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?

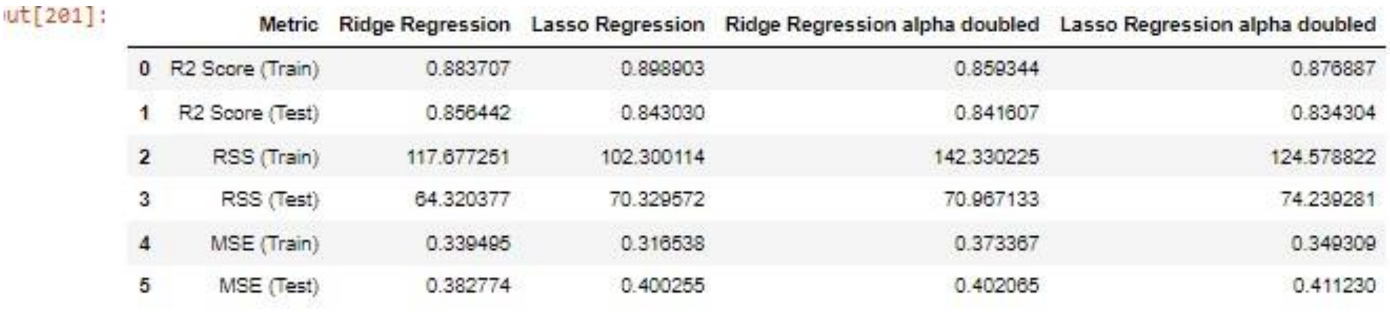
Answer:-

The optimum values of alpha for given Housing data set are

- Ridge Regression model optimal alpha is 500

- Lasso Regression model optimal alpha is 0.01

After performing the Ridge and Lasso with doubled alpha the comparison metrics matrix is furnished below



```
ut[201]:
```

	Metric	Ridge Regression	Lasso Regression	Ridge Regression alpha doubled	Lasso Regression alpha doubled
0	R2 Score (Train)	0.883707	0.898903	0.859344	0.876887
1	R2 Score (Test)	0.856442	0.843030	0.841607	0.834304
2	RSS (Train)	117.677251	102.300114	142.330225	124.578822
3	RSS (Test)	64.320377	70.329572	70.967133	74.239281
4	MSE (Train)	0.339495	0.316538	0.373367	0.349309
5	MSE (Test)	0.382774	0.400255	0.402065	0.411230

Inferences from METRICS:

1. alpha optimum : The Ridge model explains 86% and Lasso model 84% on test data

2. alpha doubled : The Ridge model explains 84% and Lasso model 83% on test data

3. alpha optimum & alpha doubled : The RSS is optimized in ridge compared to Lasso model and increased with alpha.

4. alpha optimum & alpha doubled : Mean squared error of test data in Ridge model is lesser than Lasso model and increased with alpha.

Inference:

From inferences we can learn that for best value of Regularization factor or Optimum Penalty value Ridge and lasso Models will give best results and prediction accuracy will reduce with change of alpha.

Significant Variables:

Analyzing the both models with doubled alpha the most significant variables are furnished below:

1. The selling price of houses positively corelated with Ground living Area , OverallQuality, and Location North Ridge , North Ridge _heights.

2. The selling price of houses Negatively corelated with Near positive off-site feature--park, greenbelt, etc.,basement height is between 90 - 99 inches,Houses having the kitchen quality good , average.

Inference:

There is a no change in significant variables with change of alpha to double how ever it is clearly visible Reduction in coefficints magnitude implies that when alpha increases to inf coefficients tend to close zero in Ridge and becomes Zero in Lasso.

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:-

After analyzing inferences we can conclude that Ridge Model will perform better for prediction of target variable. As most of variables has correlation with target variable. The Ridge model R2-score is high and RSS , RMSE values are low compared to Lasso.

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:

The five most important predictor variables obtained after performing the Lasso Regression to predict Housing prices are GrLivArea, OverallQual, Condition2_PosN, Neighborhood_NoRidge, GarageCars.

The five most important predictor variables after building the Lasso Regression model excluding the five most important predictor variables are 2ndFlrSF, 1stFlrSF, RoofMatl_WdShngl, BsmtQual_Gd, KitchenQual_TA, KitchenQual_Gd.

Question 4

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?

Answer:

Robustness is the ability of a model to perform well on new and unseen data, not just on the data it was trained on. A robust model can handle different types of noise, variations, and uncertainties in the data, and generalize well to different scenarios and domains.

Making a predictive model more robust to outliers is crucial for improving its accuracy and generalization to real-world data. Outliers are extreme data points that significantly differ from the rest of the data and can negatively impact the model's performance.

The following techniques are used to enhance the Model robustness

1. Data Preprocessing: Handling outliers methods include Z-score, modified Z-score, and Interquartile Range (IQR). And Truncate or Winsorize. Utilize robust statistical measures like median and percentile instead of the mean and standard deviation, which are sensitive to outliers.

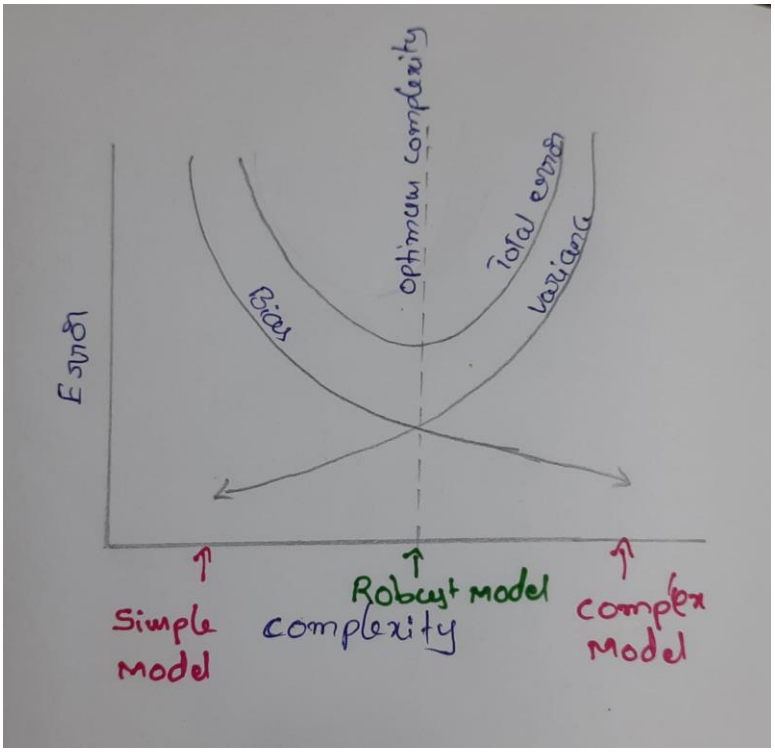
2. Data Transformation: Apply log transformation to skewed features, which can reduce the effect of extreme values. The Box-Cox transformation can stabilize the variance and handle outliers by applying a power transformation.

3. Model Regularization: Using of regularization techniques like L1 (Lasso) and L2 (Ridge) regularization, which penalize large coefficients, making the model less sensitive to extreme values.

4 . Cross-Validation: Cross-validation techniques like stratified k-fold or leave-one-out cross-validation to ensure that the model generalizes well to different data subsets, including those containing outliers.

5. Domain Knowledge: Leverage domain knowledge to understand the data and the potential reasons for outliers. In some cases, outliers might be legitimate data points that need special consideration.

A robust model should have low variance and bias and same can be explained by bias and variance trade off. Robust model total error is less hence the accuracy is more compared to simple and complex models.


optimum complexity
Total error
variance
Bias
Error
Simple Model
Robust model
complexity
complex Model