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 for the ridge is 0.2. The most important predictor variables when alpha is 0.2 is given below

	Features	Coefficient
47	GrLivArea	0.8783
42	OverallQual	0.5042
43	OverallCond	0.4033
46	TotalBsmtSF	0.3889
26	BsmtFullBath_3	0.3165

The most important predictor variables when alpha is doubled i.e 0.4 is given below

	Features	Coefficient
47	GrLivArea	0.8490
42	OverallQual	0.5058
43	OverallCond	0.3996
46	TotalBsmtSF	0.3691
26	BsmtFullBath_3	0.2677

The predictor variable coefficient has been reduced. As we already know, if the hyperparameter increases, the model coefficient decreases and regularization increases. The order of the variable has been changed.

The optimal value of alpha for lasso is 0.0002. The most important predictor variables when alpha is 0.0002 is given below

	Features	Coefficient
181	GrLivArea	0.7459
173	OverallQual	0.4125
174	OverallCond	0.3405
178	TotalBsmtSF	0.3201
172	LotArea	0.1449
46	Neighborhood_StoneBr	0.1314
176	BsmtFinSF1	0.1109
30	Neighborhood_Crawfor	0.1102
170	SaleCondition_Partial	0.1028
125	BsmtFullBath_3	0.1025

The most important predictor variables when alpha is doubled i.e 0.0004 is given

	Features	Coefficient
181	GrLivArea	0.7334
173	OverallQual	0.4489
174	OverallCond	0.3357
178	TotalBsmtSF	0.3213
172	LotArea	0.1452
176	BsmtFinSF1	0.1305
46	Neighborhood_StoneBr	0.1104
30	Neighborhood_Crawfor	0.1049
170	SaleCondition_Partial	0.1012
67	Exterior1st_BrkFace	0.0932

The predictor variable coefficient has been reduced. As we already know, if the hyperparameter increases, the model coefficient decreases and regularization increases. The order of the variable has been changed.

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?

Answer:

The optimal value for ridge and lasso is 0.01 and 0.0002. In this case, we can choose Lasso, because it makes some of the model coefficients to 0, thus resulting in model selection easier. Lasso is chosen particularly when the number of coefficients is very large.

	Metric	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.930222	0.940131
1	R2 Score (Test)	0.889529	0.892477
2	RSS (Train)	10.776076	9.245751
3	RSS (Test)	6.728331	6.548748
4	MSE (Train)	0.104016	0.096348
5	MSE (Test)	0.125381	0.123696

When comparing ridge and lasso's R square, RSS, MSE values we can conclude lasso is best since R square(test, train =) is higher, RSS, MSE (train, test) is lower.

Question 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?

Answer:

The top five predictor variables in the lasso model.

	Features	Coefficient
181	GrLivArea	0.7459
173	OverallQual	0.4125
174	OverallCond	0.3405
178	TotalBsmtSF	0.3201
172	LotArea	0.1449

The second top-five predictor variable(below figure) after deleting top five variable(above figure)

	Features	Coefficient
175	1stFlrSF	0.7618
176	2ndFlrSF	0.4359
46	Neighborhood_StoneBr	0.1591
173	BsmtFinSF1	0.1494
30	Neighborhood_Crawfor	0.1344

Question 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?

Answer:

Model accuracy is used to determine whether a model is efficient or not. Model accuracy tells about how the relationship or pattern is identified between the variables in a dataset.

If the accuracy(R square) of the train and test data is almost the same (the difference can be < 5%) we can say that model is robust and generalizable with respect to accuracy.