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

- Optimal value for Ridge regression = **6**
- Lasso Regression = **0.0001**
- If Optimal value for Ridge regression is doubled (alpha = 12), the R2 score in train data decreases but test data increases.
- Predictor variables selected for alpha = 12 (Ridge) are –
  'OverallQual', 'PoolQC_Not_applicable', '1stFlrSF', '2ndFlrSF', 'Condition1_Norm', 'Neighborhood_Somerst', 'LotConfig_CulDSac', 'CentralAir', 'Neighborhood_NridgHt', 'BsmtFullBath'],
- If Optimal value for Ridge regression is doubled (alpha = 0.0002), the R2 score in train data decreases but test data increases here too.
- Predictor variables selected for alpha = 12 (Ridge) are –
  'PoolQC_Not_applicable', 'OverallQual', '1stFlrSF', '2ndFlrSF', 'Condition2_Norm', 'Neighborhood_Somerst', 'SaleType_New', 'Neighborhood_NridgHt', 'LotConfig_CulDSac', 'CentralAir'

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?

Based on R2 score on test data, Ridge regression is performing slightly better in this scenario.

We can choose Ridge regression as the final 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?

If first important variables are not available, Lasso regression model was rebuilt using remaining variables.

Now the 10 important predictors are –

'Utilities', 'TotRmsAbvGrd', 'Neighborhood_NridgHt', 'FullBath',

'Neighborhood_Somerst', 'Exterior2nd_Wd Sdng', 'LotConfig_CulDSac',

'GarageType_Not_applicable', 'GarageArea', 'KitchenQual'

But with this, R2 value dropped about 5% in both train and test data.

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?

The final model explained 88% of data variance. It implies the accuracy of the model is reasonably good.

If there is significant difference between test and train, there is chance of **overfitting** (Test R2 score is much less than train R2 score). Whereas, overall low R2 score in train data means it is an underfit model.

In our case, The R2 score for both the train and test was very close. This tells us the model was a good fit. It is robust and generalisable.