Advanced regression assignment: House Price Prediction

(Subjective questions answered)

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:

a. For Lasso regression: Optimum value of alpha is 0.0001

R Squared values with optimum alpha value are:

R squared for train set	0.9276836216285358
R squared for test set	0.822955707487891

R squared scores start diminishing if we increase alpha value. The double of 0.0001 is still a small value & the drop in R squared is evident but not high. With double the alpha value (0.0002) the R squared values are as follows:

R squared for train set	0.9264019238306473
R squared for test set	0.8254168021556487

Changes in top 5 variables by highest coefficients:

Top 5 Variables with alpha = 0.0001 (Variable name: Coefficient value)	Top 5 Variables with alpha = 0.0002 (Variable name: Coefficient value)
LotConfig_FR2: 1.065221	LotConfig_FR2: 1.070432
LandContour_HLS: 0.630579	MoSold: 0.413054
LotShape_Reg: 0.509908	LandContour_HLS: 0.397454
BsmtHalfBath: 0.419807	EnclosedPorch: 0.367173
MoSold: 0.417513	BsmtHalfBath: 0.323126

Code for above is included at the end of Python Jupiter notebook.

b. For Ridge regression: Optimum value of alpha is 0.5

R Squared values with optimum alpha value are:

R squared for train set	0.9242607095451041
R squared for test set	0.8540061507930531

R squared scores start diminishing if we increase alpha value. The double of 0.5 is still a small difference in alpha value & the drop in R squared is evident but not high. With double the alpha value (1) the R squared values are as follows:

R squared for train set	0.9197869691086553
R squared for test set	0.8641671717762796

Changes in top 5 variables by highest coefficients:

Top 5 Variables with alpha = 0.5 (Variable name: Coefficient value)	Top 5 Variables with alpha = 1 (Variable name: Coefficient value)
LotConfig_FR2: 1.035325	LotConfig_FR2: 0.968523
MoSold: 0.446635	MoSold: 0.452255
LandContour_HLS: 0.435145	EnclosedPorch: 0.376201
BsmtHalfBath: 0.407055	BsmtHalfBath: 0.348239
EnclosedPorch: 0.380384	AgeYearBuilt: 0.323998

Code for above is included at the end of Python Jupiter 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:

Optimum lambada value (alpha) for ridge : 0.5

Optimum lambada value (alpha) for lasso : 0.0001

I will choose to go ahead with lasso regression because it helps in feature elimination by it's inherent nature. This will be useful more as the number of features is high.

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:

Following are the five most important predictor variables by coefficient values in lasso regression using the optimum alpha value.

(*Variable name: Coefficient value*)

1. LotConfig_FR2: 1.065221

2. LandContour_HLS: 0.630579

3. LotShape_Reg: 0.509908

4. BsmtHalfBath: 0.419807

5. MoSold: 0.417513

If we drop these variables from data & create a new model from the beginning, the top five predictors will be as follows:

(*Variable name: Coefficient value*)

-LotConfig_Inside: 0.788170

-RoofStyle_Shed: 0.533050

-AgeYearRemodAdd: 0.450218

-Condition1_Norm: 0.433033

-BedroomAbvGr: 0.405841

New R squared is Train: 0.93, Test: 0.83 which is good score. (Working done in a separate file, not a part of assignment python notebook)

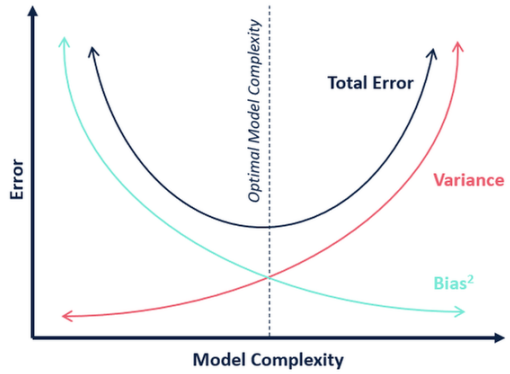
Question 4

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: To keep the model robust & generalizable it is important to keep it as simple as possible. The balance between bias & variance needs to be checked to avoid making model too complex or too simple, which in both cases makes it less effective.

Simple models are generalizable, but if a model becomes too simple then it will have high bias. Complex model has less bias, but high variance & is likely to perform worse on unseen test data. (**Continued on next page….**)

(**Answer 4 continued….**) The following image sourced from internet shows the optimum point of complexity, which is the lowest point on Total Error curve.



Steps that can be taken to make the model robust & generalizable:

1. Choosing the correct model type depending on the requirement
2. Using hyperparameters & their tuning to control the model overfitting (finding the optimum value of hyperparameters. E.g. Alpha in the current assignment)