ADVANCE REGRESSION ASSIGNMENT

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:- The optimal value of alpha for ridge and lasso for the given assignment obtained is 4 and 50. When we double the value of alpha in Lasso regression the coefficient value tend to go towards zero. Even in ridge regression the impact of shrinkage grows and value tend to go towards zero. The most important predictor variables are : OverallQual_9, OverallQual_10

- SaleCondition_Alloca
- Neighborhood_StoneBr
- Functional_Mod
- Neighborhood_Crawfor
- Neighborhood_MeadowV
- KitchenQual_TA
- KitchenQual_Gd
- SaleCondition_Partial
- Exterior1st_BrkFace
- Functional_Typ
- OverallQual_8
- Condition1_RRAe
- KitchenQual_Fa
- MSSubClass_160
- BsmtExposure_Gd
- MSZoning_FV
- Neighborhood_NoRidge
- Neighborhood_BrkSide
- OverallCond_8
- MasVnrType_Stone
- Exterior1st_CemntBd
- OverallCond_3
- BsmtFinType1_GLQ
- Neighborhood_Edwards
- OverallCond_5
- Condition1_Norm
- BsmtQual_TA
- GarageCond_Fa

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:- After finding the optimal value for ridge and lasso regression model and analysing the coefficient value I came to a conclusion that lasso regression model has a better coefficient value as most of the variable tend towards zero. This means that feature selection is helpful for me in this case hence I will opt for lasso regression instead.

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:- The 5 most important predictor variable that I consider and would remove from the dataset will be: 11stFlrSF-----------First Floor square feet 2. GrLivArea-----------Above grade (ground) living area square feet 3. Street_Pave---------Pave road access to property 4. RoofMatl_Metal------Roof material_Metal 5. RoofStyle_Shed------Type of roof(Shed)

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: The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for

datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy

predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant

to the dataset need to be retained. Those outliers which it does not

make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis