Advanced Regression Assignment.

Shameer Sheik

Note: Please limit your answers to less than 500 words per question.

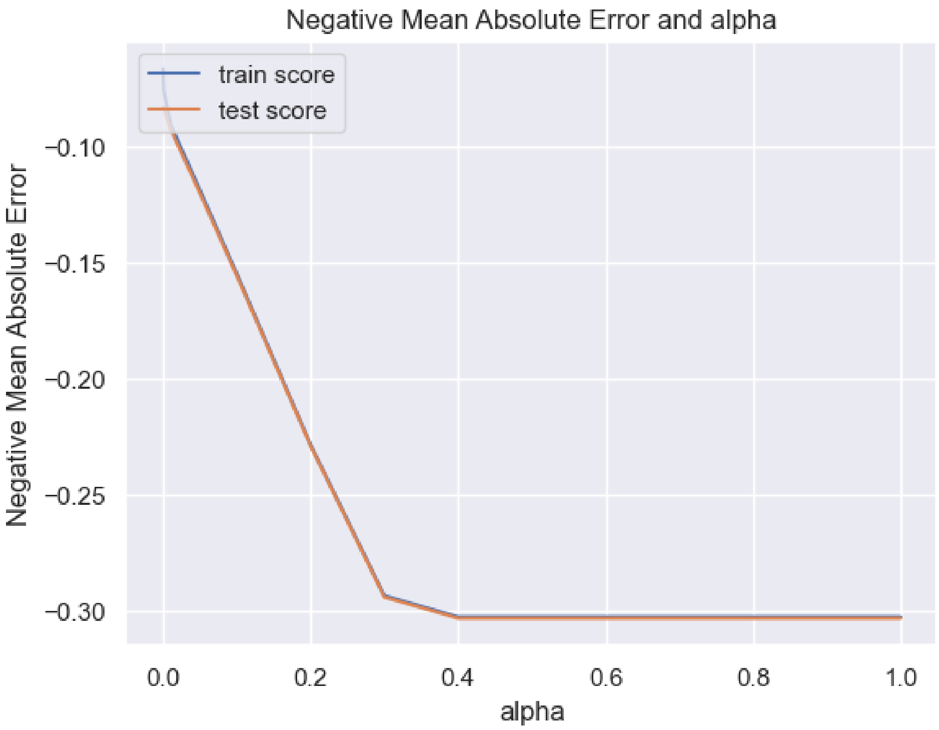
Question 1:

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:

Lasso Regression:

Optimum value of alpha is 0.01.

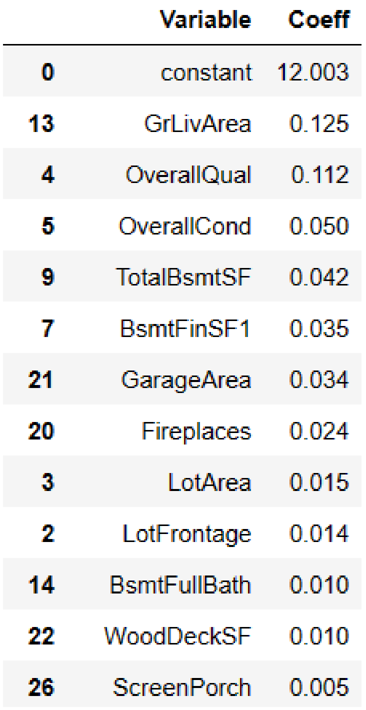


From the above graph we can see that the Negative Mean Absolute Error is quite low at alpha = 0.4 and stabilises thereafter, but we will choose a low value of alpha to balance the trade-off between Bias-Variance and to get the coefficients of smallest of features.

At alpha = 0.01, even the smallest of negative coefficients that have some predictive power towards 'SalePrice' have been generated.

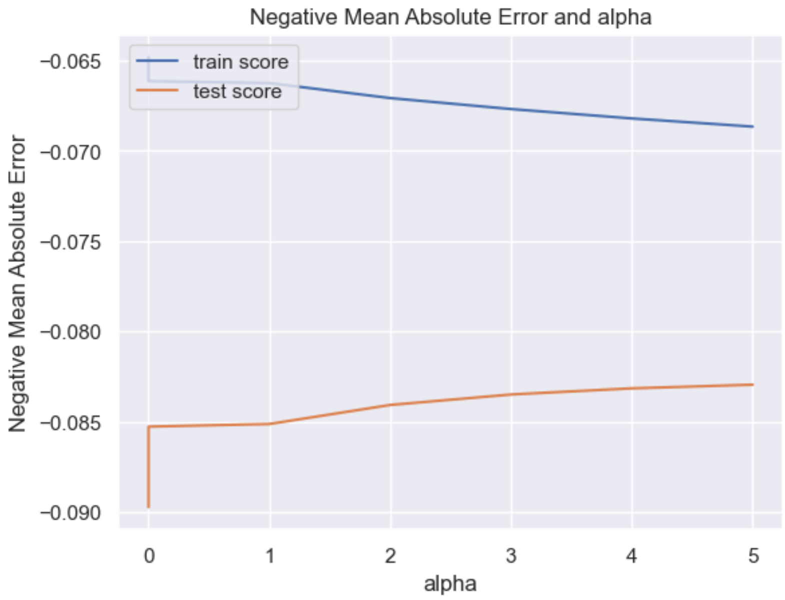
The advantage of this technique is clearly visible here as Lasso brings the coefficients of insignificant features to zero.

Prediction accuracy with train and test set is 0.8854 and 0.8894, RMSE is 0.1257



	Variable	Coeff
0	constant	12.003
13	GrLivArea	0.125
4	OverallQual	0.112
5	OverallCond	0.050
9	TotalBsmtSF	0.042
7	BsmtFinSF1	0.035
21	GarageArea	0.034
20	Fireplaces	0.024
3	LotArea	0.015
2	LotFrontage	0.014
14	BsmtFullBath	0.010
22	WoodDeckSF	0.010
26	ScreenPorch	0.005

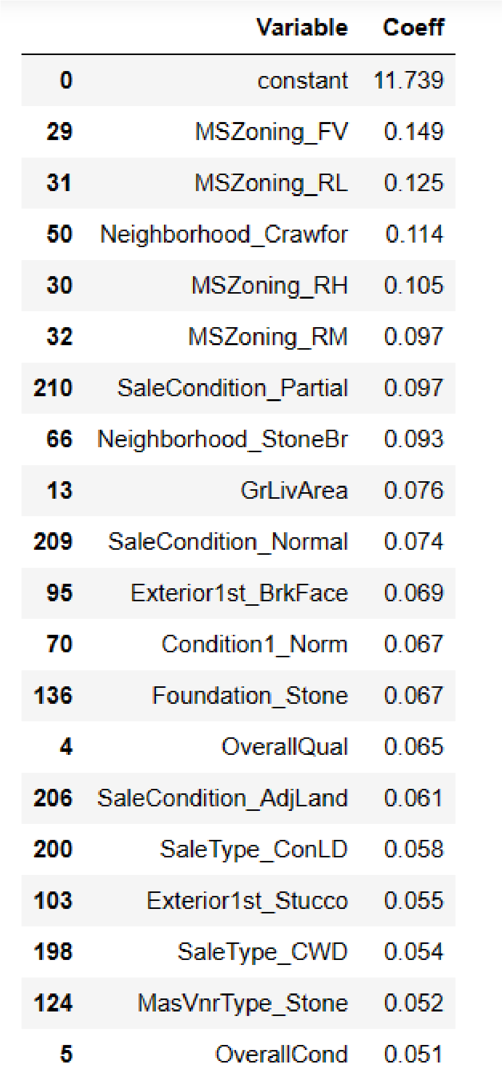
Ridge Regression:



Negative Mean Absolute Error stabilises at alpha = 2, we will choose this as optimal value for further analysis.

At alpha = 2,

Prediction accuracy with train and test set is 0.9364 and 0.9077, RMSE is 0.1148



	Variable	Coeff
0	constant	11.739
29	MSZoning_FV	0.149
31	MSZoning_RL	0.125
50	Neighborhood_Crawfor	0.114
30	MSZoning_RH	0.105
32	MSZoning_RM	0.097
210	SaleCondition_Partial	0.097
66	Neighborhood_StoneBr	0.093
13	GrLivArea	0.076
209	SaleCondition_Normal	0.074
95	Exterior1st_BrkFace	0.069
70	Condition1_Norm	0.067
136	Foundation_Stone	0.067
4	OverallQual	0.065
206	SaleCondition_AdjLand	0.061
200	SaleType_ConLD	0.058
103	Exterior1st_Stucco	0.055
198	SaleType_CWD	0.054
124	MasVnrType_Stone	0.052
5	OverallCond	0.051

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:

Though the model performance by Ridge Regression was better in terms of R2 values of Train and Test, it is better to use Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables. It removes unwanted features from the model with-out affecting the model accuracy.

It is always advisable to use simple yet robust model. Equation can be formulated using the features and coefficients obtained by Lasso.

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:

These 16 variables obtained from Lasso Regression can be concluded to have the strong effect on the SalePrice as per their coefficient values.

['GrLivArea', 'OverallQual', 'OverallCond', 'TotalBsmtSF', 'GarageArea', 'BsmtFinSF1', 'Fireplaces', 'LotArea', 'LotFrontage', 'BsmtFullBath', 'Foundation_PConc', 'OpenPorchSF', 'FullBath', 'ScreenPorch', 'WoodDeckSF']

The equation derived from the model developed is below:

Log(Y) = C + 0.125(x1) + 0.112(x2) + 0.050(x3) + 0.042(x4) + 0.035(x5) + 0.034(x6) + 0.024(x7) + 0.015(x8) + 0.014(x9) + 0.010(x10) + 0.010(x11) + 0.005(x12) - 0.007(x13) - 0.007(x14) - 0.008(x15) - 0.095(x16) + Error term (RSS + alpha * (sum of absolute value of coefficients).

Suggestions for Surprise Housing is to keep a check on these predictors affecting the price of the house.

The higher values of positive coefficients suggest a high sale value.

Some of those features/variables are:

Feature Description

- GrLivArea - Above grade (ground) living area square feet.
- OverallQual - Rates the overall material and finish of the house.
- OverallCond - Rates the overall condition of the house.
- TotalBsmtSF - Total square feet of basement area.
- GarageArea - Size of garage in square feet.

The higher values of negative coefficients suggest a decrease in sale value.

Some of those features/variables are:

Feature Description

- PropAge - Age of the property at the time of selling.
- MSSubClass - Identifies the type of dwelling involved in the sale.

When the market value of the property is lower than the Predicted Sale Price, it’s the time to buy.

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:

To make model robust and generalisable 3 features are required:

1. Model accuracy should be > 70-75%: In our case its coming 88%(Train) and 88%(Test) which is correct.
2. P-value of all the features is < 0.05
3. VIF of all the features are < 5

Thus we are sure that model is robust and generalisable