Advanced Regression Assignment

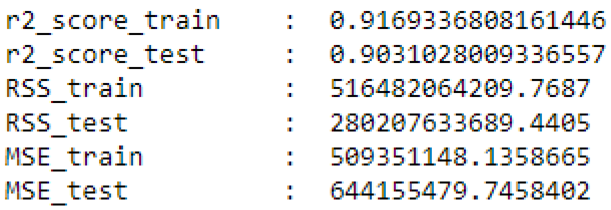
Subjective Questions

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

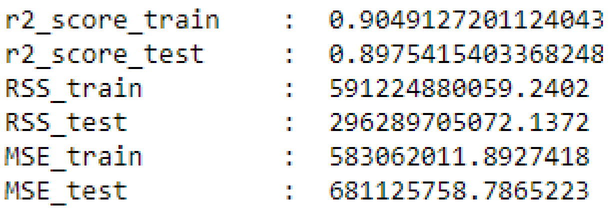
Ridge Regression:

- Optimal Value of alpha = **5.0**



```
r2_score_train     :   0.9169336808161446
r2_score_test      :   0.9031028009336557
RSS_train          :   516482064209.7687
RSS_test           :   280207633689.4405
MSE_train          :   509351148.1358665
MSE_test           :   644155479.7458402
```

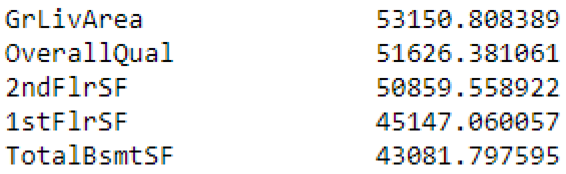
- On doubling alpha (i.e., alpha = **10.0**)



```
r2_score_train     :   0.9049127201124043
r2_score_test      :   0.8975415403368248
RSS_train          :   591224880059.2402
RSS_test           :   296289705072.1372
MSE_train          :   583062011.8927418
MSE_test           :   681125758.7865223
```

  We notice that the R^2 value of both train and test have gone down marginally while the RSS and MSE have increased

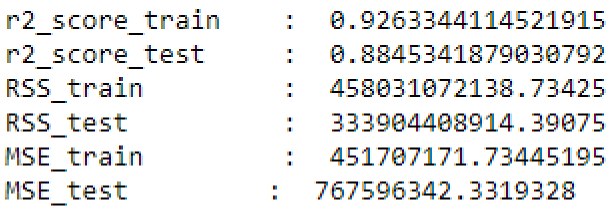
- The most important predictor variables are



```
GrLivArea                 53150.808389
OverallQual               51626.381061
2ndFlrSF                  50859.558922
1stFlrSF                  45147.060057
TotalBsmtSF               43081.797595
```

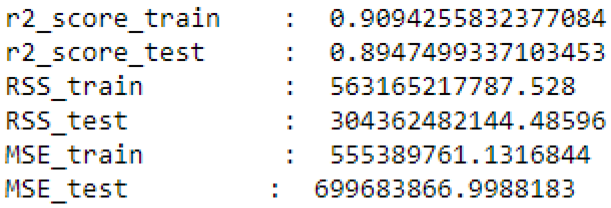
Lasso Regression:

- Optimal Value of alpha = **100.0**



```
r2_score_train     :   0.9263344114521915
r2_score_test      :   0.8845341879030792
RSS_train          :   458031072138.73425
RSS_test           :   333904408914.39075
MSE_train          :   451707171.73445195
MSE_test         :   767596342.3319328
```

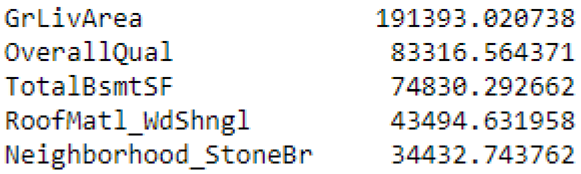
- On doubling alpha (i.e., alpha = **200.0**)



```
r2_score_train     :  0.9094255832377084
r2_score_test      :  0.8947499337103453
RSS_train          :  563165217787.528
RSS_test           :  304362482144.48596
MSE_train          :  555389761.1316844
MSE_test         :  699683866.9988183
```

  We notice that, while the R^2 value of the training data has gone down, the same value for the test data has gone up marginally.

- The most important predictor variables are



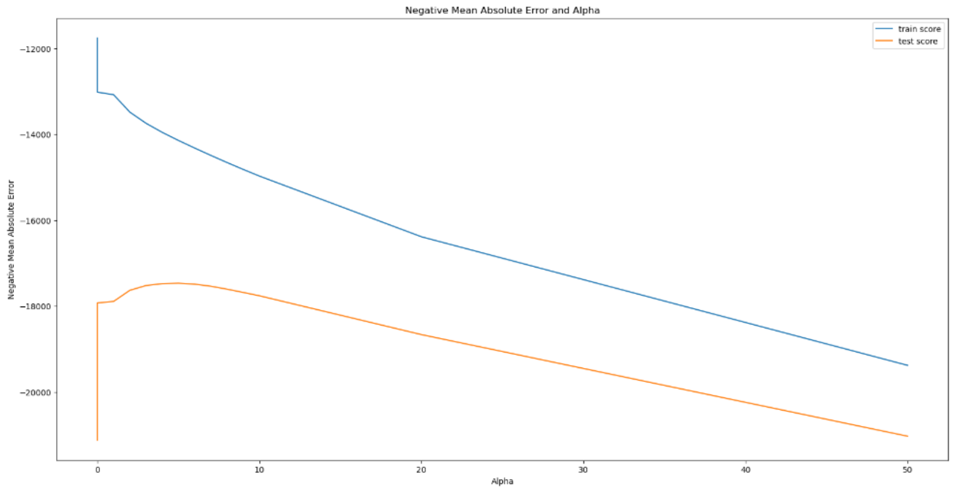
```
GrLivArea                191393.020738
OverallQual               83316.564371
TotalBsmtSF               74830.292662
RoofMatl_WdShngl          43494.631958
Neighborhood_StoneBr      34432.743762
```

Question 2

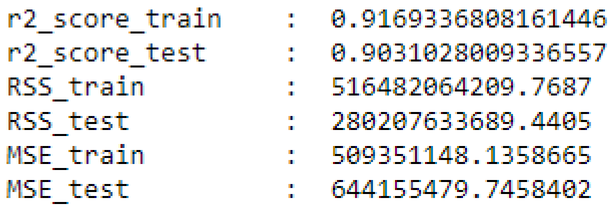
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

We can draw conclusions with the help of the following graphs

Ridge



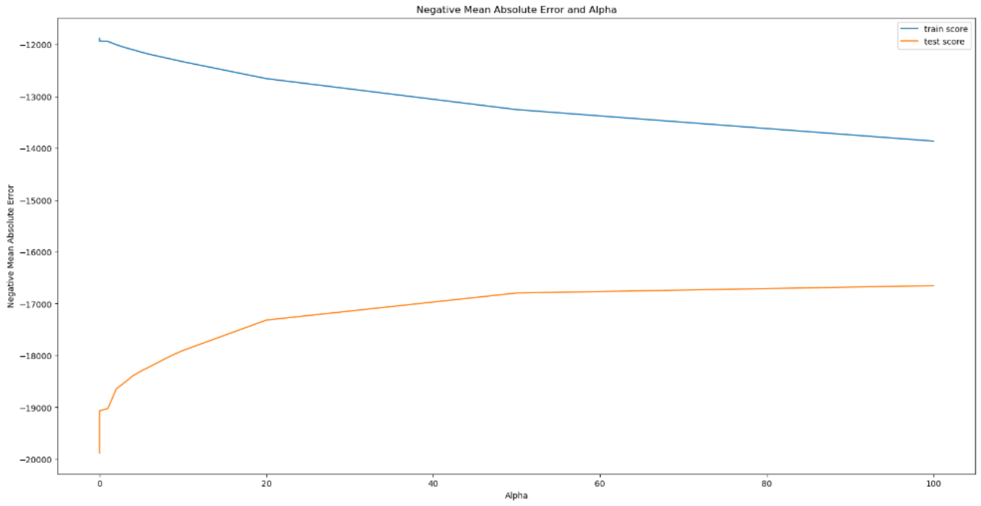
We notice that the Negative Mean absolute errors are least simultaneously at around alpha = 5



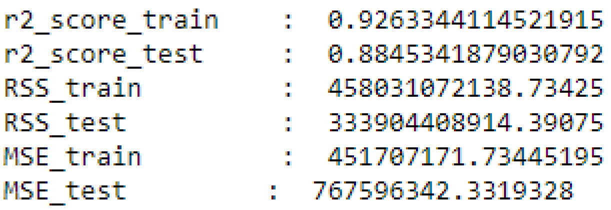
```
r2_score_train      :  0.9169336808161446
r2_score_test       :  0.9031028009336557
RSS_train           :  516482064209.7687
RSS_test            :  280207633689.4405
MSE_train           :  509351148.1358665
MSE_test            :  644155479.7458402
```

Both R^2 values of Train and Test data are around 0.9, which can be considered an acceptable value.

Lasso



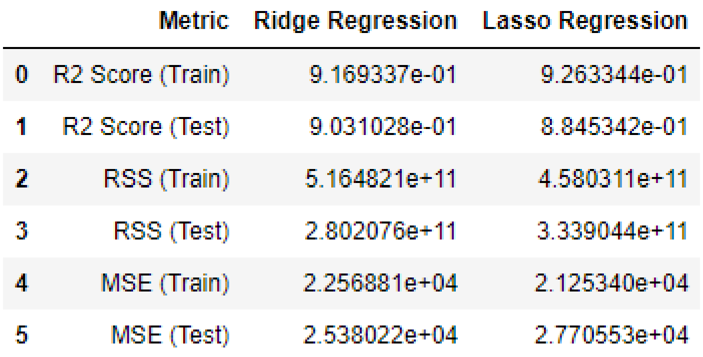
We notice that as the value of alpha increases, the error term increases on the test data and decreases on the training data. This would indicate that around alpha = 100, we are able to strike a pretty good balance to represent the trends of the data.



```
r2_score_train    :  0.9263344114521915
r2_score_test     :  0.8845341879030792
RSS_train         :  458031072138.73425
RSS_test          :  333904408914.39075
MSE_train         :  451707171.73445195
MSE_test        :  767596342.3319328
```

The R^2 values of Train and Test data are around 0.92 and 0.88, which can be considered an acceptable value.

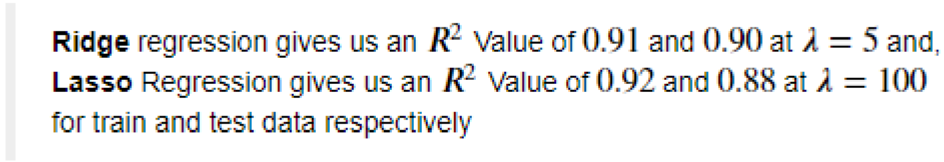
On comparing the two models, it would seem that there is only a subtle difference.



	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.169337e-01	9.263344e-01
1	R2 Score (Test)	9.031028e-01	8.845342e-01
2	RSS (Train)	5.164821e+11	4.580311e+11
3	RSS (Test)	2.802076e+11	3.339044e+11
4	MSE (Train)	2.256881e+04	2.125340e+04
5	MSE (Test)	2.538022e+04	2.770553e+04

The decision of which model can be chosen can be done based on the business requirements.

Although Lasso Regression has a smaller R^2 Value on unseen data when compared to Ridge Regression, it also has much lesser variables and hence lesser model complexity.



> **Ridge** regression gives us an R^2 Value of 0.91 and 0.90 at $\lambda = 5$ and,
> **Lasso** Regression gives us an R^2 Value of 0.92 and 0.88 at $\lambda = 100$
> for train and test data respectively

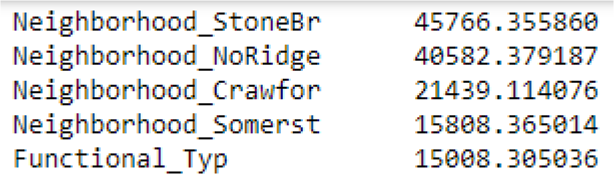
Question 3

After building the model, you realised that the five most important predictor variables in the lasso model is not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

After removing the current top 5 predictor variables and re-creating the models, we observe the following.

Lasso:

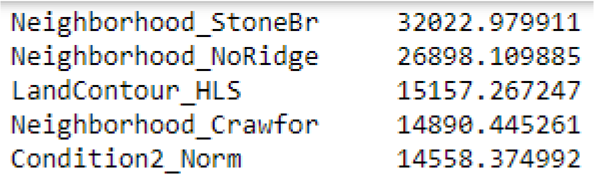
New top 5 variables of Lasso are:



```
Neighborhood_StoneBr      45766.355860
Neighborhood_NoRidge      40582.379187
Neighborhood_Crawfor      21439.114076
Neighborhood_Somerst      15808.365014
Functional_Typ            15008.305036
```

Ridge:

New top 5 variables of Ridge are:



```
Neighborhood_StoneBr     32022.979911
Neighborhood_NoRidge     26898.109885
LandContour_HLS          15157.267247
Neighborhood_Crawfor     14890.445261
Condition2_Norm          14558.374992
```

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

To ensure that a regression model is robust and generalizable, we can start by collecting a sufficient amount of data that represents the problem. We should choose relevant features that have a strong impact on the target variable. Performing exploratory data analysis by addressing missing values properly and also imputing values wherever possible**.** This involves pre-processing the dataset.

Normalizing or standardizing the numerical features to bring them to a similar scale can prevent certain features from dominating the model's learning process due to their larger magnitudes. Utilize cross-validation techniques, such as k-fold cross-validation. This helps evaluate the model's ability to generalize to unseen data.

Regularization techniques like Lasso or Ridge regularization can prevent overfitting and improve the model's generalization capability.

The model's performance should be evaluated on a separate validation set or through deployment in real-world scenarios. This evaluation provides insights into the model's generalization capability beyond the training data. We split the dataset into training and testing sets or use techniques like k-fold cross-validation to obtain reliable performance estimates. Additionally, we use evaluation metrics such as mean squared error (MSE), mean absolute error (MAE), or R-squared to quantify the model's accuracy.

The implications of building a robust and generalizable regression model are improved accuracy and reliability. A robust model is less prone to overfitting, where it becomes too specialized in the training data and fails to generalize well to new, unseen data.

A robust and generalizable model can provide more accurate predictions when applied to new data points or future scenarios. It can handle variations and outliers effectively, allowing for more reliable insights and decision-making.