**Assignment – Part 2**

**By Surya Chandra**

**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:**
The Optimal values of Lamba are 7 & 0.001 respectively for Ridge & Lasso.
The r2 score for the above optimal values of lambda is given below

| r2 Scores | | | |
|---|---|---|---|
| | Liner | Ridge (Alpha 7) | Lasso (Alpha 0.001) |
| Train | 0.958650722 | 0.92675992 | 0.907104932 |
| Test | 0.847768711 | 0.860563695 | 0.862628258 |
| | | Ridge (Alpha 14) | Lasso (Alpha 0.002) |
| Train | | 0.919357179 | 0.883559427 |
| Test | | 0.859924264 | 0.847359603 |

When the value of Lambda is doubled, the r2 score varied as mentioned in above table. After value is doubled, the below factors became important:

- constant
- OverallQual_9
- OverallCond_3
- GarageCars

Increasing the lambda reduced the r2 score in both Ridge & Lasso.

**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:**

Lamba values of 7 & 0.001 are ideal for Ridge & lasso respectively.

The r2 score for the alpha 7 & 0.001 is as follows.

| r2 Scores | | |
|---|---|---|
| Model Data | Ridge (Alpha 7) | Lasso (Alpha 0.001) |
| Train | 0.92675992 | 0.907104932 |
| Test | 0.860563695 | 0.862628258 |

As we know the higher the $R^2$ score, the better the model fits the data. So, We could simply choose the model with the highest $R^2$ score on the test set (in my case, the Lasso model), But it's also important to consider the difference between the training and test scores. Because we want to understand how good our model generalize well with new, unseen data.

The Lasso and Ridge models have smaller differences between their training and test scores, However the Lasso model has a slightly higher test score, Hence Lasso would be the best choice here.

**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 the top 5 features are dropped, below are the next most important predictors

- `Neighborhood_NridgHt`
- `MSSubClass_30`
- `SaleCondition_Partial`
- `GarageCars`
- `OverallQual_8`

**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:**

A model can be said as robust & generalizable when it has good accuracy on both training & Test data. Such model does not overfit or underfit on test/unseen data.

However, generalization comes with a cost of accuracy. A model that's too complex might have high accuracy on the training data but perform poorly on new data, which will lead to overfitting. Similarly, a model that's too simple might not capture all the patterns in the training data, leading to underfitting.

To modify the model complexity, we add penalty term to cost function. By adding it we try to generalize model complexity & make a best model. Incorrect term of penalty would lead to overfilling & underfitting. Model too complex will lead to overfitting & model that's too simple will lead to underfitting problems.

**Implications:**

The accuracy will increase steadily as we make the model more robust & generalize. i.e., by slowly varying the penalty term until we find the best fit. Once the best fit model is identified, Changes on test data/unseen data or outliers will have lesser impact on such model's accuracy. If accuracy is not maintained, we will end up with Overfitting or Underfitting conditions.