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: Ridge Regression: when the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term decrease ad the train error is showing increasing trend when value of alpha increases. When the value of alpha is 2 the test error is minimal so we can have the alpha value as 2 for ridge regression.

Lasso Regression: If we assume very small value as 0.01, when we increase the value of alpha the model try to penalise more and try to make most of the coefficients value zero. When we double the value of alpha for our ridge regression it will lead to increase in errors. Similarly, when we increase the value of alpha for Lasso to penalise more our model and more coefficient of the variable will be reduced to zero, when we increase the value of r2 square also decreases.

The most important variable after the changes has been implemented for ridge regression are as follows:

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_5toneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

The most important variable after the changes has been implemented for lasso regression are as follows:

1.GrLivArea

2.OverallQual

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: It is necessary and very important step to regularise coefficients and improve the prediction

With the decrease in variance.

Ridge Regression: This method uses a tuning parameter lambda, as the penalty is square of magnitude of coefficients, which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the

coefficients that have greater values will be penalised. As we increase the value of lambda, the variance in model is dropped and bias remains constant. Ridge includes all the variables in the final 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?

Answer: The 5 most important predictor variables that will be excluded are:

1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmt5F
5. GarageArea

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: The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable.It can be also understood using the Bias- Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. It can be inferred that a robust and generalised model will perform equally well on both training and test data. I.e the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs extremely well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model. It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.