Advanced Regression Assignment

Question 1

What is the optimal value of alpha for ridge and lasso regression?

Ans: The optimal value of alpha for Ridge Regression is 1 and for Lasso Regression it is 0.0001

What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?

Ans: In both lasso and ridge regression models, the effect of doubling alpha has pushed the coefficients closer to 0 (for both positive and negative coefficients). A few of the predictor variables have changed the order of importance. In case of the ridge regression, the first three predictor variables are in the same position where as in lasso most of the important predictor variables retain the order.

What will be the most important predictor variables after the change is implemented?

Ans:

For Ridge regression:

The first three predictor variables remained the same even after doubling alpha, however there is a slight rearrangement of order as the importance decreases.

Sl No	Best value of alpha =1	Double the value of alpha =2
1	Exterior1st_BrkComm	Exterior1st_BrkComm
2	HouseStyle_2.5Fin	HouseStyle_2.5Fin
3	MSSubClass_2-1/2 STORY ALL AGES	MSSubClass_2-1/2 STORY ALL AGES
4	MSSubClass_DUPLEX - ALL STYLES AND AGES	MSSubClass_1-1/2 STORY FINISHED ALL AGES
5	MSSubClass_2-STORY 1945 & OLDER	MSSubClass_DUPLEX - ALL STYLES AND AGES

For Lasso :

The first five predictor variables remained the same even after doubling the value of alpha, however there is a slight rearrangement of order as the importance decreases.

Sl No	Best value of alpha =0.0001	Double the value of alpha=0.0002
1	Exterior1st_BrkComm	Exterior1st_BrkComm
2	LandSlope_Sev	LandSlope_Sev
3	HouseStyle_2.5Fin	HouseStyle_2.5Fin
4	SaleType_CWD	SaleType_CWD
5	ExterCond_Po	ExterCond_Po

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?

Ans:

Ridge regression works best when the model has many predictor variables that can affect the target variable. Where as Lasso works best when the model has few predictor variable of which a few have to be selected.

I would choose ridge regression over lasso for this problem as there are many variables that can play a role in fixing a sale price for the house. Moreover in this problem the mean square error is lower in ridge regression compared to lasso regression and the r2score is higher for ridge regression than lasso.

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?

Ans:

1. MSSubClass_2-1/2 STORY ALL AGES
2. Functional_Maj2
3. MSSubClass_2 FAMILY CONVERSION - ALL STYLES AND AGES
4. MSSubClass_2-STORY 1945 & OLDER
5. MSSubClass_1-STORY W/FINISHED ATTIC ALL AGES

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?

Ans:

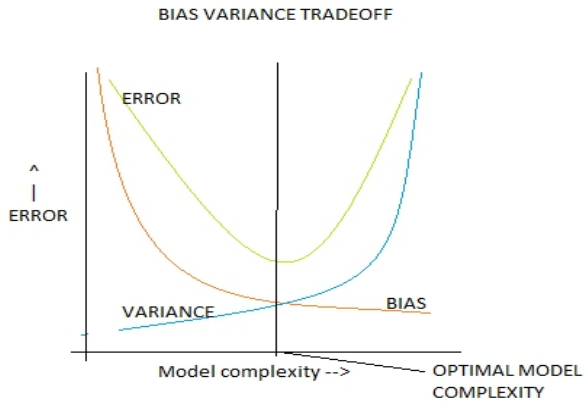
A model is said to be generalisable if it has not been overfitting the training data. In other words it does not memorize the training data. When the model comes across data that is different from what it has been trained on, it should still give a reasonable response with acceptable error. The model is considered robust if its results are consistently accurate even if some of the input variables changed.

A model can be made generalizable and robust by striking a balance between overfitting, underfitting and accuracy. Regularization is one such technique that reduces the overfitting by penalizing the coefficients that are large. The model has to be at that complexity level that can recognise the

underlying patterns but generalised enough not to memorize the training data. Simpler models are more robust and generalisable.

The implications of keeping a model robust and generalisable does bring down the accuracy score on the training data but will be more consistent on the test set. This is because we compromise on the complexity of the model to make it more generalisable.

The below figure shows the bias and variance trade off with respect to the model complexity. The optimal complexity is when the model is having enough bias to be generalised and enough variance where the model gives the least error.



The below figure shows the bias and variance trade off with respect to the regularization parameter. The optimal value is when the model is having enough bias to be generalised and enough variance where the model gives the least error. The cross validation set is used to minimize the error for the model while choosing the best parameter