Assignment 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: -

In predictive modelling, the optimal value of alpha is that value where the variance is very low at the cost of some bias and it approaches the Optimum model complexity. In our model, the Optimal value of Alpha in Ridge Regression is 2.7 and the Optimal Value of Alpha in Lasso Regression is 0.001.

There is a slight difference in r2 value in the model when the alpha values are doubled in case of ridge regression and lasso regression. In the case of the lasso regression, the value of r2 on the training set slightly decreases and in the case of the test set, the value slightly increases. In the case of the ridge regression, r2 value on training data slightly decreased but in the case of the test set, the value slightly increases. This means that if we double the value of alpha there are not any major changes. If we want to get major changes, we need to increase by 10 times or more.

The most important predictor variable is Neighborhood_Somerst in both ridge regression and lasso regression. it is because it persists a large value of the coefficient i.e., 0.436266 in case of lasso regression and 0.436079 in case of ridge 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: -

The optimal values i.e., determined during the assignment is 2.7 in case of ridge regression and 0.001 in case of lasso regression. Although there are no major changes in predictor variables values and r2 scores are also around the same. There are some advantages to use lasso regression over the ridge regression: -

- It helps to eliminate the features / predictor variables
- It makes the model robust
- It produces the sparse outputs
- It can produce many solutions to the same problem

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/features are as follows: -

- Neighborhood_Somerst
- SaleCondition_AdjLand
- Neighborhood_Veenker
- Neighborhood_NridgHt
- KitchenQual

If we exclude these variables/features from the incoming data, we have to follow the following steps: -

1. First drop these columns from both Train and test data
2. Now using GridSearchCV method finds the optimal alpha for the model
3. Built the model with the optimal alpha find by the above method
4. Now we will get the top 5 features for the optimal value of alpha as follows:
    a. SaleCondition_Normal
    b. GarageQual
    c. SaleType_Con
    d. BsmtFinSF1
    e. Foundation_Woo

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 the model is robust and generalized by the following methods: -

1. If the model needs to be robust and generalized then it should not be impacted by any outliers.
2. If the model needs to be generalized and robust then the test and training accuracy should be nearly equal or test accuracy should be greater than the training accuracy.
3. If the model needs to be generalized and robust then it should be accurate to the dataset other than the training dataset.
4. If the model needs to be generalized and robust then the weightage should not give to the outliers so that it is very close to the generalized model.

The accuracy of the model on data other than training is also high in the case of the generalized and robust model. If there are no outliers or the weightage of the outliers is less, then the accuracy is also high. Generally, the confidence interval for the model is 3 standard deviations and it helps normalize the prediction made by the model. If the model is not robust, it cannot be used for prediction analysis.