Upgrad IIITB EPGP in ML & AI

Assignment- Advanced Regression on House Price Prediction

- Vignesh G

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?

Ans:

The optimal values of alpha for Ridge and Lasso are:

Lasso = 0.0001

Ridge= 5

Changes in the model after doubling alpha values for Ridge and Lasso:

Lasso = 0.0002

Ridge =10

- R-Squared for Ridge and Lasso varies slightly after change in Alpha:
  - R-Squared for Ridge Train changes from 0**.89 to 0.90.**
  - R-Squared for Ridge Test doesn’t change much and remains at **0.88.**
  - R-Squared for Lasso Train changes from **0.90 to 0.91**
  - R-Squared for Lasso Test changes from **0.885 to 0.886**
  - RSS for both Ridge and Lasso Test remained almost the same.
  - RMSE improved for train set from **0.36 to 0.34** but remained the same for the Test set at around **0.35.**

The Top most Important predictor variables are provided below:

GrLivArea	**- 0.31468**
OverallQual	**- 0.108634**
RoofMatl_WdShngl	**- 0.080339**
BsmtFinSF1	**- 0.07354**
Neighborhood_NoRidge	**- 0.055804**
MasVnrArea	**- 0.054163**
OverallCond	**- 0.053385**
Neighborhood_NridgHt	**- 0.048224**

Neighborhood_StoneBr	**- 0.041702**
TotalBsmtSF	**- 0.035579**

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?

Ans:

The optimal value for Alpha for Ridge and Lasso is:

Lasso = 0.001

Ridge = 5

The R2 values for Ridge and Lasso are:

Ridge Train: 0.892578

Ridge Test: 0.881025

Lasso Train: 0.900947

Lasso Test: 0.885886

Root Mean Square Error (RMSE):

Ridge : 0.035854

Lasso : 0.035114

Mean Squared Error (MSE):

Ridge : 0.001286

Lasso : 0.001233

The model performance for Lasso is better in all aspects with better R2 value and low error for both test and train. We consider Lasso as our best model since it helps in feature elimination and better overall accuracy and reduces overfitting in our data. The top five features are provided below:

GrLivArea	**- 0.31468**
OverallQual	**- 0.108634**
RoofMatl_WdShngl	**- 0.080339**
BsmtFinSF1	**- 0.07354**
Neighborhood_NoRidge	**- 0.055804**

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?

The most important predictor variable now would be the following 5 features.

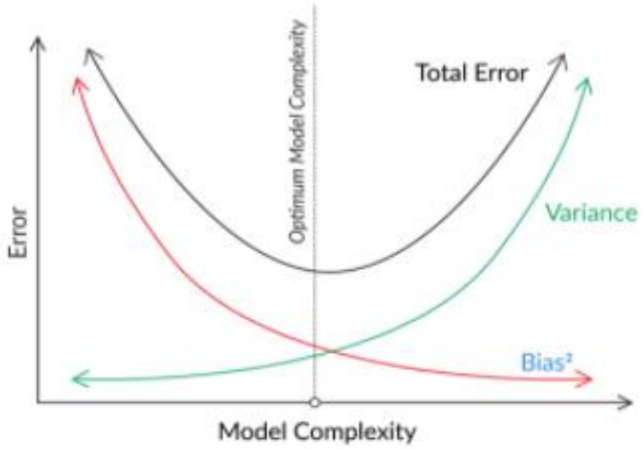
1stFlrSF	0.232522
2ndFlrSF	0.204512
TotalBsmtSF	0.168758
MasVnrArea	0.092785
OverallCond	0.075731

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?

Ans:

As per Occum's Razor, given the copeting theories and explanations, the simplest ones should be preferred.



- Simpler models are usually more 'generic' and are applicable.
- Simpler models require fewer training samples for effective training than more complex ones and easier to train.
- The more robust and generalisable the model will perform equally well on both training and test data.

We can make the model simple but not to simple so that it will affect the overall accuracy.

Making the model simple leads to Bias-Variance Trade off as shown in the above image.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High Variance makes accuracy in the training data remarkably good but when it is implemented in test/unseen data it will not perform as good as in the training data.

Thus, accuracy of the model can be maintained by keeping the balance between bias and variance as it minimizes the total error. For such cases Regularization can be used to have a balance between bias-variance and keeping the model simple.