Problem Statement - Part II

Advanced Regression Assignment

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:

The optimal value of alpha for ridge: 0.4

The optimal value of alpha for lasso: 0.0002

Top 10 predictors in ridge before doubling the alpha value:

Top 10 predictors in Ridge before doubling the alpha value

	Ridge
OverallQual	0.6297
1stFlrSF	0.5859
LotArea	0.3595
OverallCond	0.2605
2ndFlrSF	0.1609
LowQualFinSF	0.0970
BsmtFinSF1	0.0715
TotalBsmtSF	0.0497
BsmtFinSF2	0.0039
BsmtUnfSF	-0.0446

Top 10 predictors in Lasso before doubling the alpha value

	Lasso
GrLivArea	1.0022
OverallQual	0.6528
LotArea	0.3482
1stFlrSF	0.3322
GarageCars	0.3269
OverallCond	0.2614
BsmtFullBath	0.1738
Neighborhood_Crawfor	0.1654
LandContour_Low	0.1451
TotRmsAbvGrd	0.1411

Top 10 predictors in ridge and lasso after doubling the alpha value:

Top 10 predictors in Ridge after doubling the alpha value

	Ridge
OverallQual	0.6253
1stFlrSF	0.5277
GrLivArea	0.5067
GarageCars	0.3435
LotArea	0.3182
OverallCond	0.2566
TotRmsAbvGrd	0.1893
Neighborhood_Crawfor	0.1709
2ndFlrSF	0.1650
BsmtFullBath	0.1624

Top 10 predictors in lasso after doubling the alpha value

	Lasso
GrLivArea	0.9797
OverallQual	0.6708
GarageCars	0.3278
1stFlrSF	0.3167
LotArea	0.2826
OverallCond	0.2587
BsmtFullBath	0.1818
Neighborhood_Crawfor	0.1621
TotRmsAbvGrd	0.1534
LandContour_Low	0.1287

Observation: There is change in the top 5 variables in both ridge and lasso after doubling. Also, in the case of lasso regression, there are toatlly 11 variables got eliminated whereas, before doubling 9 variables got eliminated.

Coding part – Included in the jupyter 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?

We have to choose the alpha value with more care. If the value of alpha is very less it leads to overfitting and if the value of alpha is high it leads to underfitting.

In the case of Ridge regression, the coefficients value will be moved towards '0' but will not become '0'. In Lasso regression, the coefficients will be shrink to '0', which in turn says, that the particular feature can be excluded.

We applied the optimal alpha value obtained in Ridge and Lasso and build a model. Though the accuracy remains same in both, we are going ahead with Lasso regression.

We can choose **Lasso regression** as there is a **feature elimination**. Out of 50 features selected by RFE, Lasso chooses 41 features thereby reducing the model complexity and thus making the model highly robust.

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?

After dropping the top 5 important predictor variables the following five predictors has been now identified as important from lasso regression.

	Lasso
TotalBsmtSF	1.5999
TotRmsAbvGrd	0.5932
2ndFlrSF	0.4048
OverallCond	0.3132
LandContour_HLS	0.2045

Coding part – Included in the jupyter notebook.

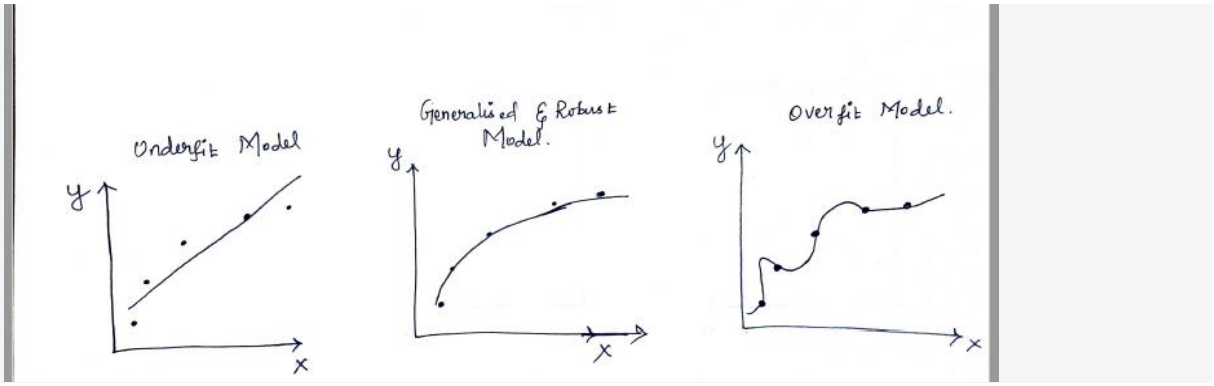
Question 4

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?

A model is said to be robust and generalisable only when model should be able to perform well on the unseen data. The model should not be too complex. If it is complex, then the model will have high variance and low bias. More the model is complex, more the overfitting. If the model has been trained on very less features (simple model) then there will be chances of underfitting (high bias and low variance). In both the cases, the total error will be high. So that our model should compromise bias and variance, in that case, it performs well on the unseen data.

The train and test data accuracy should be more or less the same. There is should not more difference say, train accuracy – 98% and test accuracy – 50% is not acceptable. This shows that the model has been overfitted and is not able to perform well on the unseen data.

In brief, we can say that a model is robust and generalisable only when it performs well on the unseen dataset and the train and test accuracy is more or less equal and the model should not be more complex.


Underfit Model
y
x
Generalised & Robust Model.
y
x
Overfit Model.
y
x