Advanced Linear regression

Assignment Part-2

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 variable after the change is implemented?

Answer: The optimal value of alpha for the ridge regression came out to be 0.1. For Lasso regression, the optimal value came out to be 0.0001.

For Lasso Regression, when we choose an alpha value twice that of current value, the R2_Score decreases slightly. But one of the distinct differences is the reduction in the number of the model co-efficients. When the alpha value was 0.0001, the number of co-efficients were 100, but upon doubling to 0.0002, the number of co-efficients reduced to 80.

For Ridge Regression, upon increasing the alpha value form 0.1 to 0.2, I found that the R2_score of train dataset reduced drastically from 82% to 80%. But the test R2_score remain same at 82%. This shows that the model underfitted.

In both Lasso and Ridge scenario, upon increasing the Alpha value, the value of co-efficient reduced.

After changing the alpha value, for both Lasso and Ridge regression, the most prominent variable remains the same, which is GrLivArea. But the co-efficient value corresponding to it reduced after increasing the alpha value.

Question 2:

You have determined the optimal value of the lambda for ridge and lasso regression during the assignment. Now which one will you choose to apply and why?

Answer: For Ridge regression, the optimal alpha value came out to be 0.1. For Lasso regression, the optimal alpha value came out to be 0.0001. But the performance of the model with their optimal alpha value was different. For Lasso regression, the R2_score for train and test dataset was slightly higher than the Ridge regression R2 score. But, the main reason for this is, the Lasso regression had close to 80 variables under them, but Ridge had just 15 variables (We used RFE Feature selection technique) and we still achieved close to 82% R2_score. Having just 15 variables makes it more interpretable and hence easy for us to explain the effects of variables on the Sale Price to the business.

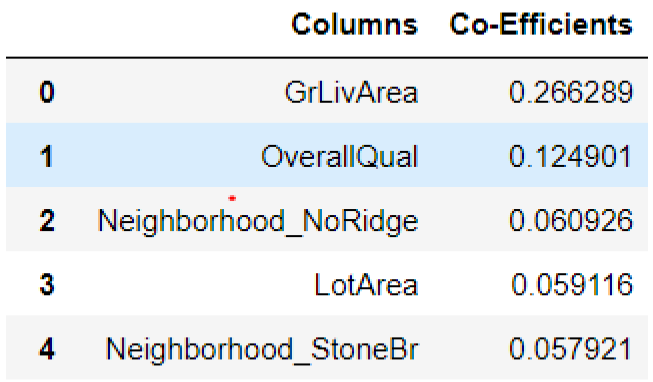
Hence, I would prefer to go ahead with the Ridge regression coupled with Feature selection technique like RFE. But without the feature selection technique like RFE, Lasso would be preferred as Ridge would have all the 150 odd variables in their equation, but Lasso would have reduced that to 80 variables (Lasso pulls the non-significant co-efficient to 0).

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: Before the exclusion of the top 5 most significant predictor variables, below are the list of them:

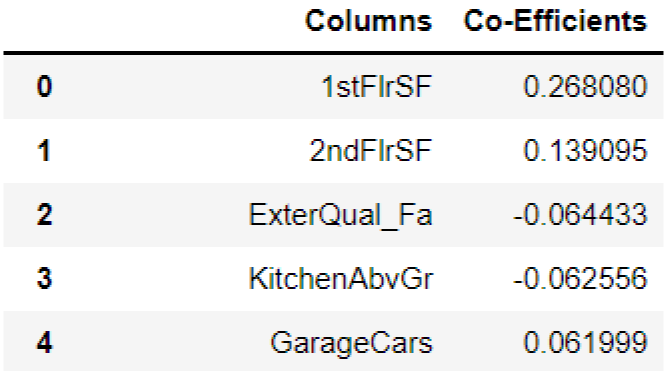
1. GrLivArea
2. OverallQual
3. Neighborhood_NoRidge
4. LotArea
5. Neighborhood_StoneBr



	Columns	Co-Efficients
0	GrLivArea	0.266289
1	OverallQual	0.124901
2	Neighborhood_NoRidge	0.060926
3	LotArea	0.059116
4	Neighborhood_StoneBr	0.057921

After removing the top 5 most significant predictor variables, and rebuilding the model, we find that the below 5 variables are the most significant predictor variables

1. 1stFlrSF
2. 2stFlrSF
3. ExterQual_Fa
4. KitchenAbvGr
5. GarageCars



	Columns	Co-Efficients
0	1stFlrSF	0.268080
1	2ndFlrSF	0.139095
2	ExterQual_Fa	-0.064433
3	KitchenAbvGr	-0.062556
4	GarageCars	0.061999

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: To make sure that the model is more robust and generalizable, we need to make sure that we add the some more weight to the regularization term i.e increase the lambda so that the training dataset would not memorize the training data, and hence in turn have higher variance.

But increasing the regularization term would lead to decrease in the accuracy. This is because, the regularization term will add more weight to the error term, resulting in increase of the error term, and hence in the accuracy of the training dataset. But the accuracy of the test data set has a higher probability of faring better than the accuracy of the test data set without regularization, as the model is now more robust and is less prone to slight deviation in unseen data.