Problem-Statement-II

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:

   In the case of ridge regression, we observe a negative mean error. When we plot the alpha curve, we can see that as the alpha value increases from 0, the error term decreases, and the learning error reduces. Specifically, as the alpha value increases, the test error decreases. The smallest test error occurs when the alpha value is set to 2, so we decide to use an alpha value of 2 for the ridge regression.

   For lasso regression, I chose to set the alpha value very small at 0.01. As we increase the alpha value, the model tends to penalize more and aims to force many coefficient values to become zero. Initially, there is a negative mean error, and the alpha value is 0.4.

   When we double the alpha value for ridge regression, making it equal to 10, the model uses more points on the curve and attempts to simplify the model. However, it's important not to oversimplify and try to fit every profile in the dataset. As seen from the graph, increasing alpha to 10 results in more errors for both testing and training.

   Similarly, when we increase the alpha value for lasso regression, we are trying to penalize the model more, and this leads to many variable coefficients dropping to zero. This ultimately leads to an improved R-squared value.

   The most significant changes observed after applying Ridge Regression are as follows:

   1. MSZoning_FV
   2. MSZoning_RL
   3. Neighborhood_Crawfor
   4. MSZoning_RH
   5. MSZoning_RM

6. Discount Status_Partial
7. Neighborhood_StoneBr
8. GrLiv Area
9. Discount Status_Normal
10. Exterior1st_BrkFace

For Lasso Regression, the most important changes are:

1. GrLiv Area
2. OverallCond
3. OverallQual
4. Total BsmtSF
5. BsmtFinSF1
6. GarageCars
7. KitchenQual
8. YearBuilt
9. HeatingQC
10. ExterQual

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 reason for selecting Lasso regression over Ridge regression is that, based on the R-squared (R2) score on the test dataset, Lasso regression slightly outperforms Ridge regression for this particular problem.

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:

   In the current Lasso regression model, the five most influential predictor variables, namely GrLivArea, OverallQual, OverallCond, TotalBsmtSF, and GarageArea, have been identified as having a significant impact on the model's performance. However, these variables are not present in the incoming data. Consequently, in the new model, we will need to exclude these five crucial predictor variables since they are unavailable, and we will need to rely on different features for prediction.

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:

The model ought to be generalized so that the test accuracy is not lower than the training score. The model should perform well on datasets other than those used during training. Excessive importance should not be given to outliers, as this could lead to lower model accuracy. To ensure this does not happen, an outlier analysis needs to be conducted, and only those outliers that are relevant to the dataset should be retained. Outliers that do not make sense should be removed from the dataset. If the model is not robust, it cannot be relied upon for predictive analysis