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?

Answer:

The optimal value of alpha for ridge and lasso regression are 10 and 0.001 respectively.

If we chose double the value of alpha for both ridge and lasso, the R2 score for both train and test sets decrease by less than 1% and RSS and RMSE increases by less than 1%.

The most important predictor variables after the change is implemented are: 'GrLivArea', 'Neighborhood_Crawfor', 'Foundation_PConc', 'OverallQual', 'Neighborhood_NridgHt', 'MSZoning_RL', 'SaleType_New', 'GarageCars', 'SaleCondition_Normal', 'MSZoning_FV'

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:

By comparing metrics of Ridge Regression and Lasso Regression, we see that R2 score in Ridge Regression is higher than that in Lasso Regression, and RSS (Residual Sum of Squares) and RMSE (Root Mean Squared Error) in Ridge Regression is lower than that in Lasso Regression. Therefore, we choose Ridge Regression to apply as as it has least RSS and RMSE and highest R2score/R-squared value amongst all the regression models.

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 five most important predictor variables now are: 'KitchenQual', 'MSZoning_FV', 'GarageCars', 'Neighborhood_NridgHt', 'Fireplaces'.

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 can make sure that a model is robust and generalisable by ensuring that our model is not highly impacted by the outliers in the training data and for this, we should carry out outlier analysis and should retain only those that are relevant to the model. The implication of the same for the accuracy of the model is that the test accuracy is not much lesser than the training accuracy. This also helps in improving the accuracy of the predictions made by the model. Confidence Intervals within 3-5 standard deviations might be useful at times to make the model robust.