Problem Statement - 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 1:

As per the calculations done in the iPython Notebook:

- Optimal value of Alpha for Ridge Regression is: 35
- Optimal value of Alpha for Ridge Regression is: 0.009

If we double the Alpha values for Ridge Regression, then:

- For Alpha = 35

Prediction Score on Train Data: 0.9126276726469724

Prediction Score on Test Data: 0.8851734907756774

And top 15 variable coefficients are:

(0.142, 'OverallQual'),

(-0.136, 'MSZoning_C (all)'),

(-0.132, 'Neighborhood_Edwards'),

(0.129, 'Neighborhood_Crawfor'),

(0.102, 'BsmtFullBath'),

(0.102, 'Neighborhood_NridgHt'),

(0.101, 'Condition1_Norm'),

(0.094, 'GarageCars'),

(-0.094, 'MSSubClass_30'),

(0.092, 'OverallCond'),

(-0.092, 'LandContour_Bnk'),

(0.092, 'Neighborhood_Somerst'),

(-0.087, 'RoofMatl_ClyTile'),

(-0.085, 'Neighborhood_IDOTRR'),

(0.084, 'MSZoning_RL'

- For Alpha = 70

Train r2 score: 0.9043686670613059

Test r2 score: 0.8817211453392724

And top 15 variable coefficients are:

(0.142, 'OverallQual'),

(-0.096, 'Neighborhood_Edwards'),

(0.091, 'OverallCond'),

(0.086, 'BsmtFullBath'),

(0.086, 'Neighborhood_Crawfor'),

(0.084, 'Condition1_Norm'),

(-0.078, 'MSZoning_C (all)'),

(0.075, 'GarageCars'),

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 optimal lambda value in case of Ridge and Lasso is below:
  - Ridge - 10
  - Lasso – 0.0004
- The Mean Squared Error in case of Ridge and Lasso are:
  - Ridge – 0.13743
  - Lasso – 0.013556
- The mean squared error of Lasso is slightly lower than that of Ridge
- Also since Lasso helps in feature reduction (as coefficient value of one of the feature become 0), Lasso has a better edge over Ridge

Therefore variables predicted by the Lasso can be applied to choose significant variables for predicting the price of house

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:

Five most important predictor variables

11stFlrSF----------First Floor square feet

GrLivArea----------Above grade (ground) living area square feet

Street_Pave---------Pave road access to property

RoofMatl_Metal------Roof material_Metal

RoofStyle_Shed------Type of roof(Shed)

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 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.