Problem Statement - Part II

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?

Answer1:

The optimal alpha value for ridge regression is 10, and for lasso regression, it's 100. When we double the alpha values for both ridge and lasso (i.e., 20 and 200), notable changes occur:

In ridge regression, coefficient values increase with higher alpha values, while the R2 score of the training data drops significantly from 0.807 to 0.45.

In lasso regression, increasing alpha leads to the removal of more features from the model. Although there's a slight decrease of 1% in the R2 score for both the test and train data.

The top features identified after these changes include Neighborhood_NoRidge, Neighborhood_NridgHt, OverallQual, and Neighborhood_Veenkar.

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:

We'll opt for Lasso regression because it offers feature selection capabilities. By eliminating unnecessary features without compromising model accuracy, Lasso helps create a more generalized, simpler, and accurate model.

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:

The initial top 5 features were Neighborhood_NoRidge, Neighborhood_NridgHt, 2ndFlrSF, OverallQual, and Neighborhood_Veenker. After excluding these features, the model's accuracy decreased from 80% and 81% to 55% and 58%. The new top 5 features are 1stFlrSF, MSSubClass_90, MSSubClass_120, TotalBsmtSF, and HouseStyle_1Story.

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 4:

To ensure the model is robust and generalizable, three key criteria must be met:

The model accuracy should exceed 70-75%. In our case, it achieves 80% on the training set and 81% on the test set, meeting this criterion effectively.

All feature p-values should be less than 0.05, indicating statistical significance.

The VIF (Variance Inflation Factor) for all features should be less than 5, indicating that multicollinearity is not a significant issue.

Meeting these criteria provides confidence that the model is robust and generalizable.