Problem Statement - Part II:

Question 1

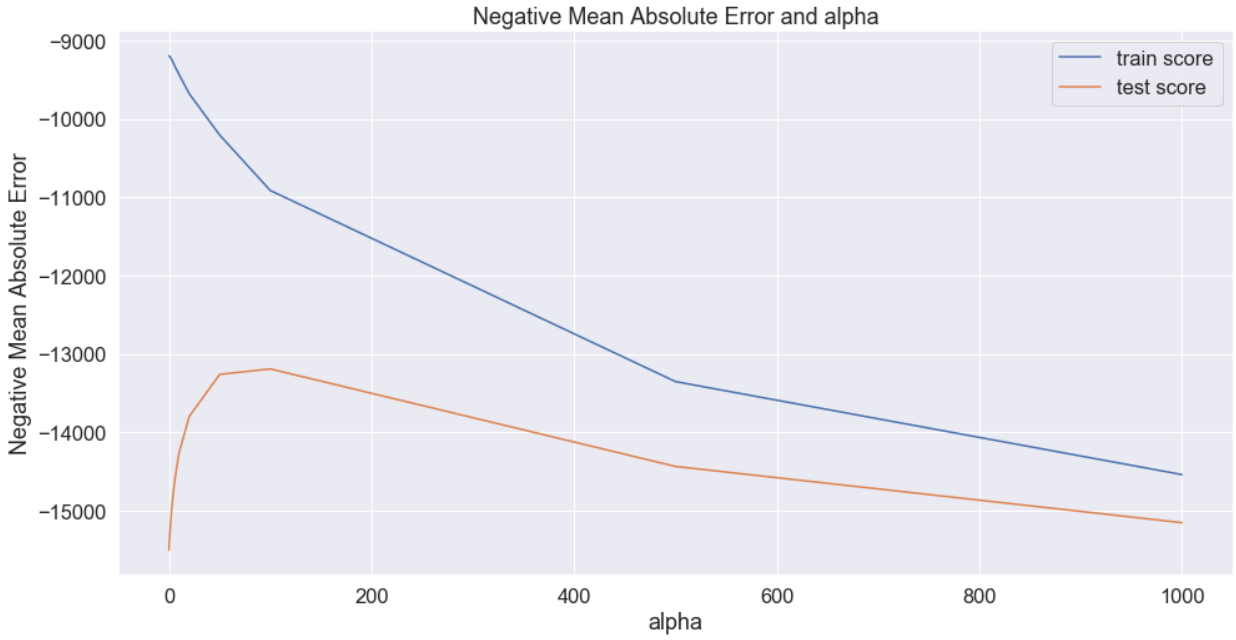
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

- The optimal value of alpha for ridge regression=6 and lasso regression=100
- When the value of alpha is doubled for ridge regression=12:
  - The train R-squared value decreased slightly = 0.93; with alpha =6, this value was 0.94
  - The test R-squared value remained nearly same as earlier =0.92
- When the value of alpha is doubled for lasso regression=100:
  - The train R-squared value decreased slightly = 0.92; with alpha =200, this value was 0.93
  - The test R-squared value decreased slightly=0.91; with alpha =200, this value was 0.92
- Some of the most important predictor variables after change is implemented are:
  - Exterior1st_CemntBd, Exterior2nd_CemntBd, GarageType_Detchd, Exterior1st_CBlock, Exterior2nd_CBlock, ExterCond_Fa, Heating_Grav,etc.

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?

- The optimal value of alpha for ridge regression=6 and lasso regression=100
- Lasso Regression is beneficial in variable selection, i.e., helps in feature shrinkage (as the coefficient value of most variables is nearly 0), Lasso offers a better edge over Ridge. hence, the variables predicted using Lasso will be more effective in predicting the price of the house.
- The train R-squared value with alpha=100 is 0.93 in case of Lasso Regression.
- The test R-squared value with alpha =100 is 0.92 in case of Lasso Regression.



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?

- The five most important predictor variables after excluding the initial variables are : Neighborhood_BrkSide, LotShape_IR3, Neighborhood_BrDale, Neighborhood_IDOTRR, Neighborhood_MeadowV.
- The above variables were derived based on removing the variables : Exterior1st_CemntBd,Exterior2nd_CmentBd,Exterior1st_CBlock,Exterior2nd_CBlock,ExterCond_Fa as described in detail in the python notebook.

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?

A model can be considered as robust if the target variable is consistently accurate, which means despite of a sudden change in independent features, the training set will not be impacted drastically. The model is considered as generalisable if it does not lead to overfitting of the training data.

There is an important relationship between the complexity of a model and its usefulness in a learning context because of the following reasons:

• Simpler models are usually more generic and are more widely applicable (are generalizable)

• Simpler models require fewer training samples for effective training than the more complex ones.

Regularization is a process used to create an optimally complex model, i.e. a model which is as simple as possible while performing well on the training data.

Through regularization, the algorithm designer tries to strike the delicate balance between keeping the model simple, yet not making it too naive to be of any use. The regression does not account for model complexity - it only tries to minimize the error (e.g. MSE), although if it may result in arbitrarily complex coefficients. On the other hand, in regularized regression, the objective function has two parts - the error term and the regularization term.

Thus, in terms of Accuracy, a robust and generalisable model will perform equally well on both training and test data implying that the accuracy will not get much affected for train and test data.

While creating the best model for any problem statement, we end up choosing from a set of models which would give us the least test error. Hence, the test error, and not only the training error, needs to be estimated in order to select the best model. This can be done in the following two ways.

1. Use metrics which take into account both model fit and simplicity. They penalise the model for being too complex (i.e. for overfitting), and thus are more representative of the unseen ‘test error’.
2. Estimate the test error via a validation set or a cross-validation approach. In validation set approach, we find the test error by training the model on training set and fitting on an unseen validation set while in n-fold cross-validation approach, we take the mean of errors generated by training the model on all folds except the kth fold and testing the model on the kth fold where k varies from 1 to n.

Namrata Bhartiya