Assignment Part – 2

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.

1) Optimal value of lambda for ridge regression is 4.

2) Optimal value of lambda for lasso regression is 100.

On increasing value of alpha from 4 to 8 in ridge the R2 score on both the train and test is decreased.

On increasing value of alpha from 100 to 200 in lasso the R2 score on both the train and test is decreased.

Predictors are same but the coefficient of these predictors have decreased as we can see in below table.

	Ridge2	Ridge	Lasso	Lasso20
MSSubClass	-12656.245105	-12834.921067	-13409.845389	-14816.413469
LotArea	19451.983104	23398.163773	20300.647276	12081.856149
OverallQual	38068.057962	43363.490324	60829.174997	65844.696639
OverallCond	19629.402665	25693.256758	28350.056750	19419.584064
YearBuilt	10557.753862	13638.730249	23626.549166	20313.937776
YearRemodAdd	9480.659963	8118.736230	7496.113845	9268.263619
MasVnrArea	19376.446439	19692.135399	14668.273646	9431.188173
BsmtFinSF1	33635.621608	39232.301012	32089.007545	33195.376015
BsmtFinSF2	3189.033280	5444.546627	0.000000	0.000000
BsmtUnfSF	14954.039046	15095.632114	0.000000	0.000000

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. As we can see the value of R2 score of Ridge on both train and test data is more than in the Lasso regression, hence for this problem Ridge regression will be good over the lasso regression.

3. 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?

Ans. The top 5 variables in the present model are:

GrLivArea: Above grade (ground) living area square feet.

1stFlrSF: First Floor square feet.

YearBuilt: Original construction date.

MasVnrArea: Masonry veneer area in square feet.

LotConfig: Lot configuration.

ScreenPorch: Screen porch area in square feet.

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

Now the most important variables are

GrLivArea: Above grade (ground) living area square feet

1stFlrSF: First Floor square feet

YearBuilt: Original construction date

MasVnrArea: Masonry veneer area in square feet

LotConfig: Lot configuration

ScreenPorch: Screen porch area in square feet

YearRemodAdd: Remodel date (same as construction date if no remodelling or additions)

4. 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. The Value of R2 score in ridge regression in train data is 0.9358, test data is 0.9141 and percent difference is 2.32%.

The Value of R2 score in lasso regression in train data is 0.9305, test data is 0.9189 and percent difference is 1.25%.

Hence model is robust and generalizable because it is performing good in both train and test data. The percentage difference between the train data and test data for both model is less than 5% so the accuracy of model is also good.