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?

Ans : Optimal value of alpha in Ridge: 4

Optimal value of alpha in Lasso: 0.001

If we double the values of alpha in both the cases:

1. There is a slight increase in train R square of Lasso while decrease in test R square of Ridge.
2. There is a dip in train RSS of Lasso while slight increase in Test RSS of ridge.
3. GrLivArea became the top priority variable in Lasso which was in 3rd position initially.
4. Top 3 variables in Ridge remained same.

Most important variables in Ridge after the change is implemented:

RoofMatl_CompShg	0.217453
RoofMatl_Tar&Grv	0.144053
RoofMatl_WdShngl	0.109205
MSZoning_RL	0.108055
RoofMatl_WdShake	0.087417
MSZoning_RM	0.080832
GrLivArea	0.074089
OverallQual	0.060978
MSZoning_FV	0.054610
RoofMatl_Membran	0.049942

Most important variables in Lasso after the change is implemented:

GrLivArea	0.137907
RoofMatl_CompShg	0.127549
RoofMatl_Tar&Grv	0.083241
OverallQual	0.077443
RoofMatl_WdShngl	0.066577
RoofMatl_WdShake	0.052293
GarageCars	0.038981
OverallCond	0.038798
TotalBsmtSF	0.026962
BsmtFullBath	0.025107

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?

Ans: Optimal value of alpha in Ridge : 4

Optimal value of alpha in Lasso : 0.001

For Ridge: R square (Train) is 0.956367 and R square (Test) is 0.869207 -> diff is 0.08716

For Lasso: R square (Train) is 0.950555 and R square (Test) is 0.872445 -> diff is 0.07811

We can see that Lasso has lesser difference and also Lasso helps in feature reduction by making the coefficients zero, we can say Lasso is better in this case.

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?

Ans: Top 5 variables in current Lasso model are:

1. RoofMatl_CompShg
2. RoofMatl_Tar&Grv
3. GrLivArea
4. RoofMatl_WdShngl
5. RoofMatl_WdShake

After removing these from the dataset and rebuilding the model, the top 5 variables are now:

1. 2ndFlrSF
2. 1stFlrSF
3. MSZoning_RL
4. OverallQual
5. MSZoning_RM

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?

Ans: A model will be robust and generalizable if it has following criteria:

1. It is not complex.
2. It should have low variance meaning when the training set changes, the model should not change drastically.
3. It should not overfit meaning that model should not memorize the training set as it will perform very poor when it is tested with test data set.
4. It should have low bias as well.

For accuracy, very complex models will have high accuracy and also high variance. On decreasing the variance, there is chance of increase in bias which will lead to reduction in accuracy. So we need to find a balance between variance and bias or complexity and accuracy through regularization techniques.