Problem statement-II

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?

Best Alpha value for Ridge is 4.0,
Best Alpha value for Lasso is 0.0001

If we double the alpha values Rsquare value
when we increase the value of alpha for lasso we try to penalize more our model and more coefficient of the variable will reduced to zero, when we increase the value of our r2 square also decreases.

For Ridge the RSquared value has gone down for both test and train value.
For lasso RSquared for train gone down but for test it remained same.
We can see all the metrics comparisons here

Metrics	Ridge(Train)	Ridge (Test)	Lasso(Train)	Lasso (Test)	Ridge twice alpha(Train)	Ridge twice alpha (Test)	Lasso twice alpha(Train)
RMS	0.034125	0.044332	0.033241	0.045434	0.035398	0.044057	0.035042
RSquared	0.906239	0.831413	0.911034	0.822928	0.899114	0.833494	0.901130
RSS	1.186647	0.858836	1.125961	0.902059	1.276832	0.848235	1.251308
MSE	0.001165	0.001965	0.001105	0.002064	0.001253	0.001941	0.001228
RMSE	0.034125	0.044332	0.033241	0.045434	0.035398	0.044057	0.035042

For Best Alpha

Top 5 positive features for Ridge: ['GrLivArea', 'FullBath_3', 'Neighborhood_NoRidge', 'OverallQual_10', 'OverallQual_9']

Top 5 negative features for Ridge: ['ExterQual_Fa', 'BsmtQual_TA', 'KitchenQual_Gd', 'BsmtQual_Fa', 'KitchenQual_TA']

Top 5 positive features for Lasso: ['GrLivArea', 'LotArea', 'FullBath_3', 'OverallQual_9', 'OverallQual_10']

Top 5 negative features for Lasso: ['Fireplaces_3', 'BsmtQual_No Basement', 'ExterQual_Fa', 'BsmtQual_TA', 'BsmtQual_Fa']

When Alpha is doubled

Top 5 positive features for Ridge: ['GrLivArea', 'FullBath_3', 'Neighborhood_NoRidge', 'OverallQual_10', 'OverallQual_9']

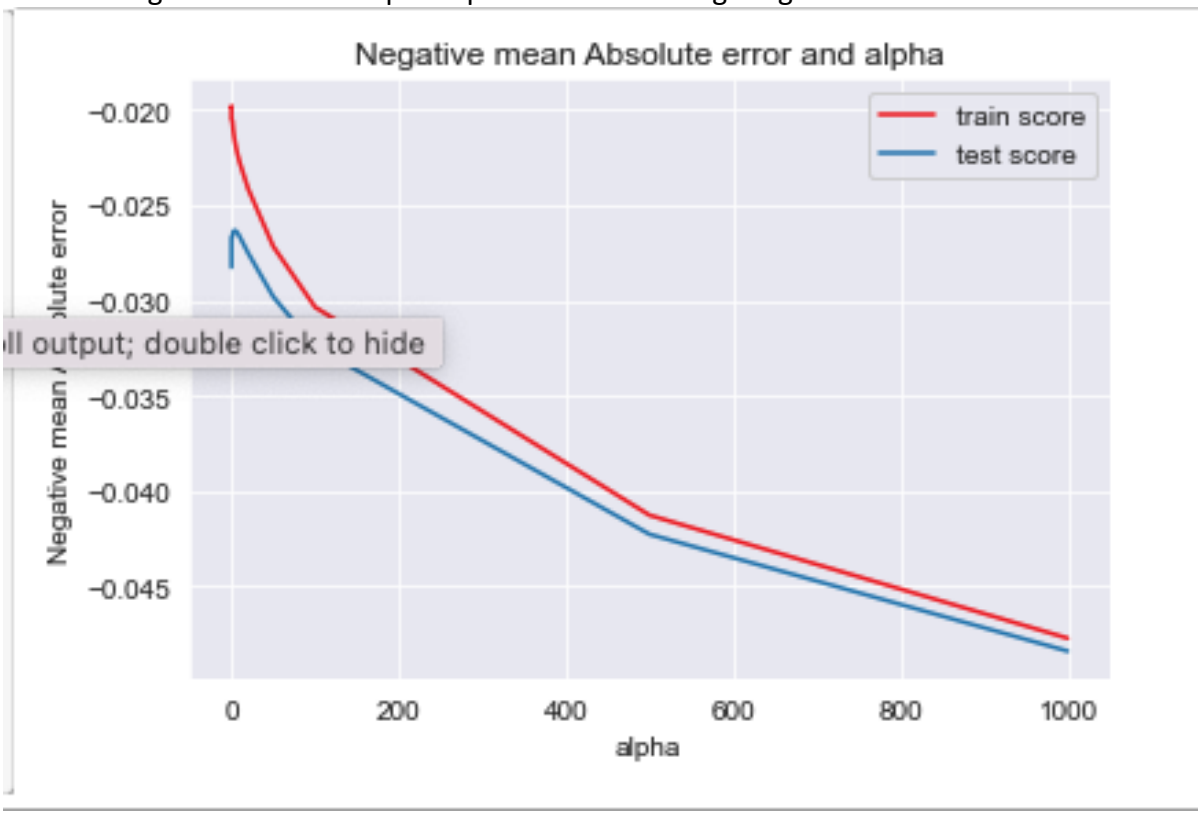
Top 5 negative features for Ridge: ['KitchenQual_Gd', 'ExterQual_Fa', 'BsmtQual_TA', 'FullBath_1', 'KitchenQual_TA']

Top 5 positive features for Lasso: ['GrLivArea', 'FullBath_3', 'OverallQual_9', 'OverallQual_10', 'Neighborhood_NoRidge']

Top 5 negative features for Lasso: ['BsmtQual_No Basement', 'Fireplaces_3', 'KitchenQual_TA', 'KitchenQual_Gd', 'KitchenQual_Fa']

In the case of ridge regression:- When we plot the curve between negative mean absolute error and alpha we see that as the value of alpha increase from 0 the error term increases initially and then starts to decrease for test data and the train error is showing increasing trend when value of alpha increases .when the value of alpha is 9 the test error is minimum so we

decided to go with value of alpha equal to 4 for our ridge regression.



For lasso regression when we increase the value of alpha the model try to penalize more and try to make most of the coefficient value zero. Initially it came as 0.0001 in negative mean absolute error and alpha.

Similarly

Q2. 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?

- Ridge regression, uses square of magnitude of coefficients which is identified by cross validation. Residual sum or squares should be small
  - The penalty is lambda times sum of squares of the coefficients, hence the coefficients that have greater values gets penalized.
  - As we increase the value of lambda the variance in model is dropped and bias remains constant.
  - Ridge regression includes all variables in final model unlike Lasso Regression.
- Lasso regression, uses absolute value of magnitude of coefficients which is identified by cross validation.

- As the lambda value increases Lasso shrinks the coefficient towards zero and it make the variables exactly equal to 0.
- Lasso also does variable selection.
- When lambda value is small it performs simple linear regression and as lambda value increases, shrinkage takes place and variables with 0 value are neglected by the model.

Q3. 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?

- Top 5 positive features for Ridge: ['TotalBsmtSF', 'BsmtFinSF1', 'MasVnrArea', 'TotRmsAbvGrd_10', 'LotArea']
- Top 5 negative features for Ridge: ['FullBath_1', 'FullBath_2', 'ExterQual_Fa', 'OverallQual_5', 'KitchenQual_Gd']

Q4. 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?

- To make the model robust and generalizable trade of is between bias and variance.
- We want the bias to be low so model is not remembering the training data
- Variance needs to be high so model can predict for data which is close to training data.
- To achieve this we penalized the the model for using to many features or higher degree polynomials,