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 to 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 in ridge and lasso regression is the value of the hyperparameter at which the model has the highest score based on the give score function. This gives the model a cutoff point where the model has low bias and variance hence avoiding the chances of Overfitting and Underfitting. In the assignment, the optimal values achieved for alpha are as:

Ridge: 10 **Lasso:** .0005

As the value of alpha increases, the model complexity reduces. Though higher values of alpha reduce overfitting, significantly high values can cause underfitting as well. Increasing alpha penalizes the coefficients more and makes them tend to zero in case of Ridge regression and equal to zero in Lasso regression.

Top ten predictors using Regularization before doubling alpha:

Ridge

	Features	rfe_support	rfe_ranking	Coefficient
5	Neighborhood_NridgHt	True	1	0.0762
0	OverallQual	True	1	0.0760
6	Condition2_Norm	True	1	0.0749
2	MSZoning_RH	True	1	0.0603
3	MSZoning_RL	True	1	0.0437
1	MSZoning_FV	True	1	0.0351
12	Heating_GasW	True	1	0.0272
11	Heating_GasA	True	1	0.0234
8	Condition2_RRNn	True	1	0.0149
4	MSZoning_RM	True	1	0.0085

Lasso

	Features	rfe_support	rfe_ranking	Coefficient
1	GrLivArea	True	1	0.112911
5	Neighborhood_Crawfor	True	1	0.098857
7	Neighborhood_NridgHt	True	1	0.093159
14	SaleCondition_Partial	True	1	0.089870
0	OverallQual	True	1	0.080050
13	SaleCondition_Normal	True	1	0.074469
8	Neighborhood_Somerst	True	1	0.069393
3	MSZoning_RH	True	1	0.067597
11	Exterior1st_BrkFace	True	1	0.061808
4	MSZoning_RL	True	1	0.028896

Top ten predictors using Regularization after doubling alpha:

Ridge

	Features	rfe_support	rfe_ranking	Coefficient
0	OverallQual	True	1	0.0769
5	Neighborhood_NridgHt	True	1	0.0588
6	Condition2_Norm	True	1	0.0492
2	MSZoning_RH	True	1	0.0335
3	MSZoning_RL	True	1	0.0315
1	MSZoning_FV	True	1	0.0248
12	Heating_GasW	True	1	0.0148
11	Heating_GasA	True	1	0.0135
8	Condition2_RRNn	True	1	0.0067
13	Heating_Wall	True	1	0.0041

Lasso

	Features	rfe_support	rfe_ranking	Coefficient
3	GrLivArea	True	1	0.115013
0	OverallQual	True	1	0.085866
14	SaleCondition_Partial	True	1	0.080480
7	Neighborhood_Crawfor	True	1	0.071098
8	Neighborhood_NridgHt	True	1	0.070227
13	SaleCondition_Normal	True	1	0.060079
9	Neighborhood_Somerst	True	1	0.054047
1	OverallCond	True	1	0.050115
12	Exterior1st_BrkFace	True	1	0.039565
2	BsmtFinSF1	True	1	0.024573

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 metrics after building the model using regularization methods are as:

Metrics	**Lasso**	**Ridge**
R2score – train	0.9248	0.9231
R2 score – test	0.9206	0.9169
Mean error square	0.01258	0.0131

- The Mean Squared Error of Lasso is slightly lower than that of Ridge.
- R2 score for Lasso is slightly better than that of Ridge.

Based on the above outcome, Lasso can be applied to choose significant variables for predicting the price of a house. Lasso also helps in feature selection, so the business can focus on the more significant variables.

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:

After dropping the five most important predictor variables and rebuilding the model using lasso, the five most important variables are as:

	Features	rfe_support	rfe_ranking	Coefficient
4	MSZoning_RH	True	1	0.067597
5	MSZoning_RL	True	1	0.028896
0	1stFlrSF	True	1	0.009608
3	MSZoning_FV	True	1	0.003657
1	2ndFlrSF	True	1	0.000000

Note: The code has been written in the Jupiter notebook.

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:

Using regularization, we can make model robust and generalizable.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

Also, making a model simple leads to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

Hyperparameter restricts model to not become overfitted by learning all the data points. Because if a model is overfitted it performs well on training data but gives poor results on testing data or unseen data. So, it cannot be robust.

In validation we should use cross validation, as if we perform validation on testing data by building several models model may try to seek peek the pattern in test data and becomes not generalized.

By making model robust and generalizable, it performs well on the unseen data.

So, it is important in regularization to tune the hyperparameter to get a value which makes model to have good accuracy and robust.