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:

Optimal values of ridge: 3. Optimal value of lasso: 50.

When the alpha value was doubled, I found out that the R2 score for both training and test accuracy for Ridge Regression model decreased by a small amount, and the training and test accuracy for Lasso Regression model decreased by a small amount.

Doubling the alpha for the Ridge Regression Model caused the value of each beta coefficient to decrease for each of the features but doubling the alpha for the Lasso Regression Model caused fewer of the beta coefficients to be feature selected.

After doubling the alpha value, the categorical variable feature of "Neighborhood" (such as Neighborhood_NridgHt, Neighborhood_StoneBr, Neighborhood_NoRidge, Neighborhood_Somerst) remained as the top feature. On top of this, the continuous feature columns such as 2ndFlrSF and ExterQual the top features. The most important features remained the same before and after doubling the alpha value for the Ridge Regression Model.

After doubling the alpha value, the categorical variable feature of "Neighborhood" (such as Neighborhood_NridgHt, Neighborhood_StoneBr, Neighborhood_NoRidge, Neighborhood_Somerst) remained as the top feature. On top of this, the continuous feature columns such as 2ndFlrSF and ExterQual, the top features. The most important features remained the same before and after doubling the alpha value for the Lasso Regression Model but the number of features who became 0 after doubling the alpha increased.

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:

I will choose the Lasso Regression due to the large number of feature variables which I had, and Lasso Regression performs Feature Selection. As I had not performed too much EDA and I am not an expert in this field and thus, lacking the domain knowledge, I do not know what is the optimal number of variables of RFE.

I would like to use the Lasso Regression so that Lasso Regression will automatically penalize the less important features, and remove highly-correlated features by selecting one and shrinking the other less significant feature variables to zero.

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:

After removing the top five predictor variables in the lasso regression model, the new most important predictor variables are now 'Neighborhood_Gilbert', 'Neighborhood_IDOTRR', 'Neighborhood_Edwards', 'Neighborhood_SWISU', and 'HouseStyle_2.5Fin'.

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:

In this case, I believe that a more robust and generalizable model can be achieved through better Feature Engineering and Feature Selection. Due to the large number of features, I selected Lasso Regression and allowed Lasso Regression to do feature selection on its own. I would like to attempt a more robust and generalizable by going through careful exploratory data analysis, consulting with subject matter experts, and do more feature selection and feature engineering.

The implications of a model that is robust and generalizable will mean that the model trained by our training data set will be well-fitted to the data set from the test set. For instance, a well-fitted model will become more robust and generalizable by achieving a higher R2 score (for example, about 0.85) on both training and test data sets. On top of these, the RSS and MSE values for both training and test datasets should be as low as desired.