SUBJECTIVE QUESTIONS

Q1. 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?

A1. The optimal value of alpha obtained for :-

1. Ridge Regression - 4.0
2. Lasso Regression - 0.001

Changes observed after doubling the values of alpha for :-

1. Ridge Regression -
    - Test R2 score - 0.854 to 0.847
    - Train R2 score - 0.812 to 0.806
2. Lasso Regression -
    - Test R2 score - 0.857 to 0.853
    - Train R2 score - 0.813 to 0.808

After the change is implemented, the most important variables for :-

1. Ridge Regression with coeff. -
    - OverallCond_7 - 1.17725322
    - OverallQual_Excellent - 0.5212771
    - House_age - 0.43249520
    - OverallQual_7 - 0.42401815
    - Neighborhood_OldTown - 0.353725

2. Lasso Regression with coeff. -
    - OverallCond_7 - 1.51277922
    - OverallQual_Excellent - 0.64570820
    - OverallQual_7 - 0.4396171
    - House_age - 0.40665814
    - Neighborhood_NridgHt - 0.313527

Q2. 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?

A2. The optimal value of alpha for Ridge Regression is 4.0 with test R2 score of 0.854 and for Lasso Regression is 0.001 with test R2 score of 0.857.
I will choose Ridge Regression to apply because although the R2 score is somewhat lower for Ridge than for Lasso but in Lasso Regression, no feature elimination takes place in this case and moreover, the computational cost of Lasso is higher than that of Ridge Regression.

Hence, I'll prefer Ridge Regression over Lasso in this case as the small difference in the R2 scores can be ignored without any issue.

Q3. 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?

A3. Initially, the five most important predictor variables in Lasso model were :-

- OverallCond_7
- OverallQual_Excellent
- OverallQual_7
- House_age
- Neighborhood_NridgHt

After removing these variables, a new Lasso model was created which gave the following top 5 predictor variables with coeff. :-

- OverallQual_8 - 0.6336551
- MSSubClass_120 - 0.5209050
- GrLivArea - 0.49040519
- OverallCond_Excellent - 0.32049113
- Neighborhood_OldTown - 0.240140

Q4. 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?

A4. According to Occam's Razor,

- Simple models are generic and can be generalised by being used widely
- Simple models require less training data and have less number of features as compared to a complex model
- Simple models are more robust as they do not change drastically with changes in the training dataset
- Simple models have low variance and high bias while complex models have low bias and high variance
- Complex models tend to overfit due to which they can not be generalised and considered robust

We can make sure that a model is robust and generic by applying regularisation in most cases.

Regularisation ensures that the model is penalised for each feature it trains on in case it gives a high error value. It keeps a balance between keeping the model simple and also that the model is not too naive to learn and underfit on the data.

Accuracy of a simple, robust and generic model leads to a bias variance trade off in which a simple model is most likely to remain unchanged even if the data points are added or removed from the existing dataset. On the other hand, a complex model will change for even a minute change in the training set and is most likely to overfit the training data.

It is important to ensure a simple model yet not too naive model is used for making predictions, otherwise the model may overfit (in case the model is taken as a complex one) or underfit (in case the model is too naive to understand the data patterns) the data set which may lead to low testing accuracy.