Problem Statement-Part ll

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?**

Ans: For ridge regression, when we plot negative mean absolute error and alpha, value of alpha is increasing from 0 and error term decreases and the train error is showing increasing trend when value of alpha increases. when the value of alpha is 9 the test error is minimum so we decided to go with value of alpha equal to 9 for our ridge regression.

For lasso regression I have decided to keep very small value that is 0.0005, when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero.

The model will apply more values to the curve and attempt to become more generalised when the alpha for ridge regression is doubled. This results in the model becoming simpler and thinking that it can fit all of the data in the data set. Similar to how we penalise our model more when we increase the value of alpha for the lasso, more coefficients of the variable will be reduced to zero.

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

1. GrLivArea
2. MSZoning_RL
3. MSZoning_RM
4. OverallQal
5. MSZoning_FV
6. TotalBsmtSF
7. OverallCond
8. Foundation_PConc
9. GarageCars
10. BsmtFinSF1

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?**

Ans: It is important to regularize coefficients and improve the prediction accuracy.

Ridge regression, which uses cross validation to identify the penalty is square of magnitude of coefficients, uses a tuning parameter called lambda. By applying the penalty, the residual sum or squares should be minimal. The coefficients with higher values are penalised because the penalty is equal to lambda times the sum of the squares of the coefficients. The variance in the model is lost as

we raise the value of lambda, while bias stays constant. In contrast to Lasso Regression, Ridge Regression includes all variables in the final model.

When performing a lasso regression, the lambda tuning parameter is used as the penalty, which is the absolute magnitude of the coefficients as determined by cross validation. As the lambda value rises, Lasso shrinks the coefficient in the direction of zero, bringing the variables exactly to zero. Lasso performs variable selection as well.

Simple linear regression is performed when lambda value is small. As lambda value increases, shrinkage occurs, and variables with 0 value are ignored by the model.

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?

Ans: The 5 most important predictor variables are:

1. GrLivArea
2. MSZoning_RL
3. MSZoning_RM
4. OverallQal
5. MSZoning_FV

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?

Ans: The model should be as simple as possible because this will increase its robustness and generalizability while reducing accuracy. The Bias-Variance trade-off can also be used to understand it. The bias increases with model complexity while decreasing variance and increasing generalizability. It implies that a robust and generalizable model will perform admirably on both training and test data, i.e., the accuracy does not significantly change for training and test data.