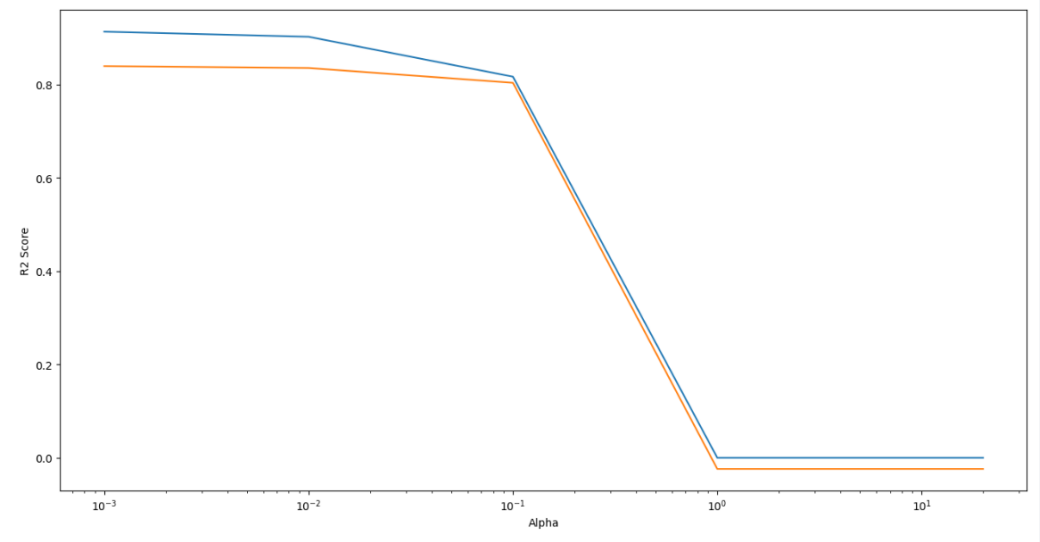
Part 2 - Subjective Questions

Question 1

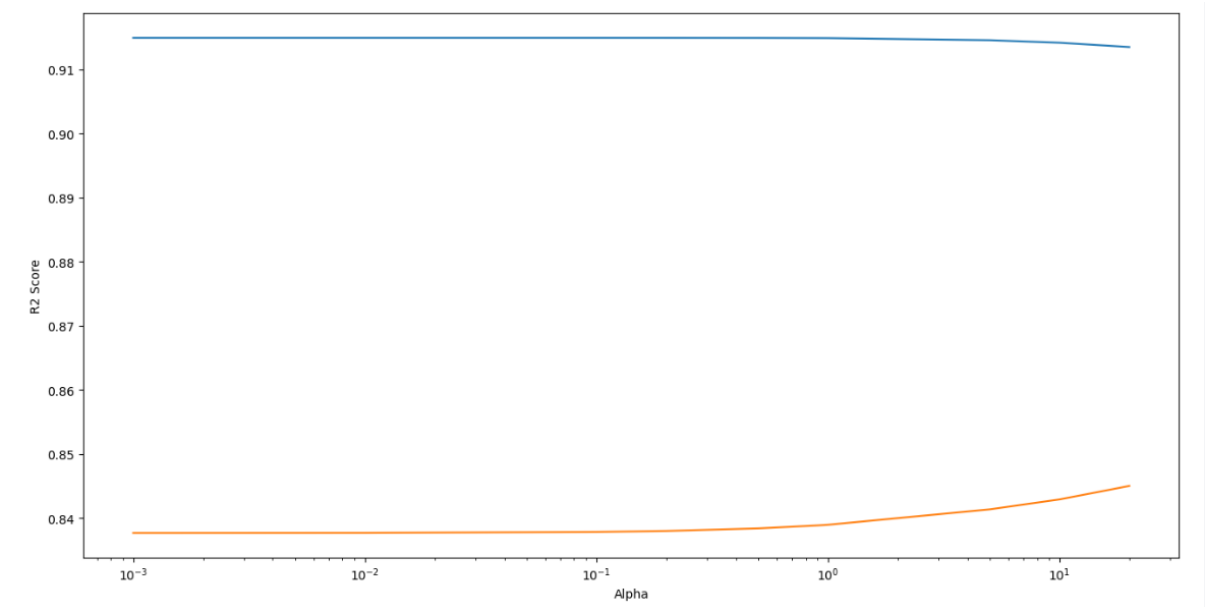
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

Optimal Value of alpha for lasso regression = 0.001



Optimal Value of alpha for ridge regression = 20.0



In Lasso Regression, Train R2 Score = 0.9112 , Test R2 Score = 0.8949 , And Upon Doubling the alpha value, Train R2 score = 0.9101 , Test R2 score = 0.8985

In Ridge Regression, Train R2 Score = 0.9113 , Test R2 Score =0.8943 , And Upon Doubling the alpha value, Train R2 score =0.9103, Test R2 Score = 0.8963

If we choose double the value of alpha for both ridge and lasso regression, model complexity will have a greater contribution to the cost. Because the minimum cost hypothesis is selected,this means that higher λ will bias the selection toward models with lower complexity. For Ridge regression – The coefficient values will increase as alpha increases. For Lasso Regression – More features are removed when alpha value is increased.

Most important predictor variables after the change is implemented are :- #BsmtUnfSF, BsmtFinType2, LandSlope, BsmtFinType1, BsmtQual, OverallQual, OverallCond, BsmtFinSF1

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 : We will choose Lasso regression here because the number of predictor variables are less here and also it shrinks the insignificant predictor variables to zero and remove them from the model without affecting the model accuracy. It makes the model simple and accurate. The values of r2_score on test data are almost similar in both Ridge and Lasso regression but because Lasso has some advantages so We would select Lasso Regression here.

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 new top 5 predictors would be :- TotalBsmtSF, , 1stFlrSF, HouseStyle_1.5Unf, 2ndFlrSF, OverallCond, PoolArea, LandSlope, BsmtFinSF1 and OverallQual.

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:

To make sure that a model is robust and generalisable, we should make sure that the model should not be impacted by outliers in the training data, test accuracy should be greater than the training score, The model should be accurate for new test datasets instead of the training data sets that we would be using and the correlation should be handled well and the VIF of all the features should less than 5.

The implications of model being robust and generalisable on the accuracy would be that the outliers treatment should be done in the training data very well. This will increase the accuracy. The multicollinearity should be handled well with

removing the highly related variables from training and performing VIF. Also we need to choose the right alpha value for model building using Ridge or Lasso Regression.