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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

From the ridge lasso regression following observations are made

1. The optimal value of alpha for lasso regression is 0.0001.
2. The optimal value of alpha for Ridge regression is 5.

When alpha value is doubled, there is reduction in R2 score and beta coefficients. Below image shows a comparison table of the same. In the table below Lasso and Ridge represent optimal values of alpha, Lasso_1 and Ridge_1 represent values after doubling the alpha value. The second table shows the comparative values of coefficient for the first 20 predictors.

	Linear	Lasso	Ridge	Lasso_1	Ridge_1
0	8.435222e-01	0.838933	0.834500	0.833505	0.825774
1	-3.164997e+23	0.823391	0.834811	0.829243	0.823720

	Linear coefficient	Lasso coefficient	Ridge coefficient	Lasso1 coefficient	Ridge1 coefficient
0	-0.026117	-0.021632	-0.019490	-0.020094	-0.018515
1	-0.017566	-0.017531	-0.013851	-0.015837	-0.011579
2	-0.035582	-0.000000	-0.000809	-0.000000	0.005749
3	0.094295	0.019256	0.024259	0.000000	0.016819
4	-0.004812	-0.000000	-0.001842	0.000000	-0.000621
5	-0.013150	-0.000000	-0.003215	0.000000	0.000564
6	-0.072336	-0.000000	-0.009966	-0.000000	-0.005739
7	-0.000172	-0.000000	0.000317	0.000000	0.001113
8	0.012781	0.007912	0.012849	0.005783	0.013197
9	-0.014497	-0.001616	-0.006833	-0.000000	-0.005457
10	-0.156013	-0.054356	-0.025978	-0.000000	-0.013511
11	-0.016757	-0.018712	-0.019754	-0.021394	-0.019622
12	-0.029498	-0.028876	-0.026426	-0.026920	-0.022828
13	0.152406	0.159691	0.115576	0.162990	0.095198
14	0.060042	0.040257	0.036997	0.025649	0.027440
15	0.016168	0.000000	0.000663	0.000000	0.001953
16	-0.001701	0.004110	0.008695	0.006703	0.011404
17	0.016257	0.014325	0.022388	0.011993	0.023315
18	0.054994	0.056362	0.047470	0.042932	0.035372
19	-0.000697	-0.000000	-0.002322	-0.000000	-0.002895

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:

I will choose lasso regression. As some coefficients will become zero in lasso regression and insignificant features will be eliminated automatically, so we do not need to do feature selection by other methods, which is more time consuming if number of features are more.

Question 3

After building the model, you realized 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:

As per lasso regression, the features which are having greatest coefficients are important variables. Below are the five most important coefficients in the current model.

Feature	**Description**
GrLivArea	Above grade (ground) living area square feet
OverallQual	overall material and finish of the house
MasVnrArea	Masonry veneer type
BsmtQual	the height of the basement
GarageCars	Size of garage in car capacity

After removing the above five variables and again training the model we get below five most important variables. These are the most important variables chosen as there coefficient are high compare to other variables.

Feature	**Description**
1stFlrSF	First Floor square feet
2ndFlrSF	Second floor square feet
Condition2	Proximity to various conditions (if more than one is present)
TotalBsmtSF	Total square feet of basement area
GarageArea	Size of garage in square feet

Question 4

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

Answer:

The model is robust when there is not much difference in performance of the trained model when there is change in a data. To make the model more robust and generalizable we need to take care that it doesn't overfit.

If the model is too complex then the model will be more accurate on trained data but there will be high variance in the model, and if it is so simple or underfit then there is large bias in the model. Both the conditions are not ideal for the model so we need to find the optimal point between high variance and high bias so that model can be more generalizable and robust. There are various techniques to do it such as cross validation, ridge & lasso regression.