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?

Answer:

Optimal value for alpha for ridge is 0.4 in my model.

Optimal value for alpha for Lasso Regression is 0.0001 in my model

After using double value of this, that is 0.8 for Ridge and 0.0002 for Lasso respectively, I get below coefficients as shown in the image below.

Please note that the coefficients in Ridge will try to tend to zero.

In Lasso, three features coefficients have been removed by making them zero.

	LinearRegression	Ridge	Lasso
YearBuilt	0.1874	0.188555	0.190313
BsmtFinSF1	0.1306	0.118155	0.105387
TotalBsmtSF	0.1829	0.172862	0.164653
GrLivArea	0.5384	0.504298	0.521078
KitchenAbvGr	-0.1678	-0.153321	-0.151260
GarageArea	0.1271	0.138813	0.132486
MSSubClass_ONE_HALF_STORY_UNFINISHED_ALL_AGES	0.0378	0.030076	0.015435
MSSubClass_TWO_STORY_PUD_1946_NEWER	-0.0682	-0.067444	-0.062683
MSZoning_FV	0.0271	0.026082	0.019282
Condition2_PosN	-0.4110	-0.217164	-0.186285
OverallQual_Very_Poor	-0.1129	-0.063872	-0.000000
RoofMatl_Membran	0.1473	0.082020	0.000000
RoofMatl_WdShngl	0.0888	0.078800	0.044631
Exterior1st_BrkComm	-0.1830	-0.133772	-0.085049
Functional_Sev	-0.1913	-0.103115	-0.000000

	LinearRegression	Ridge	Lasso
Train_R2_Score	0.827051	0.822897	0.817207
Test_R2_Score	0.772742	0.787589	0.787267
Train_RSS	2.936099	3.006607	3.103218
Test_RSS	1.739829	1.626162	1.628629
Train_MSE	0.002876	0.002945	0.003039
Test_MSE	0.003963	0.003704	0.003710

The important predictor will still be GrLivArea

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?

Answer:

When we use optimal value for lambda for ridge and lasso, we get below metrics. We can note the the difference between Train R2 Score and Test R2 Score is less in the case of Lasso. Hence I would use the Lasso Regression.

	LinearRegression	Ridge	Lasso
Train_R2_Score	0.827051	0.825468	0.823791
Test_R2_Score	0.772742	0.784030	0.782237
Train_RSS	2.936099	2.962973	2.991440
Test_RSS	1.739829	1.653415	1.667141
Train_MSE	0.002876	0.002902	0.002930
Test_MSE	0.003963	0.003766	0.003798

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?

Answer:

When optimal lambda is used for ridge and lasso, I got the below predictor variables as five most important variables.

1. **GrLivArea –** The SalePrice of the property increases by 0.529698 times when the value of GrLivArea increases.
2. **Condition2_PosN – The SalePrice of the property decreases by 0.298 times as there is negative correlation with this feature.**
3. **YearBuilt** – The SalePrice of the property will be high for the new properties. The YearBuilt will indicate the age and is having a positive correlation with value 0.1889 times with SalePrice
4. **TotalBsmtSF** – The SalePrice will increase by 0.174 times with increase in Basement squarefeet area.
5. **KitchenAbvGr** – This field indicates Kitches above grade. The model shows this as a negatively correlated feature with SalePrice by 0.159 times. With increase in the value of **KitchenAbvGr**, there is decrease in Sale Price of the property.

The above mentioned predictor variables have to be removed and the model has to be re-built.

The table below is shown as a reference with coefficient details using optimal value of lambda for lasso and ridge.

	LinearRegression	Ridge	Lasso
YearBuilt	0.1874	0.188115	0.188936
BsmtFinSF1	0.1306	0.123234	0.117430
TotalBsmtSF	0.1829	0.177137	0.174338
GrLivArea	0.5384	0.520009	0.529698
KitchenAbvGr	-0.1678	-0.160104	-0.159390
GarageArea	0.1271	0.133540	0.130324
MSSubClass_ONE_HALF_STORY_UNFINISHED_ALL_AGES	0.0378	0.033620	0.026741
MSSubClass_TWO_STORY_PUD_1946_NEWER	-0.0682	-0.067857	-0.065341
MSZoning_FV	0.0271	0.026558	0.023104
Condition2_PosN	-0.4110	-0.285603	-0.298759
OverallQual_Very_Poor	-0.1129	-0.081384	-0.011212
RoofMatl_Membran	0.1473	0.105356	0.045468
RoofMatl_WdShngl	0.0888	0.083641	0.066638
Exterior1st_BrkComm	-0.1830	-0.154464	-0.133997
Functional_Sev	-0.1913	-0.134473	-0.088123

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?

Answer:

1. The model is said to be robust and generalizable if R2 score is close to 100%. In my case, it is 82% for Train data and 77% for the test data.

   It says that the model will be predicting Sale Price accurately for 82% of the data.

2. Also the difference between R2 Score of Train data and R2 Score if TEST data should be very less.
3. The MSE (Mean Square Error) value should be as as less as possible. The lesser the value, the better is the model. In my case, it is 0.002 to 0.003, which is good.

The metrics for the developed model with optimal value of lambda is shown below for reference:

	LinearRegression	Ridge	Lasso
Train_R2_Score	0.827051	0.825468	0.823791
Test_R2_Score	0.772742	0.784030	0.782237
Train_RSS	2.936099	2.962973	2.991440
Test_RSS	1.739829	1.653415	1.667141
Train_MSE	0.002876	0.002902	0.002930
Test_MSE	0.003963	0.003766	0.003798