Advanced Regression – Case Study - II

Question-1

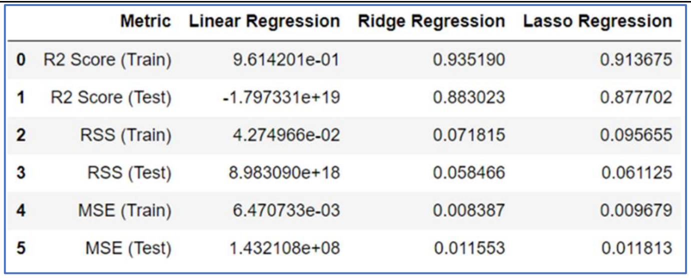
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

The Optimal Value of alpha for the Ridge and Lasso Regression are as follows:

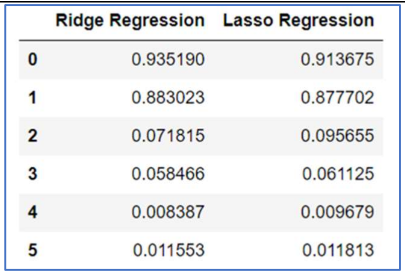
1. Ridge Regression : 10
2. Lasso Regression : 0.0001

Optimal value Ridge – 10, Lasso – 0.0001



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	9.614201e-01	0.935190	0.913675
1	R2 Score (Test)	-1.797331e+19	0.883023	0.877702
2	RSS (Train)	4.274966e-02	0.071815	0.095655
3	RSS (Test)	8.983090e+18	0.058466	0.061125
4	MSE (Train)	6.470733e-03	0.008387	0.009679
5	MSE (Test)	1.432108e+08	0.011553	0.011813

Optimal Value Ridge-20, Lasso-0.0002



	Ridge Regression	Lasso Regression
0	0.935190	0.913675
1	0.883023	0.877702
2	0.071815	0.095655
3	0.058466	0.061125
4	0.008387	0.009679
5	0.011553	0.011813

The Metrics for both Ridge and Lasso have remain same, when we double the optimal values.

With Alpha value doubled, the important predictor variables, after the change is implemented, are

1. CentralAir_Y,
2. GrLivArea,
3. OverallQual_9,
4. Neighborhood_Somerst,
5. TotalBsmtSF,
6. OverallQual_8,
7. Neighborhood_Crawfor,
8. Functional_Typ,
9. Condition1_Norm,
10. SaleCondition_Normal

Please note : The above have been derived after apply the changes in the Case Study Notebook, under Section "Case Study – Part 2" and "Question: 1"

<u>Question-2</u>

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?

<u>Answer:</u>

Prefer to select Lasso Regression.

We have noticed that both Ridge and Lasso Regression are performing well in terms of predicting the SalePrice.

However, Ridge regression does have one obvious disadvantage. It would include all the predictors in the final model. This may not affect the accuracy of the predictions but can make model interpretation challenging when the number of predictors is very large.

The number of feature variables is very large (300+) and the data may have unrelated or noisy variables, we may not want to keep such variables in the model. Lasso regression helps us here by performing feature selection.

<u>Question-3</u>

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?

<u>Answer:</u>

After dropping five most important predictor variables in the lasso model, the five most important predictor variables now are

1. MSZoning_FV
2. 2ndFlrSF
3. 1stFlrSF
4. TotalBsmtSF,
5. Functional_Typ

Please note : The above have been derived after apply the changes in the Case Study Notebook, under Section “Case Study – Part 2” and “Question: 3”

<u>Question-4</u>

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?

<u>Answer:</u>

The model development cycle goes through various stages, starting from data collation, exploration, analysis to model building.

Robust and Generalizable: We can say a model is robust and generalisable, when the model performance is not much affected with any change / variation in the data set provided.

- A generalizable model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.
    - In fact, We have noticed in the case study for Surprise Housing, that removing the top 5 features or doubling the alpha value did not have a substantial impact on the model performance.
- To make sure a model is robust and generalizable, we have to take care it is not too complex and doesn't overfit.
    - Overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will perform well and identify all the patterns of a training data, but fail to demonstrate similar performance with the unseen test data.

Accuracy : When we increase the complexity of a model, the accuracy increases. But we may end up overfitting the model on the training data set and most likely the model accuracy will drop when we run with some un-seen data.

Following are few areas which we need to focus when it comes accuracy:

- Add more data - More the data, better the chances of getting better accuracy.
    - As stated in the case study Part 1 Summary, Surprise Housing can get better accuracy if they include / collate the external factors affecting the Sale Price of the Houses, e.g. economic stability, Housing Loan Interest rates, Mortage availability, geo-political status, employment / un employment statistics..etc.
- Handle the Data : Missing values and Outliers
    - In our case study, we have treated the missing values based on the data definition from data dictionary and have also treated the outliers for the numeric variables.
- Feature Engineering - Understand the data and apply the changes, e.g. create derived columns for variables by grouping them in bins (e.g. we applied the same in our case study and grouped the age of the house based on the year build, year sold, year remodelled...so on.
- Feature Selection - find out the best subset of attributes that better explains the relationship of independent variabl es with the target variable.

In short,

- Building complexity into the model may lead to overfitting and negatively impact the model from being robust and generalizable.
- Accuracy of model can be improved by quality of data, feature selection techniques, feature engineering (handling outliers, missing data...), using regularization techniques.
- Ridge and Lasso regression techniques help in creating balance between model accuracy and complexity.