Assignment

Advanced Regression Subjective Questions

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

   Answer: In Ridge Regression after plotting the curve between negative mean absolute error & alpha we observe that the value of alpha does increases from zero. The error term was decreasing the train error is showing increasing trend when alpha value increases. The test error is minimum when the alpha value is 2, therefore I have decided to go with alpha value is equal to 2 for Ridge regression.
   And in case of lasso regression 0.01 is the value we have considered. The try to penalize high and try to make most of the coefficients to zero when we increase the value of alpha. When we double these values, the model performance remains same in both the cases.

   The most important variables after the changes has been implemented for ridge regression are:

   - MSZoning_FV
   - MSZoning_RL
   - Neighborhood_Crawfor
   - MSZoning_RH
   - MSZoning_RM
   - SaleCondition_Partial
   - Neighborhood_StoneBr
   - GrLivArea
   - SaleCondition_Normal
   - Exterior1st_BrkFace

   The most important variable after the changes has been implemented for lasso regression are:

   - GrLivArea
   - OverallQual
   - OverallCond
   - TotalBsmtSF
   - BsmtFinSF1
   - GarageArea

- Fireplaces
- LotArea
- LotFrontage

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: In my assignment I have observed that Lasso regression performance was slightly lesser on the given data set when compared to Ridge regression.
   I still decide to choose Lasso regression model for applying finally since Lasso regression helps with feature elimination and our dataset has over 80+ columns so feature elimination was an advantage realizing the most important predictor and it will perform better when the predicator has large coefficients.

   And our Final R2 values for Lasso regression model were 0.89 and 0.88 respectively on the train and test data respectively.

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: Those 5 most important predictor variables that will be excluded are as follows:
   - GrLivArea
   - OverallQual
   - OverallCond
   - TotalBsmtSF
   - GarageArea

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: Even though the accuracy will decrease the model should be as simple as possible. The model should be more Robust and more Generalized as this type of model will perform better with the unseen data.
   The simpler the model more bias but less variance and can be more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data.
   Bias: Bias is error when the model when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in the model when model tries to over learn from the data. High variance means model performs exceptionally well on training data but performs very poor on testing data as it was unseen data for the model.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.