**APPENDIX 1**

50 COMPANIES DATA REGRESSION ANALYSIS

INTERNSHIP PROJECT REPORT

**NAME OF THE CANDIDATE: LIYAKHAT YOUSUF MOGAL**

**DEPARTMENT OF**: Artificial intelligence and Data science

**B. S. Abdur Rahman Crescent Institute Of Science And Technology**

**Vandalur, Chennai**

September, 2022

**APPENDIX 2**

**Table of Contents**

CHAPTER NO. PG NO.

1. Introduction

1.1 Intro on Machine learning 3

1.2 Regression and Types 4

1.3 Problem Statement 5

1.4 Approach to Problem 5

1.5 Packages in Python 6

2. In-depth Intuition and Practical Implementation

2.1 Data description and Preprocessing 7

2.2 Data visualization techniques 7

2.3 Linear Regression 10

2.4 Ridge Regression 11

2.5 Lasso Regression 12

2.6 Decision Tree Regressor 13

2.7 Random Forest Regressor 14

2.8 Support Vector Regressor 15

3. Conclusion 16

**Introduction**

**Intro on Machine Learning**

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

The iterative aspect of machine learning is important because as models are exposed to new data, they are able to independently adapt. They learn from previous computations to produce reliable, repeatable decisions and results.

**Applications of Machine Learning**

[**Financial services**](https://www.sas.com/en_in/industry/banking.html)**:** Banks and other businesses in the financial industry use machine learning technology for two key purposes: to identify important insights in data, and prevent fraud. The insights can identify investment opportunities, or help investors know when to trade. Data mining can also identify clients with high-risk profiles, or use cyber surveillance to pinpoint warning signs of fraud.

### [Government](https://www.sas.com/en_in/industry/government.html): Government agencies such as public safety and utilities have a particular need for machine learning since they have multiple sources of data that can be mined for insights. Analysing sensor data, for example, identifies ways to increase efficiency and save money. Machine learning can also help detect [fraud and minimize identity theft](https://www.sas.com/en_in/insights/data-management/cloud-computing.html).

### [Health care](https://www.sas.com/en_in/industry/health-care.html): Machine learning is a fast-growing trend in the health care industry, thanks to the advent of wearable devices and sensors that can use data to assess a patient's health in real time. The technology can also help medical experts analyze data to identify trends or red flags that may lead to improved diagnoses and treatment.

### [Transportation](https://www.sas.com/en_in/industry/travel-transportation.html): Analyzing data to identify patterns and trends is key to the transportation industry, which relies on making routes more efficient and predicting potential problems to increase profitability. The data analysis and modeling aspects of machine learning are important tools to delivery companies, public transportation and other transportation organizations.

**Regression and Types**

### In [statistical modeling](https://en.wikipedia.org/wiki/Statistical_model), regression analysis is a set of statistical processes for [estimating](https://en.wikipedia.org/wiki/Estimation_theory) the relationships between a [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more [independent variables](https://en.wikipedia.org/wiki/Independent_variable) (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is [linear regression](https://en.wikipedia.org/wiki/Linear_regression), in which one finds the line (or a more complex [linear combination](https://en.wikipedia.org/wiki/Linear_combination)) that most closely fits the data according to a specific mathematical criterion.

### Types of Regressions:

### Linear Regression: Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression*.*

### Ridge Regression: Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

### Lasso Regression: ****Lasso regression**** is a type of [linear regression](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/)that uses [shrinkage](https://www.statisticshowto.com/shrinkage-estimator/). Shrinkage is where data values are shrunk towards a central point, like the [mean](https://www.statisticshowto.com/mean/). The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).  Lasso stands for Least Absolute Shrinkage and Selection Operation.

### Decision Tree Regression: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

### Random Forest Regression: **Random Forest Regression** is a supervised learning algorithm that uses **ensemble learning** method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

### Support Vector Regression: Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

**Problem Statement**

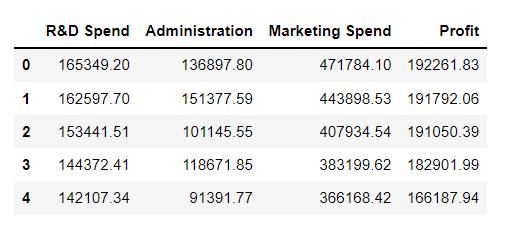
### In the given dataset, R&D Spend, Administration Cost and Marketing Spend of 50 Companies are given along with the profit earned. The target is to prepare an ML model which can predict the profit value of a company if the value of its R&D Spend, Administration Cost and Marketing Spend are given.

### Given dataset is about the 50 startups, each of these companies have recorded their R&D spent, Administration Cost, Marketing and their overall profits earned.

### Tasks to be done

* **Construct Different Regression algorithms**
* **Divide the data into train set and test set**
* **Calculate different regression metrics**
* **Choose the best model**

**The above problem statement, on applying various regression techniques to the dataset 50 companies and select the best model out of the applied models.**

****

**Approach on the problem**

**The above problem can be solved in various possible methods by using the regression algorithms. Regression algorithms used for the above problem can be given as Linear, Ridge, Lasso, Decision Tree, Random Forest, Support Vector Regression. Out of these models applied, the model with better results can be the best suited for the dataset.**

**The above Dataset contains four features**

* **R&D Spend** - Cost spent on Research and Development
* **Administration Cost** - Cost spent on support of business, indirectly related to specific product or service
* **Marketing Spend** - Cost spent on the marketing of products or services provided by the companies
* **Profit** - Overall profit earned by the companies

**Regression techniques can only be applied for the continuous data.**

**Packages in Python**

Packages in python used for machine learning are pandas, numpy, scipy and sklearn. Pandas and numpy packages are used for creating datasets or query processing on the datasets just like sql commands. Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

**Installing Pandas:**

**pip install pandas**

**Installing Numpy:**

**pip install numpy**

**Installing Scikit learn:**

**pip install scikit-learn**

**In­-depth Intuition and Practical Implementation**

**Data description and Preprocessing**

Data Description is the process of looking at the data. Description analysis gives a picture of the kind of data set we are going to handle. Description analysis can be done for checking the records in the data, the values in each feature and some statistical analysis. Preprocessing involves scaling the continuous values in a smaller scale ranges or assigning binary classifier for the categorical data by using One hot encoder. Preprocessing also helps in increasing the accuracy of the results.

**Here are some methods used in the project**

**head() –** The head method is useful in accessing the first five data records. It takes a positive integer to access the records.

**Info() –** info method is used for providing the information of the data set.

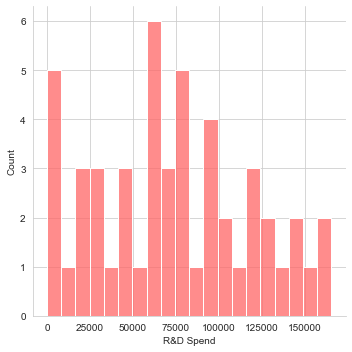
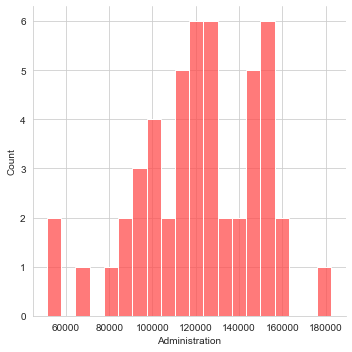
**describe() –** describe method in used for quick analysis of all the statistical methods like mean, median, variance and standard deviation etc.

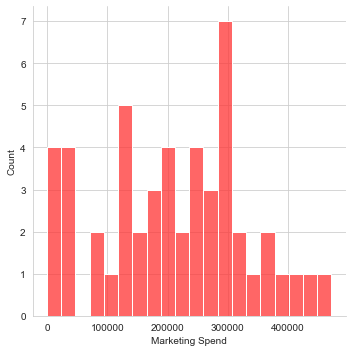
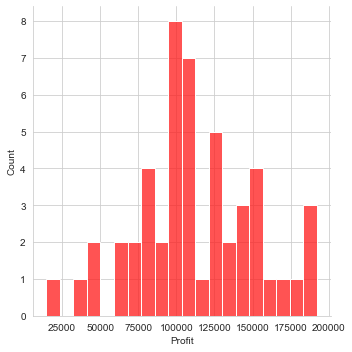
**isnull() –** this method is used to check for any null values in the cells.

**StandardScaler() –** The standard scaler uses the z score formula for scaling the data.

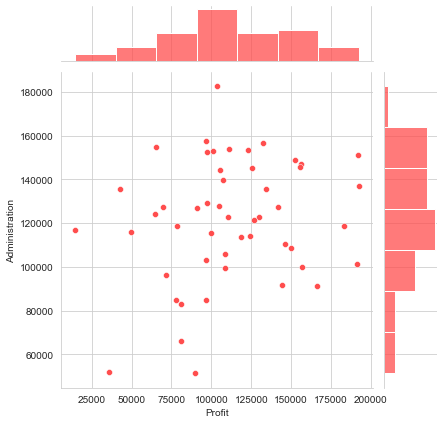
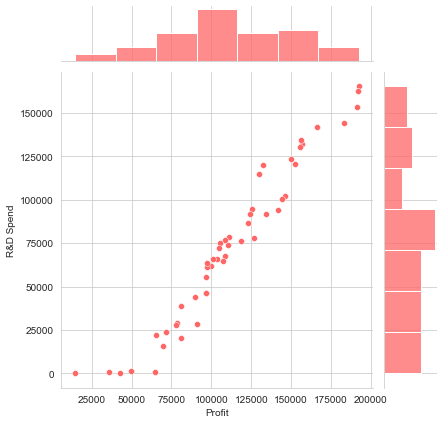
**Data visualization Techniques**

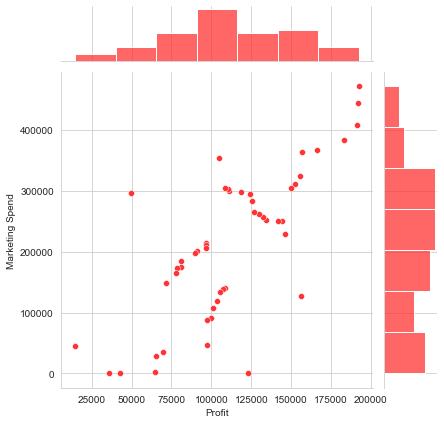
Visualizing the data set is a very important part. This process involves the analysis of trends between each feature of the data. As the data set has continuous values we can use the distribution plots and joint plot for analysing the trend for each feature. Heat map is used for correlation analysis between each feature. **These are some Distribution plots:**

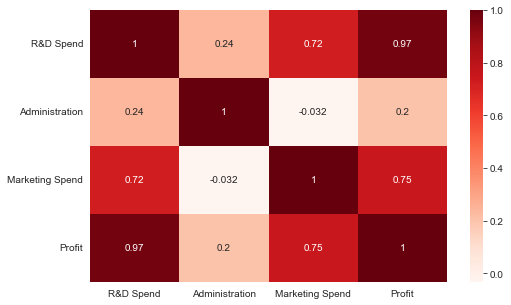
 

**These are some joint plots:**



****

**Heatmap:**

****

**From the above heat map, inferences can be taken as-**

**The above Heatmap tells us that there is a strong correlation between Profit and R&D spend, Profit and Marketing Spend. Hence we can consider that Spending in Marketing or in R&D can help in Increase of Profit.**

**Linear Regression**

Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression*.*

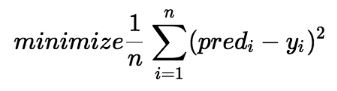
If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.

**y = β0 + β1.x1 + β2.x2 +….. + βn.xn**

Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing.

The loss function can be calculated by using mean squared error or root mean squared error.

**Mean squared error formula:**



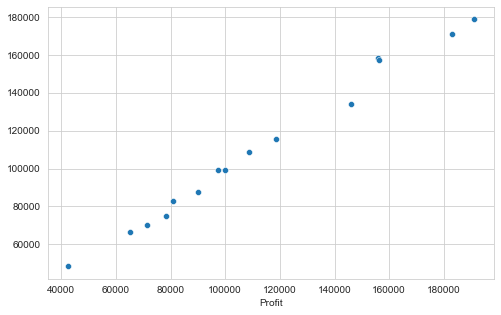
If the MSE value is nearer to zero then the model is good.

Import the linear model from sklearn.

**from sklearn.linear\_model import LinearRegression**

Splitting the data into Train and test using the train test split method from sklearn, with Test size of 30%. Random state set at 10. Fit the data into the linear model and predict the values with test split.

Plotting the graph of predictions and test sample. The **accuracy score** we get is around **97%**.



**Ridge Regression**

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

**Formula = cost function + λ∑(coefficient)2**

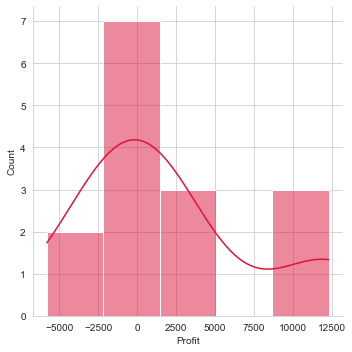
* It shrinks the parameters. Therefore, it is used to prevent multicollinearity
* It reduces the model complexity by coefficient shrinkage

We use **Grid search CV** as the model selection method, with ridge as the estimator and some random parameters for lambda values.

parameters = {'alpha': [1e-20, 1e-15, 1e-10, 1e-5, 1e-0, 0, 5, 10, 15, 20, 40, 60, 70, 80, 90, 95, 100, 101]}

Fit the data into the linear model and predict the values with test split. With cross validation of 30 samples each.

Plotting the graph of predictions and test sample. The **accuracy score** we get is around **98%**.



**Lasso Regression**

**Lasso regression** is a type of **linear regression**that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of [muticollinearity](https://www.statisticshowto.com/multicollinearity/) or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; Some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models.

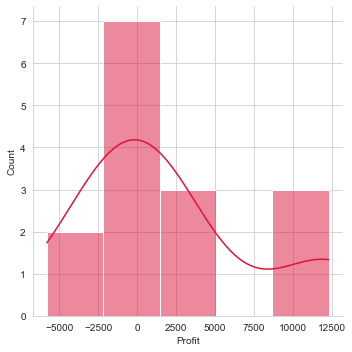
**Formula = cost function + λ∑|coefficient|**

We use **Grid search CV** as the model selection method, with ridge as the estimator and some random parameters for lambda values.

parameters = {'alpha': [1e-20, 1e-15, 1e-10, 1e-5, 1e-0, 0, 5, 10, 15, 20, 40, 60, 70, 80, 90, 95, 100, 101]}

Fit the data into the linear model and predict the values with test split. With cross validation of 30 samples each.

Plotting the graph of predictions and test sample. The **accuracy score** we get is around **97.5%**.



**Decision Tree Regression**

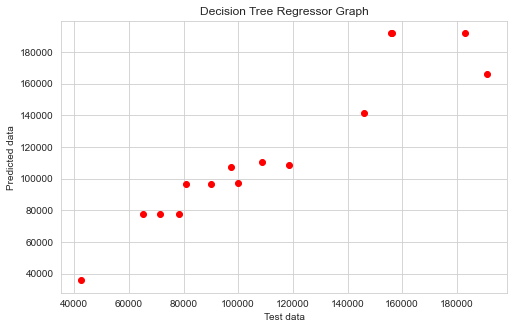
Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.

From the sklearn package import the decision tree regressor model. And passing in the max depth argument creates a tree with most possible depth, maximum the depth higher the accuracy. But the decision tree regression has overfitting problem which can be solved using the Random Forest Regression.

Splitting the data into Train and test using the train test split method from sklearn, with Test size of 30%. Random state set at 10. Fit the data into the linear model and predict the values with test split.

Plotting the graph of predictions and test sample. The **accuracy score** we get is around **92.1%**.



**Random Forest Regression**

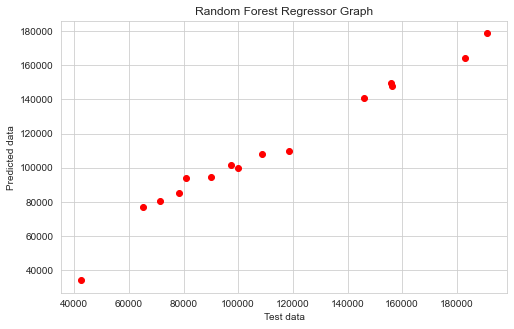
A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max\_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

Estimator argument is passed to the Random Forest Regression saying number of trees to be created. Out of which the best tree will be considered. Hence, Random Forest Regression comes under the Ensemble Technique.

Higher the Estimator value the better the accuracy results of the model will be.

Splitting the data into Train and test using the train test split method from sklearn, with Test size of 30%. Random state set at 10. Fit the data into the linear model and predict the values with test split.

Plotting the graph of predictions and test sample. The **accuracy score** we get is around **95.5%**.

****

**Support Vector Regression**

Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points. For large datasets, Linear SVR or SGD Regressor is used. Linear SVR provides a faster implementation than SVR but only considers the linear kernel. The model produced by Support Vector Regression depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to their target.

**SVM Kernal**

A kernel is a set of mathematical functions that takes data as input and transform it into the required form. These are generally used for finding a hyperplane in the higher dimensional space. The most widely used kernels include Linear, Non-Linear, Polynomial, Radial Basis Function (RBF) and Sigmoid. By default, RBF is used as the kernel. Each of these kernels are used depending on the dataset.

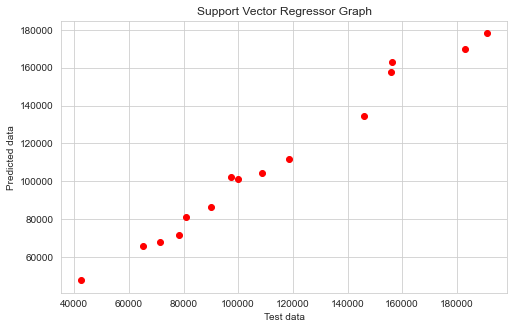
Plotting the graph of predictions and test sample. The **accuracy score** we get is around **97.5%**.

**Disadvantages of SVR**

* They are not suitable for large datasets.
* In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.
* The Decision model does not perform very well when the data set has more noise i.e. target classes are overlapping..

**Advantages of SVR**

* It is robust to outliers.
* Decision model can be easily updated.
* It has excellent generalization capability, with high prediction accuracy.
* Its implementation is easy.



**Conclusion**

Accuracy for **Linear Regression**: 97.80592308313783

Accuracy for **Ridge Regression**: 98.14913121597118

Accuracy for **Lasso Regression**: 97.8059251510803

Accuracy for **Decision Tree Regressor**: 92.94476998319507

Accuracy for **Random Forest Regressor**: 95.56356750340365

Accuracy for **SVR with Linear Kernel**: 97.50146052080812

**The above accuracy test tells us about how well the models have performed with the 50 start-up data set. The data set only has 50 records as a result some models might have failed to deliver better results like Decision Tree Regression which performed very poorly, considering Random Forest Regression has performed much better than the Decision Tree model. Looking at the other models Linear, Ridge, Lasso, SVR have almost similar results with small variations in accuracy.**

**We can say that Ridge Regression has performed well in this scenario. But we can also conclude that it depends on the model that we select on which data set the model is going to perform well.**