Assignment Part-II Subjective 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 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 1: For Ridge Regression the optimal value of alpha is 10 and for Lasso Regression the optimal value for alpha is 100.

The changes that occurred in the model when we doubled alpha are:

- The coefficients were changed because of more penalty given by the alpha value.
- The R2 scores reduced. So, the performance reduced.

The most important predictors (top 10) for Ridge doesn’t change even after the alpha was changed, so they are, OverallQual, 2ndFlrSF, GrLivArea, TotRmsAbvGrd, Fireplaces, 1stFlrSF, MasVnrArea, FullBath, GarageCars, GarageArea.

The most important predictors (top 10) for Lasso also doesn’t change even after the alpha was changed, so they are, GrLivArea, OverallQual, GarageCars, MasVnrArea, BsmtFullBath, Fireplaces, OverallCond, BsmtQual, LandSlope, KitchenQual. But there were new features which were completely discarded by the model as compared to when model’s alpha was not changed.

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 2: I will choose Lasso Regression because

- It helps Feature Selection i.e., in removing unnecessary variables from the model by bringing their coefficients to 0
- Provides better accuracy than Ridge Regression.
- Due to Regularization in place, it performs better than vanilla Linear Regression.

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 3: The five most important variables now are:

Gound Living Area (GrLivArea), Overall material and Finish of the house (OverallQual), Size of Garage in Car capacity (GarageCar), Masonry veneer area in square feet (MasVnrArea), Basement full bathrooms (BsmtFullBath).

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 4: Model can be made generalizable by:

- Monitoring the testing and training accuracy metrics of the model for underfitting and overfitting
- Applying K-fold cross validation to make sure models are tested on validation set before being tested on the actual test data
- Choosing an apt trade-off between the biasness and variance of the model
- Choosing appropriate predictors and discarding the rest to ensure models don't learn unnecessary details.
- Applying regularization to ensure models coefficients don't lead to overfit the model
- By plotting the histogram of error vs terms, and scatter plots of error we can determine if the model is taking care of Linear regression's assumption correctly or not.

These points ensure model is robust and generalizable. The implication on accuracy of model is quite significant. While if the model is not generalizable it may underfit or overfit the model, if the model is too biased on a certain assumption it may underfit resulting in poor training as well as poor testing accuracy, and if the model tries to capture all kinds of variances or details it may overfit resulting in exceptionally high training accuracy but poor testing accuracy. So, if the model follows or checks the above points then model is sure to have a good accuracy in both training and testing, while it might not be the best (due to overfitting) but enough to ensure that it is robust and generalizable.