ADVANCED REGRESSION ASSIGNMENT ON SURPRISE HOUSING

- Vinoda Varshini Rajagopal

ASSIGNMENT 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:

Optimal Value of Alpha for ridge is 10
Optimal Value of Alpha for lasso is 0.001

Changes in the model if doubling the values of alpha.

- Ridge and Lasso changes are similar for R-squared and it is given as
  - R-Squared value for train set changes from 0.91 to 0.90.
  - R-Squared Value for test set changes from 0.83 to 0.82.
- RSS value for train set has increased to 119.12 from 104.43 for ridge.
- RSS value for test set increased from 66.82 to68.24 for both ridge and lasso.
- RMSE value for train set increased from 0.30 to 0.32 for both ridge and lasso.

Important Predictor Variable after the change is implemented

These are the top 10 important predictor variable and the coefficients

- OverallQual_9 : 0.969339
- OverallQual_10 : 0.735865
- Neighborhood_StoneBr : 0.486365
- Neighborhood_NridgHt : 0.385374
- OverallQual_8 : 0.351464
- Neighborhood_NoRidge : 0.335444
- Neighborhood_Crawfor : 0.309673
- SaleType_New : 0.260206
- GrLivArea : 0.253784
- BsmtExposure_Gd : 0.233823

Note: model for this answer is done in python notebook.

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:

- In case of lasso, when lambda value increases, the co-efficents are shrinked towards zero. Penalty of lasso forces some of the coefficients estimates to be exactly equal to zero.
- Lasso performs variable selection. Models generated from the lasso are generally easier to interpret than those produced by ridge regression.
- Lasso performs better in situations where only a few among all the predictors that are used to build our model have a significant influence on the response variable.
- Higher the value of lambda in the shrinkage term, more are the model co-efficients pushed towards zero and hence more the regularization.
- So, I will choose Lasso for this model.

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 removing the top 5 important predictor variable, a new lasso model has been built.

Five important predictor variables available now with their coefficients are

- RoofMatl_WdShngl : 0.342594
- GrLivArea : 0.285835
- GarageYrBlt_2009.0 : 0.258300
- SaleType_New : 0.239836
- BsmtExposure_Gd : 0.209258

Note: model for this answer is done in python notebook.

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:

- A model is robust when its performance is not affected even there is some variation in our data.
- Generalisation refers to a model's ability to perform wee on unseen data.
- A model should not over fit in order to be robust and generalizable. A model is said to over fit when there is high variance. Such situation occurs when a model memorises all the data points.
- A simple model is good example for robust and generalizable.
- Accuracy is the percentage of correct predictions in test data. In order to have a good accuracy model, bias should be reduced.
- Achieving a model with robustness and generalizable is possible with help of bias-variance tradeoffs.
- To prevent a model from becoming too complex regularization technique is used like Ridge and Lasso.