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 1:

Alpha Value Ridge – 1.0

Alpha Value Lasso – 0.001

Making Values Double for Ridge and Lasso

Lasso's R2 square value decreased by 0.1% for both train and test data. Coefficient values decreased for some while increased for others.

Ridge's R2 square value decreased by 0.6% for Train and 1% for test data. Coefficient values decreased for some while increased for others.

Top Parameters:-

Exterior1st_AsphShn, Neighborhood_OldTown, Condition1_Norm, Condition2_PosN, Neighborhood_NridgHt

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 2:

Values of both the solution is similar but we will choose Lasso as it gives the benefit to remove unwanted features from the model with the same accuracy.

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 3:

Old Top Five Features –

Exterior1st_AsphShn, Neighborhood_OldTown, Condition1_Norm, Condition2_PosN, Neighborhood_NridgHt

New Top Five Features –

Exterior1st_BrkFace, Condition2_RRAn , SaleType_Oth, Neighborhood_NAmes,Condition2_Norm

R2 Squared value decreased by 2%.

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 4:

No Overfitting – Model should provide the similar test results when compared with the training results.

R2 Square Value – The R2 square values should be high in the range of 0 to 1 or the RSS value should not be higher which basically shows the residual error or the difference between actual value and predicted value.

Multicollinearity – Model should have less or no Multicollinearity which means the predictors should not be dependent on each other. It can be measured using VIF factor with value <5.

P Value – P values of all the predictors should be less than 0.05.