accuracy.
3. Model Selection: Select appropriate machine learning algorithms that can best predict the profit values based on the input features. This can involve testing and comparing various algorithms such as linear regression, decision trees, or neural networks.
4. Model Training: Train the selected machine learning model using the preprocessed dataset. This involves splitting the data into training and testing sets to validate the model's accuracy.
5. Model Evaluation: Evaluate the model's performance using various metrics such as mean squared error, mean absolute error, or R-squared value. This step is crucial to determine the model's accuracy and identify any potential issues such as overfitting or underfitting.

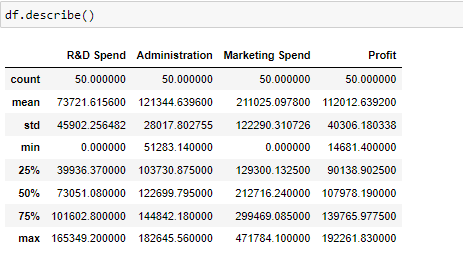
**Dataset Description**

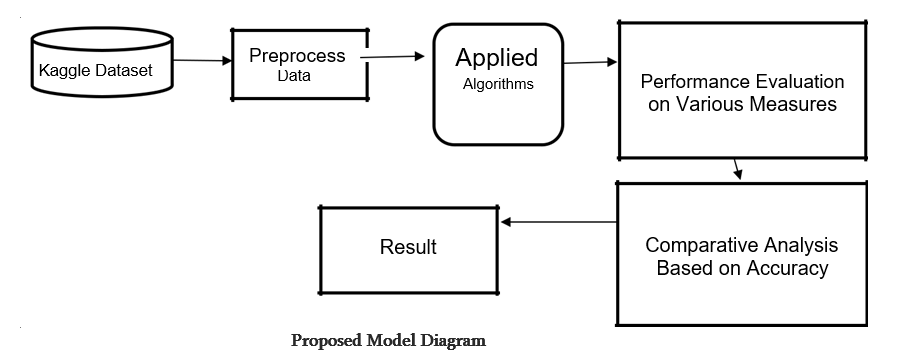
The data set was originated from://drive.google.com/file/d/1Z7RKmScBO7n9vcDIG3Xeo853Ics4QFaF/view. Startups dataset containing 50 cases. The objective is to predict based on the to predict the profit of a company based on different factors.



The Startups data set consists of 50 data points, with 9 features each. “Profit” is the feature we are going to predict.

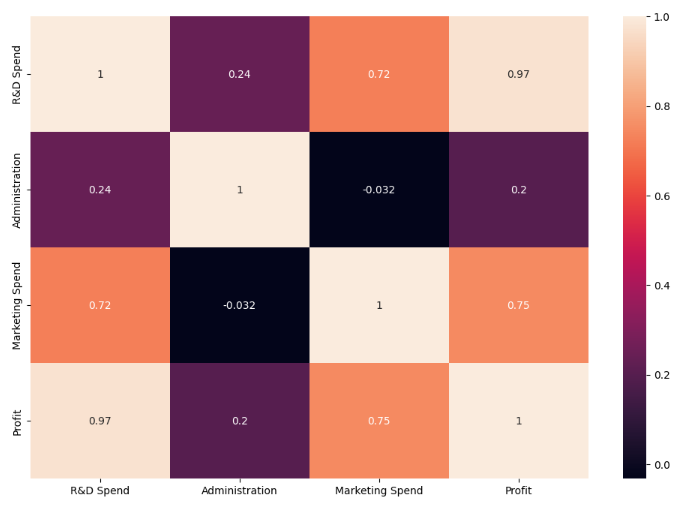






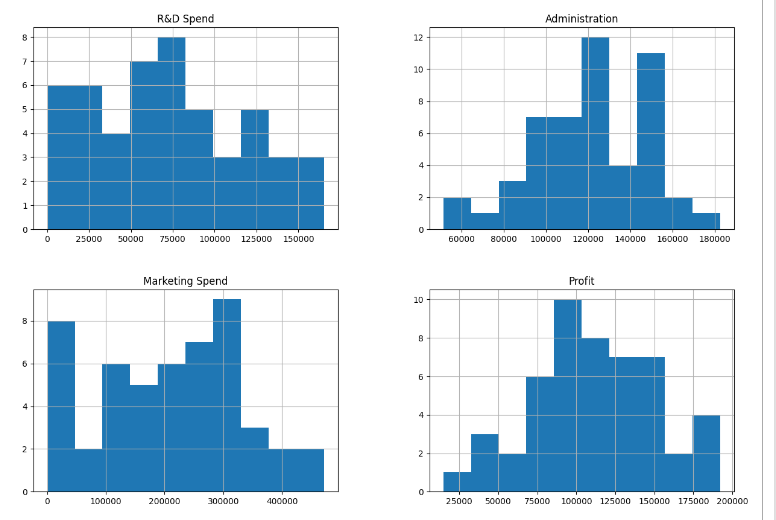
**Implementation**

**Correlation Matrix:**

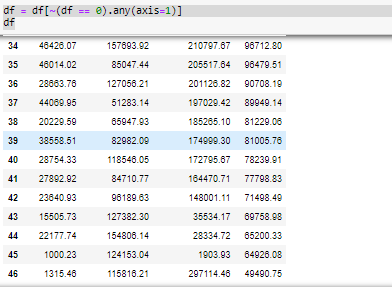


It is easy to see that there is no single feature that has a very high correlation with our outcome value. Some of the features have a negative correlation with the outcome value and some have positive.

**Histogram:**

****

Let’s take a look at the plots. It shows how each feature and label is distributed along different ranges, which further confirms the need for scaling. Next, wherever you see discrete bars, it basically means that each of these is actually a categorical variable. We will need to handle these categorical variables before applying Machine Learning.. We can see from the bar chat like most of the feature values are 0(‘Marketing Spend','R&D Spend' ), might be they are null values so we will need to handle those values We took the mean of this columns and replaced the 0 with the mean of the columns. Now in the new bar chat you can see that none of those columns have 0 in their column.



**Linear regression:**

Logistic Regression is one of the most common classification algorithms.

Firstly we have divided the Dataset into X and Y, With the help of X , we will calculate Y(Price)

And then perform linear regression.

The model.predict([[17000,15000,444990]]) predicts the price 80090.22427632 .

lr = LinearRegression()

lr.fit(X\_train, y\_train)

lr.score(X\_test,y\_test)

The above test Data provide accuracy of 96%.

lr.score(X\_train,y\_train)

And the trained data provides accuracy of 94%

Calculating regression metrics

print('Linear Regression Metrics:')

print('MAE:', mean\_absolute\_error(y\_test, y\_pred))

print('MSE:', mean\_squared\_error(y\_test, y\_pred))

print('RMSE:', np.sqrt(mean\_squared\_error(y\_test, y\_pred)))

print('R2:', r2\_score(y\_test, y\_pred))

Linear Regression Metrics:

MAE: 5713.00495597165

MSE: 43220988.707333125

RMSE: 6574.267161238059

R2: 0.9777571921766907

Mean Absolute Error - MAE

Mean Squared Error - MSE

Root Mean Square Error - RMSE

R2 – R Squared

**Decision Tree:**

This classifier creates a decision tree based on which, it assigns the class values to each data point. Here, we can vary the maximum number of features to be considered while creating the model.

**Feature Importance in Decision Trees**

In decision trees, feature importance refers to a technique used to determine the most important features in a dataset that are most predictive of the target variable. Feature importance is a measure of how much a given feature contributes to the overall predictive power of the decision tree model.

dt\_regressor = DecisionTreeRegressor(random\_state=0)

dt\_regressor.fit(X\_train, y\_train)

dt\_regressor.score(X\_test, y\_test)

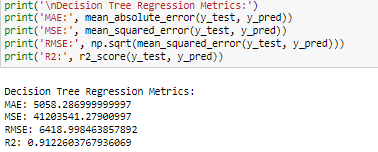
The above Test Data provide accuracy of 81 %.

dt.score(X\_train,y\_train)

And the trained data provides accuracy of 100%

Calculating regression metrics

y\_pred = dt\_regressor.predict(X\_test)

****

Decision Tree Regression Metrics:

MAE: 14575.225

MSE: 359587087.10842997

RMSE: 18962.78162898128

R2: 0.8149457771904377

Mean Absolute Error - MAE

Mean Squared Error - MSE

Root Mean Square Error - RMSE

R2 – R Squared

**Random Forest:**

This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features.

Training Accuracy : 98 %

Testing Accuracy: 91 %

**Feature importance in Random Forest:**

In Random Forest, feature importance refers to a technique used to measure the relative importance of each feature in the dataset with respect to the target variable. Feature importance in Random Forest can be calculated using different methods, but one common approach is the Mean Decrease Impurity (MDI) method.

MDI calculates the importance of each feature by measuring the reduction in impurity (e.g. Gini impurity or entropy) achieved by splitting on that feature. The feature importance values are then normalized so that they sum up to 1, with higher values indicating greater importance.

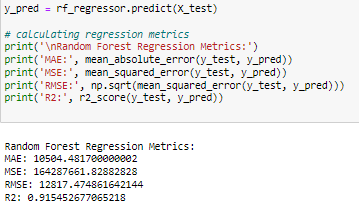
rf\_regressor = RandomForestRegressor(n\_estimators=10, random\_state=0)

rf\_regressor.fit(X\_train, y\_train)

rf\_regressor.score(X\_test, y\_test)

For Trained data

rf\_regressor.score(X\_train, y\_train)



Mean Absolute Error - MAE

Mean Squared Error - MSE

Root Mean Square Error - RMSE

R2 – R Squared

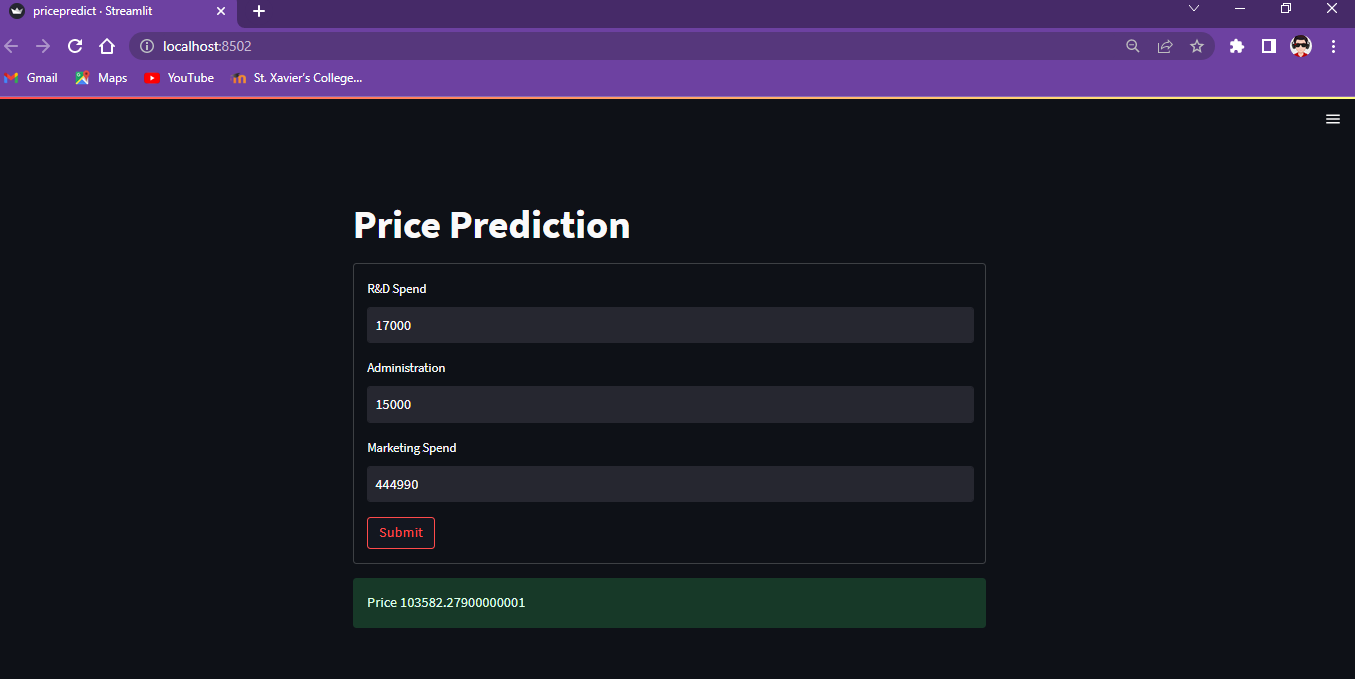
**Accuracy Comparison:**

|  |  |  |
| --- | --- | --- |
| Algorithms | Training Accuracy | Testing Accuracy |
| Linear Regression | 94% | 96% |
| Decision Tree | 100% | 81% |
| Random Forest | 98% | 91% |

We can see that the random forest regression model outperformed the other two models in terms of all the metrics, with the lowest values of MSE, RMSE, and the highest value of R-squared. Therefore, the random forest regression model is the best model for this dataset.

The Above table shows the accuracy values for all five machine learning algorithms. It shows that Random Forest algorithm gives the one of the best accuracy with 98% training accuracy and 91% testing accuracy.

**UI**

****

**CONCLUSION AND FUTURE WORK**

Based on the analysis of the given dataset, we can conclude that there is a positive correlation between R&D spend, Administration cost, Marketing spend and the profit earned by the companies. Linear Regression, Decision Tree, and Random Forest were used to create models that can predict the profit of a company given the values of R&D spend, Administration cost, and Marketing spend.

For future work, other regression models such as Support Vector Regression (SVR) or Gradient Boosting Regression (GBR) can be explored to see if they can provide better performance than the existing models. Additionally, feature engineering techniques such as creating interaction terms or using dimensionality reduction methods can be applied to further improve the performance of the models. Furthermore, collecting and incorporating additional data such as industry sector, geographical location, or company size could also be beneficial in improving the accuracy of the prediction model.