House Price Prediction

Advanced Linear Regression

By:

Tejas Kumar D A

25-Mar-2024

Assignment-based Subjective Questions

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 for ridge is alpha=4

Optimal value for Lasso is alpha=100

If we double the alpha the R2 Score slightly reduces both train and test data for both ridge & Lasso

Most Important Predictor variables are :

1. LotFrontage
2. LotArea
3. OverallQual
4. OverallCond
5. YearBuilt
6. YearRemodAdd

	Linear	Ridge	Lasso
MSSubClass	-8.142989e+03	-13092.044641	-19848.251479
LotFrontage	1.292481e+04	15917.918556	12278.141933
LotArea	4.567213e+04	22103.603820	12626.167279
OverallQual	5.687373e+04	48255.235990	76116.194439
OverallCond	4.833813e+04	26272.528631	27804.113127
YearBuilt	3.081795e+04	14253.444395	20120.071790
YearRemodAdd	2.728581e+03	7230.087900	4638.750707

Question 1 Contnd……

After the change is implemented the predictor varibales remains same but the coefficients change.

	Ridge2	Ridge	Lasso	Lasso2
MSSubClass	-12693.014313	-13092.044641	-19848.251479	-18520.615676
LotFrontage	14553.503581	15917.918556	12278.141933	7600.934233
LotArea	18042.479987	22103.603820	12626.167279	970.915141
OverallQual	41706.316085	48255.235990	76116.194439	86721.108001
OverallCond	19486.203555	26272.528631	27804.113127	16446.139542
YearBuilt	10138.147524	14253.444395	20120.071790	11952.577436
YearRemodAdd	8534.193516	7230.087900	4638.750707	7465.986652

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 R2 Score of ridge on test data is .9236 is slightly better than the R2 Score of Lasso which is 0.9206.Hence we will choose Ridge 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?

Below are the important predictors now:

1. YearRemodAdd
2. MasVnrArea
3. BsmtFinSF1
4. BsmtFinSF2
5. BsmtUnfSF

	Lasso3
MSSubClass	-22280.067709
YearRemodAdd	10801.278004
MasVnrArea	16304.020158
BsmtFinSF1	33800.136196
BsmtFinSF2	0.000000
BsmtUnfSF	0.000000
TotalBsmtSF	36599.167668

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?

Overfitting, underfitting needs to reduced to maximum extent possible using the regularisation techniques Ridge and Lasso. The model should be able to predict the unseen test data efficiently. Data should cover wide range of scenarios .EDA should be performed well which eliminates outliers etc . hyperparameters of models to find the optimal values that minimize the error or maximize the performance of your models on your data. Test data for testing should not be same as training data. Visualise the data and present them.

END