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:

For Ridge regression the optimal value of alpha is: 2.0 while for lasso regression it is 0.0001.

When the alpha value is double then

For Ridge:

Doubling the alpha value to 4.0 makes the mean train score to 0.906343 while mean test score is reduced to 0.893573

- 5 Most important predictor variables are:
    1. GrLivArea
    2. 2ndFlrSF
    3. TotalBsmtSF
    4. OverallQual_Excellent
    5. 1stFlrSF

For Lasso:

Doubling the alpha value to 0.0002 makes the mean train score to 0.905119 while mean test score becomes 0.889822.

- 5 Most important predictor variables are:
    1. GrLivArea
    2. OverallQual_Excellent
    3. TotalBsmtSF
    4. OverallCond_Fair
    5. GarageCars

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 optimal value of lambda for Ridge regression is 2.0 while for Lasso regression 0.0001.

Test score of Ridge regression is 0.895013 while for lasso it is 0.891829. Considering the better test score I will choose Ridge regression.

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:

If we remove important predictor variables in the lasso model and create another model then the new 5 important predictors variables would be now:

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual_Poor
4. BsmtFinSF1
5. OverallQual_Very Poor

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: In order to make a model robust and generalisable below operations can be performed

- Null values are replaced with median for columns having outliers and mean for continuous variables.
- The skewed columns are removed and processed to make normal
- The columns which have less correlation to target variable are removed
- When we plot the mean scores for both the training and test data against alpha, we should see a similar pattern in the curves.
- Ridge Regression:
    - Train Score : 0.909673
    - Test Score : 0.895013
- Lasso Regression:
    - Train Score: 0.909814
    - Test Score: 0.891829