Assignment Part – II

Q1) 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?

Soln: Optimal values of alpha for both ridge and lasso are: Ridge = 8 Lasso = 0.001. After doubling both the alphas i.e., 16 and 0.002

- Training R2 score decreases by a bit and test R2 score increases.
- Train RSS increases and test RSS decreases

Most important predictor variable for ridge and lasso before and after doubling the alpha is same i.e., "OverallQual_9"

```
Train RMSE =0.19874983419773665
Test RMSE =0.376907977097271
Train RSS = 64.2976217536073
Test RSS = 44.63764315737415
Ridge max col = OverallQual_9
Ridge max coef = 0.39553702105051347

Ridge_double max col = OverallQual_9
Ridge_double max coef = 0.3174582275374684

Lasso max col = OverallQual_9
Lasso max coef = 0.6831968970380966

Lasso_double max col = OverallQual_9
Lasso_double max coef = 0.7043107936780588
```

Q2) 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?

Soln: We use Lasso regression in this case because for a model having such a high number of features, feature selection becomes important and Lasso does that by equating the coefficients of many features to zero

Q3) 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?

Soln: The top 5 variables are

1) OverallQual_9,
2) OverallCond_9,
3) OverallQual_8,
4) GrLivArea,
5) Neighborhood_Crawfor.

After Removing the above five the top 5 turn out to be

1) Condition2_PosA,
2) 2ndFlrSF,
3) Exterior1st_BrkFace,
4) Functional_Typ,
5) Neighborhood_Somerst.

Q4) 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?

Soln: When the training set is altered, a model is called robust and generalizable if it does not demonstrate a significant change in performance, i.e., the model does not overfit on the training data and can handle new/unseen data adequately. A robust and generalizable model should perform equally well on training and test data when it comes to accuracy. A more generalised model is a model which does not overfit the data. To ensure our model is not overfitting the data we must regularise our model using hyperparameters. These hyperparameters will help in reducing the complexity of the model by penalising features contributing to the complexity. We add a penalty term to the cost function that increases with increasing model complexity. So, we try to bring it down and control model complexity Having a simple model will help ensure that it is robust and more generalised. But the model should not be over simplified otherwise it will underfit the data and this model will be too naive to give us a valid or accurate output. This is the scenario where model is underfitting. Accuracy of the model is defined as the ratio of number of correct predictions to the total number of input samples. The following are the implications of making a model resilient and generalizable on model accuracy: The accuracy of the model will be steadier if we make it more robust, that is, less vulnerable to outliers or changes in test data. This means that slight modifications to the test data set will not result in significant changes in accuracy values. When trying to make the model simpler by penalising the model's complexity, the accuracy on the test set will increase in the beginning. When we have made the model sufficiently basic, the accuracy will stabilise on the test data set. If the accuracy is not maintained, then the model can be underfitted or overfitted. To take care of the existing outliers in the data, we use various techniques to remove them

- Capping the values at a certain threshold
- Removing the outliers manually
- Transforming certain values (exp, log etc.)