University of Essex

School of Computer Science and Electronic Engineering

CE802 Machine Learning

Assignment: Design and Application of a Machine Learning System for a Practical Problem

Classification Comparative Study

BY: - RONAK PUNJABHAI KARMUR

STUDENT REGISTRATION NUMBER: -2202285

MSC, ARTIFICIAL INTELLIGENCE

Professor's name: Dr. Vito De Feo

word count classification: - 882

word count regression: - 650

Total Word Count: - 1595

ABSTRACT

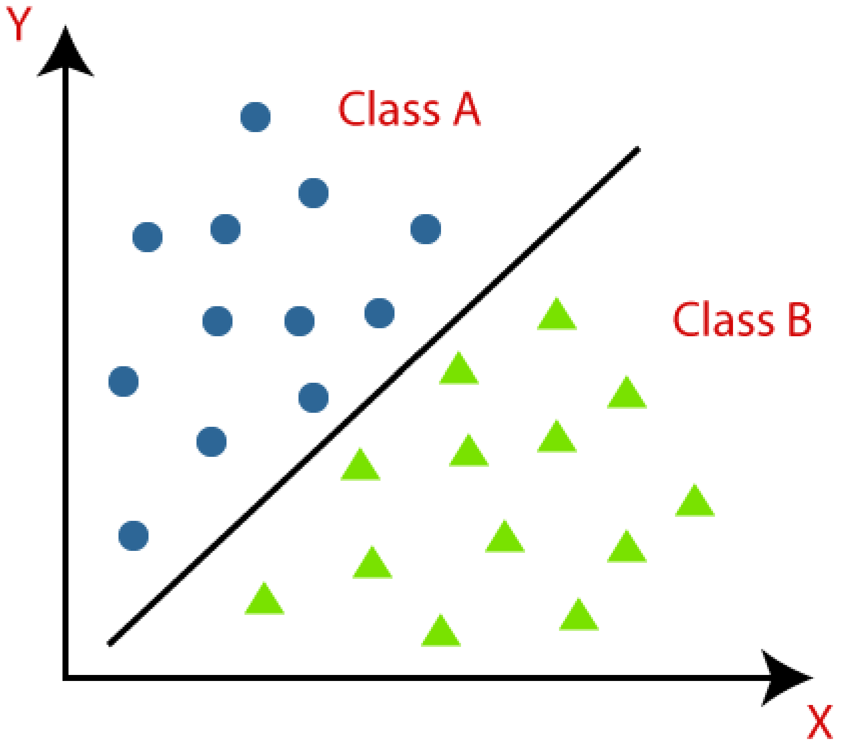
The training dataset consists of historical customer data containing each customer's features and representing labels of them either a binary class output (True/False or Yes/No) for the classification problem or target output (integer/float) for the regression. And the methods used to solve each solution to get the results to depend on each of the different datasets.

INTRODUCTION

- **Classification**

The project aims to identify the customers who can pay the electricity bill for the AENERGY aka TRUE in the class variable and the customers who cannot pay the electricity bill for the AENERGY aka FALSE in the class variable. This is a classifier problem that can be solved from the other features which have been recorded in the dataset by previous customers' historical data. The hypothesis is that Machine Learning can conveniently solve this issue if the algorithm is trained in the correct manner.

The below figure represents how the classes are in a classification problem.



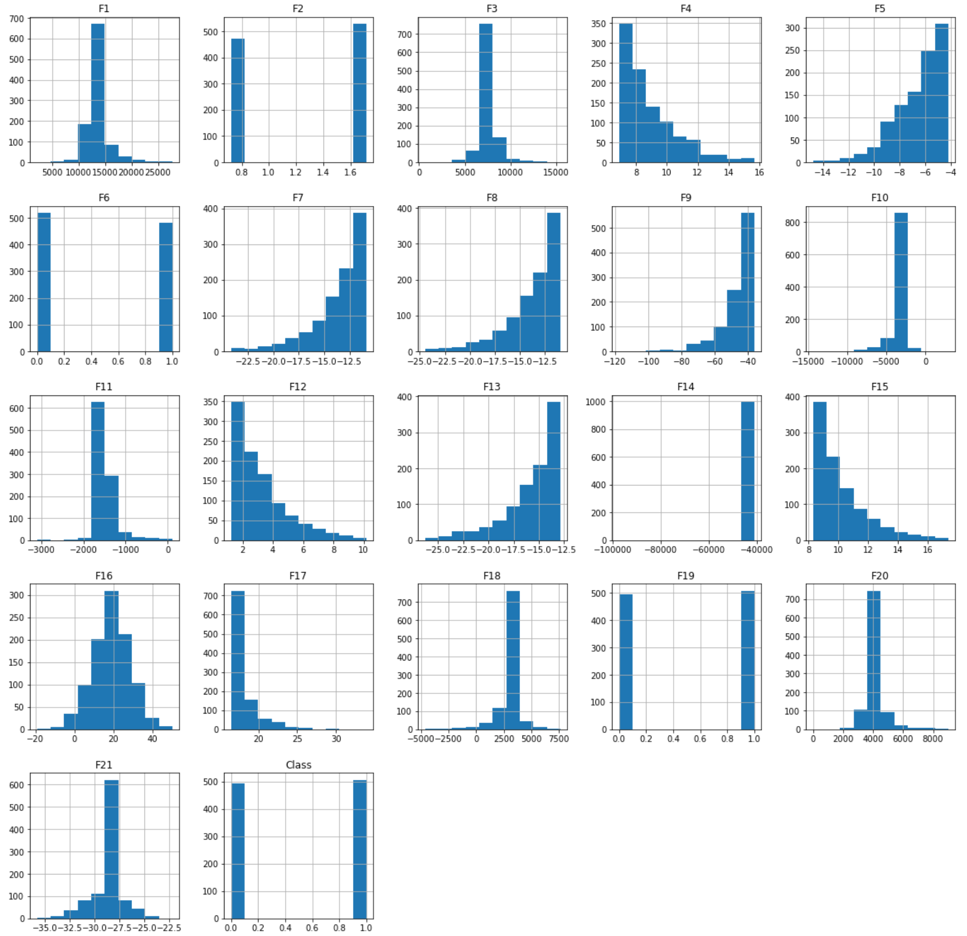
METHODS

A few of the methods used to clean the data (EDA) Exploratory data analysis to pre-processing the data to fit into the algorithm are:

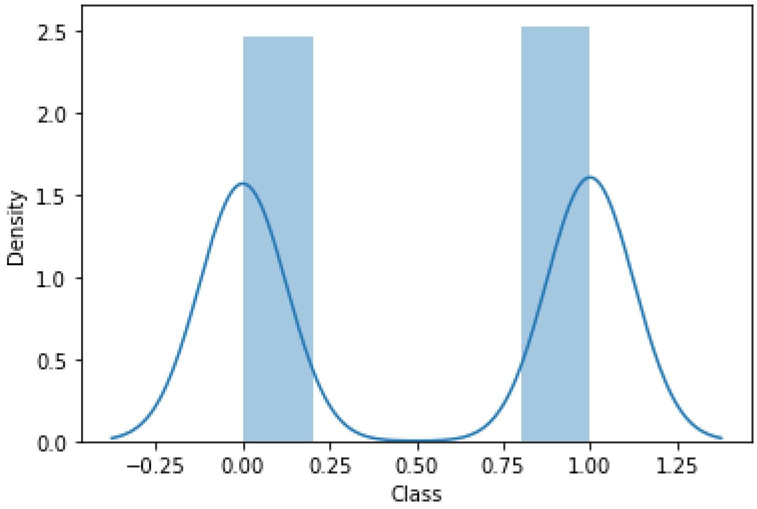
- Null Values: - Null values were dealt with by filling with mean values. EXPERIMENTAL methods used were filled with 0, removing the null indexes, etc. which resulted in lesser accuracy
- Duplicates: - The dataset did not contain any duplicates to deal with.

- Data Types: - The predictor variable which was a Boolean variable is the only variable that needed to be converted to a numerical variable to fit into the model. This was tackled by using Sklearn. preprocessing Label-Encoder which converts the True to 1 and False to 0
- Outlier: - Few of the variables containing outliers were dealt with but the accuracy dropped due to this, and from this information, we got to know that the so-called outlier is a feature of unique values for that customer who was different from the average customers.

- Skewness: - Few variables were left-skewed, few were right-skewed, and the rest were in a normal distribution. This was tackled by square root and log transform the data. But by a normal distribution, the accuracy was way worse than the original data.

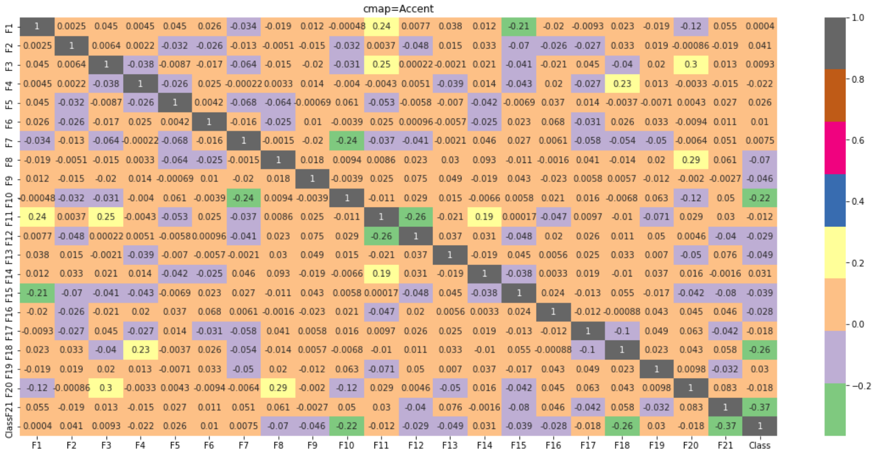


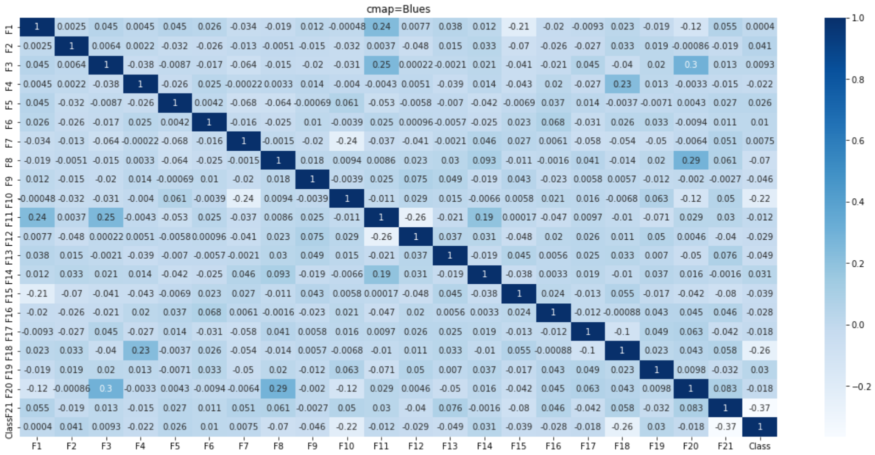
- Imbalance Data: - False and True data distribution were equally matched, so there is no need to apply up-sampling or down-sampling to fix this issue.



- Normalization: - Normalization gets a better accuracy, but whatever we predict in normalized value should be inversed. And hence no point in normalization. But we can normalize the dependent variable but the accuracy was dropped from the original value.

- Co-relation: - selecting only highly correlated attributes only negates the accuracy of the model for this dataset



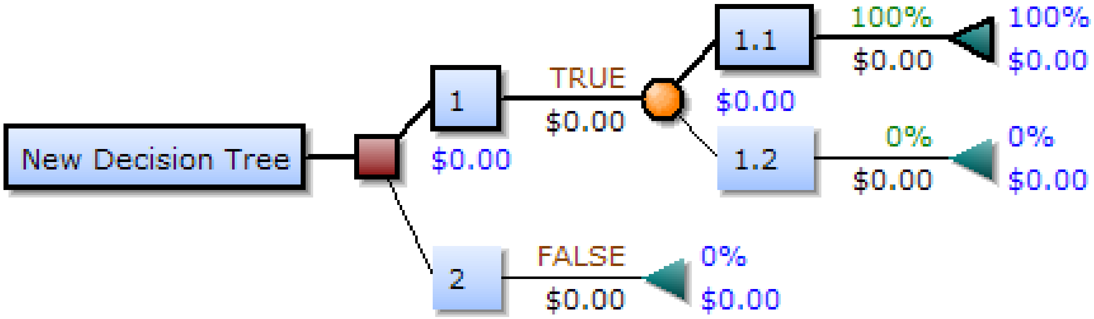


- Data Split: - Data split was performed on training data and the data was split into Training 75% and Testing 25% because to check the accuracy of the training model on testing when predicted. The Method used here to split the data was sklearn. model_selection.

ALGORITHM USED

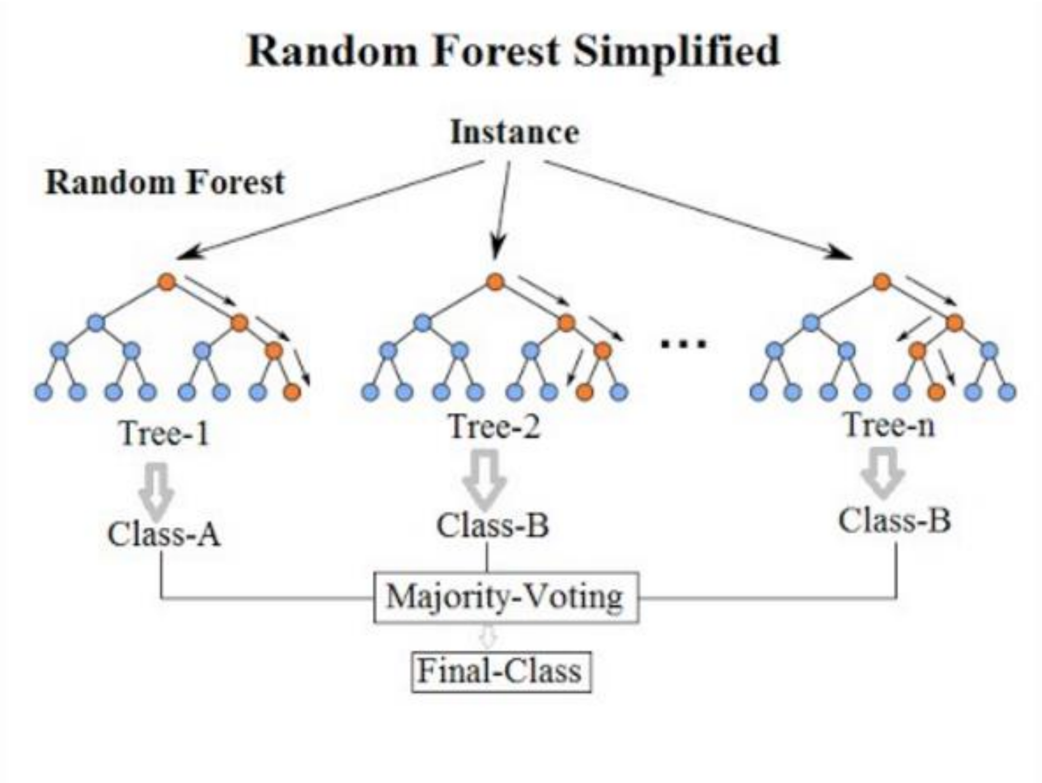
1. **Decision Tree**: **-** Decision Tree is a flow chart algorithm that is used to internal node representation attribute which can determine whether the Decision Tree consists of the parent node, child node, or branches which are basically different features, variables, values of each customer which eventually leads to the output aka end if the node either True or False.

The Below Figure represents the basic decision tree structure

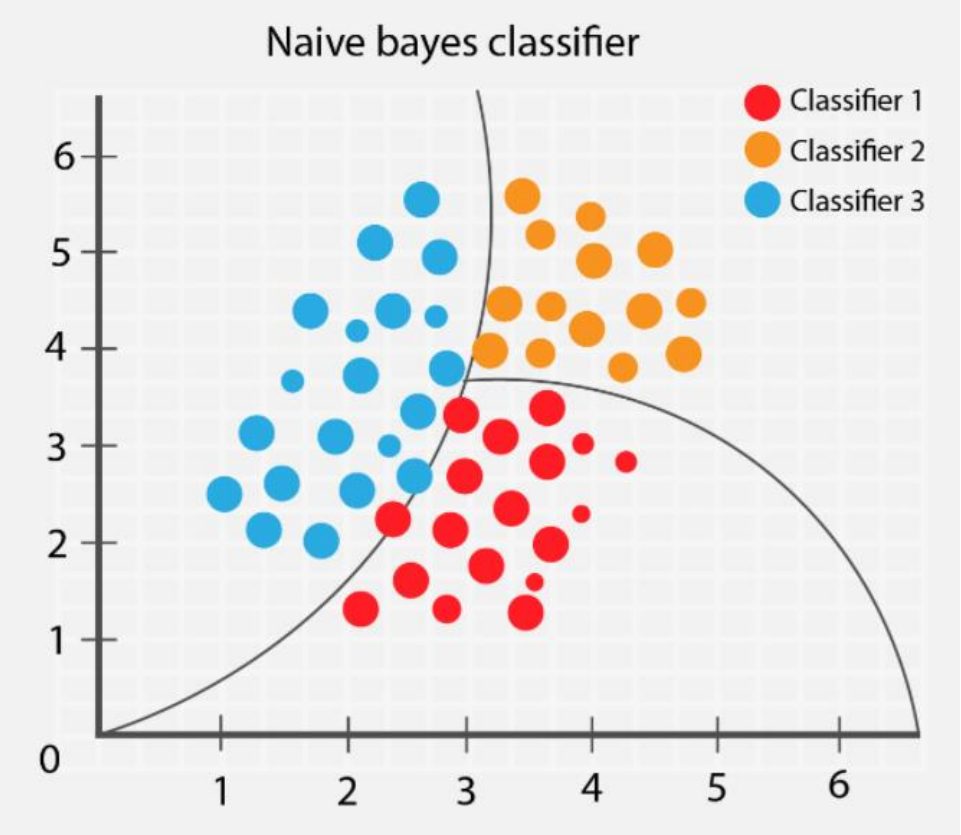


2. Random Forest: - Random Forest is one of the few algorithms which can be both used for classification and regression, here Random Forest works similarly to Decision Tree, but RF is comprised of multiple decision tree which takes or predicts the output based on a voting system. Whichever, the output gets the best voting will be selected and predicted.

The Below Diagram Represents Random Forest.



3. Naïve Bayes: - Naïve Bayes Classifier is a simple probabilistic classifying algorithm, it works on independent assumptions which use the Bayes theorem and Bayes network to predict the classification problem and solve it. It is based on prior multiplied by the likelihood of the evidence to get the maximum likelihood as a predictor.



Results

- **Decision Tree: -** The accuracy of the Decision tree algorithm model is 79.20%
    - The Classification Report for the model
        - Precision: - 0.80
        - Recall: - 0.78
        - F1-Score: - 0.79
- **Random Forest: -** The accuracy of the Random Forest algorithm model is 84.40%
    - The Classification Report for the model
        - Precision: - 0.83
        - Recall: - 0.86
        - F1-Score: - 0.85
- **Naïve Bayes: -** The accuracy of the Naïve Bayes algorithm model is 66.80%
    - The Classification Report for the model
        - Precision: - 0.66
        - Recall: - 0.69
        - F1-Score: - 0.67

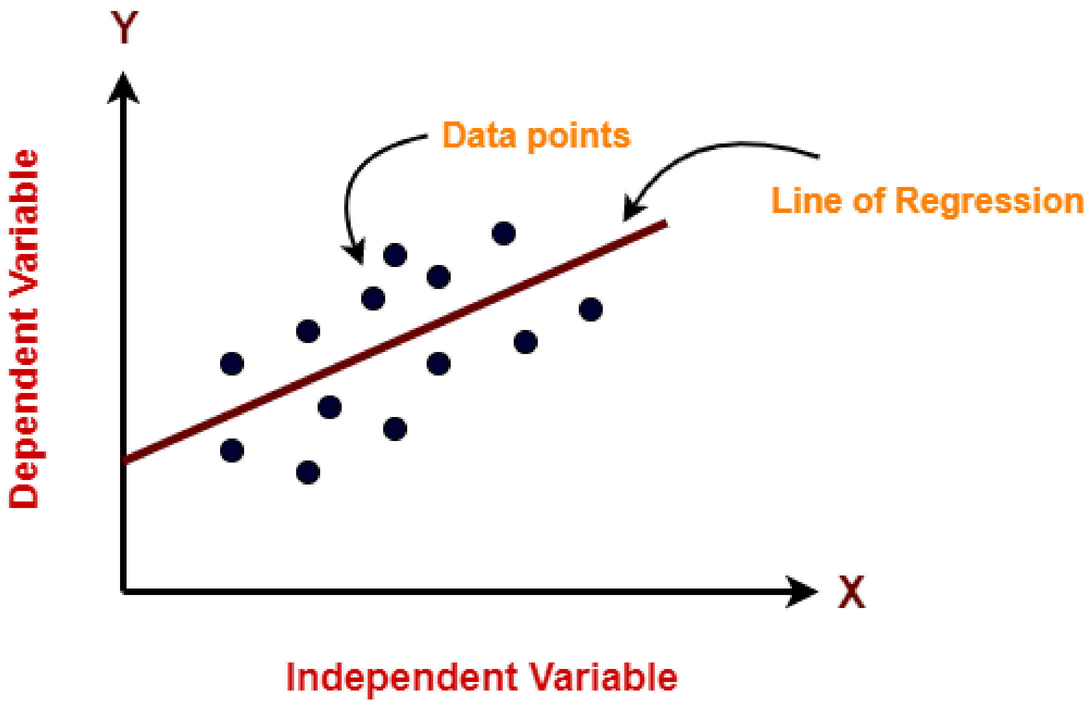
Conclusion

From the Classification Metrics that were performed on the Test set prediction, which was a split of the training dataset, we can clearly see that Random Forest had the best accuracy and confusion matrix metrics. However, other algorithms might give a better result for different kinds of algorithms and features. Random Forest was the best-performing model for this dataset to predict the customers' class labels.

- **Regression**

The project aims to predict the variation in the annual expenditure of each customer, and we are predicting this due to the increase in the price of energy cost. An ENERGY has the prehistoric customer's behavior data which can be used to predict the price of customers based on the attribute and the predictor are positive if the customer is going to spend more and if the customer is going to spend less than the predictor will be negative.

Regression algorithms are based on predictions like this which can be used to solve this problem. The below figure represents how the regression algorithm works.



METHODS

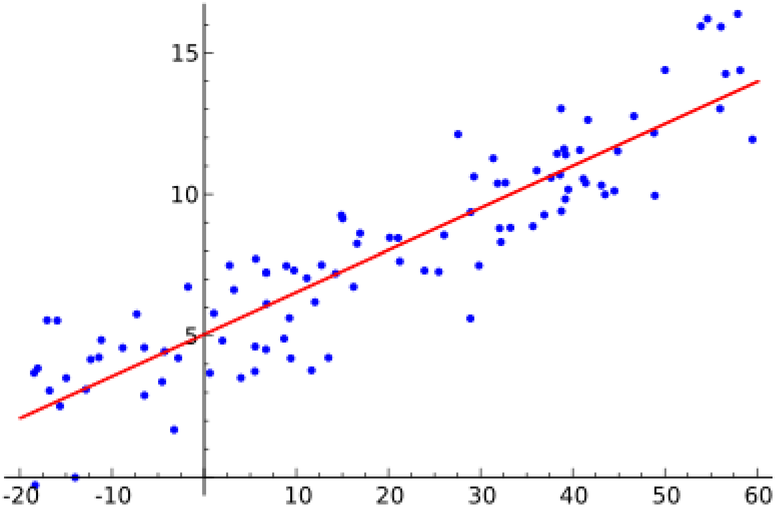
A few of the methods used to clean the data (EDA) Exploratory data analysis to pre-processing the data to fit into the algorithm are:

- Null Values: - The dataset did not contain any null values
- Duplicates: - the dataset did not contain any duplicates values
- Co-linearity: - The dataset had many features which were collinear to each other, Had to remove them and check the MSE score but the co-linearity was not affecting the dataset and the MSE score was better to keep them instead of removing them.
- Co-Relation: - Keeping only highly positive and highly negative correlated features only Increased the MSE Score which was not the best model to build
- Skewness: - Few attributes were really skewed to left and right which need to be normally distributed and by doing this the MSE dropped, and the model improved for better prediction
- Outlier: - There were many outliers same as the classification dataset but removing them is removing the unique customer data which resulted in a bad MSE score and a bad model.
- Splitting the data: - Data split was performed on training data and the data was split into Training 75% and Testing 25% because to check the accuracy of the training model on testing when predicted. The Method used here to split the data was sklearn.model_selection.

ALGORITHM USED

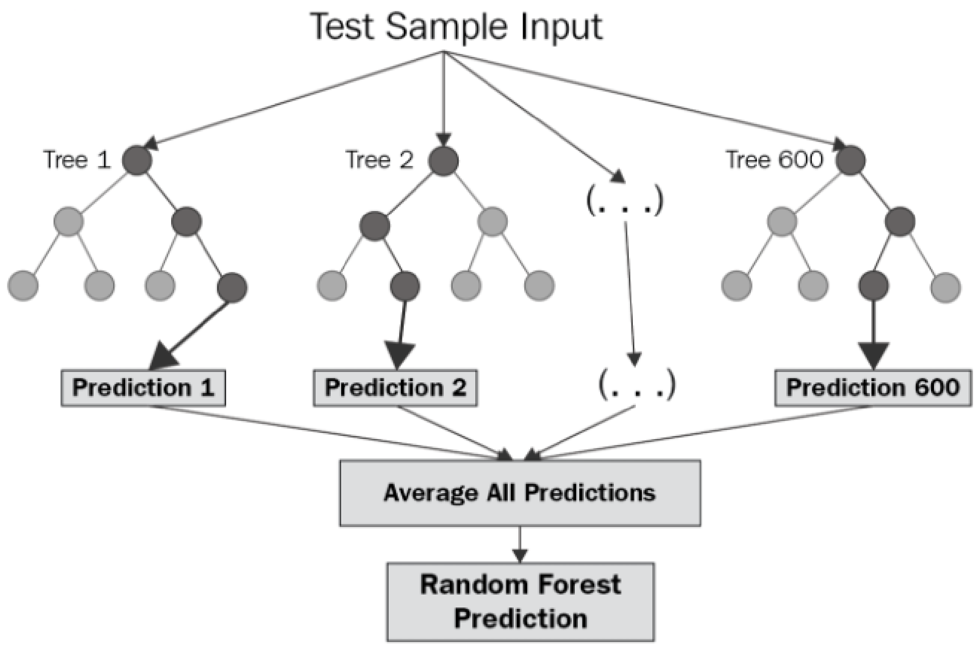
1. LINEAR REGRESSION: - Linear Regression is an algorithm that takes the relationship between the independent variable aka (predictand training features) and the dependent variable aka (predictor feature). Linear regression uses linear predictor functions that estimate the data based on unknown parameters by taking correlation of the variables/parameters.

The Below diagram represents the linear regression after prediction.



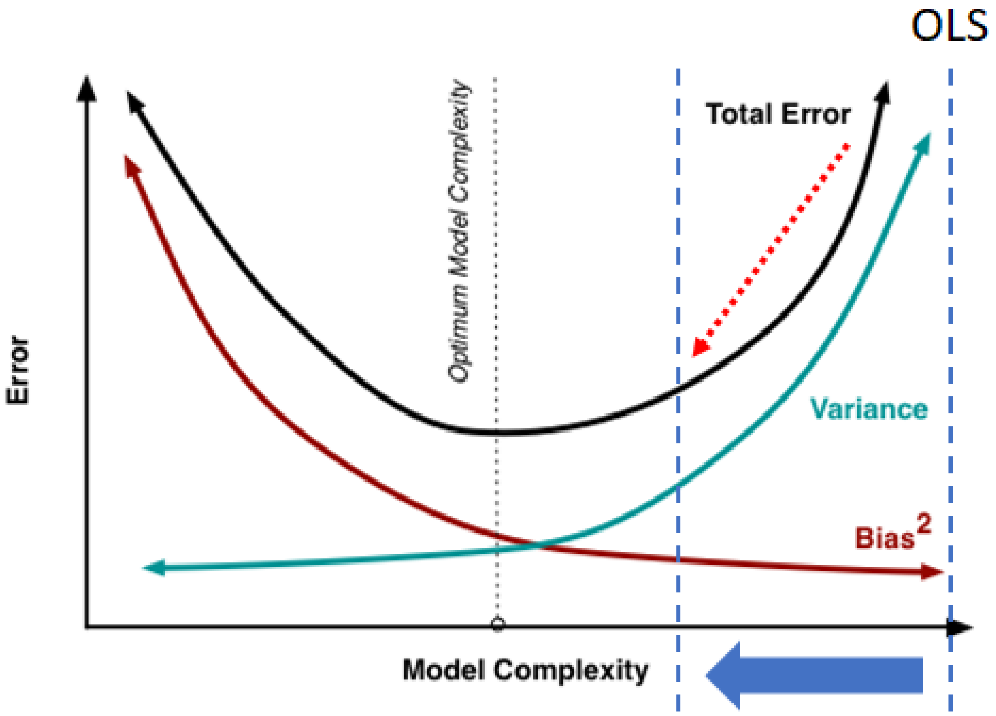
2. Random Forest Regression: - Random Forest Regressor works by creating multiple decision trees and then getting the predicted output. Classification Random Forest works by taking the voting count but for regression, the output of each decision tree and then all the output average is taken as the prediction output.

The below figure represents Random Forest Regression



3. Ridge Linear Regression: - Ridge Linear Regression works the same as multiple Linear Regression but when the dataset has multiple collinearities. Ridge regression creates parsimonious models when the dependent variable is less than multi-collinearity variables.

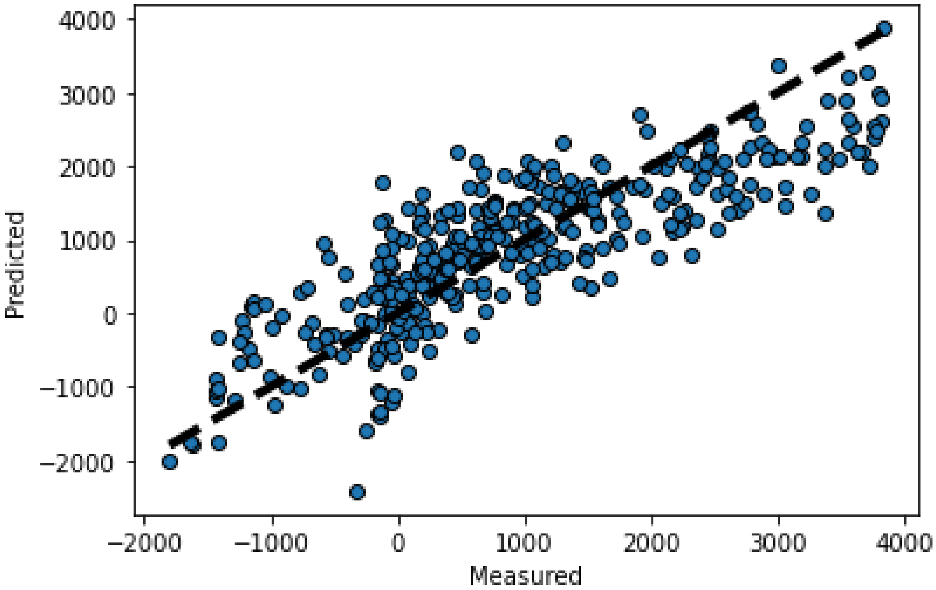
The below figure explains how Ridge Linear Regression basically is



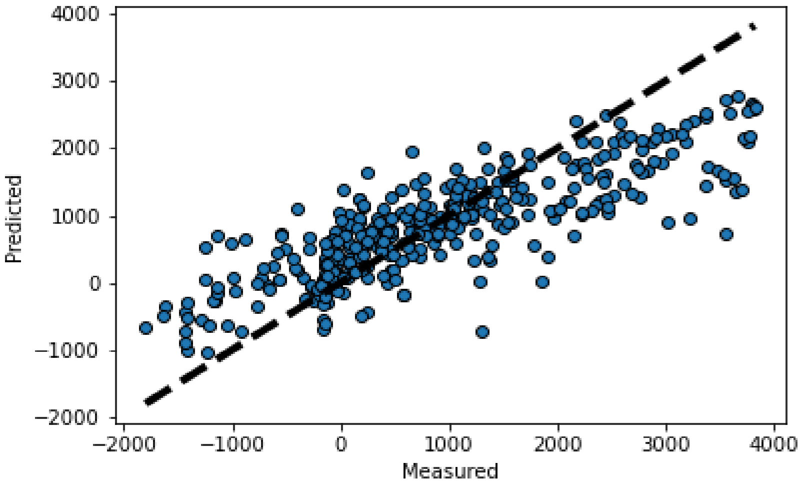
Results

- ## LINEAR REGRESSION: -

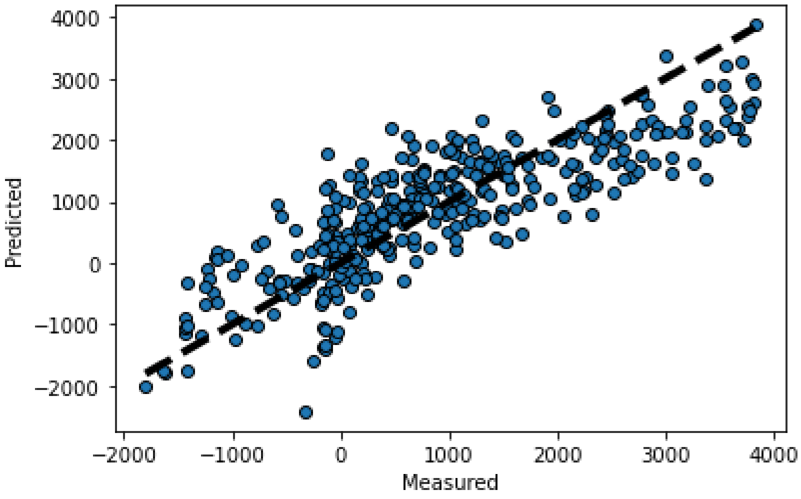
  - **Mean Squared Error: - 718.1409**
  - **Root Mean Squared Error: - 515726.4077**
  - **R2: - 0.7046**



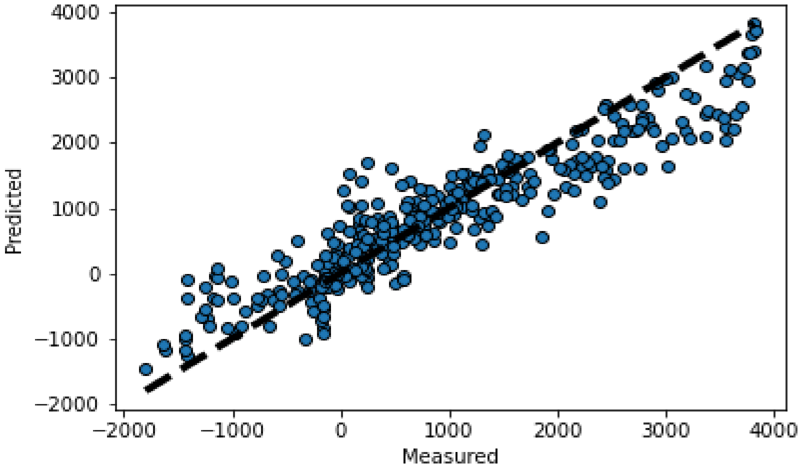
- **RANDOM FOREST REGRESSION: -**
  - **Mean Squared Error: - 764.74981**
  - **Root Mean Squared Error: - 584842.2806**
  - **R2: - 0.86670**



- **RIDGE LINEAR REGRESSION: -**
  - **Mean Squared Error: - 718.3212**
  - **Root Mean Squared Error: - 515985.4392**
  - **R2: - 0.7046**



- **LIGHTGBM: -**
  - **Mean Squared Error: - 259482.2592**
  - **Root Mean Squared Error: - 509.3940117828897**
  - **R2: - 0.7046**



CONCLUSION

From the results of the different models, we can see that Light Algorithm is performing way better than any other algorithm for this dataset and predicting the dependent variable with less over error ratio.

REFERENCES

1. **AnalysticVidya**
2. **MachineLearningMastery**
3. **GeeksforGeeks**
4. **Kaggle**
5. **Staticsticshowto**