Problem Statement - Part II

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 of alpha for ridge and lasso: 10 and 20
If the values of alpha values are doubled, then it will increase the model complexity and eventually increase the cost of execution.
The most important variable after the changes has been implemented for ridge regression are as follows:-
1. MSZoning_FV 2. MSZoning_RL 3. Neighborhood_Crawfor 4. MSZoning_RH 5. MSZoning_RM 6. SaleCondition_Partial 7. Neighborhood_StoneBr 8. GrLivArea 9. SaleCondition_Normal 10. Exterior1st_BrkFace
The most important variable after the changes has been implemented for lasso regression are as follows:-
1. GrLivArea 2. OverallQual 3. OverallCond 4. TotalBsmtSF 5. BsmtFinSF1 6. GarageArea 7. Fireplaces 8. LotArea 9. LotArea 10. LotFrontage

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:
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.
Lasso also provides features selection option, which removes unwanted features form the model without affecting the model accuracy, resulting in simplified and accurate 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?
Answer:
Those 5 most important predictor variables that will be excluded are :-
1. GrLivArea 2. OverallQual 3. OverallCond 4. TotalBsmtSF 5. GarageArea

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:

The main idea to bear in mind while generalizing the model is by using the Bias-Variance trade-off. The simpler the model, the mode the bias but with lesser variance, however, this model can be more generalizable. This implies that the model will equally perform well on both training and testing data.