Problem Statement - Part II

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?

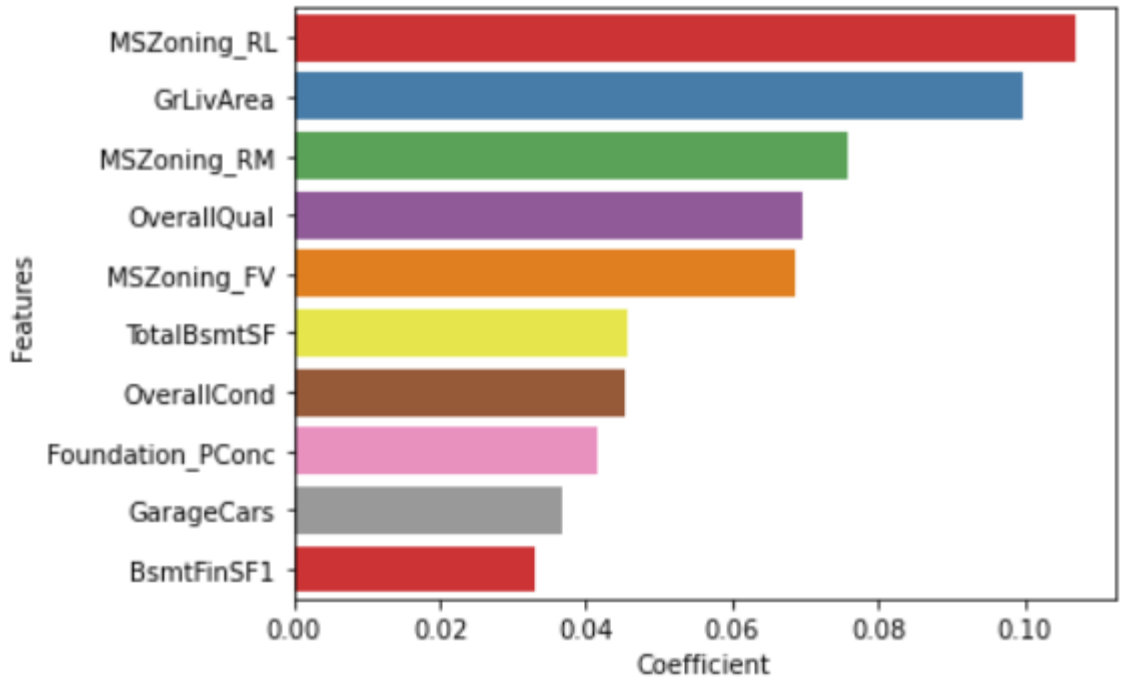
Ans: It is important to regularize coefficients and improve the prediction accuracy, also with the decrease in variance, and making the model interpretably.
Ridge regression, uses a tuning parameter called lambda as the penalty is square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small by using the penalty. The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized. As we increase the value of lambda the variance in model is dropped and bias remains constant. Ridge regression includes all variables in final model unlike Lasso Regression.

Lasso regression, uses a tuning parameter called lambda as the penalty is absolute value of magnitude of coefficients which is identified by cross validation. As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0. Lasso also does variable selection. When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

In the case of ridge regression: - When we plot the curve between negative mean absolute error and alpha we see that, as the value of alpha increase from 0, the error term decrease and the train error is shows increasing trend. When the value of alpha is 10 the test error is minimum so we decided to go with value of alpha equal to 10 for our ridge regression.

For lasso regression we decided to keep very small value that is 0.0004, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero.

When we double the value of alpha for our ridge regression, now we will take the value of alpha the model will apply more penalty on the curve and try to make the model more generalized that is, making model simpler and now thinking to fit every data of the data set.
Similarly, when we increase the value of alpha for lasso we try to penalize more, our model and more coefficient of the variable will reduce to zero. Hence based on Lasso, the factors that generally affect the price are Zoning classification, Living area square feet, Overall quality and condition of the house, Foundation type of the house, Number of cars that can be accommodated in the garage, Total basement area in square feet and the Basement finished square feet area



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?

- The optimal lambda value in case of Ridge and Lasso is as below:
  - Ridge - 10
  - Lasso - 0.0004
- The Mean Squared error in case of Ridge and Lasso are:
  - Ridge - 0.013743
  - Lasso - 0.013556
- The Mean Squared Error of Lasso is slightly lower than that of Ridge
- Also, since Lasso helps in feature reduction (as the coefficient value of one of the feature became 0), Lasso has a better edge over Ridge.

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?

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: The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and 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 i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in 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 model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of 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.

****