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 0.01. The most important predictor variables when alpha is 0.01 is given below

	Features	Coefficient
46	GrLivArea	0.9416
40	OverallQual	0.4993
23	BsmtFullBath_3	0.3537
41	OverallCond	0.3422
45	TotalBsmtSF	0.2666
1	MSZoning_FV	0.2557
39	LotArea	0.2047
2	MSZoning_RH	0.1931
43	BsmtFinSF1	0.1924
3	MSZoning_RL	0.1888

The most important predictor variables when alpha is doubled i.e 0.02 is given below

	Features	Coefficient
46	GrLivArea	0.9396
40	OverallQual	0.4996
23	BsmtFullBath_3	0.3499
41	OverallCond	0.3419
45	TotalBsmtSF	0.2666
1	MSZoning_FV	0.2536
39	LotArea	0.2046
43	BsmtFinSF1	0.1925
2	MSZoning_RH	0.1909
3	MSZoning_RL	0.1868

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
208	GrLivArea	0.7470
199	OverallQual	0.4180
200	OverallCond	0.3557
205	TotalBsmtSF	0.2802
198	LotArea	0.1517
202	BsmtFinSF1	0.1433
46	Neighborhood_StoneBr	0.1330
30	Neighborhood_Crawfor	0.1054
196	SaleCondition_Partial	0.0999
75	Exterior1st_BrkFace	0.0987

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

	Features	Coefficient
208	GrLivArea	0.7317
199	OverallQual	0.4531
200	OverallCond	0.3498
205	TotalBsmtSF	0.3084
198	LotArea	0.1523
202	BsmtFinSF1	0.1424
46	Neighborhood_StoneBr	0.1110
196	SaleCondition_Partial	0.1000
30	Neighborhood_Crawfor	0.0993
75	Exterior1st_BrkFace	0.0911

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)	0.933275	0.943027
1	R2 Score (Test)	0.886020	0.895717
2	RSS (Train)	10.304560	8.798476
3	RSS (Test)	6.942032	6.351458
4	MSE (Train)	0.101715	0.093988
5	MSE (Test)	0.127357	0.121819

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
208	GrLivArea	0.7470
199	OverallQual	0.4180
200	OverallCond	0.3557
205	TotalBsmtSF	0.2802
198	LotArea	0.1517

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

	Features	Coefficient
202	1stFlrSF	0.6558
203	2ndFlrSF	0.4430
199	BsmtFinSF1	0.2953
46	Neighborhood_StoneBr	0.1543
201	BsmtUnfSF	0.1489

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 talks 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.