Assignment Part-II (Subjective Questions)

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?**

A. Optimal Value for Ridge Regression: 10.0
Optimal Value for Lasso Regression: 1.0
When we double the value of alpha for our ridge regression, the model will apply more penalty on the curve and try to make the model more generalized i.e. making model simpler and error prone. We can also observe that when alpha is increased, we see more error in test and train data sets.
Similarly, when we double the value of alpha for lasso regression, we try to penalize our model more, and more number of coefficients of the variables would be reduced to zero.
The most important variables after the changes for ridge regression are: MSZoning_FV ,MSZoning_RL,Neighborhood_Crawfor ,MSZoning_RH ,MSZoning_RM ,SaleCondition_Partial ,Neighborhood_StoneBr ,GrLivArea ,SaleCondition_Normal ,Exterior1st_BrkFace
The most important variables after the changes for lasso regression are: GrLivArea,OverallQual,OverallCond,TotalBsmtSF,BsmtFinSF1,GarageArea,Fireplaces,Lot Area,LotArea,LotFrontage.

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?**

A. Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.
Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

In my case, I would choose Lasso regression as Lasso tends to do well if there are a small number of significant parameters and the others are close to zero.

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?

A. The 5 most important predictor variables that will be excluded would be:
1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

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?

A. The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalizable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalizable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.