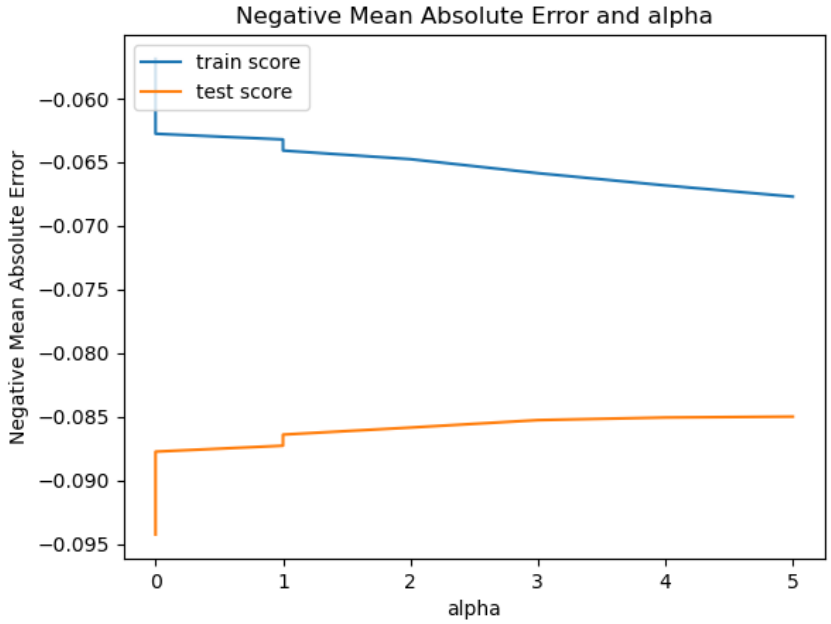
Problem Statement - Part II

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?

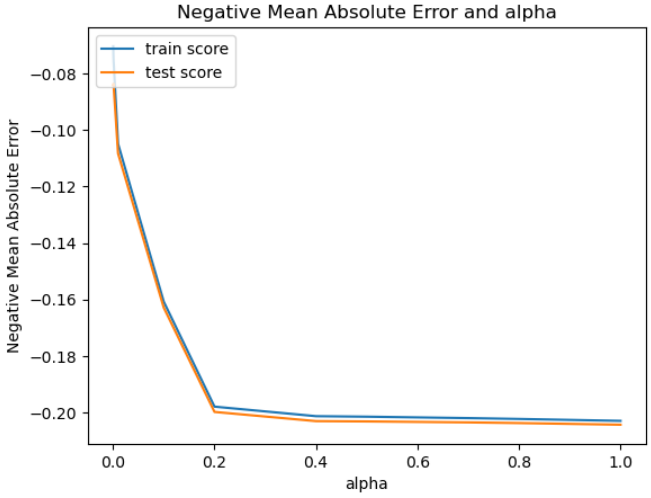
Answer:

After building the model using Ridge and Lasso regression, the optimal value, or the best value of alpha for Ridge regression is 2 whereas the optimal or the best value of alpha for Lasso regression is 0.01



For Ridge Regression: When i plot the graph between Negative Mean Absolute Error and alpha, i observed that the value of test error is minimum when alpha is 2, so i chose the optimal value of alpha for Ridge regression to be 2

For Lasso Regression: Again, same as Ridge, on plotting the graph between Negative Mean Absolute Error and alpha for lasso, I observed that at around 0.01 is the optimal value



When I double the value of alpha for Ridge regression, the penalty will increase more on the model. The complexity of the model might decrease making it simpler. But, with increase in the value of alpha, the errors on the train and test will also be increased.
When I double the value of alpha for Lasso regression, the penalty will increase, and the model will try to make more coefficients to zero.
The most important predictor variables after the change is implemented:

For Ridge regression –Top 5 features having non-zero coefficients are:

- MSZoning_FV - general zoning classification: Floating Village Residential
- MSZoning_RH - general zoning classification: Residential High Density
- MSZoning_RL - general zoning classification: Residential Low Density
- MSZoning_RM - general zoning classification: Residential Medium Density
- Street_Pave - Type of road access to property: Paved

For Lasso regression –Top 5 features having non-zero coefficients are:

- GrLivArea - Above grade (ground) living area
- OverallQual - Rates the overall material and finish of the house
- LotArea - Lot size
- BsmtFinSF1 - Type 1 finished
- TotalBsmtSF - Total square feet of basement area

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:

The optimal value of lambda for Ridge is 2 and for Lasso it is 0.01 after building the model. Among the two models, I would choose Lasso regression. As the value of lambda for Lasso is smaller compared to Ridge, this would be a better model. Lasso model does not include all features in the model. It only includes features selection by identifying features which are significant and makes other non-significant coefficients as zero. This helps in selecting the best predictor variables.

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:

After building the model, the five most important predictor variables in Lasso model are:

- GrLivArea - Above grade (ground) living area
- OverallQual - Rates the overall material and finish of the house
- LotArea - Lot size

- BsmtFinSF1 - Type 1 finished
- TotalBsmtSF - Total square feet of basement area

As we can see in the below picture, after dropping the above 5 features, the five most important predictor variables now are:

- BedroomAbvGr – Bedroom above ground
- FullBath – Basement full bathrooms
- ScreenPorch – Screen porch area
- OpenPorchSF – Open porch area
- WoodDeckSF - Wood deck area

	Variable	Coeff
0	constant	8.909
223	GrLivArea	0.177
214	OverallQual	0.114
213	LotArea	0.090
217	BsmtFinSF1	0.014
220	TotalBsmtSF	0.013
227	BedroomAbvGr	0.011
225	FullBath	0.011
236	ScreenPorch	0.010
234	OpenPorchSF	0.007
233	WoodDeckSF	0.007
219	BsmtUnfSF	0.001
235	EnclosedPorch	0.001
237	PropAge	-0.002
157	HeatingQC_TA	-0.007
165	KitchenQual_TA	-0.018
174	FireplaceQu_No Fireplace	-0.029

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:

For making any model robust and generalised, we need to balance the trade off between bias and variance (Bias-Variance trade off). We should try to make simple models which generally would have no overfitting. The accuracy or the r2 score on train and test should be same with not much difference. If there is a significant difference between the r2 score of

train and test, then that is the case of overfitting. Usually, train score would be high as model tries to memorise the data but the r2 score on test would be low and model fails to perform well on unseen data.
For simple model, bias is more, and variance is less hence making it a generalised mode. It is very important to maintain a balance between bias and variance in order to avoid overfitting and underfitting too.
A robust and generalised model usually performs equally on both the train and test data set, and there will not be much difference in their accuracy.
Below table shows the RMSE and R2 score metrics for all three regression models.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	RMSE (Train)	0.084977	0.096752	0.153266
1	RMSE (Test)	0.154157	0.137977	0.168913
2	R2 Score (Train)	0.950725	0.936123	0.839709
3	R2 Score (Test)	0.874256	0.899265	0.849031