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?

Lasso Regression – Initially alpha is kept at 0.01. When we increase the value of alpha the model tries to penalize more and tries to make most of the coefficient value zero. Initially it came as 0.4 in negative mean absolute error and alpha.

Ridge Regression - When 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 and the train error is showing increasing trend when value of alpha increases. When the value of alpha is 2 the test error is minimum so we decided to go with value of alpha equal to 2 for our ridge regression.

When we double the value of alpha for ridge regression, we will take the value of alpha equal to 10. The model will apply more penalty on the curve and try to make it more generalized without trying to fit every data of the data set. From the graph we can see that when alpha is 10, we get more error for both test and train.

Similarly, when we increase the value of alpha for lasso, we try to penalize our model.

The most important variable after the changes has been implemented for ridge regression 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

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

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

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?

[illegible]

Although the model performance was better by Ridge Regression in terms of R2 values of Train and Test, it is better to use Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables. Lasso regression would be a better option as it would help in feature elimination and the model will be more robust.

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?

The 5 most important predictor variables that will be excluded are -

1. OverallCond
2. OverallQual
3. TotalBsmtSF
4. GrLivArea
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?

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