Project Report: (Phase - 2)

Department of Computer Science and Engineering

University at Buffalo, Buffalo, NY 14260

Team: Desireddy Sai Sankeerthana — saisanke@buffalo.edu

Vamshi Jamalpur — vamshija@buffalo.edu

Algorithms :

The Algorithms we have used that are relevant to our data as follows :

1.Linear Regression

2.Random Forest

3.KNN(K-Nearest Neighbours)

4.LASSO(Least Absolute Shrinkage and Selection Operator)

5.Ridge Regrssion

6.DecisionTreeRegressor

Evaluation Metrics:

Evaluation metrics is used to measure the quality of machine learning model.It explains the performance of the model.It also can be used to compare with different alogrithms to find difference in the performance.

The evaluation metrics used in our project to measure performance are :

1. Root Mean Squared Error (RMSE)
2. Mean Absolute Error (MAE)
3. R2_Score(Coefficient of determination)

1. Root Mean Squared Error (RMSE) is value to find or to measure difference between actual value and predicted value (quality of prediction)

2. Mean Absolute Error (MAE) is absolute value to measure error between actual value and predicted value

3. R2_Score is a measure which tells variation of dependent variable that is explained using independent variable

Cross validation: To avoid over fitting and under fitting of the model we have used the cross validation

Details of the above algorithms as follows

1.Linear Regression Algorithm:

Explanation:

Linear Regression algorithm is basically used to predict value of a dependent variable based on value of independent variable.

Linear Regression is basically used for its relatively simple and easy to interpret formula that is used for prediction of dependent variable

Formula for Linear Regression is y=a+bx

From the above equation x is a independent variable whereas y is a dependentvariable.

Reference:
https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Justification:

We have chosen Linear Regression for our model/dataset .Linear Regression is statistical procedure.As per the Linear regression algorithm, a dependent variable should be contionous in data and we have variables that as contionous data which is Item_Outlet_sales and that is our dependent variable as per our dataset.

Implmentation of Linear Regression for prediction of sales helps us to understand a specific patterns of sales on bases of paticular days or for some certain times.

As we have multiple Independent variables present in our data. Also it is one of the most used regression model .Linear Regression Algorithm has been our choice to try.

Applying Linear Regression to our dataset :

Dataset is split into train data and test data and applied Linear Regression

Visualization:

Comparing actual values with predicted values for train data

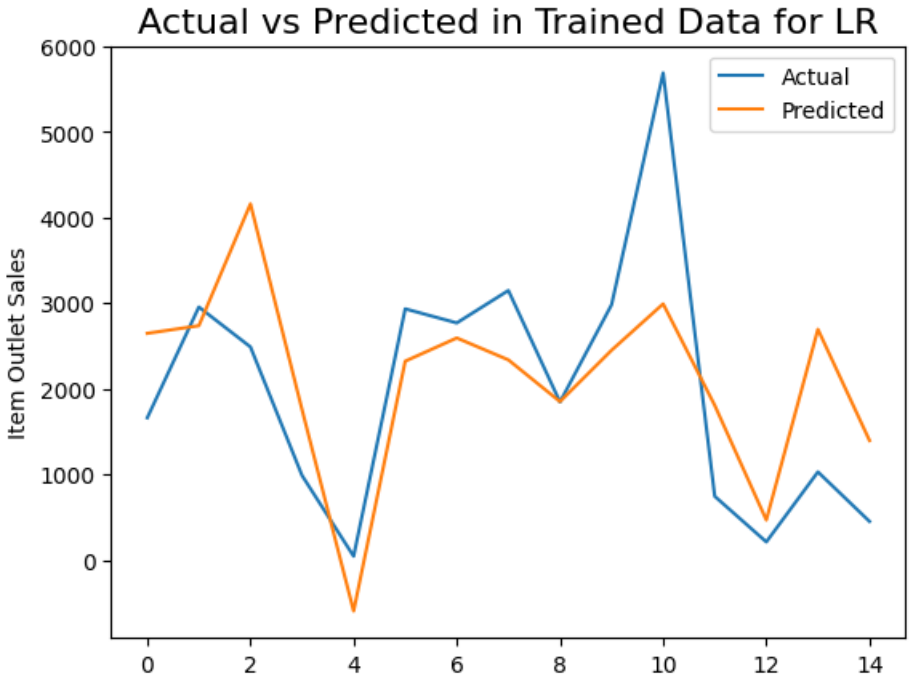

Figure1:Scatter plot for Train data

X_train and Y_train data has been fitted to Linear Regression model and fig 1 explains the predicted Y_train values and the actual Y_train values.From fig 1 we can observe the prediction of the values are closer to actual vales.

Comparing actual values with predicted values for test data

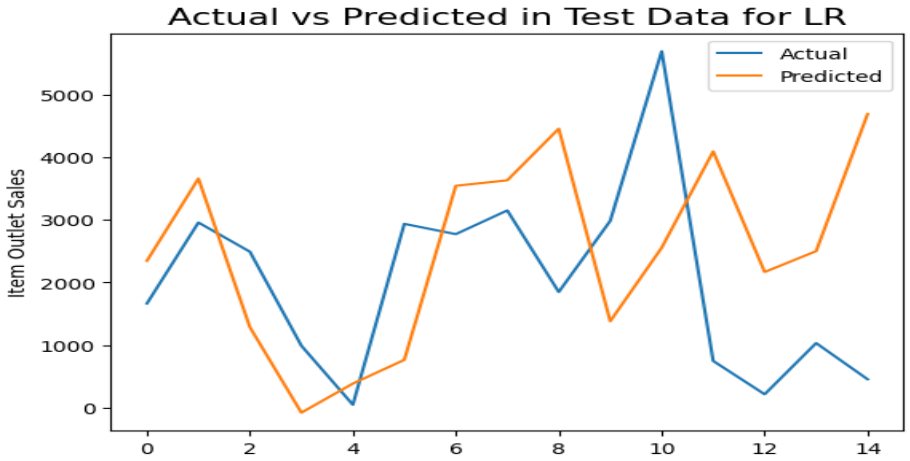

Figure2:Scatter plot for Test data

The model has learnt the data from X_train and fig 2 explains the predicted Y_test values and the actual Y_test values.From fig 2 we can observe the prediction of the test values to actual vales.

Residual Analysis for linear regression:

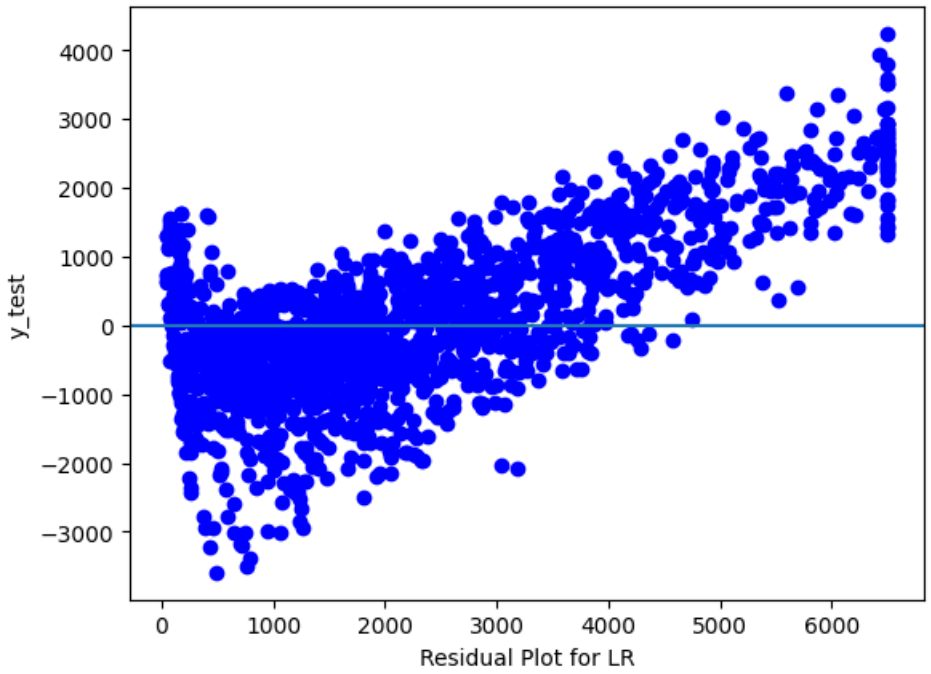

Figure3:Residual plot for Linear regression

We also used a bar graph to represent model coefficient based on Linear regression

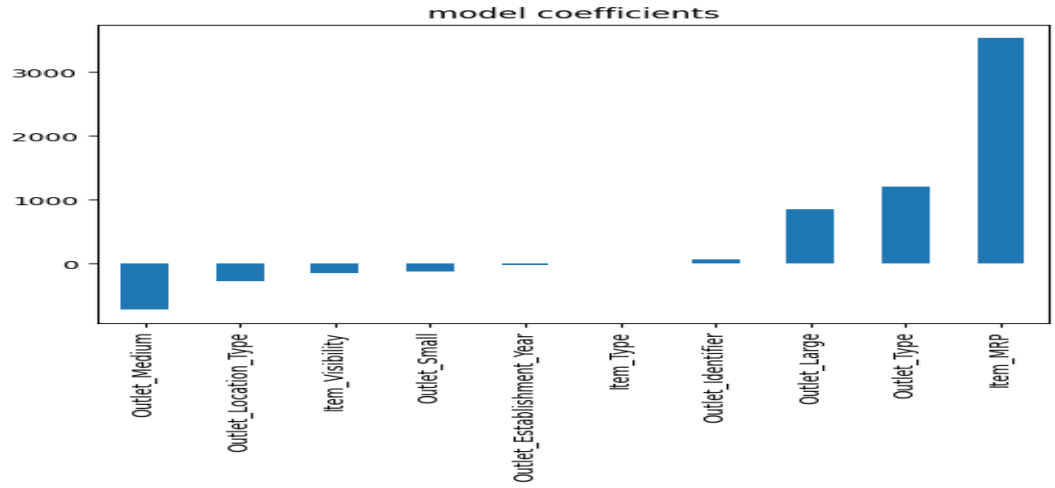

Figure 4: Bar graph for Linear Regression

Cross validation score (cv score) is value predicted between actual value and predicted value.

Evaluation metric for Linear Regression as follows:

Analysis:

For Train data values:

RMSE value: 1080.6561595041744
MAE value: 28.796335183169674
R2_Score value: 0.5514111848227321
CV Score: Mean - 1082 Standard deviation - 27.19 Min value - 1031 Max value – 1119
The RMSE Value is 1080.6561595041744
R2 score is 0.55 which means it can learn and predict the values and rsme value is 1080.

For Test data values:

RMSE value: 1140
MAE value: 29.53104437778563
R2_Score value: 0.5326526790394902

For the test data the r2_score is 0.53 we can see there is no lot change in the scores which tells we don't have any fitting problem. And we can see the rmse value is increased by 60 and the mean absolute error for the trained data is 28.7 and the test data has 29.5.

2. Random Forest Algorithm:

Explanation:

Random Forest Algorithm is basically defined as combining multiple decision trees into a single result. Random Forest Algorithm can be used for both classification and regression problems.

Random Forest primarily has three main parameters they are node size, number of tree's and number of features sampled.

Reference :

https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Justification :

We have chosen Random Forest Algorithm for our dataset because.it constructs several tree's and outputs the mean of all the tree's. By using the dataset model can learn and spilt data into different points and different number of tree's to provide mean of the results and to make prediction which helps to understand more about data and continuous values. It is a most powerful algorithm which improves the accuracy. It also evaluates importance of features in the model. It takes less time to train compared to other algorithms. It has good prediction

even with large set of data. It can also maintain accuracy when there is any data is missed.As we have multiple features in our dataset random forest might provide accurate results so we have used random forest algorithm.

Visualization:

Comparing actual values with predicted values for train data

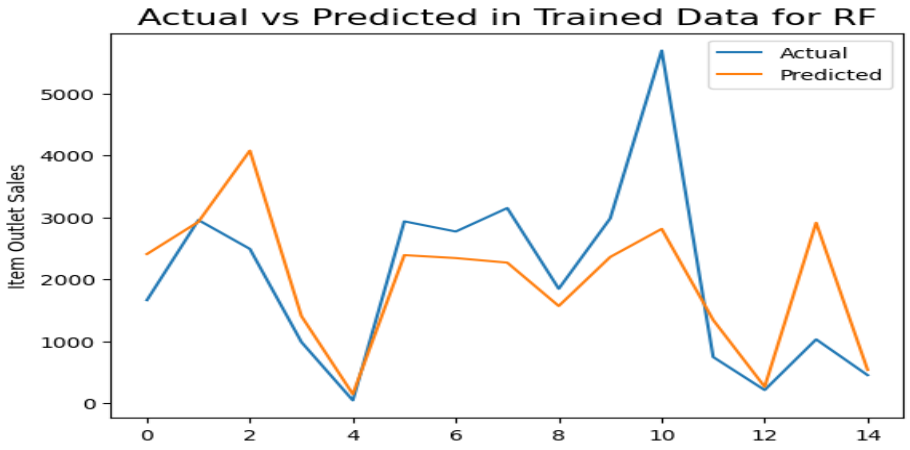

Figure 5: Scatter plot for Train data

X_train and Y_train data has been fitted to RandomForest model and fig 5 explains the predicted Y_train values and the actual Y_train values.From fig 5 we can observe the prediction of the train values to the actual vales.

Comparing actual values with predicted values for test data

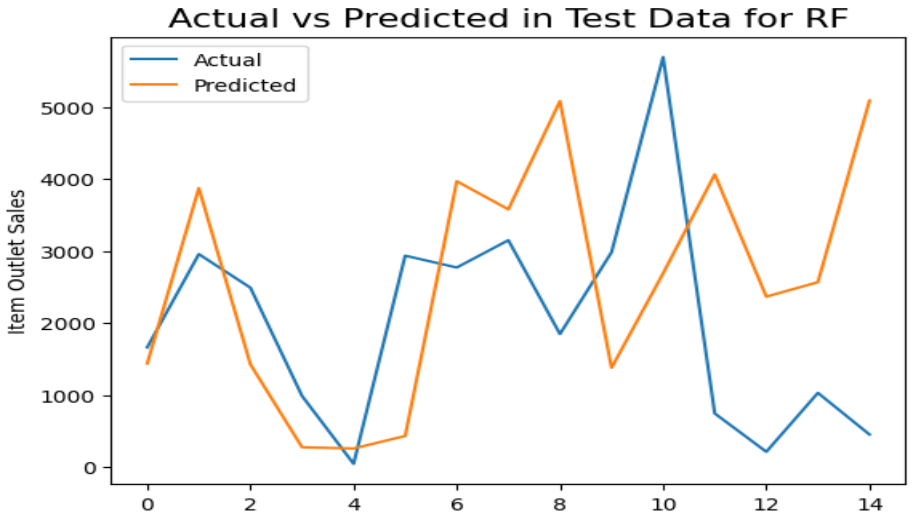

Figure 6: Scatter plot for Test data

The model has learnt the data from X_train and fig 6 explains the predicted Y_test values and the actual Y_test values.From fig 6 we can observe the prediction of the test values to actual vales.

Residual Analysis for Random Forest:

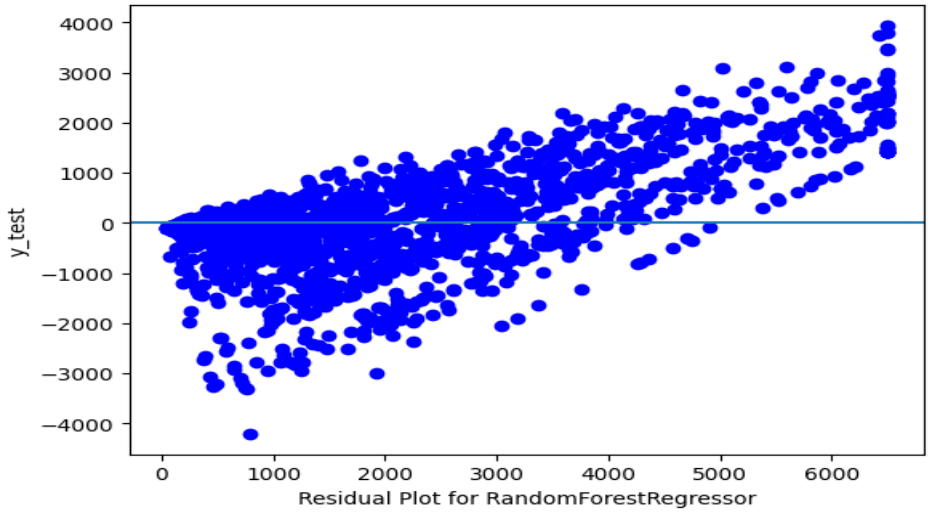

Figure 7: Residual plot for Random Forest

We also used a bar graph to represent feature importance for the data based on Random Forest

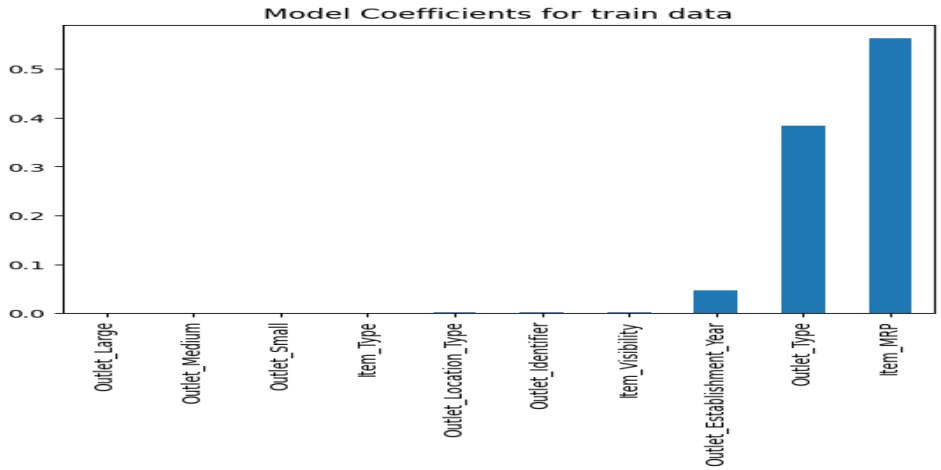

Figure 8: Bar plot for Random Forest

Evaluation metric for Random Forest as follows

Analysis:

For Train data:

RMSE value: 991.9218074715666
MAE value: 26.79167704440102
R2_Score value: 0.6220553071438422
CV Score: Mean – 1009 Standard deviation – 31.93 Min value – 953.7 Max value – 1062

For Test data values:

RMSE value: 1065
MAE value: 27.769585400290584
R2_Score value: 0.592547138563075

For the test data the r2_score is 0.59 we can see there is decrease in the r2_score but not a big difference. And we can see the rmse value is increased by more than 70 and the mean absolute error for the trained data is 26.7 and the test data has 27 which tells that there is a little error.

3. K-Nearest Neighbors Algorithm:

Explanation:

KNN main aim is to label a set of unlabled data based on k similar elements.KNN algorithm can be used on both classification and regression problems.

for Regression problems, a number with decimal points is given as output

to apply knn on real time data we need to determine how we measure closeness and determine value of k . if data is numerical we can determine closeness by caluculating distance such as euclidian etc.

Reference :

https://scikit-learn.org/stable/auto_examples/neighbors/plot_regression.html

Justification:

We have chose k-nearest neigbour algorithm to apply on our dataset because it is easy to implement and it chooses the output based on the nearest neighbours.Our data has alot of continous values by using this algorithm we can understand the range of sales and can make prediction on sales.

We have passed the average k-nearest neighbours (20) value to make a prediction

Visualization:

Comparing actual values with predicted values for train data

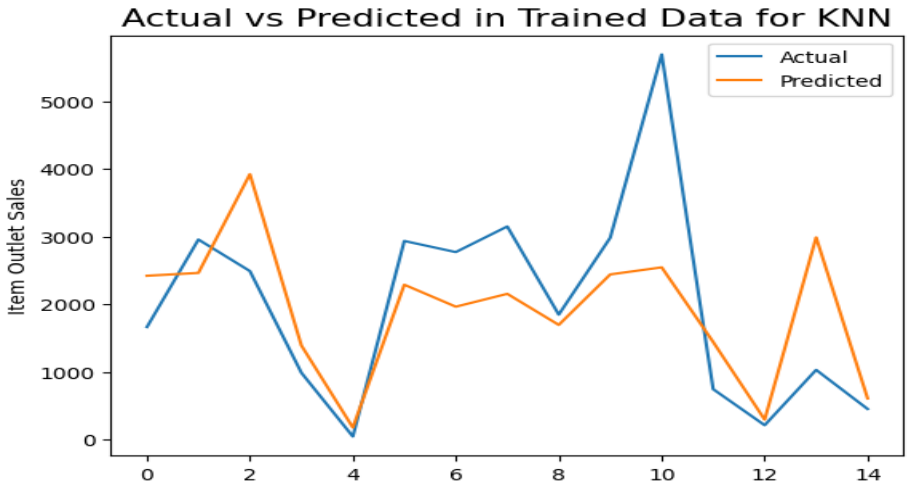

Figure 9: Scatter plot for Train data

X_train and Y_train data has been fitted to KNN model and fig 9 explains the predicted Y_train values and the actual Y_train values.From fig 9 we can observe the prediction of the train values to the actual vales.

Comparing actual values with predicted values for test data

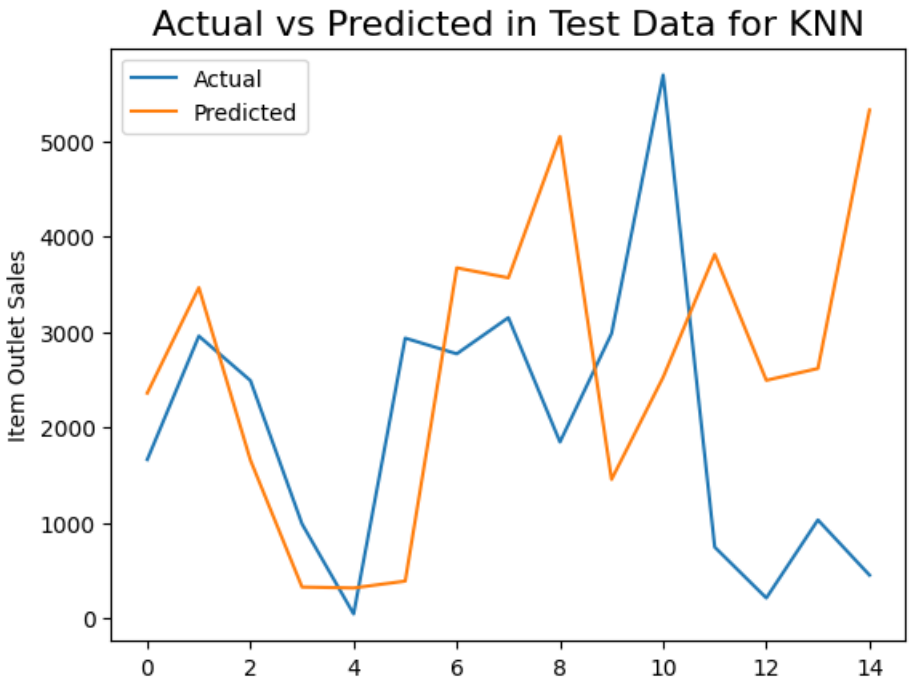

Figure 10: Scatter plot for test data

The model has learnt the data from X_train and fig 10 explains the predicted Y_test values and the actual Y_test values.From fig 10 we can observe the prediction of the test values to actual vales.

Residual Analysis for KNN:

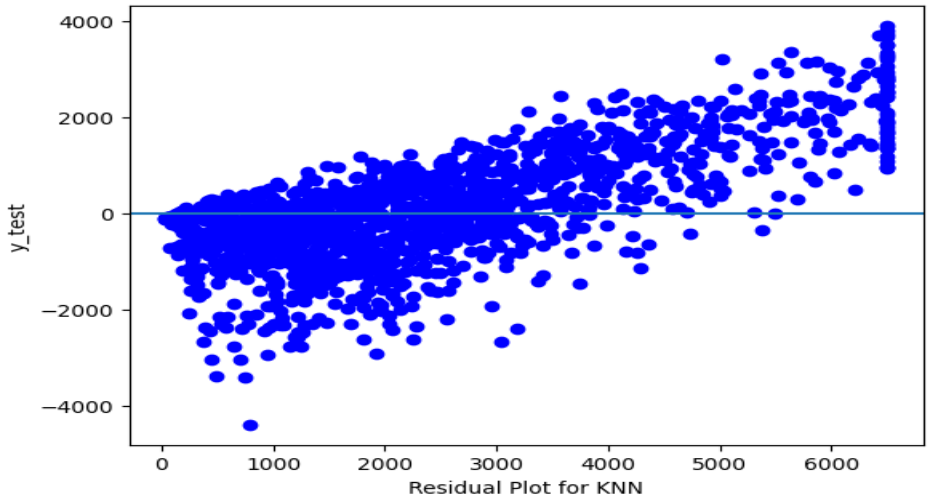

Figure 11: Residual plot for Knn

Evaluation metric for K-nearest neighbours as follows

Analysis:

For Train data:

RMSE value: 971.4962300956702
MAE value: 26.452336634136703
R2_Score value: 0.637460264227194
CV Score: Mean – 1021 Standard deviation – 31.37 Min value – 987.7 Max value – 1078

For Test data values:

RMSE value: 1075
MAE value: 27.928842006815145
R2_Score value: 0.5844978356589332

For the test data the r2_score is 0.58 we can see there is decrease in the r2_score by .4 And we can see the rmse value is increased by more than 100 and the mean absolute error for the trained data is 26.4 and the test data has 27.9 which tells that there is a little error.

4. LASSO Algorithm:

Explanation:

LASSO stands for Least Absolute Shrinkage Selection Operator. It is technique used over regression to achieve more accurate results. LASSO uses absolute coefficient values for normalization.

Mathematical formula as follows

Sum of Squares + lambda*(sum of absolute values of coefficient)

Here lambda indicates amount of shrinkage

Reference :

https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.Lasso.html

Justication :

We have used this algorithm because we have many features in our independent variable. As we know lasso algorithm selects the features. It would help us in obtaining more accurate results. It works by introducing a bias term .It is useful in the feature selection for our data. It is also helpful if we have independent variables

Visualization:

Comparing actual values with predicted values for train data

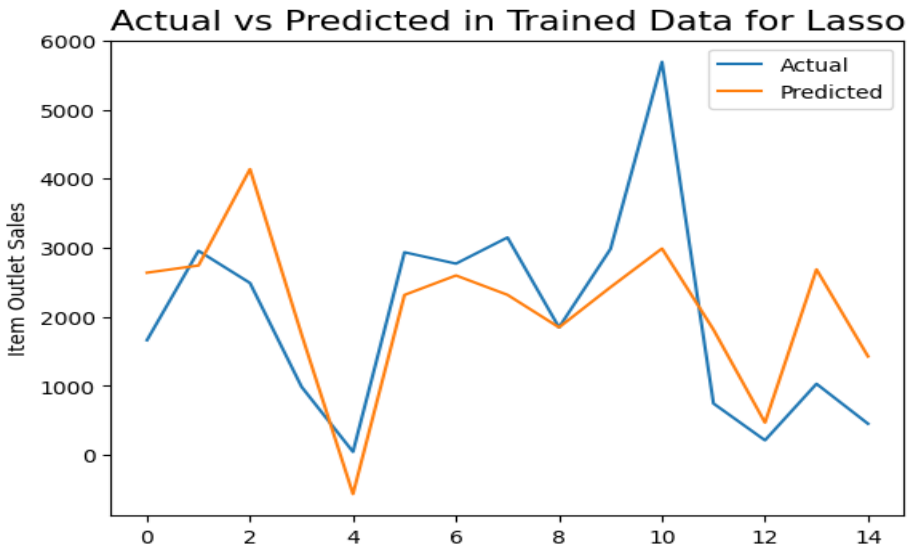

Figure 12: Scatter plot for train data

X_train and Y_train data has been fitted to LASSO model and fig 12 explains the predicted Y_train values and the actual Y_train values.From fig 12 we can observe the prediction of the train values to the actual vales.

Comparing actual values with predicted values for test data

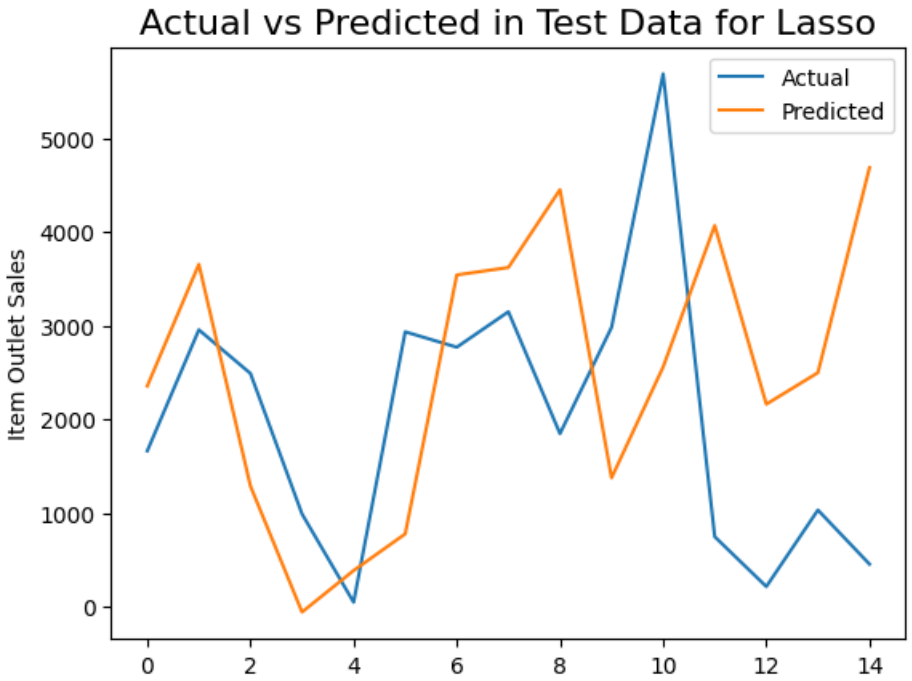

Figure 13: Scatter plot for test data

The model has learnt the data from X_train and fig 13 explains the predicted Y_test values and the actual Y_test values.From fig 13 we can observe the prediction of the test values to actual vales.

Residual Analysis for LASSO:

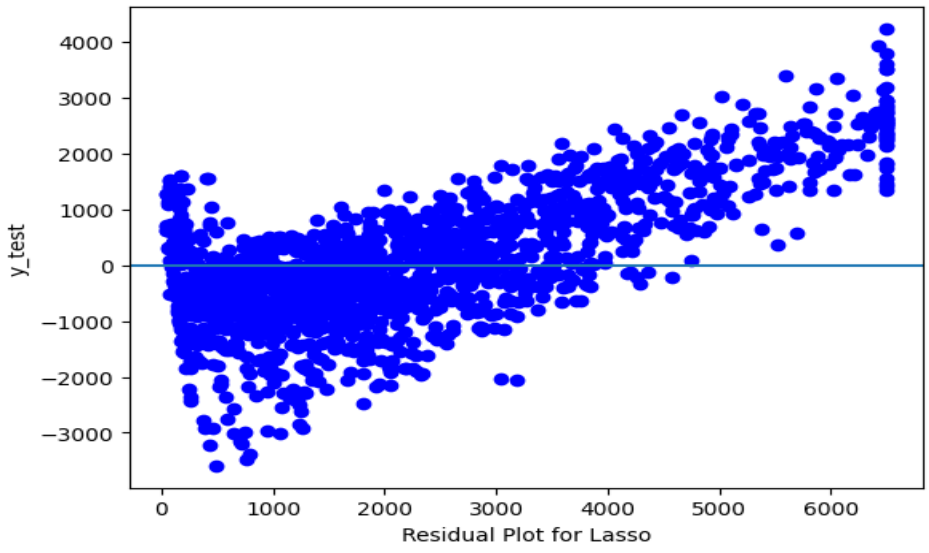

Figure 14: Residual plot for LASSO

We also used a bar graph to represent model coefficient based on LASSO

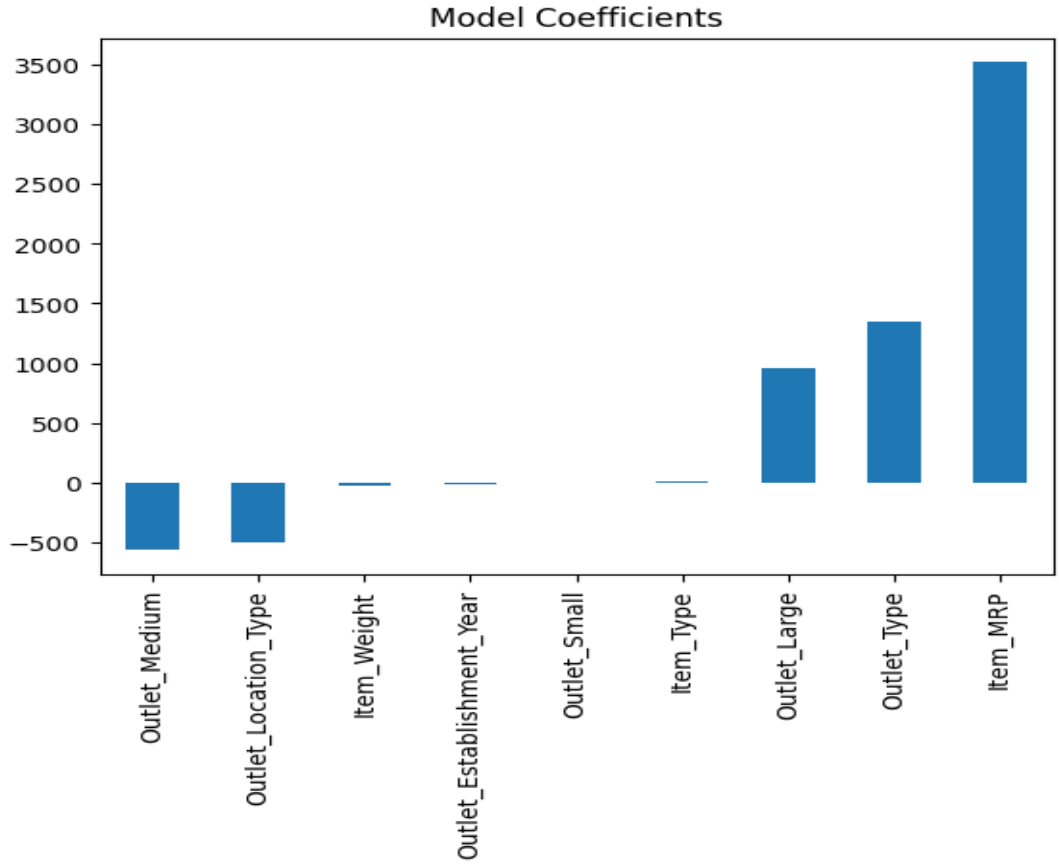

Figure 15:Bar Plot for LASSO

Evaluation metric for LASSO as follows:

Analysis:

For Train data:

RMSE value: 1085.6548812340538
MAE value: 28.865403262692258
R2_Score value: 0.5472515696975232
CV Score: Mean – 1087 Standard deviation – 27.05 Min value – 1039 Max value – 1128

For Test data values:

RMSE value: 1147
MAE value: 29.612270288848926
R2_Score value: 0.5269267945542813

For the test data the r2_score is 0.52 which is less compared to the previous models And we can see the rmse value is increased by more than 60 and the mean absolute error for the trained data is 28.4 and the test data has 29.6 which tells that there is a little error.

5.Ridge Regression Algorithm:

Explanation:

Ridge Regression is a method used to analyze any data that suffers from multicollinearity. Its aim is to reduce standard error by adding some bias in the estimates of the regression.

Multicollinearity occurs when several independent variables in model are correlated.Ridge regression has one hyperparameter which is value of alpha.

Reference:

https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.Ridge.html

Justification :

We have used Ridge Regression for our model because we have several features in our independent variable which might lead to multicollinearity to prevent it we use ridge regression.We have passed alpha value which is hyperparameter for ridge regression.advantage of ridge regression is to avoid overfitting.overfitting occurs when the trained model performs well on training data and poorly on testing data . It works by applying a penalising term(reducing the weights and biases) to overcome fitting.It increases the bias to improve variance.It can consistently performs well on both training and testing,so we have chosen this model.Also,

By using alpha value and cross validation ,the evaluation of test ,train data indicate better performance.As alpha value increases the model becomes less sensitive to variation of independent variable

Visualization:

Comparing actual values with predicted values for train data

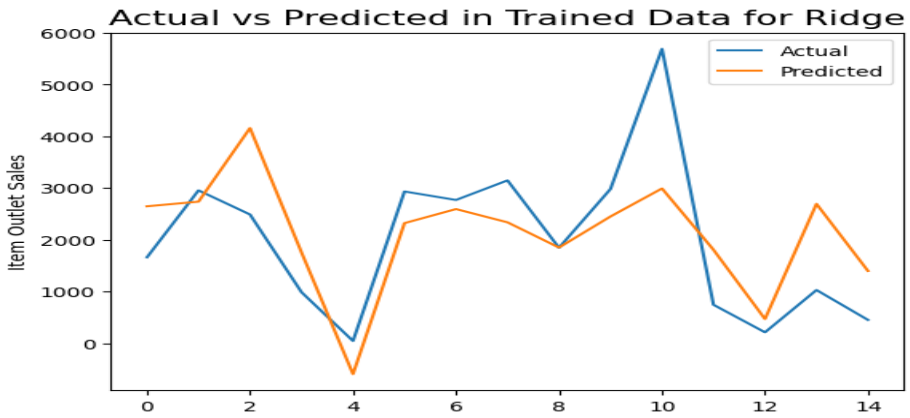

Figure 16: Scatter plot for train data

X_train and Y_train data has been fitted to Ridge Regression model and fig 16 explains the predicted Y_train values and the actual Y_train values.From fig 16 we can observe the prediction of the train values to the actual vales.

Comparing actual values with predicted values for test data

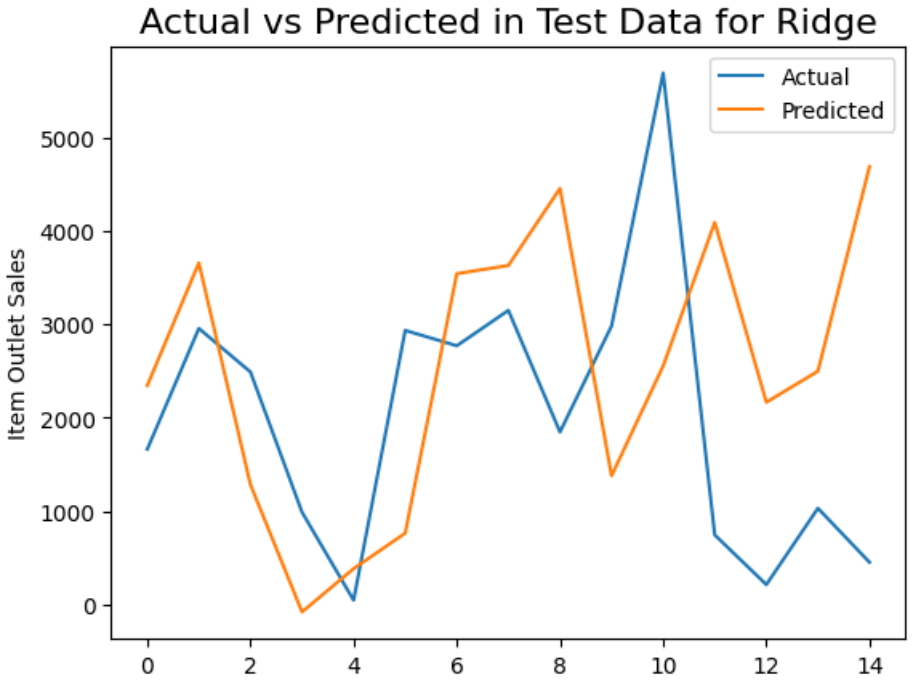

Figure 17: Scatter plot for test data

The model has learnt the data from X_train and fig 17 explains the predicted Y_test values and the actual Y_test values.From fig 17 we can observe the prediction of the test values to actual vales.

Residual Analysis for Ridge Regression:

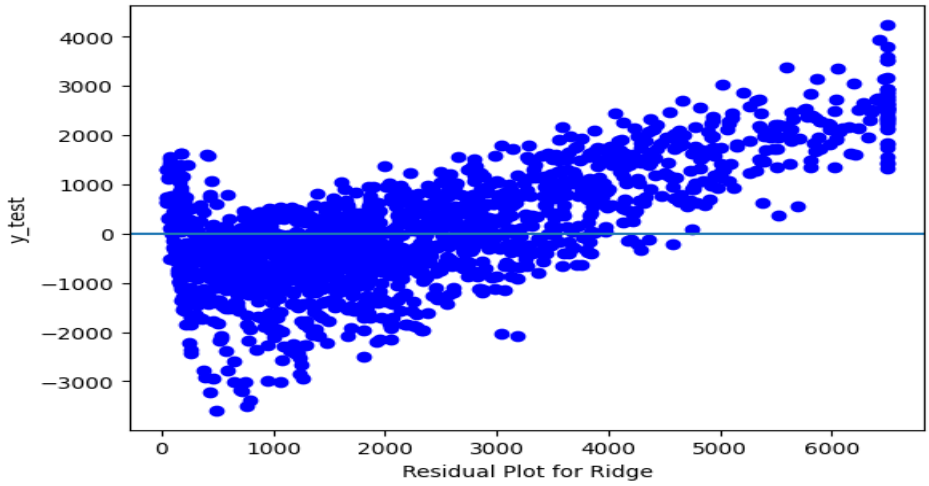

Figure 18: Residual plot for Ridge Regression

Evaluation metric for Ridge Regression as follows:

Analysis:

For Train data:

RMSE value: 1085.3396996833753
MAE value: 28.863623550599055
R2_Score value: 0.547514410571066
CV Score: Mean – 1087 Standard deviation – 26.17 Min value – 1041 Max value – 1126

For Test data values:

RMSE value: 1146
MAE value: 29.591717460528717
R2_Score value: 0.5279929954682852

For the test data the r2_score is 0.57 we can see there is decrease in the r2_score by 2 And we can see the rmse value is increased by more than 60 and the mean absolute error for the trained data is 28.8 and the test data has 29.5 .

6.DecisionTree Regression Algorithm:

Explanation:

DecisionTree Regression Algorithm is used for both classification and regression

The main goal of DecisionTree Regression Algorithm is to create a training model that can be used to predict value of target variable. It will formulate some set of rules to do prediction.A decision tree is built from rootnode and involves partition of data into subsets that contain instance with similar data.This algorithm is prone to overfitting so it better to use specified minimum number of children

DecisionTree Regression is of two types 1.Categorical DecisionTressRegression

2.Continous DecisionTreeRegression

Reference :

https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

Justification :

We have continous data in our dependent variable so we use continous decisiontree regression for it

A decisionTree consists of root node and a leaf node

Leaf node are the nodes that doesn't split.we have passed number of leaf nodes so that algorithm avoids overfitting.

Visualization:

Comparing actual values with predicted values for train data

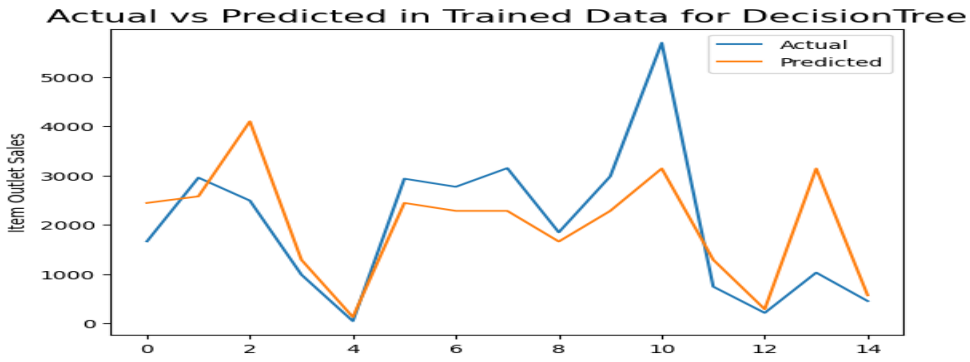

Figure 19: Scatter plot for train data

X_train and Y_train data has been fitted to DecisionTree Regression model and fig 19 explains the predicted Y_train values and the actual Y_train values.From fig 19 we can observe the prediction of the train values to the actual vales.

Comparing actual values with predicted values for test data

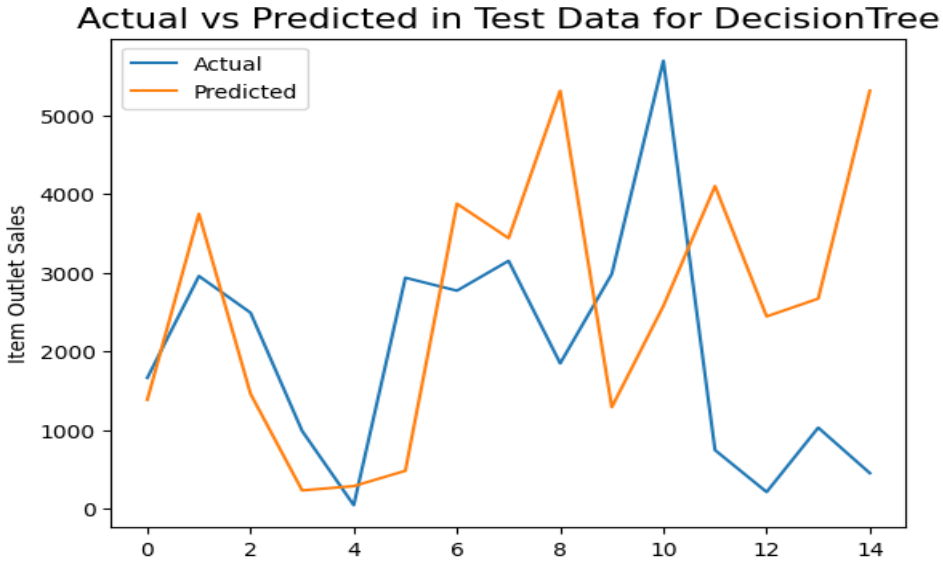

Figure 20: Scatterplot for test data

The model has learnt the data from X_train and fig 20 explains the predicted Y_test values and the actual Y_test values.From fig 20 we can observe the prediction of the test values to actual vales.

Residual Analysis for DecisionTree Regression:

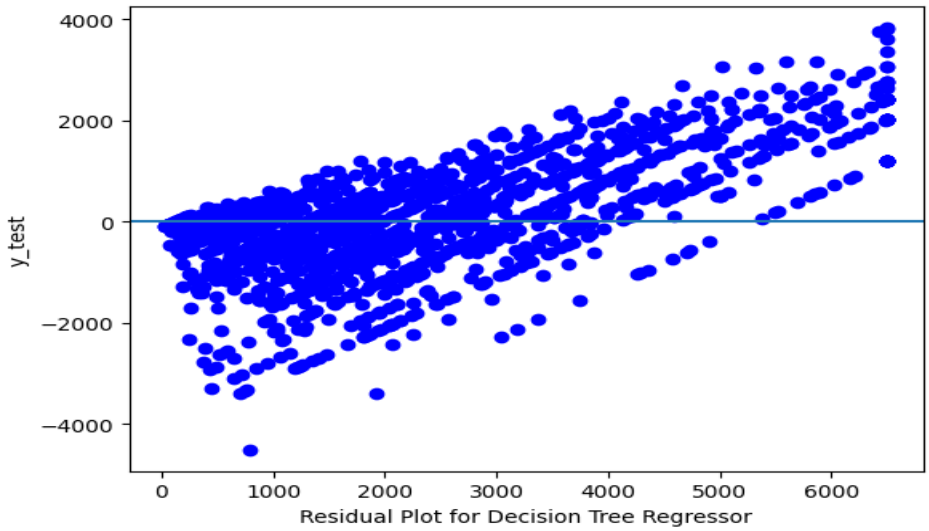

Figure 21: Residual plot for decisionTree regression

Evaluation metric for Decision Tree Regression as follows:

Analysis:

For Train data:

RMSE value: 989.4673803210671

MAE value: 26.73091732411912

R2_Score value: 0.6239233778481474

CV Score: Mean – 1012 Standard deviation – 31.86 Min value – 956.9 Max value – 1070

For Test data values:

RMSE value: 1068

MAE value: 27.750112159259455

R2_Score value: 0.5903740049447113

For the test data the r2_score is 0.590 we can see there is decrease in the r2_score by 3 And we can see the rmse value is increased by 70 and the mean absolute error for the trained data is 26.7 and the test data has 27.7 which tells that there is a little error.

Comparing RMSE values of all algorithms :

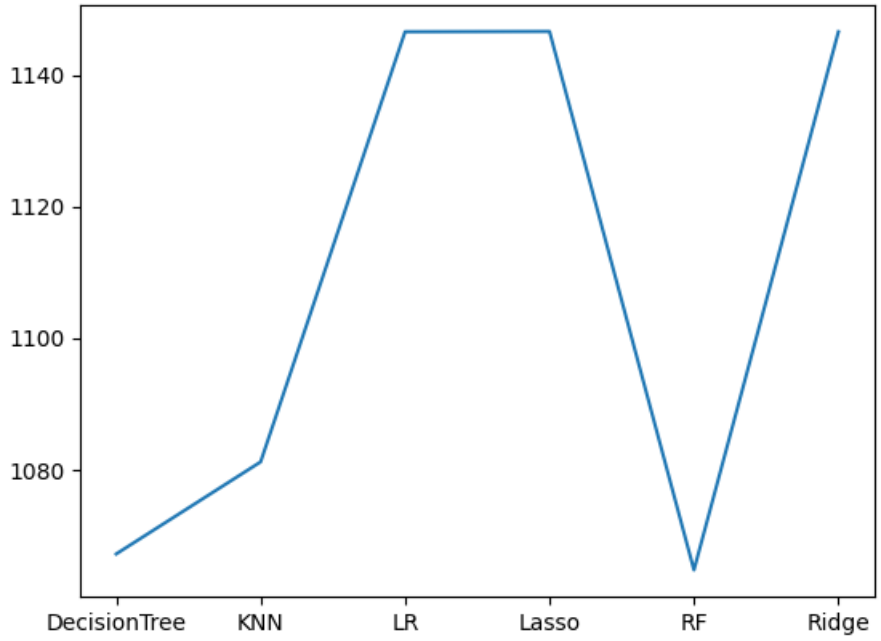

Figure 22: RMSE values Comparison

AS we know RMSE values should be less for a model. From above graph we get to know that, the least RMSE values are for Decision Tree and Random Forest.

Comparing R2Score of all the algorithms:

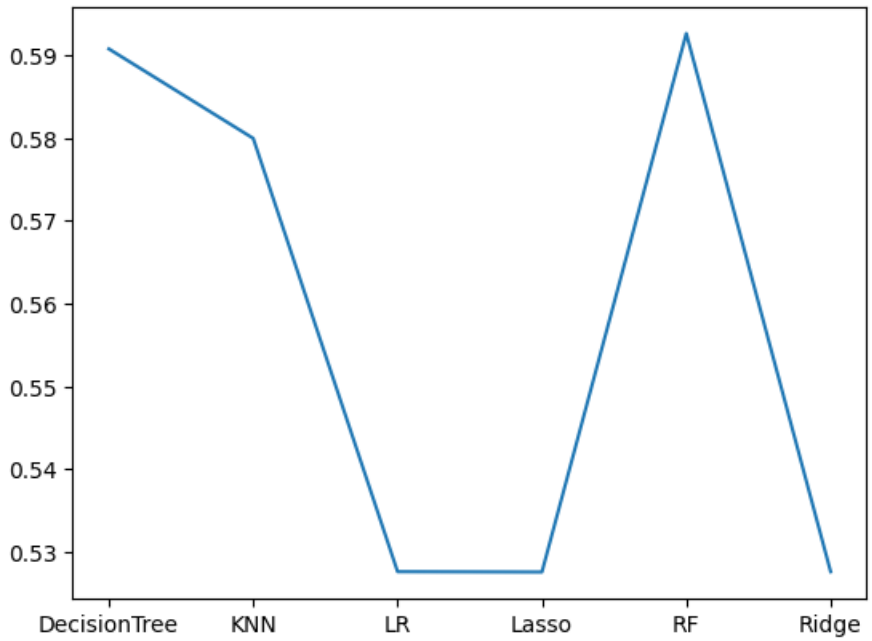

Figure 23: R2Score values Comparison

AS per above Graph,

We can see that Linear Regression, LASSO, Ridge Regression has atmost same R2_score .

We can see that Decision Tree , KNN, Random Forest has highest R2Scores.

Comparing MAE of all the algorithms

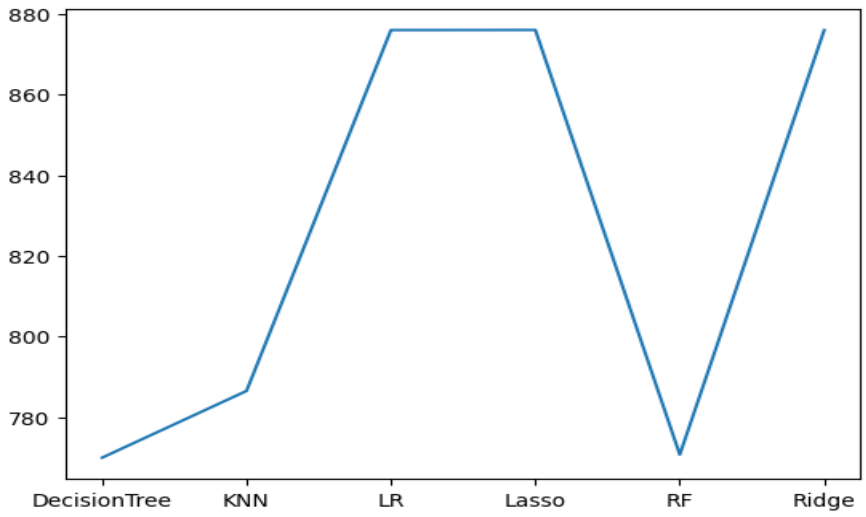

Figure 24: MAE values Comparison

As we know that MAE(mean absoluted error= predicted value – actual value) should be low for a model. As per the graph we get to know that least MAE values are for Decision tree and Random Forest.

As comparing all the models we have seen the Decision Tree Algorithm has the r2_score= 0.5903740049447113 and the Rmse 1068 for the test data and MAE value: 27.750112159259455 is and the Random Forest has the r2 score=: 0.592547138563075 and the Rmse value is 1065 and MAE value: : 27.769585400290584.and we pick Random Forest for the next phase.