Problem Statement – Part II

Question 1

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?

Answer-

- Optimal value of alpha for Ridge model = 10
- Optimal value of alpha for Lasso model = 0.001

Below are the changes in the model if we double the value of alpha for Ridge and Lasso regression-

Changes in Ridge-

- R2 score for the training set saw a slight decrease from 0.94 to 0.93.
- The R 2 score for the test set remained constant at 0.93.

Changes in Lasso-

- R2 score for the training set saw a slight decrease from 0.92 to 0.91.
- The R2 score for the test set saw a slight decrease from 0.93 to 0.91.

Most important predictor variables after the change implemented-

1. GrLivArea
2. OverallQual_8
3. OverallQual_9
4. Functional_Typ
5. Neighborhood_Crawfor
6. Exterior1st_BrkFace
7. TotalBsmtSF
8. CentralAir_Y

The above answers come on the basis of whatever coding have done in jupyter notebook

Question 2

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?

Answer-

The selection between Ridge and Lasso regression hinges on the specific needs of the task at hand.

- If the aim is to conduct feature selection from a large pool of variables, the preferred choice would be Lasso regression due to its inherent capability of reducing irrelevant features.
- Conversely, if the goal is to prevent overly large coefficients or mitigate the magnitude of coefficients, Ridge Regression becomes favorable. This method helps in controlling and minimizing the impact of variable coefficients.

Question 3

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?

Answer-

Upon discovering that the five most crucial predictor variables identified by the Lasso model are absent in the incoming data, the next step is to construct a new model. This new model will exclude these five essential predictors.

The previously highlighted top 5 predictors in the Lasso model consisted of:

- OverallQual_9
- GrLivArea
- OverallQual_8
- Neighborhood_Crawfor
- Exterior1st_BrkFace.

Upon excluding our previously identified top 5 predictors from the Lasso model, the revised top 5 predictors we got after creating another model are as follows:

- 2ndFlrSF
- Functional_Typ
- 1stFlrSF
- MSSubClass_70

- Neighborhood_Somerst

Question 4

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?

Answer-

Ensuring a model is robust and generalizable involves a few key considerations.

- Robustness: A robust model remains stable despite variations in the data. It means the model's performance doesn't fluctuate significantly with changes in the dataset.
- Generalizability: A generalizable model adapts well to new, unseen data from the same data distribution used to create the model. It's essential for the model to perform consistently with new data.

To attain robustness and generalizability, preventing overfitting is crucial. Overfitting occurs when a model learns from noise or specific details in the training data, resulting in a high variance. This leads to an inability to recognize patterns in new, unseen data.

To prevent overfitting, the model should not be excessively complex. An overly complex model is prone to high variance and may not generalize well.

Considering accuracy, a highly complex model might achieve a high accuracy on the training data, but it's less likely to perform well on new data. Balancing model accuracy and complexity is crucial. Reducing variance (by simplifying the model) may introduce some bias, which might slightly reduce accuracy.

In essence, finding a balance between accuracy and complexity is vital. Techniques like Ridge Regression and Lasso, known as regularization methods, help strike this balance by controlling model complexity and improving its generalizability.