Subjective Question and Answers

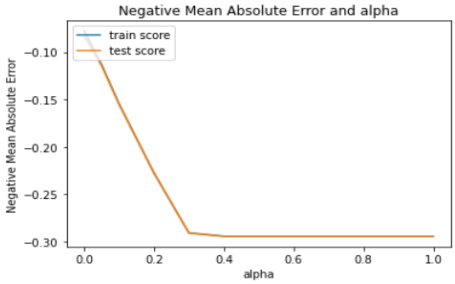
Question1:

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:

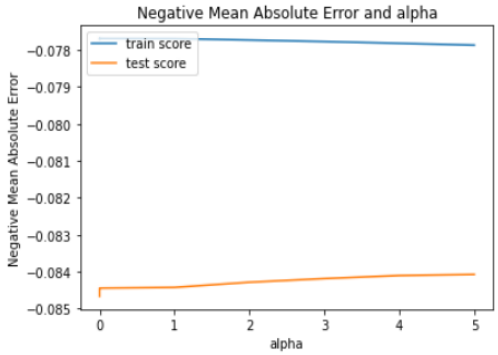
The optimal value of alpha for Ridge regression is 2 and for Lasso regression is 0.4.

Lasso Chart:



In the above chart, increase in alpha has also increased the negative mean absolute error for both train and test data, but after alpha =4 the values became equal and there is no difference.

Ridge Chart:



In the above ridge chart, increase in alpha has also increased the negative mean absolute error for both train and test data, unlike Lasso it increases gradually along with alpha from the point alpha =0 to alpha=2, after that it has stabilized.

Lasso Important predictor variables are listed below:

```
'LotFrontage', 'BsmtUnfSF', 'OverallCond', 'OverallQual', 'BsmtFullBath
', 'LotArea', 'IsRemodelled', 'MSSubClass', 'GrLivArea', 'Neighborhood_
CollgCr', 'BuiltOrRemodelAge', 'd_HeatingQC', 'OpenPorchSF', 'd_LotShap
e', 'd_BsmtExposure', '2ndFlrSF', 'WoodDeckSF', 'd_GarageFinish', 'd_Ex
terQual', 'FullBath', 'Fireplaces', 'GarageCars', 'MasVnrArea', 'd_Fenc
e', 'OldOrNewGarage', 'Neighborhood_Gilbert', 'd_LotConfig', 'd_Kitchen
```

```
Qual', 'HalfBath', 'MSZoning_RL', 'd_BldgType', 'd_BsmtFinType1', 'Bedr
oomAbvGr'
```

Ridge Important predictor variables are listed below:

```
d_BsmtExposure, BsmtUnfSF, d_BsmtFinType1, d_ExterQual, LotFrontage,
LotArea, OverallCond, Neighborhood_CollgCr, Neighborhood_BrDale,
BsmtFullBath, OverallQual, d_BsmtQual, d_HeatingQC,
d_GarageFinish, Neighborhood_ClearCr, MSSubClass,
Neighborhood_BrkSide,GrLivArea, 2ndFlrSF, Neighborhood_Crawfor,
OpenPorchSF, WoodDeckSF, d_Fence, IsRemodelled,
GarageCars,BuiltOrRemodelAge, MSZoning_RM, Fireplaces, d_LotShape,
MSZoning_FV, TotalBsmtSF
```

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?

Answer:

In linear regression it is always advised to use simple and still robust model. So I would choose Lasso regression. It is because during our analysis we found that the Lasso assigned zero value for insignificant features and given the leverage to us to use the predictive variables. Unlike Lasso, in Reidge regression, though increase in the value of lambda has dropped the variance in the model, yet it has included all variables in the final model.

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?

Answer:

LotFrontage
BsmtUnfSF
OverallCond
OverallQual
BsmtFullBath

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?

Answer:

The model should be as simple as possible. Though the simplicity comes with decrease in accuracy, but still, it will be more robust and generalisable.

So, the simpler model the more bias however, less variance and more generalisable. Therefore, the implication will be that the model will perform equally well with both test and training data and accuracy does not change for training and test data.

Bias is actually an error in the model, when it is low, this means that the model is weak to learn data. On the other hand, when it is high it means it is unable to understand the details in the data. This will be clearly visible when the model performs poor in training and test data.

Variance is also an error in the model, when it is high, it means the model performs exceptionally well in the training data however, performs poor in the unseen test data. Therefore, it is important to have a balance between Bias and Variance in order to avoid underfitting and overfitting of model.