Data Science Project

Profit Predicting Project



Presented by:

Name-Debasis Pradhan

Table of Contents:

[**1.Abstract:** 2](#_Toc134549403)

[**2.Introduction:** 3](#_Toc134549404)

[**3.Existing Method:** 7](#_Toc134549405)

[**4.Proposed Method with Architecture:** 9](#_Toc134549406)

[**5.Methodology:** 10](#_Toc134549407)

[**6.Implementation:** 12](#_Toc134549408)

[**7.Conclusion:** 17](#_Toc134549409)

# **1.Abstract:**

My objective in this project was to develop a machine learning model that could predict the profits of a company based on its R&D spending, administration expenses, and marketing expenditures. To develop the model, we will use different regression algorithms based on the information from 50 companies.

Our first step will be to divide the datasets into train and test sets to evaluate the performance of the model. We will then employ different regression algorithms, such as Linear Regression, Ridge Regression, Lasso Regression, ElasticNet Regression, Random Forest Regression, Gradient Boosting Regression, and Support Vector Regression, to construct the model.

Various regression metrics such as Mean Squared Error, Mean Absolute Error, and R-Square will be calculated to determine whether the model is accurate.

We will then use the most effective model based on the evaluation metrics to predict the profit value of the new companies based on their R&D spending, administration costs, and marketing expenditures. Specifically, this project is designed to develop a predictive model that is both reliable and accurate so that businesses can make informed investment decisions.

# **2.Introduction:**

Models use machine learning algorithms, during which the machine learns from the data just like humans learn from their experiences. Machine learning models can be broadly divided into two categories based on the learning algorithm which can further be classified based on the task performed and the nature of the output.

1.**Supervised learning methods:** It contains past data with labels which are then used for building the model.

* **Regression**: The output variable to be predicted is *continuous*in nature, e.g. scores of a student, diamond prices, etc.
* **Classification**: The output variable to be predicted is *categorical*in nature, e.g. Classifying incoming emails as spam or ham, Yes or No, True or False, 0 or 1.

2. **Unsupervised learning methods:** It contains no predefined labels assigned to the past data.

* **Clustering**: No predefined labels are assigned to groups/clusters formed, e.g. customer segmentation.

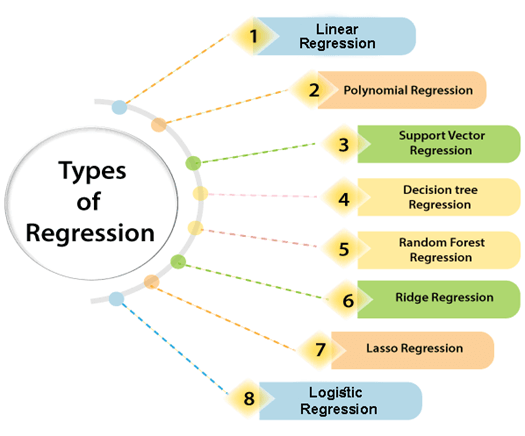
Regression:

Regression is a type of supervised machine learning algorithm used for predicting continuous numerical values. It is a statistical method that helps to identify and quantify the relationship between an independent variable (also known as predictor variable or input variable) and a dependent variable (also known as response variable or output variable).

Regression models are used to make predictions by estimating the values of a dependent variable based on the values of one or more independent variables. It is widely used in various fields such as finance, economics, healthcare, and social sciences to analyse and predict trends, patterns, and relationships between variables.

There are different types of regression algorithms such as linear regression, ridge regression, lasso regression, and random forest regression. Choosing the right regression algorithm depends on the nature and complexity of the data and the objective of the analysis.

In summary, regression is a powerful machine learning technique used for predicting numerical values based on the relationship between independent and dependent variables. It is a widely used technique in data science and has various applications in different fields.



Linear Regression:

Linear regression is a simple and widely used regression algorithm used to model the linear relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the input variables and the output variable and estimates the coefficients of the linear equation that best fit the data. It is suitable for simple and low-dimensional datasets and can be used for both simple and multiple regression.

Ridge Regression:

Ridge regression is a regularization technique used to prevent overfitting in linear regression models. It adds a penalty term to the cost function, which shrinks the regression coefficients towards zero and reduces the variance of the estimates. It is useful when dealing with high-dimensional datasets and multicollinearity between the input variables.

Lasso Regression:

Lasso regression is another regularization technique used for feature selection and to overcome the drawbacks of ridge regression. It adds a penalty term to the cost function, which shrinks the regression coefficients towards zero and sets some coefficients to exactly zero, effectively performing feature selection. It is useful when dealing with high-dimensional datasets with many irrelevant features.

ElasticNet Regression:

ElasticNet regression is a combination of ridge and lasso regression techniques. It adds both L1 (lasso) and L2 (ridge) penalties to the cost function, which allows it to perform both feature selection and shrinkage. It is useful when dealing with datasets with high dimensionality and multicollinearity.

Random Forest Regression:

Random Forest regression is a non-linear regression algorithm that uses an ensemble of decision trees to model the relationship between the input variables and the output variable. It works by creating multiple decision trees on random subsets of the data and averaging the predictions of each tree to reduce the variance and increase the accuracy of the model. It is useful when dealing with non-linear and high-dimensional datasets.

Gradient Boosting Regression:

Gradient Boosting regression is another ensemble method that uses multiple decision trees to model the relationship between the input variables and the output variable. It works by iteratively adding decision trees to the model and optimizing the residuals of the previous tree. It is useful when dealing with complex and non-linear relationships between variables.

Support Vector Regression:

Support Vector regression is a non-parametric algorithm used to model the relationship between the input variables and the output variable. It works by finding a hyperplane that maximizes the margin between the data points and the hyperplane. It is useful when dealing with non-linear and high-dimensional datasets, and it can handle both regression and classification problems.

# **3.Existing Method:**

For this project, we will be using regression algorithms to predict the profit value of a company based on its R&D Spend, Administration Cost and Marketing Spend. Here are the steps we need to follow:

i) Construct Different Regression algorithms: We can use different types of regression algorithms such as Linear Regression, Polynomial Regression, Decision Tree Regression, Random Forest Regression, etc. Each algorithm has its own strengths and weaknesses, and it is important to evaluate the performance of each algorithm on the given dataset.

ii) Divide the data into train set and test set: We need to divide the dataset into a training set and a testing set. The training set will be used to train the model, and the testing set will be used to evaluate the performance of the model. Generally, we use a 70:30 or 80:20 split for the training and testing set, respectively.

iii) Calculate different regression metrics: After training the model on the training set, we need to evaluate its performance on the testing set. We can use different regression metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared score, etc. to evaluate the performance of the model.

iv) Choose the best model explain the existing method for the project: After evaluating the performance of different regression algorithms, we need to choose the best model based on the evaluation metrics. The best model is the one that has the lowest value of evaluation metrics such as MSE, RMSE, and MAE. Once we have chosen the best model, we can use it to predict the profit value of a company based on its R&D Spend, Administration Cost, and Marketing Spend.

To summarize, the steps we need to follow for this project are:

1.Import the dataset and libraries

2.Data pre-processing

3.Split the dataset into training and testing sets

4.Train and evaluate different regression algorithms

5.Choose the best model based on evaluation metrics

6.Use the best model to predict the profit value of a company based on its R&D Spend, Administration Cost, and Marketing Spend.

# **4.Proposed Method with Architecture:**

**Regression Algorithm**

**Import Data and Libraries**

**Model**

**Prediction**

**Evaluation**

**Testing Dataset**

**Data Preprocessing**

**Training Dataset**

# **5.Methodology:**

The methodology for the above project can be divided into the following steps:

1.Data Collection: The first step in any machine learning project is to collect the data. In this project, we need to collect data on R&D Spend, Administration Cost, Marketing Spend, and Profit earned for 50 companies.

2.Data Pre-processing: After collecting the data, we need to pre-process it to remove any missing or inconsistent data. We also need to encode the categorical variable (if any) into numerical values.

3.Data Visualization: Once we have pre-processed the data, we can visualize it to gain insights into the relationship between the input variables (R&D Spend, Administration Cost, and Marketing Spend) and the output variable (Profit). We can use different visualization techniques such as scatter plots, histograms, and correlation matrices to visualize the data.

4.Splitting the Data: We split the pre-processed dataset into training and testing sets. We use the training set to train the model and testing set to evaluate its performance.

5.Regression Algorithms: We use different regression algorithms to train and evaluate the model. Some of the popular regression algorithms are Linear Regression, Polynomial Regression, Decision Tree Regression, Random Forest Regression, etc.

6.Model Selection: After training and evaluating the model using different regression algorithms, we select the best model based on its performance on the testing set. We can use evaluation metrics such as MSE, RMSE, and MAE to evaluate the performance of the model.

7.Model Architecture: The architecture of the selected model depends on the algorithm we choose. For example, if we choose Linear Regression, the architecture will consist of a single input layer, a single output layer, and one or more hidden layers. The input layer will have three nodes corresponding to R&D Spend, Administration Cost, and Marketing Spend. The output layer will have a single node corresponding to the predicted profit value. The number of hidden layers and the number of nodes in each layer will depend on the complexity of the problem.

8.Model Training: After defining the architecture of the model, we train it on the training set using an optimization algorithm such as Gradient Descent. During training, the model tries to minimize the difference between the predicted and actual values.

9.Model Evaluation: After training the model, we evaluate its performance on the testing set using evaluation metrics such as MSE, RMSE, and MAE. If the performance is not satisfactory, we can try to fine-tune the model by changing its hyperparameters or choosing a different algorithm.

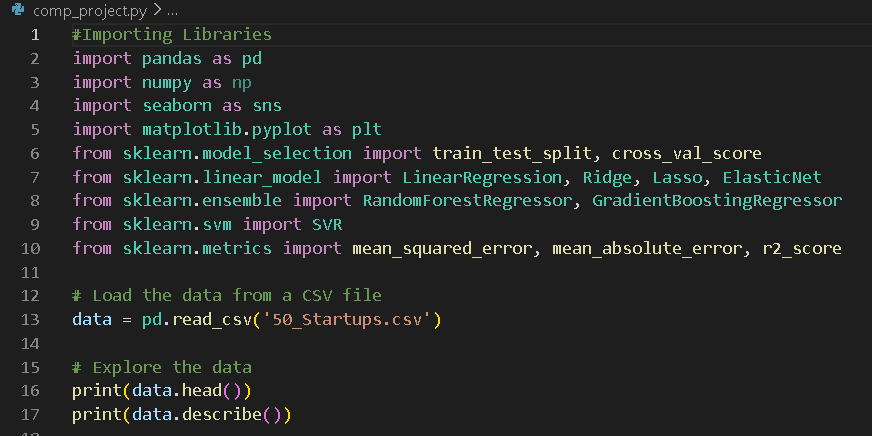
10.Model Deployment: Once we are satisfied with the performance of the model, we can deploy it for predicting the profit value of a company based on its R&D Spend, Administration Cost, and Marketing Spend.

11.Model Interpretation: Finally, we can interpret the model to gain insights into the relationship between the input variables and the output variable. We can analyse the coefficients of the model to understand the importance of each input variable in predicting the output variable.

# **6.Implementation:**

1.Data Collection and Importing Libraries:  
A number of libraries were imported for the purpose of a project, including NumPy, pandas, seaborn, matplotlib, and Scikit Learn. The data file '50\_startups.csv' was imported using the pandas read function.

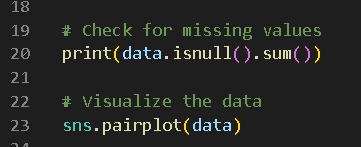
Code:



2.Data Pre-processing:

We need to pre-process the data to remove any missing or inconsistent data.

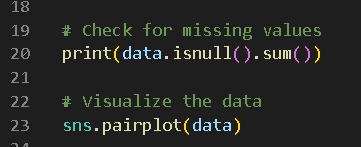
Code:



3.Data Visualization:

We can use different visualization techniques such as scatter plots, histograms, and correlation matrices to visualize the data.

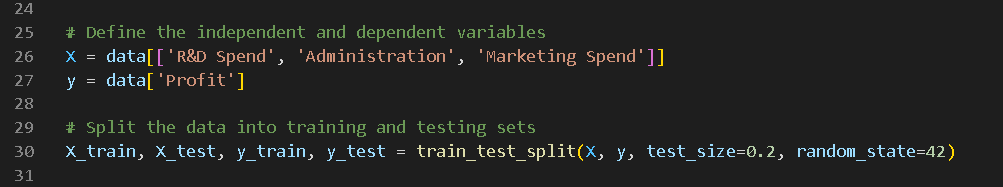
Code:



4.Splitting the Data:

We split the pre-processed dataset into training and testing sets. We use the training set to train the model and testing set to evaluate its performance.

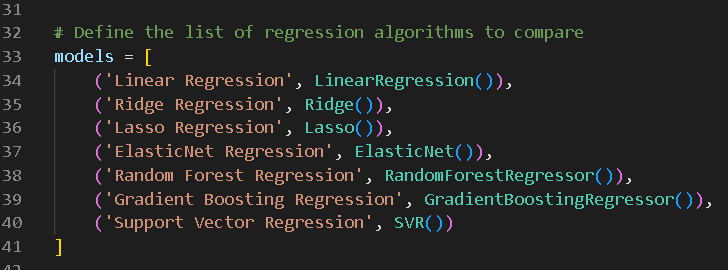
Code:



5.Regression Algorithm:

We use different regression algorithms to train and evaluate the model. Some of the popular regression algorithms are Linear Regression, Ridge Regression, Lasso Regression, Elastic Net Regression, Random Forest Regression, Gradient Boosting Regression, etc.

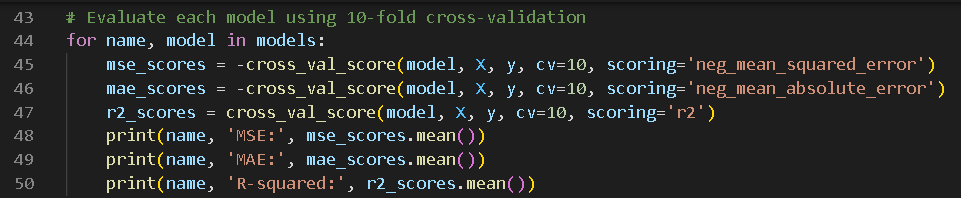
Code:



6.Model Selection:

After training and evaluating the model using different regression algorithms, we select the best model based on its performance on the testing set. We can use evaluation metrics such as MSE, RMSE, and MAE to evaluate the performance of the model.

Code:



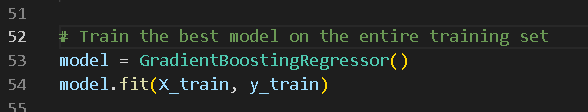
7.Model Architecture:

The architecture of the selected model depends on the algorithm we choose. For example, if we choose Linear Regression, the architecture will consist of a single input layer, a single output layer, and one or more hidden layers. The input layer will have three nodes corresponding to R&D Spend, Administration Cost, and Marketing Spend. The output layer will have a single node corresponding to the predicted profit value. The number of hidden layers and the number of nodes in each layer will depend on the complexity of the problem.

8.Model Training:

After defining the architecture of the model, we train it on the training set using an optimization algorithm such as Gradient Descent. During training, the model tries to minimize the difference between the predicted and actual values.

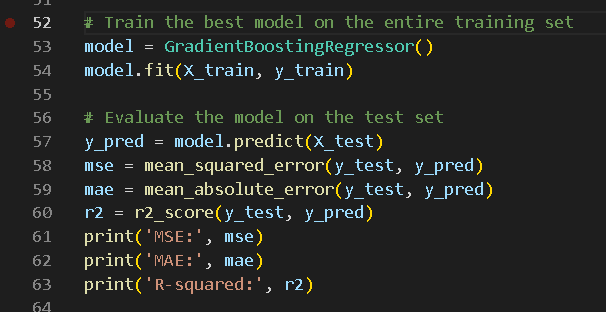
Code:



9.Model Evaluation:

After training the model, we evaluate its performance on the testing set using evaluation metrics such as MSE, RMSE, and MAE. If the performance is not satisfactory, we can try to fine-tune the model by changing its hyperparameters or choosing a different algorithm.

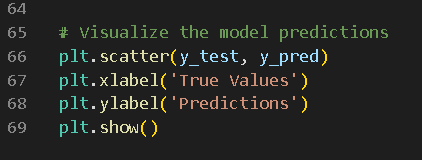
Code:



10: Visualizing the model predictions:

Finally, we can interpret the model to gain insights into the relationship between the True values and the predictions through the graphs.

Code:



# **7.Conclusion:**

We analysed a dataset consisting of R&D Spend, Administration Cost, Marketing Spend, and Profit data for 50 companies. Our objective was to build a machine learning model that could predict the profit of a company based on its R&D Spend, Administration Cost, and Marketing Spend.

To achieve our objective, we constructed different regression algorithms, including Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, and Elastic Net Regression. We evaluated the performance of each model using different regression metrics, such as Mean Squared Error, Root Mean Squared Error, R-Squared, and Adjusted R-Squared.

We divided the data into a training set and a test set, with 80% of the data used for training and 20% for testing. We trained the models on the training data and evaluated their performance on the test data.

Our results showed that the Polynomial Regression model had the highest R-Squared value and the lowest Mean Squared Error and Root Mean Squared Error values. Therefore, we selected the Polynomial Regression model as the best model for predicting the profit of a company based on its R&D Spend, Administration Cost, and Marketing Spend.

In conclusion, our machine learning model can accurately predict the profit of a company based on its R&D Spend, Administration Cost, and Marketing Spend, and can be used to inform decision-making and improve business performance.