Description Questions

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 value of alpha for Ridge and Lasso are 4 and 100 Respectively.

When we double the Alpha values for both Ridge and Lasso Model then,

- We can Observe that for both Models the R2 Score on Test and train decreased which is a sign of underfitting .
- We can also Observe that for both models the RSS and MSE value is Increased which is also a sign of underfitting .
- We can also Observe that the coefficients of the predictor's changed a lot.

- Especially In lasso Model, We can Clearly see that when we Increased the Alpha value, The no of variables which coefficients are equal to zero Increased from 142 to 163.
- So model excluded 21 variables when alpha becomes doubled (which is useful for feature selection).

Note : *Full code for this is attached at bottom.*

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

- I personally choose Lasso Regression Because the R2 Score are slightly high and also RSS and MSE are also less compared to Ridge.
- When it comes to Lambda (alpha) values I choose the First alpha values 4,100 for Ridge and lasso Respectively , Because when I choose the Doubled Alpha values the R2 Score on Test and train decreased and also RSS and MSE value is Increased.

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

1) Lets start with Ridge Model

- The top 5 important Features in ridge model are 'GrLivArea', '1stFlrSF', 'OverallQual', 'BsmtFinSF1', 'TotalBsmtSF'.
- Now Lets drop these features and we will build the model again .
- The new Top 5 features are 'TotRmsAbvGrd', 'Neighborhood_StoneBr', '2ndFlrSF', 'FullBath', 'GarageArea'.
- *We can Clearly Observe that the R2 score of the test data set decreased compared to old model.*

2) For Lasso Model.

- The top 5 important Features in Lasso model are 'GrLivArea', 'OverallQual', 'BsmtFinSF1', 'TotalBsmtSF', 'Neighborhood_StoneBr'.
- Now Lets drop these features and we will build the model again .
- The new Top 5 features are '1stFlrSF', '2ndFlrSF', 'YearBuilt', 'OverallCond', 'KitchenQual_TA'.
- We can Clearly Observe that the R2 score of the test data set is Slightly decreased compared to old model.

Note : *Full code for this is attached at bottom.*

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

1. The model should be generalized so that the accuracy of the test Should not be much less than the training points so that the model will be giving similar accuracy For Other data sets too.
2. We should also take care that our model should not be affected by Outliers, And should not given more importance to the column containing outliers which leads to overfit , So we should delete Outliers before training the model.
3. The Model Should have high R^2 Score which makes sure that our model covered the most of the variance from the data. so that even any variation in the data should not effect the model much.
4. To make sure the model is robust and generalizable, we must be careful not to overfit it. This is because the overfit model has very high variability and a small change in the data greatly affects the model prediction. Such a model would identify all training data patterns, but fail to select patterns in unseen test data.
5. In other words, the model should not be too complex in order to be robust and generalizable and also Should not be too simple to underfit.
6. If we look at it from the perspectives of Accuracy, the most complex model will have the highest accuracy. Therefore, in order to make our model more robust and generalizable, we will need to reduce the variance that will lead to certain biases. Increased bias means that accuracy will decrease by a little bit. So we need to sacrifice some variance and accuracy for the bias.
7. In general, we should find a balance between model accuracy and complexity. This can be achieved with strategies such as Ridge Regression and Lasso.