Assignment Part II – 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?

Ans:
We got optimal value of alpha for
Ridge = **1.0**
Lasso = **0.0001**
Top 5 predictor variables for with current alpha values, when Ridge = **1.0**

	Co-Efficient
OverallQual_9	0.362559
MSZoning_RL	0.301544
MSZoning_RH	0.284979
MSZoning_FV	0.245181
MSZoning_RM	0.224300

Top 5 predictor variables when Lasso = 0.0001

MSZoning_RL	0.397898
MSZoning_RH	0.391060
OverallQual_9	0.370370
MSZoning_FV	0.353722
MSZoning_RM	0.319172

Changes in the model after doubling the value of alpha for both Ridge and Lasso are ,

- We observed that though accuracy and RSS almost remain same, however there is a slight increase on mean squared error, and some features are replaced with new ones.
- Features highlighted in Yellow are removed and which are highlighted in green are newly added.
- There are no changes in terms of predictor variable in the case of **Lasso**

Top 5 predictors
With Ridge_double = **2,**

OverallQual_9	0.329849
MSZoning_RL	0.232590
MSZoning_RH	0.205927
OverallQual_8	0.191166
Neighborhood_Crawfor	0.179342

Lasso_double = **0.0002**

OverallQual_9	0.365278
MSZoning_RL	0.330301
MSZoning_RH	0.314614
MSZoning_FV	0.281567
MSZoning_RM	0.250498

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?

Ans:
Lasso regression would be a better option it would help in feature elimination and the model will be more robust. From the given assignment, the following are the main reasons for this decision

- No: of predictor variable is less than of Ridge regression, we did regression on top 70 features , however in the lasso model only 58 are used for prediction.
- Mean squared error is less when compared to Ridge regression.
- Test accuracy is decent enough for a generalized 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?

Ans:

After removing top 5 predictor column from both training and testing data, performed a Lasso regression with optimal alpha = 0.0001, the new predictor variables are as follows.

Neighborhood_Crawfor
OverallCond_9
Heating_GasW
GarageQual_Gd
GrLivArea

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?

Ans:
There are several ways to ensure that a machine learning model is robust and generalizable:
Cross-validation: Use k-fold cross-validation to evaluate the model's performance on different subsets of the data. This will give you a better idea of how well the model will perform on unseen data.
Regularization: Use techniques such as L1 (Lasso)or L2 (Ridge) regularization to prevent overfitting and make the model more generalizable.
Early stopping: Monitor the performance of the model on a validation set during training and stop training when the performance starts to degrade.

Test on a diverse set of data: Make sure to test the model on a diverse set of data, including examples that are different from those in the training set in terms of distribution and variability.

If a model has high generalization, it means that it will perform well on new data, making accurate predictions or classifications. In other words, a model with high generalization will have a low test error.

The implications of poor generalization for the accuracy of the model are that it may have a high training accuracy, but it will not perform well on new data, which is the main purpose of the model, to generalize to new unseen data. This is a problem because it means that the model is overfitting, and it has learned the noise in the training data.