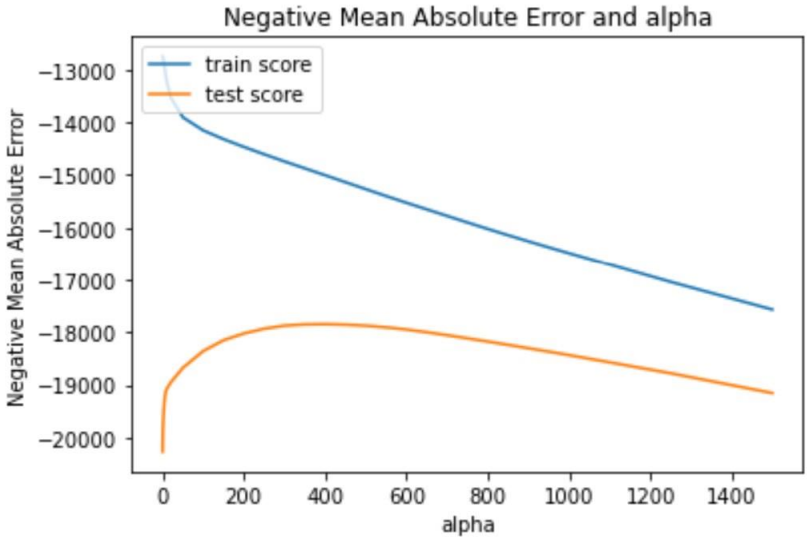
Advanced Regression Part II

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:

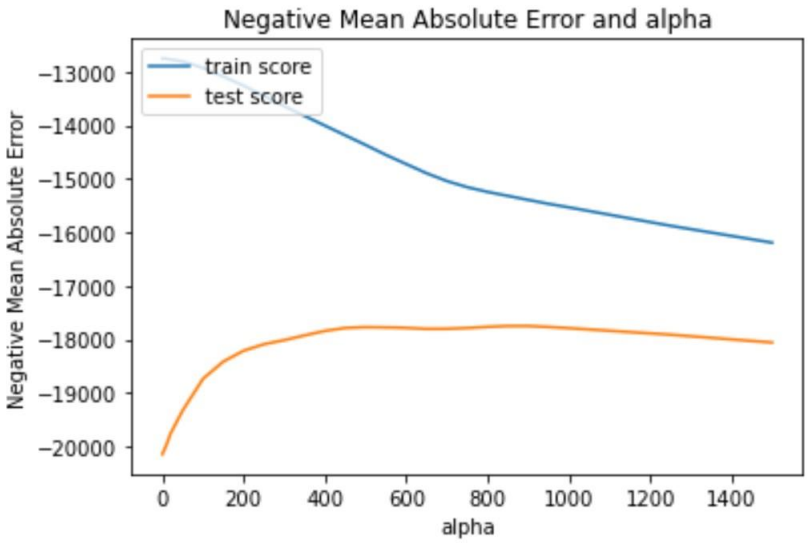
Based on the graphs observed with gridsearch cross fold validation with ridge and lasso as estimators based on neg_mean_absolute_error scoring, a common pattern in both the regularization techniques is that the test score keeps increasing until a point and then it reduces where as train score has a strictly decreasing curve.
Below are the observations based on the model:

a. Ridge estimator alpha vs negative mean absolute error



From the above graph, it is easy to interpret that test score showed increasing trend until 400 and then started dropping. Optimal value gridsearch showed is also 400

b. Lasso estimator alpha vs negative mean absolute error



Similarly for lasso, test scores showed non decreasing trend until 800-900 range and then it started dropping.
Best alpha returned by gridsearch was 850

When alphas are doubled, both test & train scores dropped in ridge & lasso with an increase in MSE and also an increase in RSS for both test & train data

	Metric	Ridge_400	Ridge_800	Lasso_850	Lasso_1700
0	R2 train	0.895034517	0.880388690	0.892668589	0.876582289
1	R2 test	0.847891829	0.841740371	0.854510997	0.846407736
2	RSS train	593872544273.961303711	676735543680.676391602	607258467576.655517578	698271355849.285400391
3	RSS test	294172313979.214721680	306069036864.929443359	281371057266.170043945	297042502135.672973633
4	MSE train	545337506.220350146	621428414.766461372	557629446.810519338	641204183.516331911
5	MSE test	810392049.529517174	843165390.812477827	775126879.521129608	818298903.955021977

Before the change, most important features(top 10) are as follows:

a. Ridge:

OverallQual, GrLivArea, Neighborhood_NridgHt, TotRmsAbvGrd, 1stFlrSF, KitchenQual, GarageCars, Neighborhood_StoneBr, Neighborhood_NoRidge, ExterQual

b. Lasso:

GrLivArea, OverallQual, Neighborhood_NridgHt, GarageCars, SaleType_New, BldgTypeFloorsMap, KitchenQual, Neighborhood_StoneBr, Neighborhood_NoRidge, BsmtFinType1

After this change, most important features(top 10) are as follows:

a. Ridge:

OverallQual, GrLivArea, Neighborhood_NridgHt, KitchenQual, TotRmsAbvGrd,

1stFlrSF, GarageCars, ExterQual, GarageArea, Neighborhood_NoRidge

b. Lasso:

GrLivArea, OverallQual, GarageCars, Neighborhood_NridgHt, KitchenQual, ExterQual, SaleType_New, BldgTypeFloorsMap, FireplaceQu, Neighborhood_StoneBr

Out of top 10 features, 7-8 remained same even after doubling alpha for both ridge & lasso.

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:

	Metric	Ridge	Lasso
0	R2 train	0.895034517	0.892668589
1	R2 test	0.847891829	0.854510997
2	RSS train	593872544273.961303711	607258467576.655517578
3	RSS test	294172313979.214721680	281371057266.170043945
4	MSE train	545337506.220350146	557629446.810519338
5	MSE test	810392049.529517174	775126879.521129608

Based on metrics comparison for ridge and lasso, lasso has a good test score and MSE value also lasso does feature selection. Therefore for the current data & problem Lasso model is preferred.

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:

After dropping the top 5 features ('GrLivArea', 'OverallQual', 'Neighborhood_NridgHt', 'GarageCars', 'SaleType_New')

And building a new model with lasso(model code is available in py notebook), the new 5 important features for lasso are: Neighborhood_StoneBr, SaleCondition_Partial, RoofMatl_WdShngl, Neighborhood_Crawfor, Neighborhood_NoRidge

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:

Bias v/s variance trade-off plays an important role in explaining a model. Bias stands for how generalizable a model is and variance is defined as the change in the model when the data changes. A good model should always be generalizable for new data at the same time it should not vary much with any new unseen data. Which means it should have low bias, low variance.

Overfitting: when a model is complex enough to understand whole training data, it doesn't perform very well on test data this problem is called overfitting.

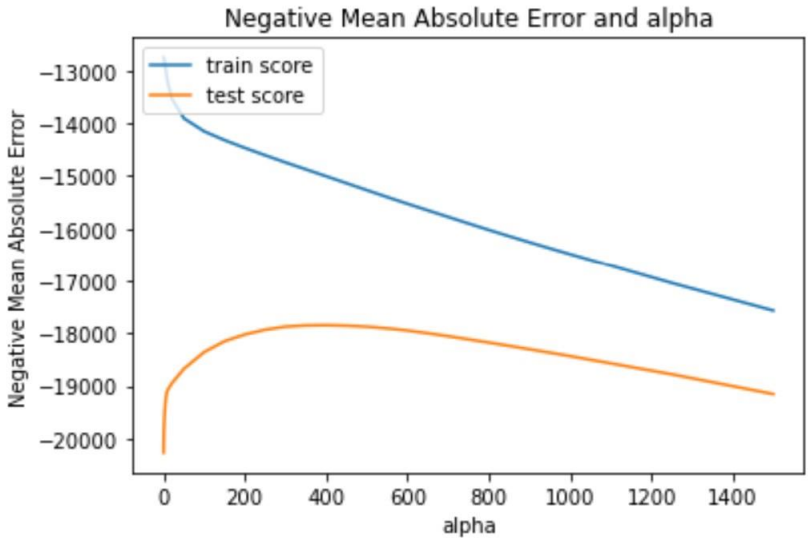
More accuracy implies model should be more complex so some bias needs to be there so that it is more generalizable.

To avoid overfitting, regularization techniques are used.

Regularization: it is the process of penalizing models for using features. So this would add an extra alpha coefficient

Ridge and lasso are two regularization techniques for a linear model.

When model becomes generalizable, it performs better on test data while it drops it's performance in test data.



With change in alpha(increase in generalization) until a point, test score increases and it falls after a certain point. So any value until this point with good train & test score is considered as a good alpha value, this decision should be taken based on business priorities by considering test-train scores trade-off.