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:

Ridge Regression:
When we plot the curve between negative mean absolute error (MAE) and alpha, we observe the following:

- As the value of alpha increases from 0, the error term decreases.
- The training error shows an increasing trend as alpha increases.
- At an alpha value of 2, the test error is minimized. Therefore, we chose alpha = 2 for our ridge regression model.

Lasso Regression:
For lasso regression, we decided on a very small alpha value of 0.01. When alpha increases, the model penalizes more, driving most coefficient values towards zero. Initially, the negative MAE was 0.4 for this small alpha value.

Effect of Increasing Alpha:

- **Ridge Regression:** Doubling the alpha value to 10 increases the penalty, making the model more generalized and simpler, avoiding overfitting. The graph shows higher error for both test and train datasets at alpha = 10.
- **Lasso Regression:** Increasing alpha penalizes the model more, reducing more coefficients to zero. As alpha increases, the R^2 value also decreases.

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. 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 improving prediction accuracy, reducing variance, and enhancing model interpretability.

Ridge Regression:

Ridge regression uses a tuning parameter called lambda, which acts as a penalty term proportional to the square of the magnitude of coefficients. This lambda value is determined through cross-validation. The goal is to minimize the residual sum of squares while incorporating this penalty. Specifically, the penalty is lambda times the sum of squares of the coefficients, meaning coefficients with larger values are penalized more heavily. As lambda increases, the model's variance decreases while the bias remains constant. Unlike lasso regression, ridge regression retains all variables in the final model.

Lasso Regression:

Lasso regression also employs a tuning parameter called lambda, but its penalty is the absolute value of the magnitude of the coefficients, determined through cross-validation. As the lambda value increases, lasso regression shrinks the coefficients towards zero, effectively performing variable selection by setting some coefficients to exactly zero. When lambda is small, the model performs similarly to simple linear regression. As lambda increases, the model applies more shrinkage, and variables with coefficients reduced to zero are excluded from the model.

From our analysis the MSE is lesser in case of Lasso and hence we should use that and it also helps in feature reduction.

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: Those 5 most important predictor variables that will be excluded are :-

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: 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.