Murali B

Project Notes - 2

Contents Page number  
  
Output:   
  
1.1 Shape of train set 3  
1.2 Shape of the test set 3  
1.3 OLS Regression results 4  
1.6 Train set accuracy 5  
1.7 Test set accuracy 5  
1.8 LDA model train set 5  
1.9 LDA model test set 6  
1.10 Train model accuracy 6  
1.11 Test model accuracy 6  
1.12 Model accuracy details 6  
1.13 LDA model coefficient 6  
1.14 Linear regression model train set 7  
1.15 Linear regression model test set 7  
1.16 Logistic regression train set 7  
1.17 Logistic regression test set 8  
1.18 KNN model train set 8  
1.19 RMSE value test set 9  
1.20 RMSE value 9  
1.21 RMSE value plot 9  
1.22 Train model accuracy 10  
1.23 Test model accuracy 10  
1.24 KNN Regression 10  
1.25 Naïve Bayes train set 10  
1.26 Naïve Bayes test set 11  
1.27 SVM train set 11  
1.28 SVM test set 11  
1.29 Decision tree train set 12  
1.30 Decision tree test set 13  
1.31 Model accuracy train set 13  
1.32 Model accuracy test set 13  
1.33 Decision tree Regression 14  
1.34 Feature Importance 14  
1.35 Feature importance 14  
1.36 Boosting train set 15  
1.37 Boosting test set 16  
1.38 Bagging train set 17  
1.39 Bagging test set 17  
1.40 Random forest train set 18  
1.41 Random forest test set 19  
1.42 Feature importance 20  
1.43 Best parameters 21  
1.44 LDA with Gridsearch CV 21  
1.45 LDA with GridSearch CV 21  
1.46 Best parameters 21  
1.47 KNN model train set 22  
1.48 KNN model test set 22  
1.49 Naïve Bayes train set 22  
1.50 Naïve Bayes test set 23  
1.51 Gradient Boosting train set 23  
1.52 Gradient Boosting test set 24  
1.53 Best parameters 25  
1.54 Decision tree train set 25  
1.55 Decision tree test set 25  
1.56 Random Forest train set 26  
1.57 Random Forest test set 26  
1.58 Bagging test set 27  
1.59 Bagging train set 28  
1.60 Linear Regression train set 29  
1.61 Linear Regression test set 30  
1.62 Lasso regression 30  
1.63 Best parameters 31  
1.64 Ridge Regression 31  
1.65 Best parameters 31

Project Notes – 2

**1). Model building and interpretation.**

a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)

Train and Test split

  
Output 1.1 Shape of train set   
  
We see that there are 15129 observations with 22 observations.

  
Output 1.2 Shape of test set

From the above output, we see 6484 observations in test dataset.

OLS Regression Results : A screenshot of a computer

Description automatically generated with low confidence  
Output 1.3 OLS Regression results  
  
From the given output, the model shown is a OLS regression with 15129 observations 15106 residuals applied using least squares method and has r-square value is 0.702 Adj r squared values is 0.702. 6 variables have negative coefficients whereas all others have positive coefficients. Even P value shows only for 7 variables. OLS Regression is basically a linear regression method used to model the relationship between a dependent variable and one or more independent variables.

Head of the train set :   
A picture containing text, font, number, line

Description automatically generated  
Table 1.4 Head of train set

Head of test set :   
A picture containing text, font, screenshot, white

Description automatically generated  
Table 1.5 Head of test set

LDA Model :   
  
  
LDA model is basically a dimensionality reduction technique which maximizes the separation between the classes.

Performance Matrix on train data set :   
  
Output 1.6   
  
The Accuracy of performance matrix on train set is 11.46%  
  
  
Performance Matrix on test data set :   
  
Output 1.7   
  
The Accuracy of performance matrix on test set is 0.86%  
  
  
LDA model Train set:   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.8 LDA model Train set  
  
LDA model on train set shows as RMSE value as 200909.89   
  
  
LDA model Test :   
  
A picture containing text, screenshot, font, line

Description automatically generated   
Output 1.9 LDA model test  
  
LDA model on test set shows as RMSE value as 224146.84

Train model accuracy :   
  
Output 1.10   
  
Test model accuracy:   
  
  
Output 1.11   
  
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.12 Model accuracy details  
  
  
LDA model coefficient :   
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.13 LDA model coefficient   
  
From the output, we see that LDA model has negative coefficient values.

Linear Regression Model Train set :   
  
A picture containing text, screenshot, font, line

Description automatically generated  
  
Output 1.14 Linear regression model train set   
  
We can observe that Linear Regression model on train set has RMSE value as 200909.89 but LDA model on train set has RMSE value as 200909.89.

Linear Regression Model Test set :   
  
A screenshot of a test

Description automatically generated with medium confidence  
  
Output 1.15 Linear Regression Model Test set  
  
RMSE value on Liner Regression model test data shows as 224146.84 , LDA model test data has RMSE values as 224146.84.

Logistic Regression :   
  
It is a statistical technique which models the relationship between the independent variables and the probability of the binary outcome.   
  
  
Logistic Regression on Train set :   
  
A screenshot of a computer

Description automatically generated with low confidence  
  
Output 1.16 Logistic Regression on Train set  
  
We see Logistic Regression on train set shows RMSE value as 391333.00 , 200909.89 for Linear Regression model and 200909.89 for LDA model on train set.   
  
Logistic Regression on Test set :   
A screenshot of a test

Description automatically generated with low confidence  
  
Output 1.17 Logistic Regression on Test set   
  
We see Logistic Regression on test set shows RMSE value as 391029.89 , 224146.84 for Linear Regression model and 224146.84 for LDA model on train set.

KNN model :   
  
K-Nearest neighbors which is used for classification and regression.   
  
Train set :   
  
A screenshot of a calculator

Description automatically generated with low confidence  
  
Output 1.18 KNN model train set

RMSE value shows as 0 MAE value as 0 R-Square as 1.0 for KNN train set , RMSE as 391333.00 MAE as 230618.24 for Logistic Regression train set , RMSE as 200909.89 MAE as 108613.73 for Linear Regression Model train set , RMSE as 200909.89 MAE as 108613.73 for LDA model train set.

RMSE value   
  
Test set :   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.19 RMSE value of test set.   
  
RMSE value shows as 318538.70 R-Square as 0.24 for KNN test set , RMSE as 391029.89 for Logistic Regression test set , RMSE as 224146.84 for Linear Regression Model test set , RMSE as 224146.84 for LDA model test set.

RMSE values :   
A picture containing text, font, screenshot, number

Description automatically generated  
  
Output 1.20 RMSE values   
  
We can see RMSE value for k from 1 to 10 form the above output.

RMSE value plot :   
A picture containing text, screenshot, plot, line

Description automatically generated  
Graph 1.21 RMSE value plot  
  
From the graph, we see there is an downward trend of RMSE value for K value from 1 to 10.   
  
  
Train model accuracy:   
  
Output 1.22

KNN model accuracy score on train set as 100%  
  
Test model accuracy:  
  
Output 1.23   
  
KNN model accuracy score on test set as 100%

kNN Regression :   
  
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.24 KNN Regression

Naive Bayes :   
  
It is an algorithm based on Bayes theorem used for classification tasks , text classification and spam filtering.

Train set :   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.25 Naive Bayes train set  
  
RMSE values shows as 256901.47 MAE value as 157262.208 for Naïve Bayes model train set, RMSE value shows as 0 R-Square as 1.0 MAE as 0 for KNN train set , RMSE as 391333.00 MAE as 230618.24 for Logistic Regression train set , RMSE as 200909.89 MAE value as 108613.73 for Linear Regression Model train set , RMSE as 200909.89 MAE as 108613.73 for LDA model train set.

Test set :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.26 Naive Bayes test set  
  
RMSE values shows as 300553.50 MAE value as 189452.84 for Naïve Bayes model test set, RMSE value shows as 318538.70 R-Square as 0.24 MAE as 197163.91 for KNN test set , RMSE as 391029.89 MAE as 227012.91 for Logistic Regression test set , RMSE as 224146.84 MAE value as 1230123.89 for Linear Regression Model test set , RMSE as 224146.84 MAE as 1230123.89 for LDA model test set.   
  
  
SVM :   
  
This algorithm is used for classification and regression tasks and used to find the best hyperplane that separates the data into different classes.

Train set :   
  
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.27 SVM train set   
  
RMSE value shows as 413965.68 MAE value as 242070.33 for SVM model train set , RMSE values shows as 256901.47 MAE value as 157262.208 for Naïve Bayes model train set, RMSE value shows as 0 R-Square as 1.0 MAE as 0 for KNN train set , RMSE as 391333.00 MAE as 230618.24 for Logistic Regression train set , RMSE as 200909.89 MAE value as 108613.73 for Linear Regression Model train set , RMSE as 200909.89 MAE as 108613.73 for LDA model train set.

Test set :   
  
A screenshot of a graph

Description automatically generated with low confidence  
  
Output 1.28 SVM test set  
  
RMSE value shows as 412959.69 MAE value as 238810.25 for SVM model test set, RMSE values shows as 300553.50 MAE value as 189452.84 for Naïve Bayes model test set, RMSE value shows as 318538.70 R-Square as 0.24 MAE as 197163.91 for KNN test set , RMSE as 391029.89 MAE as 227012.91 for Logistic Regression test set , RMSE as 224146.84 MAE value as 1230123.89 for Linear Regression Model test set , RMSE as 224146.84 MAE as 1230123.89 for LDA model test set.   
  
  
  
Decision Tree :   
  
It is a machine learning algorithm used for both classification and regression tasks.   
  
Model scores details :   
  
Train set :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.29 Decision tree train set  
  
RMSE value shows as 0 MAE value as 0 for Decision tree model train set, RMSE value shows as 413965.68 MAE value as 242070.33 for SVM model train set , RMSE values shows as 256901.47 MAE value as 157262.208 for Naïve Bayes model train set, RMSE value shows as 0 R-Square as 1.0 MAE as 0 for KNN train set , RMSE as 391333.00 MAE as 230618.24 for Logistic Regression train set , RMSE as 200909.89 MAE value as 108613.73 for Linear Regression Model train set , RMSE as 200909.89 MAE as 108613.73 for LDA model train set.

Test set :   
  
A picture containing text, font, number, software

Description automatically generated  
Output 1.30 Decision tree test set  
  
RMSE value shows as 412959.69 MAE value as 238810.25 for SVM model test set, RMSE values shows as 300553.50 MAE value as 189452.84 for Naïve Bayes model test set, RMSE value shows as 318538.70 R-Square as 0.24 MAE as 197163.91 for KNN test set , RMSE as 391029.89 MAE as 227012.91 for Logistic Regression test set , RMSE as 224146.84 MAE value as 1230123.89 for Linear Regression Model test set , RMSE as 224146.84 MAE as 1230123.89 for LDA model test set. RMSE value shows as 185940.23 MAE value as 101114.45 for Decision tree model test set.   
  
  
Decision Tree Regressor :   
  
Decision tree is an algorithm which splits the data into different branches based on the features values.   
  


Train model accuracy:   
  
Output 1.31  
  
Decision tree model accuracy on train set is 69.39%  
  
  
Test model accuracy:   
  
Output 1.32  
  
Decision tree model accuracy on test set is 62.9%

Decision Tree Regression :   
  
A picture containing text, screenshot, font, line

Description automatically generated   
Output 1.33 Decision Tree Regression

Feature Importance :   
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.34 Feature Importance  
  
A screenshot of a computer code

Description automatically generated with low confidence  
Output 1.35 Feature Importance   
  
From the above output, we can see top 5 important features are quality, living\_measure, lat , long and coast.

c.Interpretation of the model(s)   
  
Feature importance was mainly performed to reduce the overall dimensions, but there was no improvement found in the modeling accuracy. Each of the outlier strategy was evaluated for the various algorithms, and the best performing algorithm was subjected to hyper parameter tuning. We can look at the co-efficient and interpret values based on positive and negative. variables co-efficient is positive meaning for example every increase in value of living\_measure there is price increase we can apply the logic in other variable which is positive and whichever variable co-efficient is negative for example if there are negative co-efficient variable tends to lose the price or can say there is decrease in price. P - Values whichever variable has a higher value then 0.05 that is insignificant and the variables which are under or equal to 0.05 are significant values and important variables for prediction on our data.

**2). Model Tuning and business implication**a.Ensemble modelling, wherever applicable   
  
Boosting :   
  
It is a machine learning method which basically combines multiple decision trees to create a strong predictive model.   
  
  
  
Model scores :   
  
Train set :   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.36 Boosting train set  
  
RMSE value shows as 413965.68 MAE value as 242070.33 for SVM model train set , RMSE values shows as 256901.47 MAE value as 157262.208 for Naïve Bayes model train set, RMSE value shows as 0 R-Square as 1.0 MAE as 0 for KNN train set , RMSE as 391333.00 MAE as 230618.24 for Logistic Regression train set , RMSE as 200909.89 MAE value as 108613.73 for Linear Regression Model train set , RMSE as 200909.89 MAE as 108613.73 for LDA model train set. RMSE value shows as 0 MAE value as 0 for Decision tree model train set. RMSE value shows as 112532.22 MAE value as 71421.87 for Boosting model train set.

Test set :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.37 Boosting test set  
  
RMSE value shows as 412959.69 MAE value as 238810.25 for SVM model test set, RMSE values shows as 300553.50 MAE value as 189452.84 for Naïve Bayes model test set, RMSE value shows as 318538.70 R-Square as 0.24 MAE as 197163.91 for KNN test set , RMSE as 391029.89 MAE as 227012.91 for Logistic Regression test set , RMSE as 224146.84 MAE value as 1230123.89 for Linear Regression Model test set , RMSE as 224146.84 MAE as 1230123.89 for LDA model test set. RMSE value shows as 185940.23 MAE value as 101114.45 for Decision tree model test set. RMSE value shows as 129410.93 MAE value as 76841.13 for Boosting model test set.

Bagging :   
  
Bagging is basically a technique which combines multiple models trained on different subsets of the training data to make predictions.   
  
  
Model scores :   
  
Train set :   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.38 Bagging train set  
  
RMSE value shows as 59141.25 MAE value as 30000.51 for Bagging model train set, RMSE value shows as 413965.68 MAE value as 242070.33 for SVM model train set , RMSE values shows as 256901.47 MAE value as 157262.208 for Naïve Bayes model train set, RMSE value shows as 0 R-Square as 1.0 MAE as 0 for KNN train set , RMSE as 391333.00 MAE as 230618.24 for Logistic Regression train set , RMSE as 200909.89 MAE value as 108613.73 for Linear Regression Model train set , RMSE as 200909.89 MAE as 108613.73 for LDA model train set. RMSE value shows as 0 MAE value as 0 for Decision tree model train set. RMSE value shows as 112532.22 MAE value as 71421.87 for Boosting model train set.   
  
Test set :   
  
A screenshot of a graph

Description automatically generated with low confidence   
Output 1.39 Bagging test set  
  
RMSE value shows as 133156.39 MAE value as 74624.16 for Bagging model test set, RMSE value shows as 412959.69 MAE value as 238810.25 for SVM model test set, RMSE values shows as 300553.50 MAE value as 189452.84 for Naïve Bayes model test set, RMSE value shows as 318538.70 R-Square as 0.24 MAE as 197163.91 for KNN test set , RMSE as 391029.89 MAE as 227012.91 for Logistic Regression test set , RMSE as 224146.84 MAE value as 1230123.89 for Linear Regression Model test set , RMSE as 224146.84 MAE as 1230123.89 for LDA model test set. RMSE value shows as 185940.23 MAE value as 101114.45 for Decision tree model test set. RMSE value shows as 129410.93 MAE value as 76841.13 for Boosting model test set.

Random forest :   
  
It is a method which combines multiple decision trees to create a robust and predictive model used for classification and regression tasks.   
  
Model scores :   
  
Train set :   
  
A picture containing text, screenshot, number, software

Description automatically generated  
Output 1.40 Random forest train set  
  
RMSE value shows as 50426.82 MAE value as 26215.09 for Random forest model train set, RMSE value shows as 59141.25 MAE value as 30000.51 for Bagging model train set, RMSE value shows as 413965.68 MAE value as 242070.33 for SVM model train set , RMSE values shows as 256901.47 MAE value as 157262.208 for Naïve Bayes model train set, RMSE value shows as 0 R-Square as 1.0 MAE as 0 for KNN train set , RMSE as 391333.00 MAE as 230618.24 for Logistic Regression train set , RMSE as 200909.89 MAE value as 108613.73 for Linear Regression Model train set , RMSE as 200909.89 MAE as 108613.73 for LDA model train set. RMSE value shows as 0 MAE value as 0 for Decision tree model train set. RMSE value shows as 112532.22 MAE value as 71421.87 for Boosting model train set.

Test set :   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.41 Random forest test set  
  
RMSE value shows as 126574.97 MAE value as 69493.97 for Random forest model test set, RMSE value shows as 133156.39 MAE value as 74624.16 for Bagging model test set, RMSE value shows as 412959.69 MAE value as 238810.25 for SVM model test set, RMSE values shows as 300553.50 MAE value as 189452.84 for Naïve Bayes model test set, RMSE value shows as 318538.70 R-Square as 0.24 MAE as 197163.91 for KNN test set , RMSE as 391029.89 MAE as 227012.91 for Logistic Regression test set , RMSE as 224146.84 MAE value as 1230123.89 for Linear Regression Model test set , RMSE as 224146.84 MAE as 1230123.89 for LDA model test set. RMSE value shows as 185940.23 MAE value as 101114.45 for Decision tree model test set. RMSE value shows as 129410.93 MAE value as 76841.13 for Boosting model test set.

Feature Importance :   
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.42 Feature Importance  
  
We can see top 5 most important features are quality, living\_measure , lat , long , furnished.

The best model is Random Forest Model because other models are either overfitted or have low value and based on RMSE value If the noise is small, which means model is good at predicting observed data, and if RMSE is large, this means model is failing to account for important features. These models will help us analyze the most best way to determine correct prices and areas and all the variables which are available in our data.

LDA with GridSearchCV :   
  
This method allows to systematically search for the best combination of hyperparameters for a given machine learning model.  
  
  
Linear Discriminant Analysis :   
  
   
  
Best Parameters :   
  


Output 1.43 Best Parameters

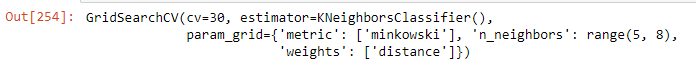
Train set :   
  
A black text on a white background

Description automatically generated with low confidence  
Output 1.44 LDA with GridSearchCV train set  
  
Train set has RMSE values as 202061.90 with MAE values as 108600.50

Test set :   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.45 LDA with GridSearchCV test set

Test set has RMSE values as 225654.15 with MAE values as 124405.89

b. Any other model tuning measures (if applicable) :   
  
KNN Model :   
  
  
  
Best Parameters :   
  
   
Output 1.46 Best parameters

Train set :   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.47 KNN Model train set   
  
KNN model train set has RMSE value as 0 , MAE as 0 and LDA model train set has RMSE values as 202061.90 and MAE value as 108600.50

Test set :   
  
A screenshot of a test

Description automatically generated with low confidence  
Output 1.48 KNN Model test set   
  
KNN model test set has RMSE value as 318389.24 , MAE as 196927.18 and LDA model test set has RMSE values as 225654.15 and MAE value as 124405.89

Naive Bayes :   
  


Train set :   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.49 Naive Bayes train set   
  
Naives Bayes train set has RMSE value as 256901.47 , MAE as 157262.20, KNN model train set has RMSE value as 0 , MAE as 0 and LDA model train set has RMSE values as 202061.90 and MAE value as 108600.50

Test set :   
  
A screenshot of a computer

Description automatically generated with low confidence  
Output 1.50 Naive Bayes test set  
  
Naives Bayes test set has RMSE value as 300553.50 , MAE as 189452.84 KNN model test set has RMSE value as 318389.24 , MAE as 196927.18 and LDA model test set has RMSE values as 225654.15 and MAE value as 124405.89

Gradient Boosting :   
  
These models have various hyperparameters that can be tuned to optimize performance such as the learning rate, number of estimators , maximum depth of the trees.   
  
A black text on a white background

Description automatically generated with low confidence

Train set :   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.51 Gradient boosting train set  
  
RMSE value shows as 50426.82 MAE value as 26215.09 for Random forest model train set, RMSE value shows as 59141.25 MAE value as 30000.51 for Bagging model train set, RMSE value shows as 413965.68 MAE value as 242070.33 for SVM model train set , RMSE values shows as 256901.47 MAE value as 157262.208 for Naïve Bayes model train set, RMSE value shows as 0 R-Square as 1.0 MAE as 0 for KNN train set , RMSE as 391333.00 MAE as 230618.24 for Logistic Regression train set , RMSE as 200909.89 MAE value as 108613.73 for Linear Regression Model train set , RMSE as 200909.89 MAE as 108613.73 for LDA model train set. RMSE value shows as 0 MAE value as 0 for Decision tree model train set. RMSE value shows as 112532.22 MAE value as 71421.87 for Boosting model train set. RMSE value shows as 151710.19 MAE value as 94597.80 for Gradient Boosting model train set.

Test set :   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.52 Gradient boosting test set  
   
RMSE value shows as 126574.97 MAE value as 69493.97 for Random forest model test set, RMSE value shows as 133156.39 MAE value as 74624.16 for Bagging model test set, RMSE value shows as 412959.69 MAE value as 238810.25 for SVM model test set, RMSE values shows as 300553.50 MAE value as 189452.84 for Naïve Bayes model test set, RMSE value shows as 318538.70 R-Square as 0.24 MAE as 197163.91 for KNN test set , RMSE as 391029.89 MAE as 227012.91 for Logistic Regression test set , RMSE as 224146.84 MAE value as 1230123.89 for Linear Regression Model test set , RMSE as 224146.84 MAE as 1230123.89 for LDA model test set. RMSE value shows as 185940.23 MAE value as 101114.45 for Decision tree model test set. RMSE value shows as 129410.93 MAE value as 76841.13 for Boosting model test set. RMSE value shows as 160496.34 MAE value as 97368.76 for Gradient Boosting model test set.

Decision Tree :   
  
A black text on a white background

Description automatically generated with low confidence

Best Parameters :   
  
   
  
Output 1.53 Best parameters

Train set :   
  
A picture containing text, screenshot, font, line

Description automatically generated  
Output 1.54 Decision Tree train set  
  
  
Naives Bayes train set has RMSE value as 256901.47 , MAE as 157262.20, KNN model train set has RMSE value as 0 , MAE as 0 and LDA model train set has RMSE values as 202061.90 and MAE value as 108600.50, Decision tree model train set has RMSE values as 213928.86 MAE values as 103063.81

Test set :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.55 Decision Tree test set   
  
Naives Bayes test set has RMSE value as 300553.50 , MAE as 189452.84 KNN model test set has RMSE value as 318389.24 , MAE as 196927.18 and LDA model test set has RMSE values as 225654.15 and MAE value as 124405.89 , Decision tree model test set has RMSE values as 215437.11 MAE values as 112120.02

Random Forest :   
  
A black text on a white background

Description automatically generated with low confidence  
  
Train set :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.56 Random Forest train set  
  
  
Random Forest train set has RMSE value as 3446.20 , MAE as 53.83, Naives Bayes train set has RMSE value as 256901.47 , MAE as 157262.20, KNN model train set has RMSE value as 0 , MAE as 0 and LDA model train set has RMSE values as 202061.90 and MAE value as 108600.50, Decision tree model train set has RMSE values as 213928.86 MAE values as 103063.81   
  
Test set :   
  
A picture containing text, screenshot, font, number

Description automatically generated  
Output 1.57 Random Forest test set  
  
Random Forest test set has RMSE value as 169640.65 , MAE as 91496.61 , Naives Bayes test set has RMSE value as 300553.50 , MAE as 189452.84 KNN model test set has RMSE value as 318389.24 , MAE as 196927.18 and LDA model test set has RMSE values as 225654.15 and MAE value as 124405.89 , Decision tree model test set has RMSE values as 215437.11 MAE values as 112120.02

Bagging :   
  
A black text on a white background

Description automatically generated with low confidence

Train set :   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.58 Bagging test set  
  
  
Bagging train set RMSE values shows 54017.34 MAE value as 28128.85 , Random forest train set RMSE values shows 50426.82 MAE value as 26215.09 , Boosting train set RMSE values shows as 59141.25 MAE value as 30000.51 , SVM train set RMSE values shows 413965.68 MAE value as 242070.33 , Navies Bayes train set RMSE values shows 256901.47 MAE value as 157262.20 , KNN model train set RMSE values shows 0 MAE value as 0 R-Square value as 1.0, Logistic regression train set RMSE values shows 391333.00 MAE value as 230618.24 , Linear regression train set RMSE values shows 200909.89 MAE value as 108613.73 , LDA model train set RMSE values shows 200909.89 MAE value as 108613.73 , Decision tree train set RMSE values shows 0 MAE value as 0 , Boosting train set RMSE values shows 112532.22 MAE value as 71421.877 and Gradient Boosting train set RMSE values shows 151710.19 MAE value as 94597.80

Test set :   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.59 Bagging train set  
  
Bagging test set RMSE values shows 131693.05 MAE value as 71913.36 , Random forest test set RMSE values shows 126574.97 MAE value as 69493.97 , Boosting test set RMSE values shows as 133156.39 MAE value as 74624.16 , SVM test set RMSE values shows 412959.69 MAE value as 238810.25 , Navies Bayes test set RMSE values shows 300553.50 MAE value as 189452.84 , KNN model test set RMSE values shows 318538.70 MAE value as 197163.91 R-Square value as 0.24, Logistic regression test set RMSE values shows 391029.89 MAE value as 227012.91 , Linear regression test set RMSE values shows 224146.84 MAE value as 123013.89 , LDA model test set RMSE values shows 224146.84 MAE value as 123013.89 , Decision tree test set RMSE values shows 185940.23 MAE value as 101114.45 , Boosting test set RMSE values shows 129410.93 MAE value as 76841.13 and Gradient Boosting test set RMSE values shows 160496.34 MAE value as 97368.76

Linear Regression :   
  
A picture containing text, font, screenshot, algebra

Description automatically generated

Train set :   
  
A picture containing text, screenshot, number, font

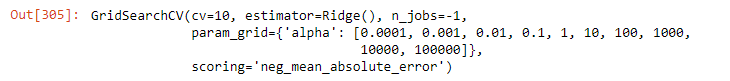
Description automatically generated  
Output 1.60 Linear Regression train set  
  
  
Linear Regression model train set RMSE values shows 200602.69 MAE value as 124864.58, Bagging train set RMSE values shows 54017.34 MAE value as 28128.85 , Random forest train set RMSE values shows 50426.82 MAE value as 26215.09 , Boosting train set RMSE values shows as 59141.25 MAE value as 30000.51 , SVM train set RMSE values shows 413965.68 MAE value as 242070.33 , Navies Bayes train set RMSE values shows 256901.47 MAE value as 157262.20 , KNN model train set RMSE values shows 0 MAE value as 0 R-Square value as 1.0, Logistic regression train set RMSE values shows 391333.00 MAE value as 230618.24 , Linear regression train set RMSE values shows 200909.89 MAE value as 108613.73 , LDA model train set RMSE values shows 200909.89 MAE value as 108613.73 , Decision tree train set RMSE values shows 0 MAE value as 0 , Boosting train set RMSE values shows 112532.22 MAE value as 71421.877 and Gradient Boosting train set RMSE values shows 151710.19 MAE value as 94597.80

Test set :   
  
A picture containing text, screenshot, number, font

Description automatically generated  
Output 1.61 Linear Regression test set  
  
Linear Regression model test set RMSE values shows 198968.62 MAE value as 124739.75, Bagging test set RMSE values shows 131693.05 MAE value as 71913.36 , Random forest test set RMSE values shows 126574.97 MAE value as 69493.97 , Boosting test set RMSE values shows as 133156.39 MAE value as 74624.16 , SVM test set RMSE values shows 412959.69 MAE value as 238810.25 , Navies Bayes test set RMSE values shows 300553.50 MAE value as 189452.84 , KNN model test set RMSE values shows 318538.70 MAE value as 197163.91 R-Square value as 0.24, Logistic regression test set RMSE values shows 391029.89 MAE value as 227012.91 , Linear regression test set RMSE values shows 224146.84 MAE value as 123013.89 , LDA model test set RMSE values shows 224146.84 MAE value as 123013.89 , Decision tree test set RMSE values shows 185940.23 MAE value as 101114.45 , Boosting test set RMSE values shows 129410.93 MAE value as 76841.13 and Gradient Boosting test set RMSE values shows 160496.34 MAE value as 97368.76

Lasso Regression :   
  
It is a linear regression technique that is used for feature selection and can help mitigate the issue of multicollinearity in regression models.   
  
  
A picture containing text, font, line, screenshot

Description automatically generated  
Output 1.62 Lasso Regression

Best estimator :   
  
  
Output 1.63 Best estimator  
  
  
  
Ridge Regression :   
  
It is a linear regression technique which uses squared values but it is similar to lasso regression.   
  
  
Output 1.64 Ridge Regression  
  
  
Best estimator :   
  
  
Output 1.65 Best estimator

c. Interpretation of the most optimum model and its implication on the business   
  
Ensemble methods are really performing better than linear models. Of all the ensemble models, Gradient boosting regressor is giving better R2 score. The important features that mostly drive the price of the property are: 'furnished', 'year\_built', 'living\_measure','quality', 'lot\_measure15', 'ceil\_measure', 'total\_area'. Best of all models is Xtreme Gradient boost which is an enhanced version of gradient boost. It includes regularization and is faster too. Going forward this model can be improved further as we don’t have much data for very high-priced houses. So, when more data comes in we can revisit our model and make necessary changes to accommodate more variation in data to deliver better results by decreasing RMSE values. For further improvisation, the datasets can be made by treating outliers in different ways and hyper tuning the ensemble models.One needs to thoroughly check its property on parameters suggested and list its price accordingly, similarly if one wants to buy house, they need to check the features suggested for the house and calculate the predicted price. The same can then be compared to listed price. Making polynomial features and improvising the model performance can also be explored further.