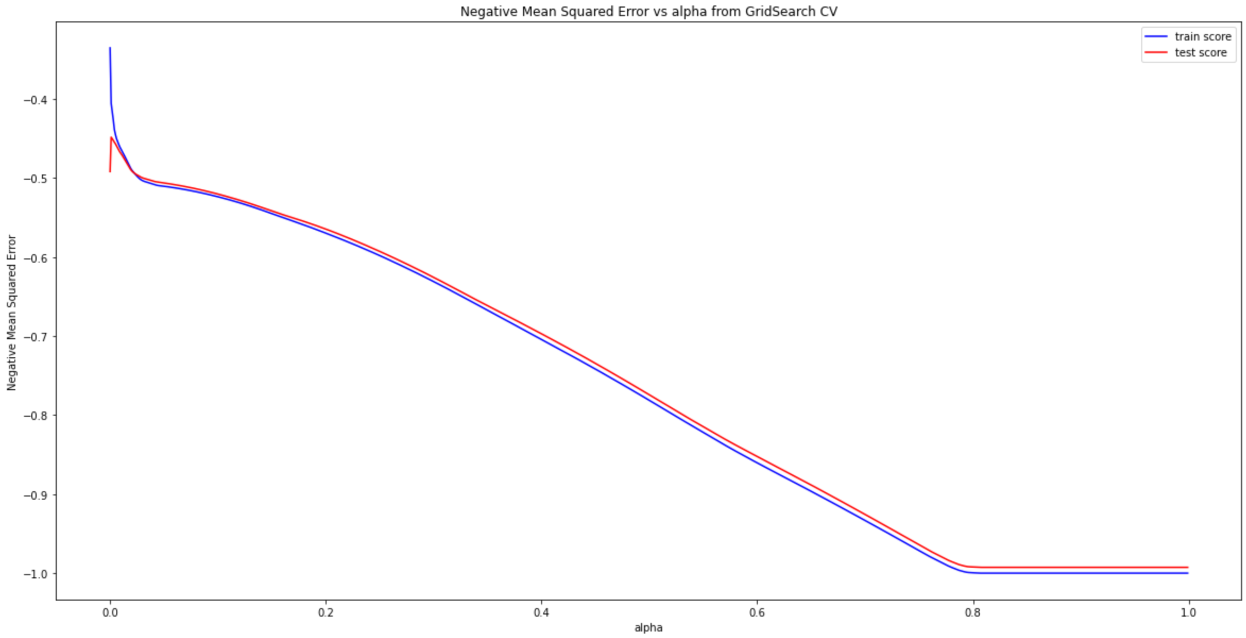
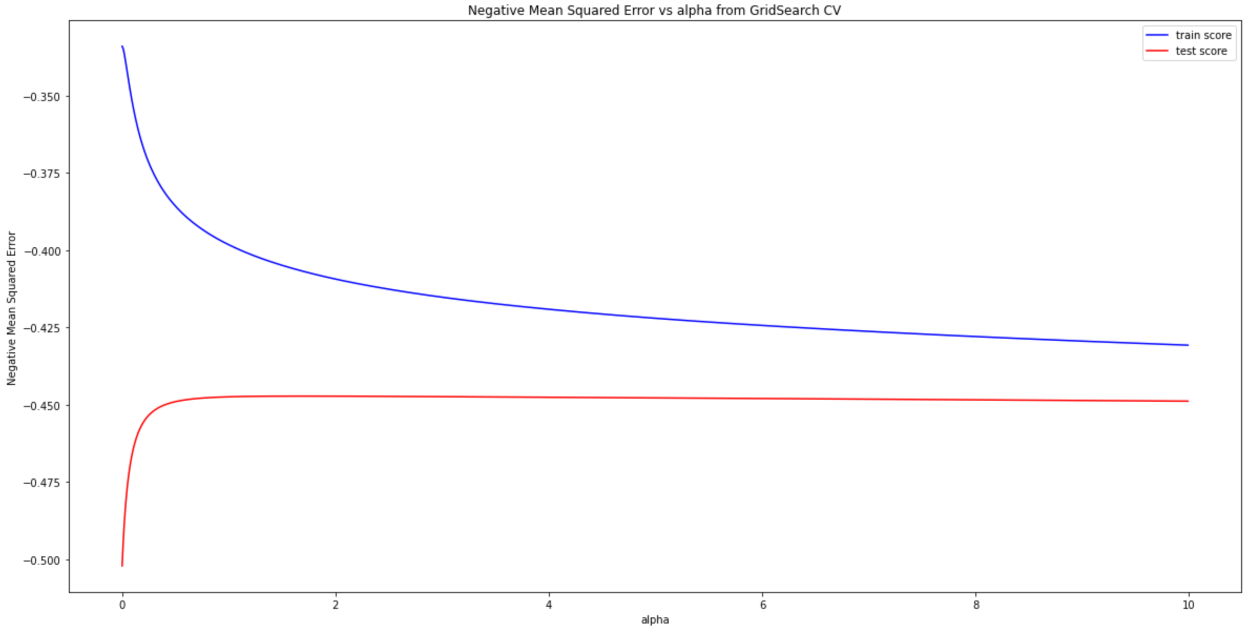
Question 1

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?

- The optimal value of alpha for ridge is **1.671**. For Lasso, optimal value of alpha is **0.0011**
- On doubling the alpha, model is getting more generalized
  - For training data, r squared decreased a bit and root mean square increased a bit for both Ridge and Lasso.
  - However for test data, it performs well with with increased r squared amd decreased root mean square value.
  - It goes by the principle that with higher values of alpha, model will have increased bias and decreased variance.





- Top 10 important variables for Ridge after doubling the optimal alpha: *Condition2_PosN, RoofMatl_WdShngl, Neighborhood_NoRidge, Neighborhood_NridgHt, GrLivArea, OverallQual, Heating_OthW, Neighborhood_Somerst, BldgType_Twnhs, MSSubClass_75*
- Top 10 important variables for Lasso after doubling the optimal alpha: *Condition2_PosN, RoofMatl_WdShngl, Neighborhood_NoRidge, Neighborhood_NridgHt, GrLivArea, OverallQual, Neighborhood_Somerst, BldgType_Twnhs, SaleType_New, LotConfig_CulDSac*

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?

Lasso uses significantly less number of independent variables for the model (only 29 out of total 50 from RFE) yet Lasso is able to explain the data relatively better. For training data, R squared and root mean squared error for both regression is comparable, merely 0.0004 lesser R2 score and 0.001 greater rmse.
Even though Ridge outperforms Lasso for the test data, I would still go for Lasso to have lesser complexity

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?

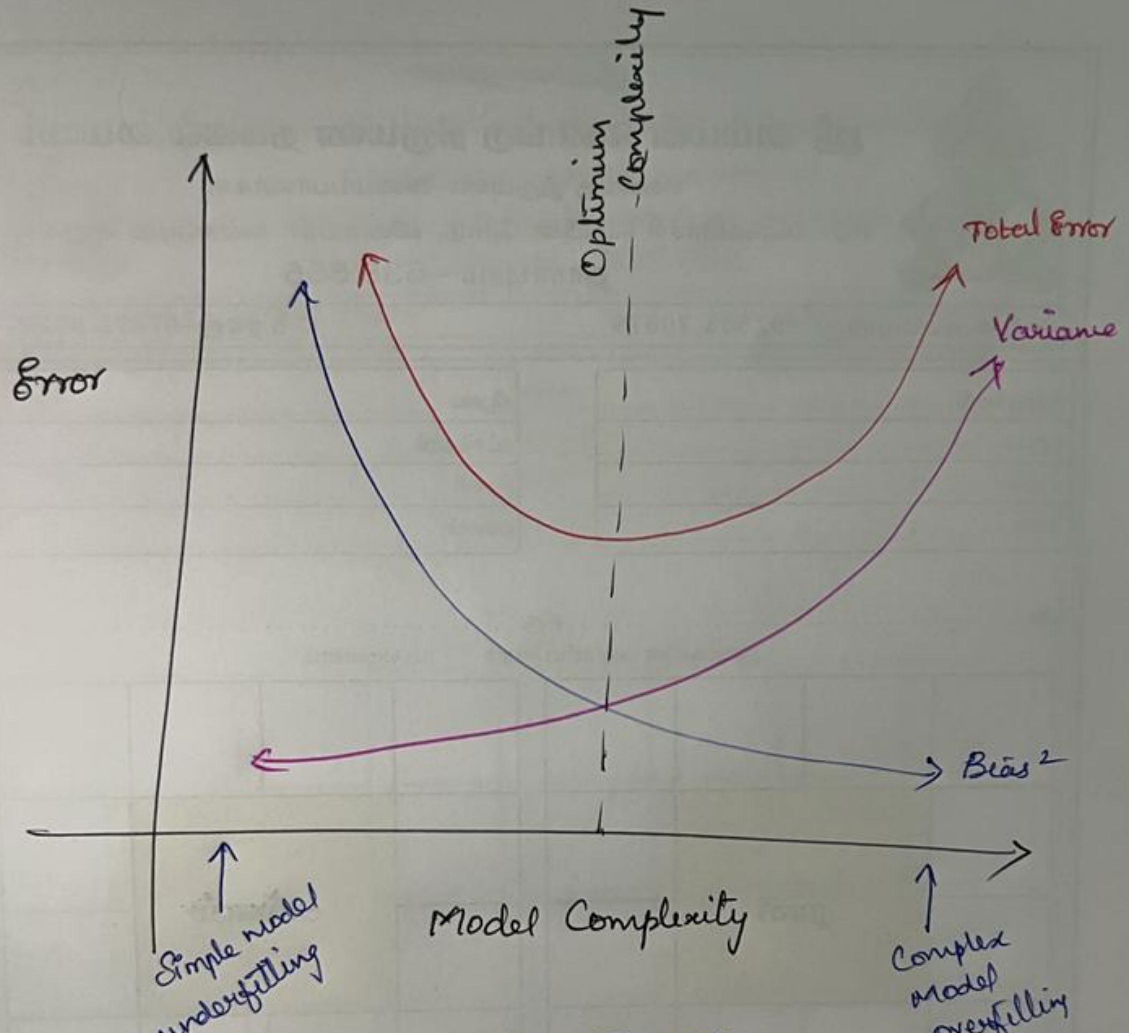
Following are the five most important variables now for the Lasso model:
OverallQual, GrLivArea, SaleType_New, Condition1_Norm, LotConfig_CulDSac
Also to note, the new optimal value of alpha is 0.0021 and the training R squared value dropped to 0.78

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?

Typical Problem:
Most often, the model developed may overfit the data. One reason may be due to the fact there are more predictor variables used. Hence instead of learning the patterns , model has also learnt the noise present in the data

Solution:
Regularization can be used to make model robust and generalizable. It essentially penalizes the magnitude of model coefficients by shrinking them towards 0. Instead of just minimizing the RSS, RSS + Penalty term (function of coefficients) is minimized.



Bias - Variance Trade off -

Regularization : reduces model complexity (thus variance) with tolerable bias to have optimum model.

Impact:

When applying regularization, we are essentially reducing model complexity significantly (thus variance) by accepting little bias. When we accept that , we have to let go of the predictive power of the model (Rsquared, rmse etc). It is basically a bias variance trade off.