Advanced Regression assignment Question's Answers

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:

Optimal alpha value for Ridge: 1.0
Optimal alpha value for Lasso: 0.001

		Train	Test
Before Double	Ridge	0.90494	0.74849
	Lasso	0.91091	0.71285
Doubled Alpha	**Ridge**	**0.90001**	**0.74849**
	Lasso	**0.90919**	**0.73333**

Observation: After double the alpha value, the R2 Score decreases little bit.

Below are the most important variables that can be considered after double the alpha value.

	Feaure	Coef
2	LotArea	0.296062
32	MSZoning_FV	0.141669
10	1stFlrSF	0.104311
8	BsmtUnfSF	0.103543
3	OverallQual	0.097404

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:

After double the Alpha value, the R2 score from Lasso is comparable to Ridge in the test set. While in the train set there are not much difference from ridge and lasso. I will choose lasso to apply for the problem. And, it also helps to reduce the overfitting.

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 R2 Score has dropped to low, after removing the variables which we chosen early, we got below variables – coefficient. These are the top 5 predictors that can be used in case of the above ones are not available.

	Feaure	Coef
12	LowQualFinSF	0.091327
5	MasVnrArea	0.087479
20	TotRmsAbvGrd	0.074291
37	LotShape_IR2	0.068836
4	OverallCond	0.067529

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:

The model has to differentiate the variance in unseen data/ test set data as well. A model with low bias is generally best to fit. Usually, we do EDA on the data before building a model such as Handling outliers, multicollinearity, Regularization, Data preprocessing, Manipulate missing values, reducing the variables, data cleansing. And feature processing is the most important one

to build a model with more accuracy. These things usually help us to build the better model. For example, in Regression algorithm – choosing the better feature using lasso & RFE.