Subjective Questions: Advanced ML

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 following are the optimal values for ridge and lasso.
Ridge regression : 3
Lasso Regression : 100

In Ridge regularisation, when the alpha value is doubled, the value of the coefficients will decrease. Increasing the value of alpha will result in increasing the penalty which in turn reduces the coefficients.

In Lasso regularisation, there is an inbuilt feature selection available. Hence, some of the feature will be 0 by default. If we still increase the alpha value, the values of the remaining coefficients will continue to decrease.

The following are the top 10 variables For Ridge with the alpha of 3

S.No	Feature	Coefficient
1	OverallQual_9	28406.66279
2	Neighborhood_StoneBr	28087.50159
3	OverallCond_3	-21491.37182
4	Neighborhood_Crawfor	21426.28981
5	MSSubClass_160	-19317.48794
6	KitchenQual_Fa	-18119.94498
7	MSZoning_FV	16452.0979
8	KitchenQual_Gd	-15275.45454
9	BsmtQual_Gd	-15241.7417
10	OverallQual_10	14593.10049

The following are the top 10 variables for Ridge with the alpha of 6

S.No	Feature	Score
1	OverallQual_9	25222.29521
2	Neighborhood_StoneBr	23536.60632
3	Neighborhood_Crawfor	18988.19709
4	OverallCond_3	-16473.75002
5	MSSubClass_160	-15095.86803
6	KitchenQual_Gd	-14121.19458
7	BsmtQual_Gd	-13973.92087
8	MSZoning_FV	12319.82539

9	KitchenQual_Fa	-12069.38493
10	KitchenQual_TA	-11704.18038

For Lasso:

The following are the top 10 important variables with the optimal value of alpha of 100

S.No	Feature	Coefficient
1	OverallQual_9	39155.50145
2	Neighborhood_StoneBr	33965.10891
3	Neighborhood_Crawfor	30209.61846
4	OverallCond_3	-20875.7671
5	MSSubClass_160	-17575.90023
6	OverallQual_8	15565.1302
7	MSZoning_FV	15391.19877
8	Neighborhood_NridgHt	14323.57183
9	OverallQual_10	13395.21605
10	KitchenQual_Gd	-12752.98201

The following are the top 10 important variables when the alpha is doubled to 200 for Lasso.

S.No	Feature`	Coefficient
1	OverallQual_9	39260.78497
2	Neighborhood_Crawfor	27851.42467
3	Neighborhood_StoneBr	26848.61126
4	OverallQual_8	15622.89391
5	MSZoning_FV	11570.21549
6	Neighborhood_NridgHt	10781.32496
7	BsmtQual_Gd	-10523.22032
8	MSSubClass_160	-10333.73219
9	KitchenQual_Gd	-9900.359441
10	KitchenQual_TA	-8434.182075

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:

The optimal value for Ridge is 3 and Lasso is 100.
We will choose Lasso over ridge due to the following reasons.

a) Mean square error is lower for lasso than Ridge
b) R2 value of Lasso is higher than ridge with lower number of features in Test dataset.
c) Lasso performs better than Ridge in unseen data.
d) Lasso has feature selection that helps in keeping the model simple by making some coefficients as 0.

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?

After removing the 5 most important variables and rebuilding the model with lasso regression, the following are the 5 most important variable.

S.No	Feature	Coefficient
1	Neighborhood_MeadowV	-39508.40067
2	OverallQual_6	-35838.06933
3	OverallQual_4	-35507.40806
4	OverallCond_8	35461.88708
5	OverallQual_5	-35341.25673

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 considered robust if it is stable.
That is , the model should perform well in the unseen data.
The model should not change when there is any change in training data.
In addition, we should have a proper trade-off between bias and Variance.
If the model is stable, there should not much change in the accuracy of the data when it is run on test and training data.