Problem Statement - 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?
Ans:

For ridge regression: -As we plot the curve between negative mean absolute error and alpha, we see that as the value of alpha increases from 0 the error term decreases, while the train error is increasing when the value of alpha increases. The test error is minimal when alpha is 2 so we decided to use this value for our ridge regression.

For lasso regression: - It has been decided to keep very small value that is 0.01, as the model penalizes more as alpha increases, and tries to make most of the coefficients zero. Alpha and negative mean absolute error came to 0.4 initially.

The model will apply more penalty on the curve that makes it easier to make it more generalized, an d the model
will not have to think about fitting every data point of the data set when we double the value of alph a for our ridge regression.

As we increase the value of alpha for lasso, more coefficients of the variable are reduced to zero, and as a result, our r2 square also decreases.

After the changes have been implemented, the most important variables for ridge regression are: -

1. MSZoning_FV
2. MSZoning_RL
3. Neighborhood_Crawfor
4. MSZoning_RH
5. MSZoning_RM
6. SaleCondition_Partial
7. Neighborhood_StoneBr
8. GrLivArea
9. SaleCondition_Normal
10. Exterior1st_BrkFace

For lasso regression, after the changes have been implemented, the following variables are important: -

1. GrLivArea
2. OverallQual3. OverallCond
4. TotalBsmtSF
5. BsmtFinSF1
6. GarageArea
7. Fireplaces
8. LotArea
9. LotArea
10. LotFrontage

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:

With a decrease in variance, and making the model interpretable, regularizing coefficients and impro ving prediction accuracy are important.

Ridge regression uses a tuning parameter called lambda as the penalty is the square of the magnitude of the error coefficients which are identified by cross-validation.
Unlike Lasso regression, Ridge regression includes all variables in the final model.
The Lasso regression technique uses a tuning parameter called lambda, which is an absolute magnitude value of coefficients that are identified by cross-validation. As lambda value increases, Lasso shrinks, so the coefficients skew towards zero, making the variables equal to 0. Lasso also selects variables.
When lambda value is low, it performs simple linear regression, whereas as lambda value increases, it performs complex regression.
The model ignores variables with 0 values as a result of shrinkage.

Q3. After building the model, you realized 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: We will exclude the following five most important predictor variables:
1. GrLivArea
2. OverallQual
3. OverallCond
4. TotalBsmtSF
5. GarageArea

Q4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Ans: Simplicity is the most practical approach, though accuracy will decrease, it will be more practical

It is robust and generalizable. Bias-variance can be viewed as a trade-off. It's simpler, the better

A model with higher bias and lower variance is more generalizable. Therefore, it would be more accurate

As a result, a robust and generalizable model will perform equally well on training and test data

For training and test data, accuracy does not differ much.

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