Advanced Regression Assignment

Q1. 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?

Ans. The optimal value for Ridge Regression is **50** after choosing top 20 variables using the RFE method, and for the Lasso Regression it's **0.001**. In regards to the changes in the model if we double the value of the alpha.

Before Doubling

	Ridge Co-Efficient
OverallQual	0.038746
GrLivArea	0.025639
TotalBsmtSF	0.017282
2ndFlrSF	0.016319
BsmtFinSF1	0.015338
1stFlrSF	0.015009
GarageQual_TA	0.007082
GarageCond_TA	0.006817
GarageQual_Gd	0.004856
GarageCond_Gd	0.003254
GarageCond_Fa	0.002777
GarageCond_Po	0.001868
BsmtFinSF2	0.001282
GarageQual_Fa	0.000854
BsmtUnfSF	0.000499
GarageQual_Po	-0.000428
LowQualFinSF	-0.001687
Electrical_Mix	-0.001910
BsmtCond_Po	-0.001910
YearBuilt	-0.023648

After Doubling

	Ridge Co-Efficient
OverallQual	0.038746
GrLivArea	0.025639
TotalBsmtSF	0.017282
2ndFlrSF	0.016319
BsmtFinSF1	0.015338
1stFlrSF	0.015009
GarageQual_TA	0.007082
GarageCond_TA	0.006817
GarageQual_Gd	0.004856
GarageCond_Gd	0.003254
GarageCond_Fa	0.002777
GarageCond_Po	0.001868
BsmtFinSF2	0.001282
GarageQual_Fa	0.000854
BsmtUnfSF	0.000499
GarageQual_Po	-0.000428
LowQualFinSF	-0.001687
Electrical_Mix	-0.001910
BsmtCond_Po	-0.001910
YearBuilt	-0.023648

There's not much difference in the values of the coefficients, we can say it's negligible. Now for Lasso Regression

Before Doubling

	Coeff Values	Variables
0	0.041049	OverallQual
1	-0.022755	YearBuilt
2	0.014656	BsmtFinSF1
3	0.000323	BsmtFinSF2
5	0.017380	TotalBsmtSF
6	0.000484	1stFlrSF
8	-0.002554	LowQualFinSF
9	0.044023	GrLivArea
10	-0.002749	BsmtCond_Po
12	0.000706	GarageQual_Fa
13	0.004369	GarageQual_Gd
15	0.007671	GarageQual_TA
16	0.000448	GarageCond_Fa
17	0.001628	GarageCond_Gd
19	0.003678	GarageCond_TA

After Doubling

	Coeff Values	Variables
0	4.136903e-02	OverallQual
1	-2.215936e-02	YearBuilt
2	1.411535e-02	BsmtFinSF1
5	1.711873e-02	TotalBsmtSF
6	4.877296e-04	1stFlrSF
8	-1.652617e-03	LowQualFinSF
9	4.339887e-02	GrLivArea
10	-1.964710e-03	BsmtCond_Po
11	-2.619995e-18	Electrical_Mix
13	3.173219e-03	GarageQual_Gd
15	6.436541e-03	GarageQual_TA
17	6.283546e-04	GarageCond_Gd
19	3.440099e-03	GarageCond_TA

There's a slight change in Lasso Regression in the number of decimal places. The number of decimal places have increased.

The important predictors after the change is implemented would be:

1. Overall Quality
2. Year built
3. BsmtFinSF1
4. TotalBsmtSF
5. 1stFlrSF

Q2 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?

Ans. The optimum value for the Ridge and Lasso Regression models are :

1. Ridge – 50
2. Lasso – 0.001

I would choose the Lasso Regression because it helps in feature reduction as well, the coefficients which are to be removed are reduced to 0. So, Lasso has an advantage over Ridge, that's why Lasso should be used.

Q3. 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?

Ans. The current most important predictors in the Lasso model are:

1. Overall Quality
2. Year built
3. BsmtFinSF1
4. TotalBsmtSF
5. 1stFlrSF

With these predictors the r2 score of the model is 86% for the training data and 77% for the Test data. After removing these important predictors the r2 score of the model dropped to 67% for the training data and 57% for the test data. The important predictors for the new model are:

1. BsmtFinSF2
2. BsmtUnfSF
3. 2ndFlrSF
4. LowQualFinSF
5. GrLivArea

Q4. 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?

Ans. Generalisation and high test scores make the model robust. The idea is to have high test scores after the prediction given that the training scores are higher than the test scores. If the training scores are high and test scores are low or they have significant difference in them, then it means that the data has been overfitted. Overall once we are training the model and we need to make sure that the test data is not seen by model in any case otherwise it will learn the data and overfit. The train-test split needs to be done before the training of the model takes place and the test data doesn't need to touched while training

the model. However, proper pre-processing steps needs to be taken before training the model or else the difference in between the data can be high.

Robustness of a model is generally not solely based on high test scores, but also depends on the assumption that the train scores are higher than the test scores. Both scores have to be high enough to be acceptable for the specific business case and expectations of the model.

Also we need to keep in mind that the test and train score are high enough to meet the business standards and the helps in solving the problem at hand. The models and the features used in solving the problem statement can also vary and the train-test score as well depending on the domain knowledge and the problem statement at hand. As per Occum's Razor rule the model does not require to be more complex than it needs to be.