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:

The optimal value of alpha we got in case of Ridge and Lasso is :

1. Ridge - 2.0
2. Lasso - 0.0001

- When we double the value of alpha in case of ridge , the R2 score remains almost the same whereas the MSE increase slightly.
- When we double the value of alpha in case of lasso there is a very big decrease in R2 value making the model bad. R2 train doesn't decrease much but R2 test falls by a large number hence the prediction power is not good
- This means the model is penalized even more and more coefficients move towards zero

Most important predictor variables for Ridge:

	Features	rfe_support	rfe_ranking	Coefficient	
0	OverallQual	True	1	0.126	
2	TotalBsmtSF	True	1	0.061	
1	YearBuilt	True	1	0.057	
18	GarageQual_TA	True	1	0.009	
16	GarageQual_Gd	True	1	0.009	
14	GarageFinish_none	True	1	0.003	
13	GarageType_none	True	1	0.003	
19	GarageQual_none	True	1	0.003	
3	YrSold	True	1	0.003	
9	Exterior2nd_Other	True	1	0.000	

Most important predictor variables for Lasso:

	Features	rfe_support	rfe_ranking	Coefficient
1	OverallQual	True	1	0.163
5	GrLivArea	True	1	0.155
3	YearBuilt	True	1	0.135
7	Square_footage	True	1	0.119
11	Neighborhood_StoneBr	True	1	0.097
2	OverallCond	True	1	0.095
4	TotalBsmtSF	True	1	0.092
0	LotArea	True	1	0.060
8	Neighborhood_Crawfor	True	1	0.050
10	Neighborhood_NridgHt	True	1	0.043

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?

- As a whole regressions values are :

  R2:
  Ridge - Train = 0.9302888220931927 , Test = 0.896700419255886
  Lasso - Train = 0.927529179409694 , Test = 0.9028162866777875

- MSE :
  Ridge - 0.00297345560908371
  Lasso - 0.0027974117165618963

- We notice the MSE of ridge is lower than that of lasso
- Also R2 difference between train and test is less in lasso than in ridge
- Lasso also helps in feature elimination as one of the coefficents have shrunk towards zero during the process
- Thus lasso has a better model compared to ridge

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?

- After dropping the 5 most important variables form the lasso model and running it again these are the variables that remain :

  1. TotalBsmtSF
  2. TotRmsAbvGrd
  3. OverallCond
  4. Total_Bathrooms
  5. LotArea

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?

- Generic models are considered more simple . Here accuracy will be affected but model overall will work on better range of scenarios
- Bias-Variance trade-off explains that the more simpler the model the more the bias which leads to lesser variance
- Complex models will have vice-versa – High variance and low bias
- Regularization which we have achieved using ridge and lasso helps in maintaining this balance by making sure the coefficients are minimized to zero.
- This will avoid overfitting of data
- It penalizes the model if it tends to get more and more complex with more predictors
- This process also decreases error or MSE to a minimum value at optimum model complexity
- Bias-Variance trade off will help make the model simple