Advance Regression Assignment

SUBJECTIVE QUESTIONS

Table of Contents

Assignment-based Subjective Questions

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?**

   A: For Ridge it is 0.1, and for Lasso it is 0.001. However, the Ridge regression doesn't seem to be solving the overfitting problem, as the R2 value for Training is 0.95 and for Test it is 0.65, where as the Lasso seems to be doing well with R2 value for Training is 0.83 and for Test it is 0.81. With this below are the important paraments for each Ridge and Lasso.

Feature	Ridge
PoolQC_Gd	-0.74675
Condition2_PosN	-0.509939
GrLivArea	0.184795
RoofMatl_WdShngl	0.17827
1stFlrSF	0.160266

Feature	Lasso
GrLivArea	0.238334
OverallQual__10	0.085111
OverallQual__9	0.078668
GarageCars	0.063517
Neighborhood_NoRidge	0.045413

Now, if we double the existing optimum value, below is what we get.

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.950711	0.762700
R2 Score (Test)	0.718287	0.743164

Feature	Ridge
PoolQC_Gd	-0.59857
Condition2_PosN	-0.42317
GrLivArea	0.173801
RoofMatl_WdShngl	0.167812
1stFlrSF	0.148463

Feature	Lasso
GrLivArea	0.145785
GarageCars	0.062703
OverallQual__9	0.042937
OverallQual__10	0.036833
FireplaceQu_Na	-0.03407

After hyperparameter value is doubled

Ridge – No change.

Lasso – Neighborhood_NoRidge is replaced with FireplaceQu_Na, rest of the variable order is reshuffled.

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?**
   A: We will use Lasso regression for this as it seems to be doing better with

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.954845	0.830478
R2 Score (Test)	0.652312	0.814258
RSS (Train)	0.555629	2.085979
RSS (Test)	1.889969	1.00966
MSE (Train)	0.023328	0.0452
MSE (Test)	0.065689	0.048012

The Ridge regression doesn't seem to be solving the overfitting issue with all the variables, where as Lasso regression while making coefficients of some variables as 0 performing well on both train and test data sets. Hence it would be wise to implement the Lasso regression on this data.

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?**
   A: The top variables in Lasso are as follows –

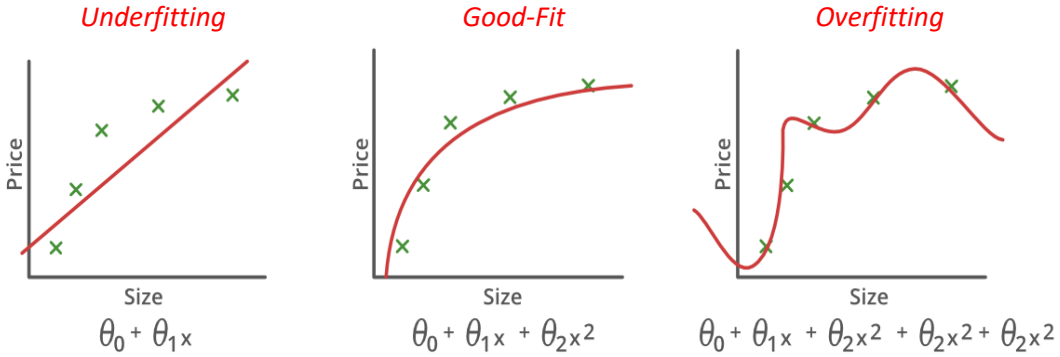
Feature	Lasso
GrLivArea	0.238334
OverallQual__10	0.085111
OverallQual__9	0.078668
GarageCars	0.063517
Neighborhood_NoRidge	0.045413

So, after removing the original variables before dummy variable creation (which are GrLivArea, OverallQual, GarageCars, Neighborhood), below are the top 5 variables we get.

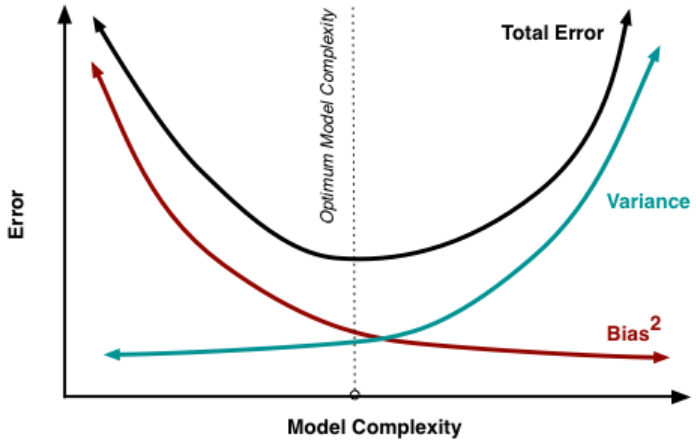
Features	Lasso
1stFlrSF	0.149781
2ndFlrSF	0.10587
GarageArea	0.087525
TotRmsAbvGrd	0.041274
BsmtExposure_Gd	0.039583

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?**

A: A model is good fit if it performs well on both train and test data set with considerable accuracy. To achieve the good accuracy model has to understand the features well, however it has downside that if it learns all the data points it may tend to overfit and might not work well on the test dataset. On the other hand, if the model doesn't learn the existing data points it may become underfit.



Another factor a good model should have the right complexity and error. So, to achieve the optimum model complexity. If model is too complex it will have high variance, which means it learns too much from the data that it considers even the noise as something to learn from. At the same time it will have low bias which might sound right but it also means that it will only fit the training data well.



In order to make model Robust and Generalized we need to strike a balance between Bias and Variance. This is done by penalizing each individual variable so that its coefficient is adjusted to the minimum. This affects the overall accuracy of the model on training data however it increases the chances to fit the model on unseen data.