Assignment Part 2

Submitted by Prasanna M P

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?

The Optimal value of alpha for **ridge is 2** and for **lasso it is 0.0001.**

Building Ridge Model by doubling the value of alpha to 4:

```
The R2 Score of the model on the test dataset for doubled alpha is
0.8259998671982054
The MSE of the model on the test dataset for doubled alpha is
0.0018622905336132816
```

Building Lasso Model by doubling the value of alpha to 0.0002

```
The R2 Score of the model on the test dataset for doubled alpha is
0.8237798637847477
The MSE of the model on the test dataset for doubled alpha is
0.0018860508105446841
```

Since the alpha value was quite small, doubling it does not do any significant change in both the models and the R2 and MSE remains almost the same. The most important predictor variables also remain the same, but the Central Air Conditioned - Yes attribute gains higher importance than Lot Area in the Lasso Model.

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:

#Removing the 5 most important predictor variables from the incoming dataset

X_test_rfe3 =
X_test_rfe2.drop(['Total_sqr_footage','GarageArea','TotRmsAbvGrd','OverallCond','LotArea'],axis=1)

X_train_rfe3 =
X_train_rfe2.drop(['Total_sqr_footage','GarageArea','TotRmsAbvGrd','OverallCond','LotArea'],axis=1)

```
The R2 Score of the model on the test dataset is 0.7330077964268464
The MSE of the model on the test dataset is 0.002857567090648254
```

```
The most important predictor variables are as follows:
```

	Lasso Co-Efficient
LotFrontage	0.146535
Total_porch_sf	0.072445
HouseStyle_2.5Unf	0.062900
HouseStyle_2.5Fin	0.050487
Neighborhood_Veenker	0.042532

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?

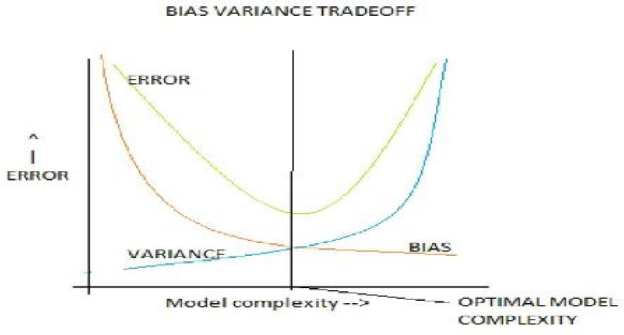
Ans:

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.