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: When we plot the curve between negative mean absolute error and alpha for ridge regression, we observe that the error term decreases as alpha increases from zero, and the train error shows an increasing trend. We chose to use a value of alpha equal to 2 for our ridge regression since the test error is lowest when alpha is equal to 2.

I have chosen to maintain a very low value for lasso regression, which is 0.01. As alpha increases, the model attempts to punish more and attempt to make the majority of the coefficient values zero. Alpha and a negative mean absolute error of 0.4 were the initial values.

The model will apply more penalty on the curve and attempt to become more generic when the value of alpha for our ridge regression is doubled. By doing this, we make the model more straightforward and stop trying to fit all of the data in the data set. According to the graph, there is more error for both the test and the train when alpha is 10 or higher.
Similar to how increasing the lasso's alpha penalises our model more and causes more coefficients of the variable to be reduced to zero, increasing the r2 square results in a reduction in both.

Following the implementation of the modifications for ridge regression, the most crucial variables are as follows:

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

Following the implementation of the modifications, the most crucial variables for lasso regression are as follows:

1. GrLivArea
2. OverallQual
3. 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: Regularizing coefficients is crucial for increasing prediction accuracy, reducing variation, and making the model understandable.

Ridge regression, which employs cross validation to identify the penalty is square of magnitude of coefficients, requires 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 when we raise the value of lambda, while bias stays constant. In contrast to Lasso Regression, Ridge Regression incorporates 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 reduces the coefficient in the direction of zero, bringing the variables exactly to zero. Lasso performs variable selection as well. When lambda is small, straightforward linear regression is performed; however, as lambda rises, shrinkage occurs and variables with a value of 0 are ignored by the model.

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: The top 5 important predictor variables that will be excluded are :-

2. GrLivArea
3. OverallQual
4. OverallCond
5. TotalBsmtSF
6. 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: The model should be as simple as feasible 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. Its accuracy implication is that a robust and generalizable model will perform similarly on both training and test data, i.e., the accuracy does not change significantly for training and test data.

Bias: When a model is unable to learn from the data, bias occurs. High bias prevents the model from learning specifics from the data. Model's performance on training and test data is subpar.

Variance: When a model tries to overlearn from the data, variance occurs. High variance indicates that the model performs remarkably well on training data since it was well trained on those data, but it performs dreadfully on testing data because that data was unknown to the model.

To prevent overfitting and underfitting of data, bias and variance must be balanced.