Advance Regression Assignment Part 2

Surprise Housing

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:

Optimal value of alpha for Ridge is: **7.0**

Optimal value of alpha for Lasso is: **0.001**

If we double the value of alpha as follows:

Optimal value of alpha for Ridge is: **14**

Optimal value of alpha for Lasso is: **0.002**

As observed the R squared for test set after doubling the alpha increases.

Ridge max col = GrLivArea

Ridge max coef = 0.30015931617448616

Ridge_double max col = GrLivArea

Ridge_double max coef = 0.2729369582474351

Lasso max col = GrLivArea

Lasso max coef = 0.3040142357443962

Lasso_double max col = GrLivArea

Lasso_double max coef = 0.317942215616150

GrLivArea is the most important predictor variable.

Also it is logical that the above ground living area will make impact on it's selling price.

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:

	Ridge Regression	Lasso Regression	Ridge Regression Double	Lasso Regression Double
Train R square	0.947078	0.946261	0.942475	0.933817
Test R square	0.899667	0.883419	0.902123	0.899258

We will use ridge regression as the R square for ridge is comparatively high.

Question 3:

After building the model, you realized 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 Model code is available in the Python file.

Top 5 features:

TotalBsmtSF 0.1918430209005055
GrLivArea 0.3040142357443962
Neighborhood_Crawfor 0.29011694290972645
Neighborhood_Somerst 0.2256070198615187
Exterior1st_BrkFace 0.2102328555223782

Top 5 features after dropping above 5 features:

OverallQual 0.18028160058430578
YearBuilt 0.1542593532101461
BsmtFinSF1 0.12430252079465438
CentralAir 0.23938927289785245
1stFlrSF 0.2299225298919935

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 robustness of a model is how effective it performs while being tested on a new (independent) dataset. Robust algorithm is the one, where the testing error is close to the training error.

A generalisable model is the one which works as good as it works on the model it is trained or tested on.

Model should be accurate to the data sets other than the dataset it is trained on.