Advanced-Regression-Assignment-Problem statement_ThaoTran

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 1

Since I only chose Lasso regression, because I have many variables and Lasso helps to reduce parameters (as some coefficient can be zero), my following answer focuses only in Lasso regression.

For Lasso, first, I create a list of values starting from 0.0001, next I applied GridSearchCV method to detect the optimum value. Using GridSearchCV, it returned me the most optimum lamdba which was 100. Then, I re-ran the Lasso with alpha = 100.

I also tried to increase the lamdba, I experienced the decreased in number of features and R2 squared decrease as the result. Because the model was getting more generalised and more simpler when we increase number of lamdba.

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 2

As described above, since Lasso help to reduce number of features, I thought it would be approriate to using when I was dealing with large number of features (I had more than 50 included dummy varibles).

The list of important predictors after applied Lasso was in below table. Please note that those predictors contains dummy variables created from categorical variables.

YearBuilt
MasVnrArea
TotalBsmtSF
Fireplaces
1.5Unf

1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl
BrkFace
HdBoard
MetalSd
Plywood
Stucco
VinylSd
None
Fa
Gd
TA
CBlock

PConc
Slab
Fa
Gd
TA
Fa
Gd
TA
Fa
Gd
TA
Attchd
BuiltIn
RFn
Unf

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 3

In my model, 5 most important variables are in my current model.

- YearBuilt
- MasVnrArea
- TotalBsmtSF
- Fireplaces
- 1.5Unf ➔ dummy variables.

However, my approach is not strictly following machine learning approach. First, I plotted box plots for all categorical variables with y axis was SalePrice. According to plots, I picked out varibles that had sale price vary among categories. In my point of view, it’s nonsense to pass variables that may not impact on dependent variables.

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 4

The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, It cannot be trusted for predictive analysis.