Advanced Regression Subjective Questions and Answers

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?

Ans: Optimal value of alpha for Ridge regression = 20 and for Lasso regression = 150

Ridge Regression:

Alpha = 40

R2 score training set	R2 score test set
0.8321657035965009	0.808870076420829

Effect: R2 scores are almost similar.

Most important predictor variables are: OverallQual, GrLivArea, GarageCars, Neighborhood_NoRidge, 2ndFlrSF, FullBath.

Lasso Regression:

Alpha = 300

R2 score training set	R2 score test set
0.8595652827495798	0.8536928723477338

Effect: R2 score remains almost same.

Most important predictor variables are: GrLivArea, OverallQual, OverallCond, Neighborhood_NoRidge, Neighborhood_NridgHt, Neighborhood_StoneBr, BsmtExposure_Gd, BsmtFullBath, Fireplaces.

Overall, we cannot observe huge change in the model after alpha is doubled.

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?

Ans:

Measure on test set	Ridge Regression	Lasso Regression
Alpha	20	150
Mean Squared Error	1020608391.0301318	823547067.8423003
No. of feature selected	272	106

MSE for Lasso is much less than Ridge regression. Also Lasso regression helps in reduction in number of predictor variables. So, it can be concluded that model using Lasso should be used as final model.

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?

Ans:

Top 5 predictors for current Lasso model	GrLivArea OverallQual, GarageCars, Neighborhood_NoRidge, Neighborhood_NridgHt

After removing these predictors, our new model becomes:

MSE on test set: 1020290423.3725954

R2 score on test set: 0.8388675630543575

Top 5 predictors now:

Condition2_PosN

TotalBsmtSF

2ndFlrSF

TotRmsAbvGrd

GarageArea

So, it can be observed that MSE has increased for the new model using Lasso.

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?

Ans: The robustness and generalizability of a model can be ensured by making the model simple but not very simple which cannot be used. Simpler models are usually more generic and can be used widely. Simpler models require fewer training samples to get trained, so that they make more error in training set, but do not fail miserably to predict on test sets. Since simple models need fewer samples and fewer features for training, they do not seem to change wildly when training set changes. Regularization can be used to make a model simpler. The regularization process ensures the balance between keeping the model simple and not making it too naïve to be of any use.

The main implication of making a model simple is Bias-Variance tradeoff. Bias refers to the accuracy and variance refers to consistency. Complex models can accurately predict when there is plenty of training dataset, but it can lead to overfitting and that's why it fails on test dataset. Complex models have low bias and high variance. So we need to strike a balance here between bias and variance since it minimizes the total error.