Advance Regression Assignment

Assignment-based subjective Questions

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?**

Answer:

- The optimal value of alpha for ridge regression is 5 and for lasso regression is 100.
- If we choose to double the value of alpha which is changing the value from 5 to 10, then in Ridge regression, the r2 score of train data is reduced from 0.94 to 0.93 that means the value of r2 either drops or there may be slight changes in the r2 value. The value of coefficients are increasing as the value of alpha increases and even the coefficients move more towards 0.
- Top predictor variables by ridge regression after the change are as follows:
    - OverallQual_9
    - Neighborhood_StoneBr
    - Neighborhood_Crawfor
    - Functional_Typ
    - BsmtQual_Gd

- In Lasso regression if we double the alpha value which is changing the value from 100 to 200, then the r2 score for train and test data drops and also the number of non zero coefficient drops. Features extracted from this algorithm is also pretty good in this algorithm.
- Significant predictors remain the same after the implementation. Only the coefficient drops for each predictor in both the regression.
- Top predictor variables by lasso regression after the change are as follows:
    - OverallQual_9
    - Neighborhood_StoneBr
    - Functional_Typ
    - OverallQual_8
    - SaleCondition_Partial

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:

We would go with the lasso regression because of the following reasons:

- R2 score for test and train is better in lasso regression than in ridge regression.
- In Lasso regression, few of the coefficients were equal to 0, which reduced the number of feature predictors and made the model less complex.
- The mean square error lasso is slightly lower than of ridge regression.
- Also since lasso helps in feature selection and so, lasso have 159 feature equal to zero whereas ridge regression has only 20 feature equal to zero.

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?**

Answer:

Current top 5 important predictor variables:

- OverallQual
- BsmtQual
- Exterior1st
- MasVnrType
- Neighborhood

After excluding 5 important variable given by lasso, here is the next 5 important variable for prediction

- Salecondition
- BsmtExposure
- MSZoning
- BsmtFinType1
- MSSubClass

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?**

Answer:

Outliers in the training data should impact on the model's robustness and generalisation. Additionally, the model should be able to be generalised so that the test accuracy is not lower than the training score itself. The model should be accurate for datasets that are different from those used during training. Outliers should not be given too much weight

so that the model's accuracy is high. The outlier analysis must be performed and only those that are relevant to the dataset must be preserved in order to verify that this is not the case. In order to remove outliers from the dataset, it is necessary to do so. The influence of outliers can be reduced using a more robust error metric. This would assist improve the model's accuracy in making predictions.