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

The optimal value of alpha for ridge regression is **8.0**, while the optimal value of alpha for lasso regression is **0.0001**.

When we increase alpha, the both regression algorithms will increase the penalty and thus reduce more coefficients toward zero.

The ten most important predictors will be 'Neighborhood_NoRidge','GrLivArea', '2ndFlrSF', 'FullBath', 'TotRmsAbvGrd', '1stFlrSF', 'GarageCars', 'Fireplaces', 'Neighborhood_NridgHt', 'BsmtExposure_Gd' in ridge regression

The ten most important predictors will be 'GrLivArea','RoofMatl_WdShngl','Neighborhood_NoRidge', 'GarageCars', 'Neighborhood_NridgHt', 'Neighborhood_StoneBr', 'BsmtExposure_Gd', '2ndFlrSF', 'Neighborhood_Somerst', 'Neighborhood_Crawfor' in lasso regression

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

We will choose lasso regression over ridge regression because lasso's r2 score is better than ridge's r2 score

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

So we have created the new model by removing top 5 features from dataset in the notebook and redoing lasso regression on it, we get TotalBsmtSF, TotRmsAbvGrd, MasVnrArea, LotArea, FullBath as 5 most predictor variables

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

We should make sure that variance in training data set is not that different with variance explained for testing dataset. The choice of lambda parameter is also an important factor in making sure model is robust. If the lambda is too high, the model might have lot of features with coefficient zero, thus might not produce best result, if the lambda is too low, the model might over learn the training data and thus will not perform well is the testing data