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?

Answer:

After performing GridSearchCV operation on the Ridge and Lasso,

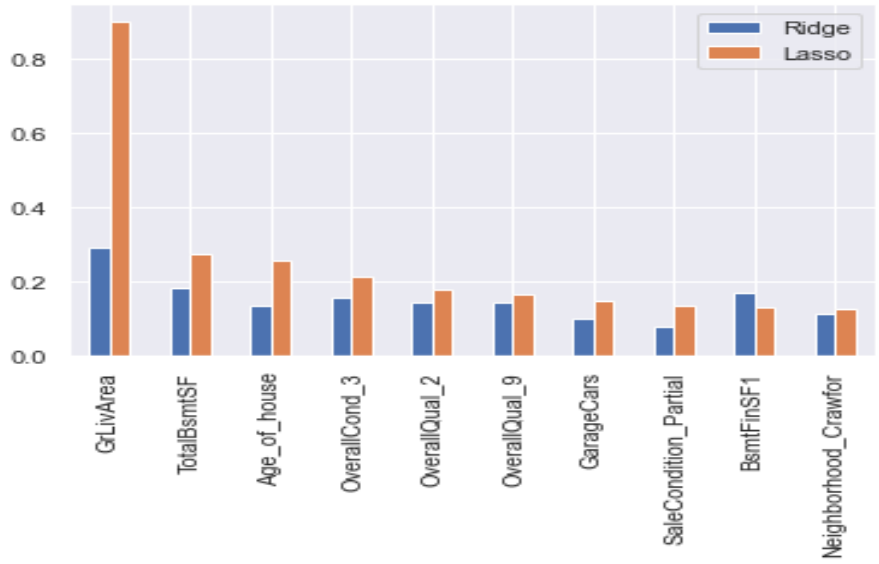
- The optimum value of alpha for Ridge regression is 2.0
- The optimum value of alpha for Lasso regression is 0.0003

If we choose double the alpha i.e,. 4.0 for Ridge and 0.0006 for Lasso

- The test data score has been increased to 0.90 for Ridge
- Train data score decreased and test data has been increased.

After doubling the alpha value the top 10 significant variables are

- 'GrLivArea'
- 'TotalBsmtSF'
- 'Age_of_house'
- 'OverallCond_3'
- 'OverallQual_2'
- 'OverallQual_9'
- 'GarageCars'
- 'SaleCondition_Partial'
- 'BsmtFinSF1'
- 'Neighborhood_Crawfor'



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 checking the metrics, we can see that Lasso regression is performing well for the given data. Also, it is removing the insignificant variables from the data by making the coefficient's zero and the model will be more robust

Here are the metrics for Lasso

R2_score train data: 0.95
R2_score test data: 0.90
RSS train data: 6.84
RSS test data: 5.45
MSE train data: 0.007
MSE test data: 0.014

Here are the metrics for Ridge

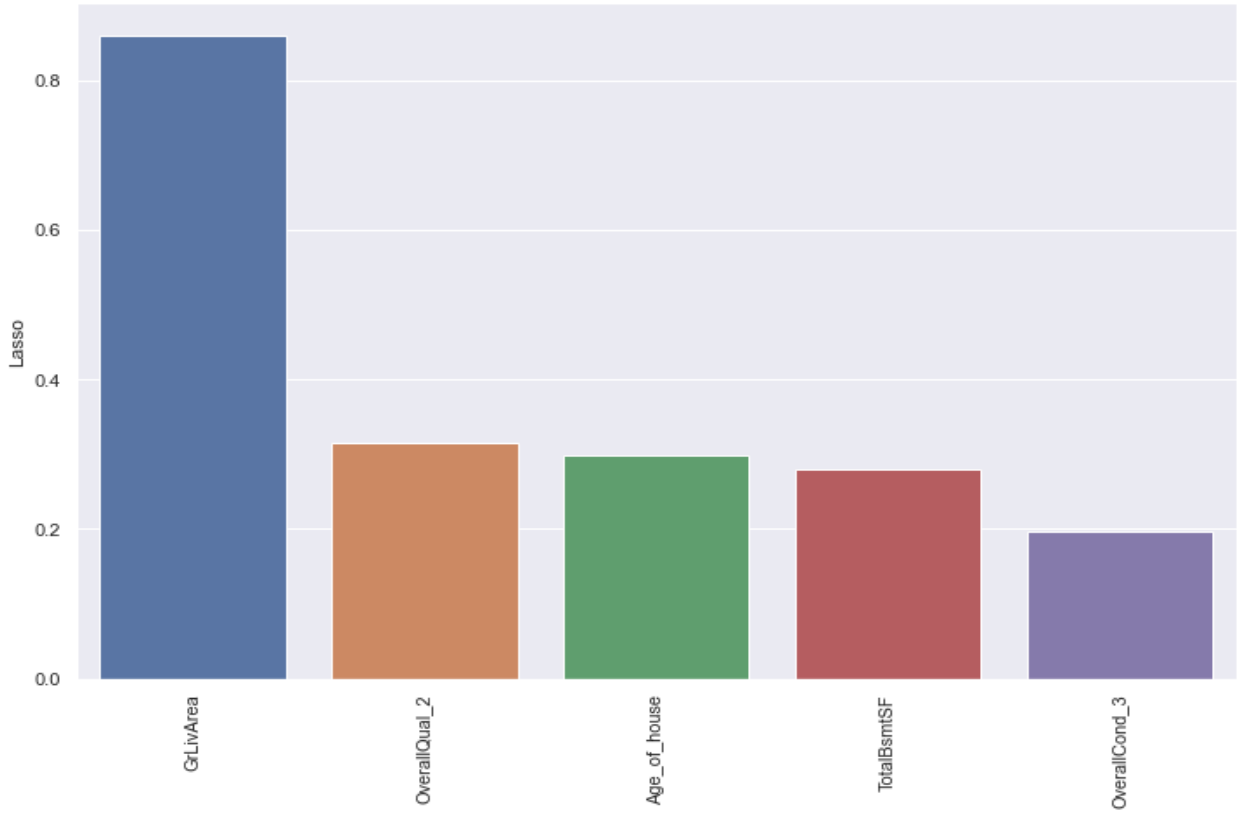
R2_score train data: 0.95
R2_score test data: 0.89
RSS train data: 6.43
RSS test data: 5.52
MSE train data: 0.007
MSE test data: 0.014

Question 3

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?

Answer:

The top 5 most important variables in lasso with optimum value are



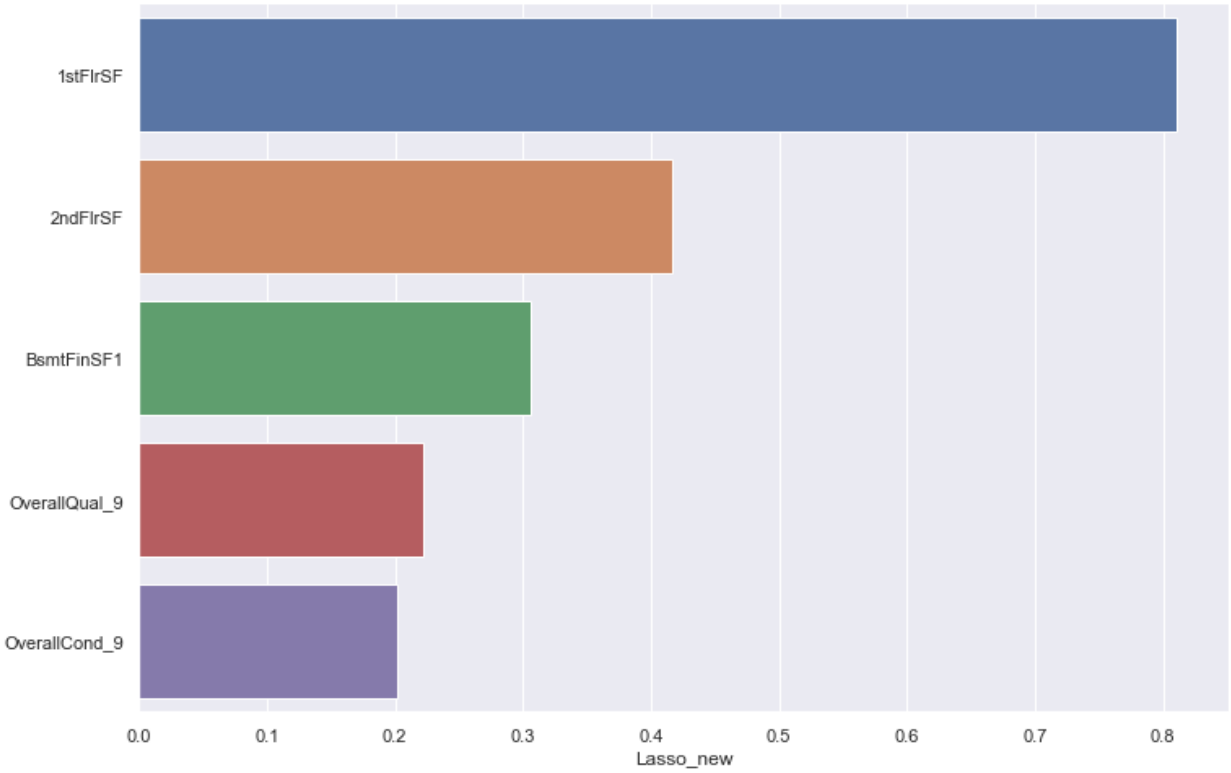
Metrics with optimum value

R2_score train data: 0.95
R2_score test data: 0.90
RSS train data: 6.84
RSS test data: 5.45
MSE train data: 0.007
MSE test data: 0.014

After removing top 5 variables, we build the model again and noticed that the metrics are

R2_score train data: 0.94
R2_score test data: 0.90
RSS train data: 7.09
RSS test data: 5.19
MSE train data: 0.008
MSE test data: 0.013

These are the top 5 variables after removing top 5 significant variables from lasso Regression model



Question 4

How can you make sure that a model is robust and generalizable? What are the imp lications of the same for the accuracy of the model and why?

Answer:

A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalizable so that the te st accuracy is not lesser than the training score. The model should be accurate for d atasets other than the ones which were used during training. Too much weightage should not be given to the outliers so that the accuracy predicted by the model is hi gh. To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. This would help in crease the accuracy of the predictions made by the model. Confidence intervals c an be used (typically 3-5 standard deviations). This would help standardize the predi ctions made by the model. If the model is not robust, it cannot be trusted for predicti ve analysis.