Advanced Regression Assignment -2

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: **1.0**

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

Suppose if the alpha value is doubled,

For Ridge: There is a slight variation in the coefficient values and the r2_score is also dropping.

For Lasso: Some features are removed from the model and there is a slight variation in the r2_score.

Top 10 Features:

1) LotConfig_FR2
2) BedroomAbvGr
3) ExterQual
4) Condition2_Norm
5) LotArea
6) HalfBath
7) LotConfig_Inside
8) GarageFinish
9) BsmtFinSF2
10) ExterCond

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:

Model was created both in Ridge and Lasso and there is no considerable difference in r2_score. However, comparing the optimal value of lambda/alpha in Lasso (0.001) is lesser than that of Ridge(1.0). Hence I would consider Lasso model over Ridge 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 dropping the first most important predictor below features and the model is rebuilt.

1)LotConfig_FR2

2)BedroomAbvGr

3)ExterQual

4)Condition2_Norm

5)LotArea

The below are the new set of top five features:

1) LotConfig_CulDSac

2) HalfBath

3) OverallCond

4) FullBath

5) LotConfig_Inside _RRNn

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 classic approach towards the assessment of any machine learning model revolves around the evaluation of its generalizability i.e. its performance on unseen test scenarios. Evaluating such models on an available non-overlapping test set is popular, yet significantly limited in its ability to explore the model's resilience to outliers and noisy data / labels (i.e. robustness).

Here are some changes you can make to your data:

1) Fix missing values and outliers: If the data has missing values and outliers can lead to inaccurate model. Outliers can affect the mean, median that we are imputing to continuous variables

2) Remove the outliers. This works if there are very few of them and its certain they're anomalies and not worth predicting

3) Feature Selection: Domain knowledge plays an important role in feature selection, additional techniques like data visualization also helps the selecting the features. Statistical parameters like p-Values, VIF can give us significant variables.

4) Algorithm selection: Choosing the right machine learning algorithm is very important to get accurate model.

5) Cross validation: To reduce overfitting user cross validation i.e. leave a sample on which you do not train the model & test the model on this sample before got to the final mode.