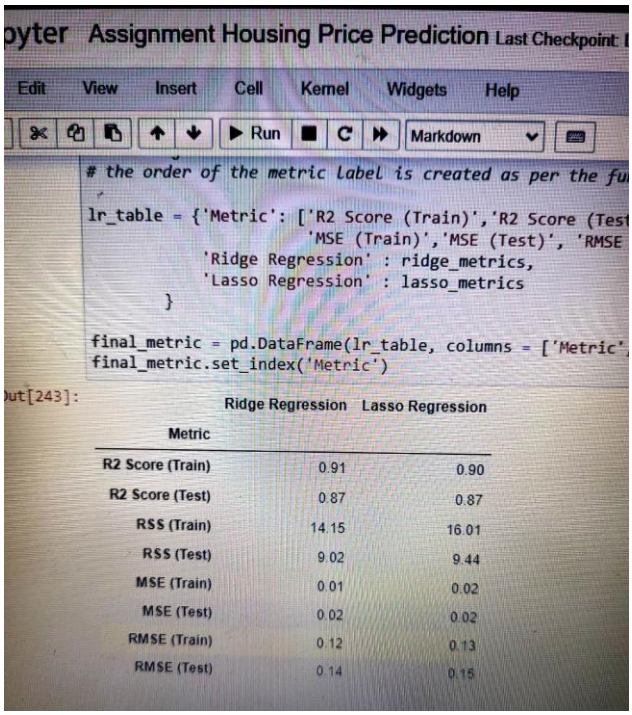
Question 1

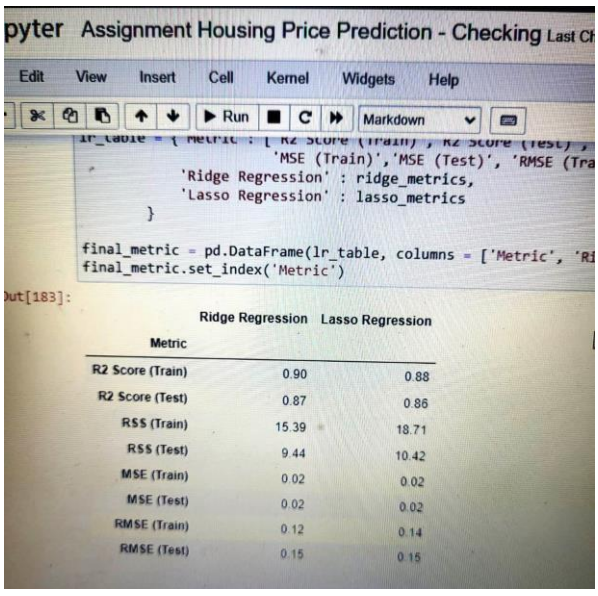
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

The optimum value of alpha for Ridge Regression was 10 and for Lasso it was 0.001 as per my model. Now say I double the alpha value for both ridge and lasso model then below are the change observed.

Change 1: Picture of Metric values for Ridge and Lasso model before and after doubling the alpha value



Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.91	0.90
R2 Score (Test)	0.87	0.87
RSS (Train)	14.15	16.01
RSS (Test)	9.02	9.44
MSE (Train)	0.01	0.02
MSE (Test)	0.02	0.02
RMSE (Train)	0.12	0.13
RMSE (Test)	0.14	0.15



Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.90	0.88
R2 Score (Test)	0.87	0.86
RSS (Train)	15.39	18.71
RSS (Test)	9.44	10.42
MSE (Train)	0.02	0.02
MSE (Test)	0.02	0.02
RMSE (Train)	0.12	0.14
RMSE (Test)	0.15	0.15

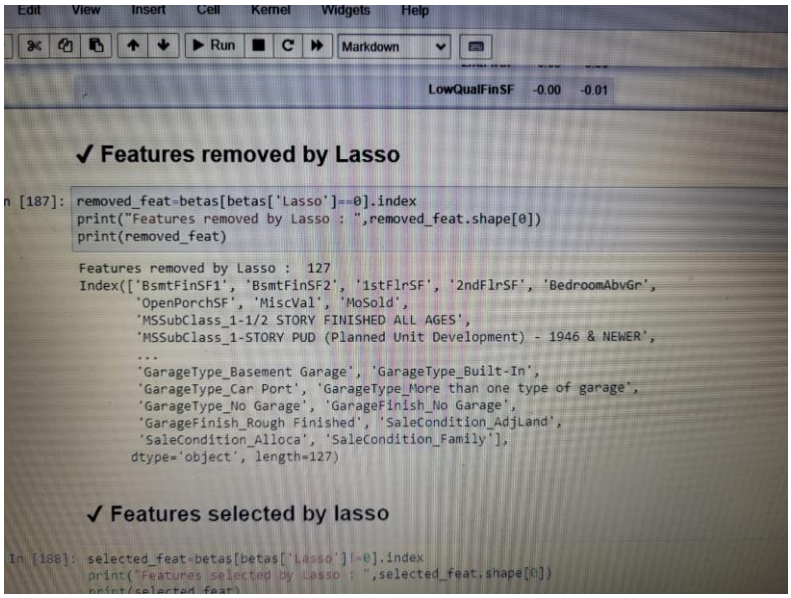
Ridge Training Data: Before changing the alpha value the R squared value for training data of my model was 0.91 which on doubling the alpha was changed to 0.90 (**Decreased**)

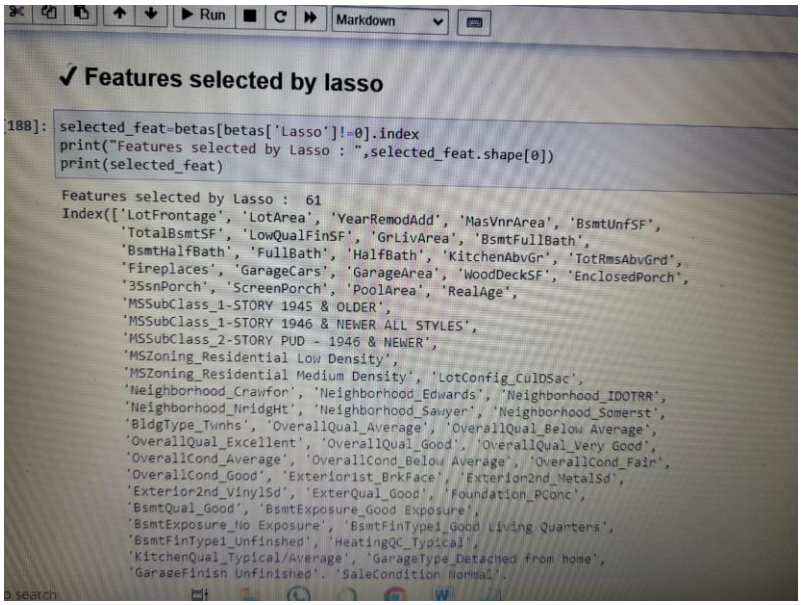
Ridge Test Data: R squared value remained unchanged as 0.87 (**No change**)

Lasso Training Data: Before changing the alpha value the R squared value for training data of my model was 0.90 which on doubling the alpha was changed to 0.88 (**Decreased**)

Lasso Test Data: Before changing the alpha value the R squared value for test data of my model was 0.87 which on doubling the alpha was changed to 0.86 (**Decreased**)

<u>Change 2:</u> Also the features removed and selected by Lasso regression was change on changing the alpha value. Previously the lasso model has removed 104 variables and has selected 84. Now on changing the alpha, 127 variables were removed and 61 were selected.





Change 3: The top 5 variables selected by Ridge and Lasso regression before and after changing the alpha value

Before changing the alpha

Ridge:
OverallQual_Excellent 1.14
OverallQual_Very Good 1.11
Neighborhood_Crawfor 1.10
Neighborhood_NridgHt 1.10
Neighborhood_Somerst 1.08
Lasso :
OverallQual_Excellent 1.21
OverallQual_Very Good 1.13
Neighborhood_Crawfor 1.12
GrLivArea 1.12
Neighborhood_NridgHt 1.10

After changing the alpha
OverallQual_Excellent 1.11
OverallQual_Very Good 1.09
Neighborhood_Crawfor 1.09
Neighborhood_NridgHt 1.08
GrLivArea 1.06

OverallQual_Excellent 1.17
GrLivArea 1.12
OverallQual_Very Good 1.12
Neighborhood_Crawfor 1.09
Neighborhood_Somerst 1.07

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?

The model we choose will depend on the objective of the problem. If we have too many features and our objective is feature selection then we will go with Lasso regression. If our motive is to reduce the coefficient magnitude then we will prefer Ridge regression. Here our objective is to select the optimum variables so we will go with Lasso regression model.

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are 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?

In this case we need to rebuild the model with those variables removed . The top 5 variables now selected by lasso regression is as below.

2ndFlrSF 1.10
Neighborhood_Somerst 1.08
1stFlrSF 1.08
Exterior1st_BrkFace 1.07
BsmtExposure_Good Exposure 1.06

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?

A model is said to be robust when the variation in the data doesn't affect the performance of the model much. Also we need to make sure that we do not complex the model design as this can lead to over fitting problem wherea slight variance change can lead to a drastic change in the prediction. An easy way to identify this issue is that the model will perform very good with nearly accurate on test data but fails when met with real data sets.

When we talk about accuracy, a too complex model will be very accurate with less bias but this might not perform well in real time . So we need to make sure that we create a balance between variance and bias which will result will a little less accurate model but it will be able to generate good result. To do this we make use of regularization technique like ridge and lasso regression .