Assignment Part-II

Questions

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 1:

The optimal alpha value for the models are as follows:

Ridge Regression – 100

Lasso Regression – 0.001

The increase in the alpha value will increase the strength of the regularization penalty thereby reducing the coefficients of the features towards 0 in case for Ridge and to zero in case of Lasso.

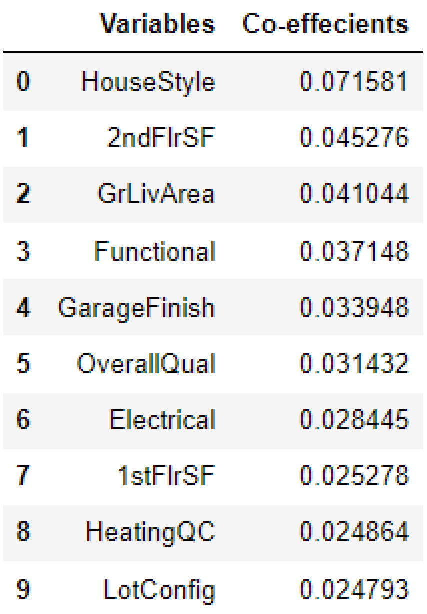
The model should become more robust towards multi-collinearity and overfitting. This can be observed in slight reduction in the accuracy measure (r2). Although accuracy has decreased but this makes the model more conservative.

If we look at the important predictor variables there is a slight change in that some variables become less impactful, and some move up the importance hierarchy.

Important Features pre and post change in alpha value:

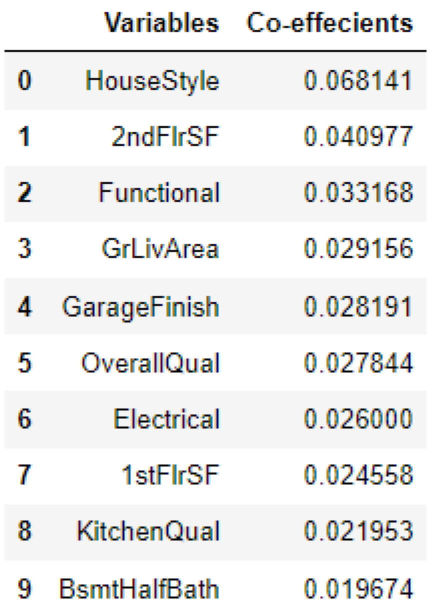
Ridge Regression Top Features:

Before doubling of Alpha



	Variables	Co-effecients
0	HouseStyle	0.071581
1	2ndFlrSF	0.045276
2	GrLivArea	0.041044
3	Functional	0.037148
4	GarageFinish	0.033948
5	OverallQual	0.031432
6	Electrical	0.028445
7	1stFlrSF	0.025278
8	HeatingQC	0.024864
9	LotConfig	0.024793

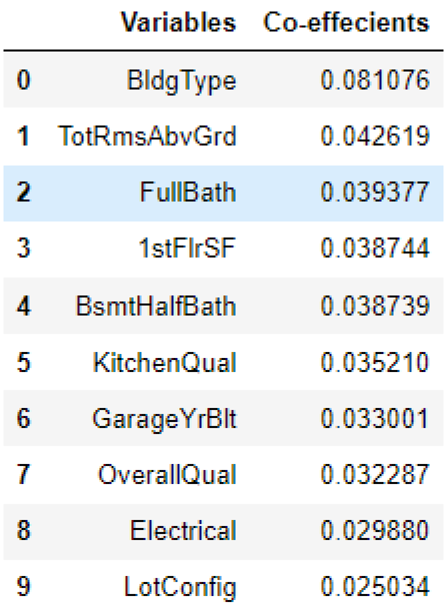
After doubling of Alpha



	Variables	Co-effecients
0	HouseStyle	0.068141
1	2ndFlrSF	0.040977
2	Functional	0.033168
3	GrLivArea	0.029156
4	GarageFinish	0.028191
5	OverallQual	0.027844
6	Electrical	0.026000
7	1stFlrSF	0.024558
8	KitchenQual	0.021953
9	BsmtHalfBath	0.019674

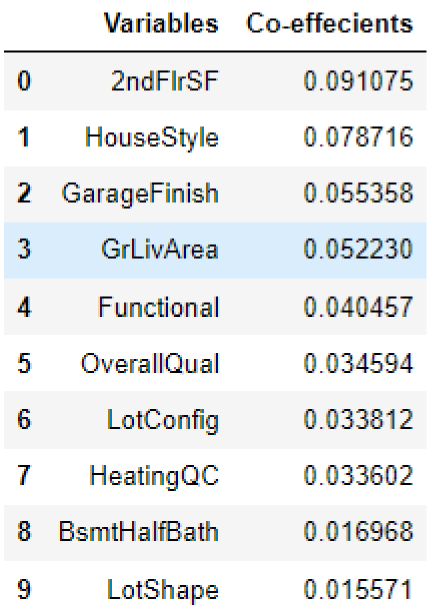
Lasso Regression Top Features:

Before doubling of Alpha



	Variables	Co-effecients
0	BldgType	0.081076
1	TotRmsAbvGrd	0.042619
2	FullBath	0.039377
3	1stFlrSF	0.038744
4	BsmtHalfBath	0.038739
5	KitchenQual	0.035210
6	GarageYrBlt	0.033001
7	OverallQual	0.032287
8	Electrical	0.029880
9	LotConfig	0.025034

After doubling of Alpha



	Variables	Co-effecients
0	2ndFlrSF	0.091075
1	HouseStyle	0.078716
2	GarageFinish	0.055358
3	GrLivArea	0.052230
4	Functional	0.040457
5	OverallQual	0.034594
6	LotConfig	0.033812
7	HeatingQC	0.033602
8	BsmtHalfBath	0.016968
9	LotShape	0.015571

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 2:

Overall accuracy for both the regularization models is pretty same however since we get same level of accuracy with fewer variables in Lasso the model complexity is somewhat reduced hence, we can go with Lasso Regression.

One major callout is that since the train and test accuracies are so close the model needs further fine tuning.

Question 3:

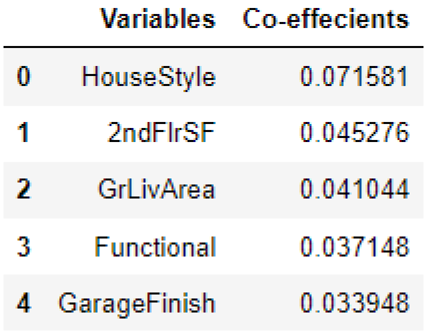
After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Answer 3:

Ridge Regression:

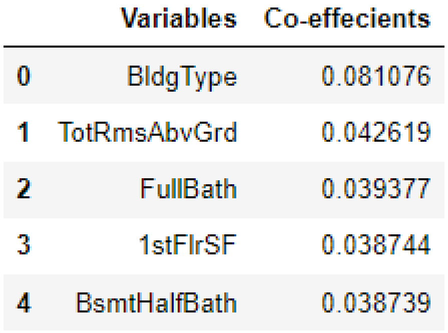
Top 5 Variables

Original Model:



	Variables	Co-effecients
0	HouseStyle	0.071581
1	2ndFlrSF	0.045276
2	GrLivArea	0.041044
3	Functional	0.037148
4	GarageFinish	0.033948

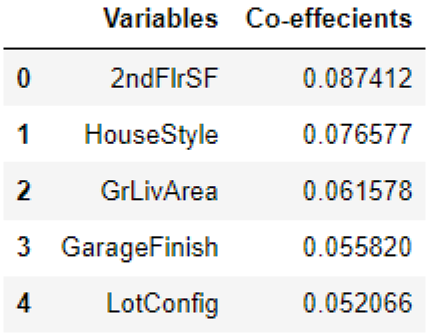
Post Exclusion of earlier top 5 variables:



	Variables	Co-effecients
0	BldgType	0.081076
1	TotRmsAbvGrd	0.042619
2	FullBath	0.039377
3	1stFlrSF	0.038744
4	BsmtHalfBath	0.038739

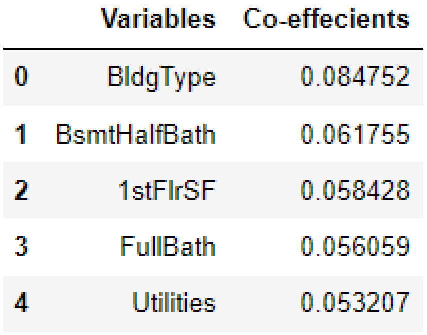
Lasso Regression:

Original Model:



	Variables	Co-effecients
0	2ndFlrSF	0.087412
1	HouseStyle	0.076577
2	GrLivArea	0.061578
3	GarageFinish	0.055820
4	LotConfig	0.052066

Post Exclusion of earlier top 5 variables:



	Variables	Co-effecients
0	BldgType	0.084752
1	BsmtHalfBath	0.061755
2	1stFlrSF	0.058428
3	FullBath	0.056059
4	Utilities	0.053207

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 4:

A robust model will perform well even if there are variation in the data giving more reliable outcomes.

A model can be made robust and generalized through the use of following strategies:

- Adding more data – Adding as much training data as possible improves model robustness as the data might become more representative covering different types of possible variations.
- Proper pre-processing – Following relevant pre-processing steps like
  - Outlier **detection** and handling
  - Feature engineering
- Cross Validation - Techniques like k-fold cross-validation helps assess model performance on different subsets of the data. This helps identify overfitting and ensures robustness.
- Ensemble Approach – Using different models and combining the output will help model robustness.
- Holdout Validation – This approach helps in understanding the model performance on unseen different data subsets and helps to evaluate test accuracy.
- Simple Model Constructs – Model constructs which are not overtly complex tend to generalize better.

Mostly improvement in robustness or generalization of a model will lead to slight reduction in accuracy on the training data as we remove features which might be collinear, and also more features might be explaining the train data more (overfitting) so removing them will adversely affect the model accuracy. This trade-off will always be at the core of model construct selection.

However, a generic and robust model will have although a slightly lower but a stable accuracy measure for variations in the data.