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 ridge regression is : Ridge(alpha = 10.0)
- The optimal value of alpha for lasso regression is : Lasso(alpha = 0.0005)
- When the value is doubled in Ridge regression , It adds a factor of sum of squares of coefficients in the optimization objective.
- As the value of alpha increases, The model complexity decreases in both ridge and lasso regression .
- It is simply the weighted sum of each data point with coefficients as the weights. This prediction is achieved by finding the optimum value of weights based on certain criteria, which depends on the type of regression algorithm being used.
- The coefficient values : ridge regression coefficient values are decreased after Lasso regression is performed as shown below,
- Ridge coefficient values :

	Features	rfe_support	rfe_ranking	Coefficient
3	TotalBsmtSF	True	1	0.0382
6	MSZoning_RL	True	1	0.0330
0	BsmtFinSF1	True	1	0.0322
4	MSZoning_FV	True	1	0.0321
9	Heating_GasA	True	1	0.0215
5	MSZoning_RH	True	1	0.0184
13	Heating_Wall	True	1	0.0107
1	BsmtFinSF2	True	1	0.0103
10	Heating_GasW	True	1	0.0039
12	Heating_OthW	True	1	0.0027

- Lasso coefficient values:

	Features	rfe_support	rfe_ranking	Coefficient
0	BsmtFinSF1	True	1	0.0374
3	TotalBsmtSF	True	1	0.0342
6	MSZoning_RL	True	1	0.0341
4	MSZoning_FV	True	1	0.0264
1	BsmtFinSF2	True	1	0.0071
2	BsmtUnfSF	True	1	-0.0000
5	MSZoning_RH	True	1	0.0000
7	MSZoning_RM	True	1	-0.0000
8	Exterior1st_BrkComm	True	1	-0.0000
9	Heating_GasA	True	1	0.0000

- The predicted outcome for any data point i is:

$$\hat{y}_i = \sum_{j=0}^{M} w_j * x_{ij}$$

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 of Ridge regression is
{'alpha': 10.0}
Score: -0.08291402566238563

The optimal value of Lasso regression is
{'alpha': 0.0005}
Score: -0.08416675572830837

By comparing both the regressions Lasso regression should be considered because the Model Complexity is decreased by comparing the both.

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?

Answer:

	Features	rfe_support	rfe_ranking	Coefficient
0	BsmtFinSF1	True	1	0.0374
3	TotalBsmtSF	True	1	0.0342
6	MSZoning_RL	True	1	0.0341
4	MSZoning_FV	True	1	0.0264
1	BsmtFinSF2	True	1	0.0071
2	BsmtUnfSF	True	1	-0.0000
5	MSZoning_RH	True	1	0.0000
7	MSZoning_RM	True	1	-0.0000
8	Exterior1st_BrkComm	True	1	-0.0000
9	Heating_GasA	True	1	0.0000

According to the model last 5 most important predictor variables are not available in the incoming data. The five most important predictor variables should be considered further are as follows:

1. BsmtUnfSF
2. MSZoning_RH
3. MSZoning_RM
4. Exterior1st_BrkComm
5. Heating_GasA

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?

Answer:

A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalisable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not given to the outliers so that the accuracy predicted by the model is high.