Problem Statement - Part II

Assignment 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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Q) What is the optimal value of alpha for ridge and lasso regression?

Ans) Ridge and Lasso Regression Model are built with optimum alpha calculated in GridSearchCV method. Optimum alpha = 9.0 for ridge and 0.0001 for lasso model.

	Optimum alpha
ridge	9.0
lasso	0.0001

Q) What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso?

	Optimum alpha		Doubled Alpha Value
ridge	9.0		18
lasso	0.0001		0.0002

Ans) R2score on training data has decreased but it has increased on testing data

Q) What will be the most important predictor variables after the change is implemented?

Ans) Predictors are same but the coefficient of these predictor has changed

Question 2

Q) 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) The r2_score of lasso is slightly higher than lasso for the test dataset so we will choose lasso regression to solve this problem

Question 3

After building the model, you realized 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) Below are the top 5 features after removing the top 5 features from top 10

1. FullBath
2. GarageArea
3. KitchenQual
4. Fireplaces
5. LotArea

Question 4

How can you make sure that a model is robust and generalizable? 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 equals the training score. The model should be accurate for datasets other than those used during training. Too much importance should not be given to the outliers so that the accuracy predicted by the model is high. To ensure this is not the case, the outliers analysis needs to be done, and only those 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.