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 1:

The optimal value of alpha is the hyperparameter which provides for an optimum trade-off between bias and variance. If the lambda is too high, it will lead to underfitting (simpler model) and if lambda is too low it will lead to overfitting (complex model). Therefore, it is important to get the optimal value for Ridge and Lasso regression.

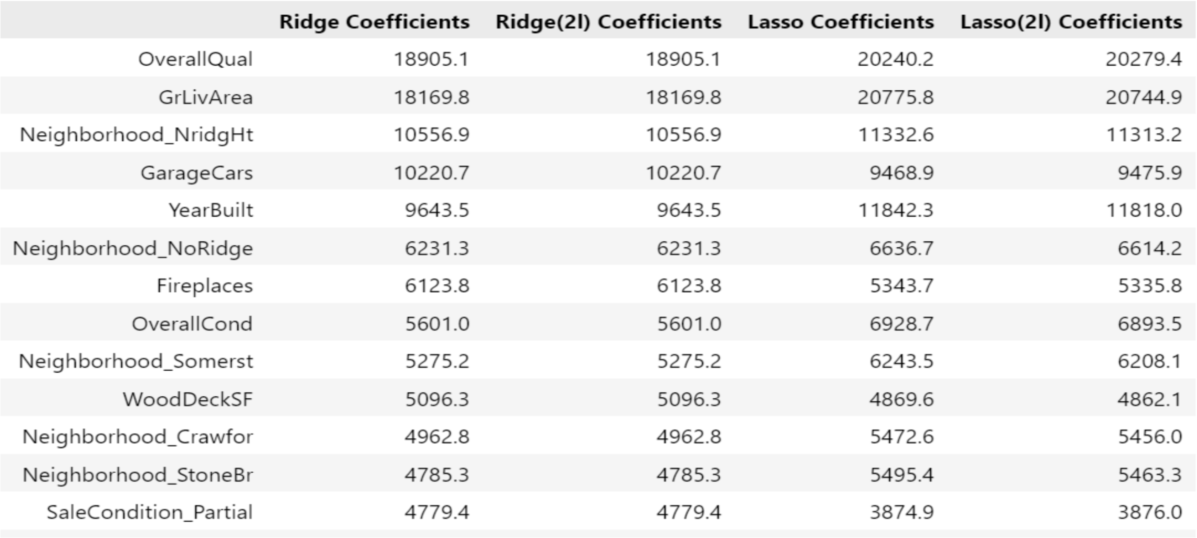
The optimal value of alpha for our current model for ridge regression is 70 and for lasso regression it's 20

-We observe that there is very minimal change in the train & test R-square for both ridge and lasso regression after lambda is doubled. The R-square value reduces marginally and the RMSE increases marginally for our model after doubling lambda.

Metrics after doubling alpha:

Regression Type	Ridge Regression (alpha = 20)	Ridge Regression (alpha = 20*2)	Lasso Regression (alpha = 70)	Lasso Regression (alpha = 70*2)
Train R-squared	0.833	0.832	0.83	0.83
Test R-squared	0.852	0.852	0.848	0.849
Test Root Mean Square Error	26546	26545	26844	26811

Coefficients in Ridge & Lasso



	Ridge Coefficients	Ridge(2l) Coefficients	Lasso Coefficients	Lasso(2l) Coefficients
OverallQual	18905.1	18905.1	20240.2	20279.4
GrLivArea	18169.8	18169.8	20775.8	20744.9
Neighborhood_NridgHt	10556.9	10556.9	11332.6	11313.2
GarageCars	10220.7	10220.7	9468.9	9475.9
YearBuilt	9643.5	9643.5	11842.3	11818.0
Neighborhood_NoRidge	6231.3	6231.3	6636.7	6614.2
Fireplaces	6123.8	6123.8	5343.7	5335.8
OverallCond	5601.0	5601.0	6928.7	6893.5
Neighborhood_Somerst	5275.2	5275.2	6243.5	6208.1
WoodDeckSF	5096.3	5096.3	4869.6	4862.1
Neighborhood_Crawfor	4962.8	4962.8	5472.6	5456.0
Neighborhood_StoneBr	4785.3	4785.3	5495.4	5463.3
SaleCondition_Partial	4779.4	4779.4	3874.9	3876.0

-For Ridge regression, the coefficients remain almost the same whereas in Lasso regression, there are some minor changes to the coefficients.

- The important predictor variables still remain the same for Lasso & Ridge regression, even after doubling lambda

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:

Metrics for Ridge and Lasso:

Regression Type	Ridge Regression	Lasso Regression
Train R-squared	0.833	0.83
Test R-squared	0.852	0.848
Test Root Mean Square Error	26546	26844

- We will choose a which model to apply based on the overall accuracy of the model, how good it is able to explain the variability of the target variable because after seeing the test and train accuracy, we can be sure that after regularization, there are overfitting or underfitting issues in the model

- We will make a decision between Ridge and Lasso based on observing the test and train R-squared values.

- We select the model which does not have much difference between train and test metrics, and also based on which of these models have a lower RMSE values.

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 3:

The model metrics after excluding the five most important predictor variables are:

- Train R-squared = 0.598
- Test R-squared = 0.605
- Test RMSE score = 43351

- After dropping the top 5 important variables, we could observe that in Lasso regression, the train & test R-square reduces from 0.8 to 0.6. This explains that we miss out on explaining about 20% variability when we remove top 5 important variables from the model.

- The RMSE values increases after removing the top 5 important variables

- The five important predictor variables now are:

1. BsmntQual	**2. FirePlaces**	**3. SaleCondition**	**4. WoodDeckSF**	**5. BsmtFinType2**

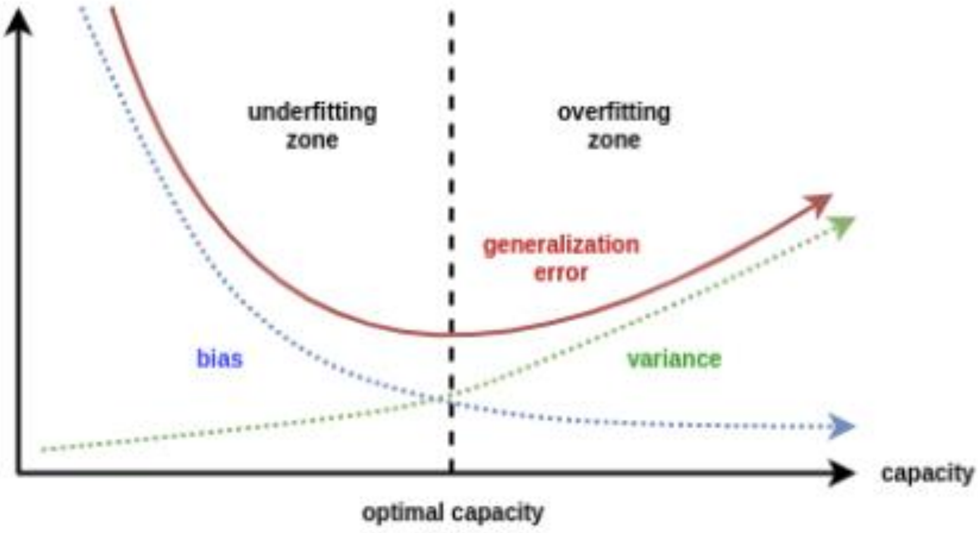
Question 4:

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 4:

We can make sure our model is robust and generalisable, based on the train & test R-square values for the model.

- There should be not a high difference between the R-square values of train and test data. If the model has high train R-square and low test R-square means that it is memorizing the data, and not able to understand the trends and generalize, changes to input will have significant impacts to the output. It has high variance and it is not robust.
- To make sure, there is no overfitting, we can use regularization but we have to choose an optimal value of lambda which can reduce the variance but this will result in decreasing the accuracy of the model because it makes the model less complex leading to reduction in variance and increase in bias.
- But this bias-variance trade-off helps us to make the model generalisable and robust
- When we try to find an optimum value of lambda to make the model more generalisable during regularization, there will be an increase in the generalisation error and decrease in accuracy of the model. This can be seen in the image below:



Source: https://towardsdatascience.com/regularization-and-geometry-c69a2365de19