Advanced Regression Assignment Part-II

Subjective Questions

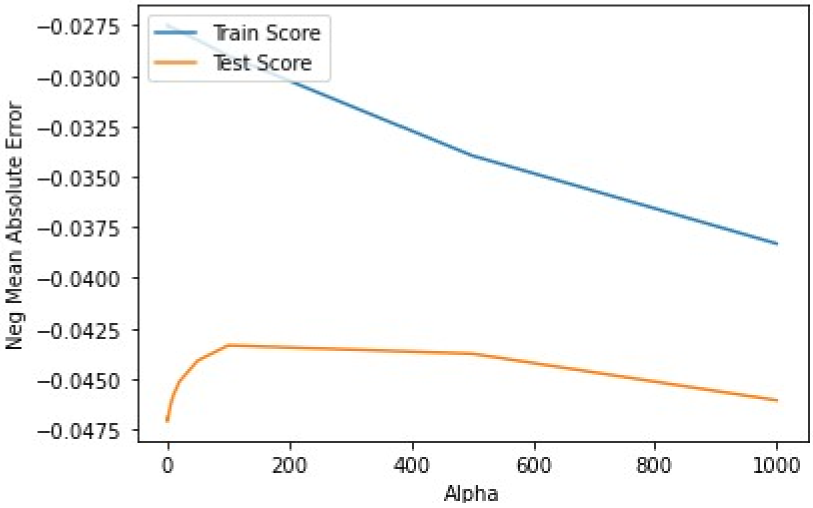
Question 1:

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

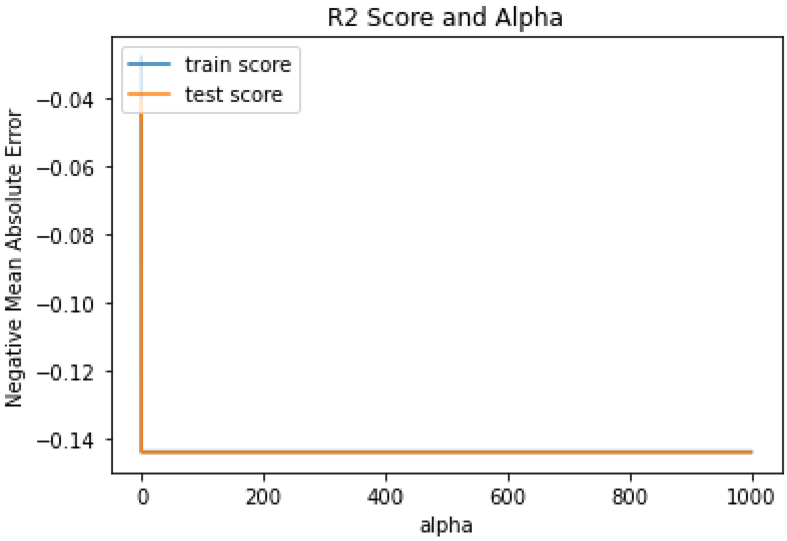
The current optimal value of alpha for ridge regression is 100. The train R2 (square) score for this is 0.95, and the test R2 (square) score is 0.9. If we **double** the value of alpha for the same parameters, we get the train R2 (square) score to be 0.94 and the test R2 (square) score to become 0.9.

The Neg Mean Absolute Error vs Alpha graph for ridge regression is given below:



The optimal value alpha for lasso regression is 0.001. The train R2 (square) score for this value of alpha is 0.94 and the test R2 (square) value is 0.91. If we **double** the value of alpha, we get the train R2 (square) score to be 0.93 and the test R2 (square) score to be 0.91.

The Neg Mean Absolute Error vs Alpha graph for lasso regression is given below:



	Optimal Alpha	**Doubled Alpha**
Ridge	Alpha = **100**	Alpha = **200**
Train R2	0.95	0.94
Test R2	0.9	0.9
Lasso	Alpha = **0.01**	Alpha = **0.02**
Train R2	0.94	0.93
Test R2	0.91	0.91

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:

Please refer the table above.

We have determined the optimal value for ridge (alpha = 100) and lasso (alpha = 0.01). We know that while ridge has a higher R2 score for the training set (0.95), its test R2 score is lower that the test R2 for lasso.

On the other hand, for our optimal value of lasso (alpha = 0.01), our train R2 score is 0.94, and its test R2 score is 0.91. Both these scores are quite good, so we should think of another factor to choose between ridge and lasso. We know that lasso regression results in a model which has lesser parameters because it forces some coefficients to become zero. Thus, the lasso model has lesser coefficients to deal with, which is better for practical purposes.

Therefore, given the values we have obtained for optimal alphas, **we choose the model with lasso regression** because it gives us lesser number of important features, while dropping the unimportant ones.

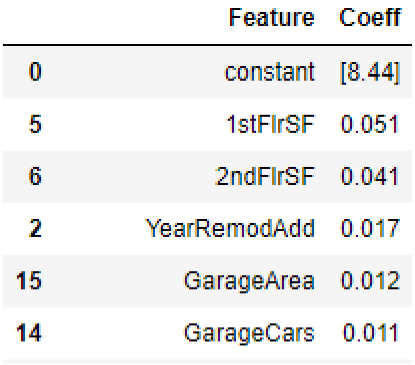
Question 3:

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?

Answer:

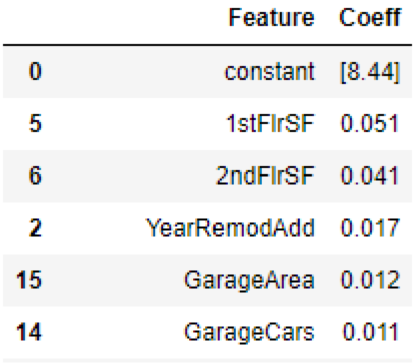
On running the same code with having the top five significant variables removed, the found the below 5 variables to be the next most significant ones:

Lasso (along with the constant):



	Feature	Coeff
0	constant	[8.44]
5	1stFlrSF	0.051
6	2ndFlrSF	0.041
2	YearRemodAdd	0.017
15	GarageArea	0.012
14	GarageCars	0.011

Ridge (along with the constant):



	Feature	Coeff
0	constant	[8.44]
5	1stFlrSF	0.051
6	2ndFlrSF	0.041
2	YearRemodAdd	0.017
15	GarageArea	0.012
14	GarageCars	0.011

Question 4:

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?

Answer:

A model is considered to be robust when a variation in the data does not affect its performance. A model is considered to be generalisable when it can adapt well to previously unseen data, but is drawn from the same distribution one uses to create the model in the first place.

There are several factors that affect the robustness and generalisability of a model. One of the most crucial factors affecting both is overfitting. One must make sure one's model does not overfit. To this end, the outliers in a dataset should be treated with care. Often times, rare outliers may be present in our dataset that affect that coefficients to a large extent but, because of their rare nature, may actually not be so important for our model. We must use our domain knowledge to perhaps drop a few of them if need be. Moreover, the test accuracy should ideally not be less than the training accuracy. If there is a huge difference between the training and test accuracy, where the former is higher than the latter, we may risk overfitting the model, which would make our model non-robust and not generalisable.

A very high accuracy is generally not considered to be a good sign. This is because a very accurate model has a high chance of overfitting the data. Neither should our model be too complex. Care must be taken to strike the right balance between a model’s complexity and accuracy.