Advanced Regression Assignment

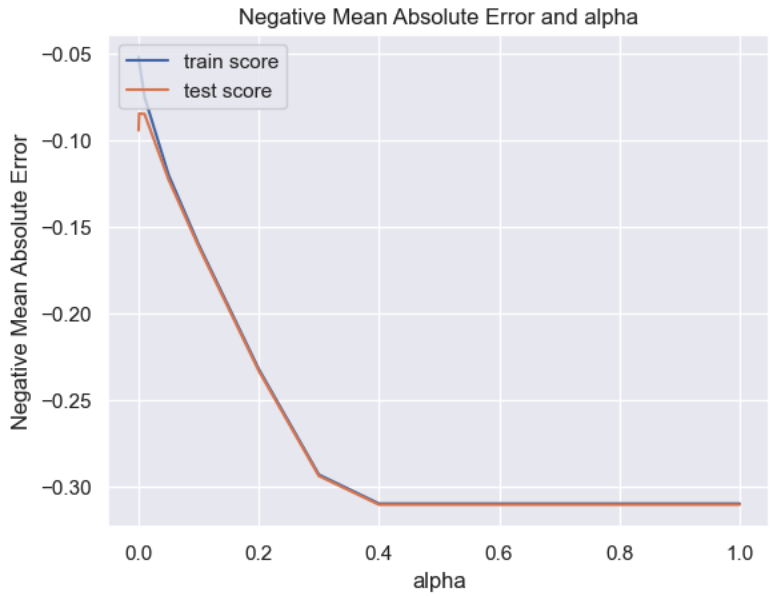
Problem Statement - Part II

1) **Question** : 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:

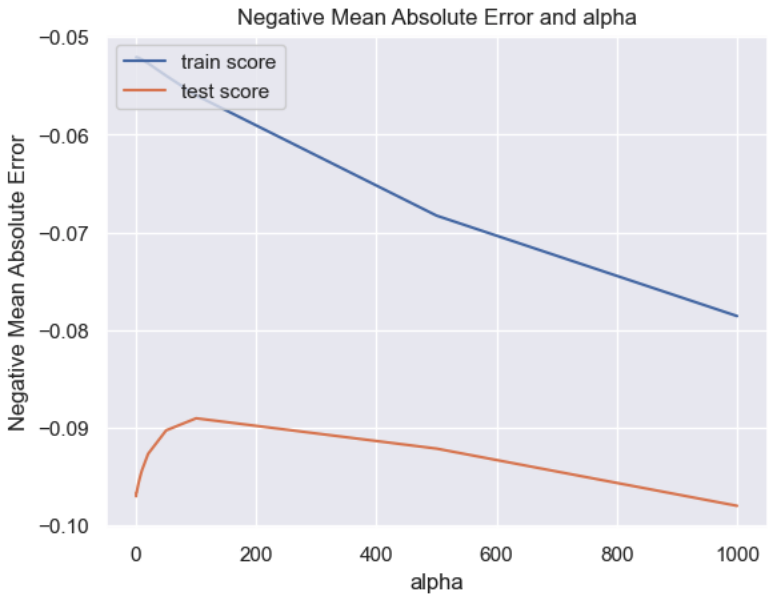
In case of Lasso regression, the test score started reducing when the alpha value is 0.01 and for higher values of alpha, the r2 scores are tending towards 0 and lesser. (refer picture below)

So selected the alpha as 0.01 here



In case of Ridge regression, the test score started reducing when the alpha value is 100 and for higher values of alpha, the error seems showing higher deviations (refer picture below)

So selected the alpha as 100 for ridge regression.



Top 10 most important predictor variables for Lasso Regression after changing alpha values are (in order of +ve impact)

GrLivArea, OverallQual, TotalBsmtSF, YearBuilt, YearRemodAdd, GarageArea, GarageCars, BsmtFinSF1, Fireplaces, OverallCond

Top 10 most important predictor variables for Ridge Regression(+ve) after changing alpha values are (in order of +ve impact)

OverallQual, GrLivArea, 1stFlrSF, TotalBsmtSF, OverallCond, GarageArea, BsmtFinSF1, 2ndFlrSF, YearRemodAdd, GarageCars

Also noticed slight change in the RMSE and r2 values. In both the cases, the coefficient of predictor variables changes.

2) **Question** : 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:

Ridge regression performs L2 regularization, i.e adds penalty equivalent to the square of the magnitude of coefficients. Whereas Lasso regression performs L1 regularization, i.e adds penalty equivalent to the absolute value of the magnitude of coefficients.

In our exercise both lasso and ridge regression gives good r2 scores and root mean square values. But there is a huge difference in the number of predictor variables suggested by ridge and lasso regressions. Ridge suggested 325 variables, whereas lasso suggested 53 variables. As per the basic rule of thumb to keep the model robust and simple, its decided to apply Lasso regression.

3) **Question** : 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 five most important predictor variables were (before removing)

'GrLivArea', 'OverallQual', 'YearBuilt', 'TotalBsmtSF', 'OverallCond'.

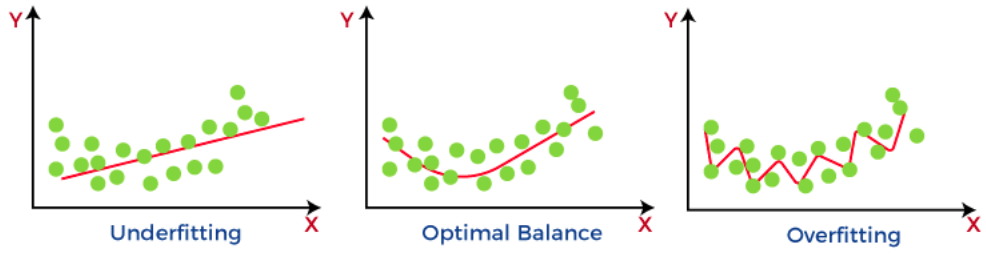
The most important variables after removing & recreating the model are

1stFlrSF, 2ndFlrSF, YearRemodAdd, GarageArea, BsmtFinSF1

4) **Question** : 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:

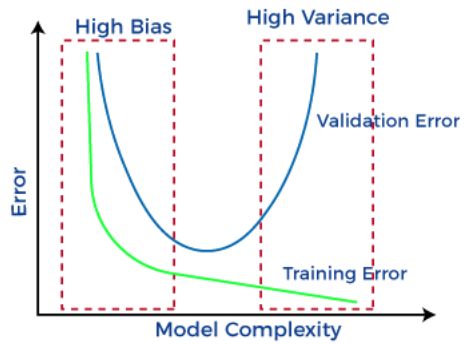
The model should be as robust, simple(but not simpler than that) and generalisable as possible. Simple models will have less number of predictor or independent variables and will be more generic and also require less assumptions and data. Generalization is a term used to describe a model's ability to forecast/react to new data. If a model is trained too well on training data, it will be unable to generalize. In this case overfitting happens and the model will not perform well on test or new data sets. We can figure out this during the testing phase. In this case the model work very well for the train data and for the test data it will give a very low r2 score for test data. The error terms will also be very high in this case.



Underfitting, Best Model, and Overfit

The implications of complexity & generalization can be understood using Bias & Variance. Bias is the error in the model; i.e in the prediction process, a contrast between the predicted values and actual values is error due to Bias. A high bias model cannot perform generalisation and will perform poor or train & test data. Variance tells how much a variable is different from its expected value. A low variance shows a little deviation in the prediction, while a high variance shows high deviation in predicting target variable. Variance has to be low for models that can generalize well.

Below picture explains the same.



error v Model Complexity

Implications of complexity on accuracy are

- When the model complexity increases, the training error decreases, and the test error increases, resulting in less accuracy in predicting unseen data.
- When the model is very complex, the gap between training and generalization/test error is very high, resulting in overfitting
- When the model is less complex, the model will have a high training error, resulting in underfitting(error in both training & test data).