Assignment Problem ll

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?

Optimal value of alpha for ridge: 2.0

Optimal value of alpha for lasso : 0.001

Ridge:

Choosing double value of alpha for ridge: 4.0

When alpha value of Ridge was doubled the R2 square value of train data decreased a little but the R2 square value of test data increased.

Alpha 2.0

R2 square train - 0.9250053162096344
R2 square test -0.8886814947404266
Alpha 4.0
R2 square train- 0.9216472788003705
R2 square test - 0.8916466959084074

Lasso:

Choosing double value of alpha for lasso: 0.01

When alpha value of Lasso was doubled the R2 square value of train data as well as test data decreased.

Alpha 0.001

R2 square train - 0.9085218650563394

R2 square test - 0.8938791238192652
Alpha 0.01
R2 square train - 0.8650489592016473
R2 square test - 0.8631193131677695

Important predicted variables:

Ridge:

OverallQual
YearBuilt
BsmtFinSF1
TotalBsmtSF
1stFlrSF
GrLivArea
MSZoning_RL
Neighborhood_Crawfor
RoofMatl_Metal

Lasso:

OverallQual
YearBuilt
BsmtFinSF1
TotalBsmtSF
1stFlrSF
GrLivArea
MSZoning_RL
Neighborhood_Crawfor

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?

If we consider the R2 square value then for train data, ridge have higher value but for test data, Lasso has higher r2 square value. If we go by definition then, Ridge and Lasso regression, uses a tuning parameter called alpha as the penalty is square of magnitude of coefficients, which is identified by cross validation. As we increase the value of alpha the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

In Lasso regression 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.

Hence, we will make use of Ridge as it considers all the variables.

Question 3 After building the model, you realised that the five most important predictor varia bles in the lasso model are not available in the incoming data. You will now have to create an other model excluding the five most important predictor variables. Which are the five most i mportant predictor variables now?

The 5 most important variables that will be exculuded are:

GrLivArea
OverallQual
YearBuilt
BsmtFinSF1
TotalBsmtSF

After excluding the important variables the new important predictor are:

RoofMatl_Metal
1stFlrSF
MSZoning_RL
Neighborhood_Somerst
2ndFlrSF

Question 4 How can you make sure that a model is robust and generalisable? What are the im plications of the same for the accuracy of the model and why?

1. The model should be as simple as possible, even if the accuracy decrease a bit but it will be more robust and generalizable. Also, the test accuracy should not be much lesser than the training score 2.
2. The simpler the model the more the bias but less variance and more generalizable. .
3. Outliers should be considered but too much importance should not be given so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained.
4. If the model is not robust, it cannot be trusted for predictive analysis.