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- Ridge regression:- When alpha is 2, this is when test error is minimum, therefore alpha equal to 2 is the optimal value for ridge regression.

Lasso regression optimal value for alpfa is 0.4

Up on doubling the value of alpha for ridge regression model will have more penalty on the curve which will make the model more generic which is high bias and less complex and same will result in more total error in test and train data.

Similarly, for lasso, model will penalize more and more coefficient of the variable will be zero the value of our R2 will also decreases.

The most important variable for ridge regression are as below:-

29	MSZoning_FV	0.149
31	MSZoning_RL	0.125
50	Neighborhood_Crawfor	0.114
30	MSZoning_RH	0.105
32	MSZoning_RM	0.097
210	SaleCondition_Partial	0.097
66	Neighborhood_StoneBr	0.093
13	GrLivArea	0.076
209	SaleCondition_Normal	0.074
95	Exterior1st_BrkFace	0.069

The most important variable for lasso regression are as below:-

x1	OverallQual	0.132
x2	GrLivArea	0.105
x3	TotalBsmtSF	0.038
x4	GarageArea	0.030
x5	BsmtFinSF1	0.029
x6	Fireplaces	0.016
x7	LotFrontage	0.004
x8	OverallCond	0.003

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?

Answer Lasso regression will be preferred choice as R2 score is very high in my case could be a case of over fitting. In case of Lasso, when alpha increases then coefficients of variable also shrink and can become zero, thus Lasso also helps in variable selection.

Q3. 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?

Answer The five most important predictor variables which will be excluded are :-

x1	OverallQual	0.132
x2	GrLivArea	0.105
x3	TotalBsmtSF	0.038
x4	GarageArea	0.030
x5	BsmtFinSF1	0.029

Q4. 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: In order to have a model robust and generalized with respect to complexity one should strive to attain the optimal model complexity, thereby reducing the total error. A simple sssmodel would usually have high bias and low variance, whereas a complex model would have low bias and high variance. In either case, the total error would be high which mean accuracy would decrease