ADVANCED REGRESSION ASSSIGNMENT

Submitted By:

Abhishek Kumar Goyal (APFE21709647)

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:

Optimal value of alpha for Ridge = 60

Optimal value of alpha for Lasso = 1000

When the values of alpha are doubled for Ridge and Lasso:

For Ridge, alpha = 120, change in metrics and coefficients of predictors are:

Updated metric values:

```
metrics
```
[113] ✓ 0.1s

...

	Metric	Alpha = 60	Alpha = 120
0	R2 Value (Train)	8.700000e-01	8.700000e-01
1	R2 Value (Test)	8.600000e-01	8.600000e-01
2	RSS Value (Train)	6.892160e+11	7.071721e+11
3	RSS Val ue (Test)	3.319774e+11	3.439322e+11
4	MSE Value (Train)	2.618754e+04	2.652647e+04
5	MSE Value (Test)	2.775336e+04	2.824866e+04

Updated beta coefficients:

	Alpha = 1000	Alpha = 2000
LotArea	5699.32	4902.49
BsmtFinSF1	7780.65	7584.50
TotalBsmtSF	8424.77	8542.40
GrLivArea	29785.85	29517.51
GarageArea	7333.45	7651.20
prop_age	-18804.56	-16993.41
OverallQual_10	10801.41	9515.64
OverallQual_8	13060.97	12205.78
OverallQual_9	14229.31	13394.57
OverallCond_6	1178.17	0.00
OverallCond_7	4241.52	2055.78
OverallCond_8	5308.53	3487.99
MasVnrType_BrkFace	0.00	0.00
MasVnrType_None	-0.00	-0.00
MasVnrType_Stone	2508.10	2096.21

For Lasso, alpha = 2000, change in metrics and coefficients of predictors are:

Updated metric values:

	Metric	Alpha = 60	Alpha = 120
0	R2 Value (Train)	8.700000e-01	8.700000e-01
1	R2 Value (Test)	8.700000e-01	8.600000e-01
2	RSS Value (Train)	6.961343e+11	7.071721e+11
3	RSS Val ue (Test)	3.298014e+11	3.439322e+11
4	MSE Value (Train)	2.631864e+04	2.652647e+04
5	MSE Value (Test)	2.766226e+04	2.824866e+04

Updated beta coefficients:

	Alpha = 1000	Alpha = 2000
LotArea	5699.32	4902.49
BsmtFinSF1	7780.65	7584.50
TotalBsmtSF	8424.77	8542.40
GrLivArea	29785.85	29517.51
GarageArea	7333.45	7651.20
prop_age	-18804.56	-16993.41
OverallQual_10	10801.41	9515.64
OverallQual_8	13060.97	12205.78
OverallQual_9	14229.31	13394.57
OverallCond_6	1178.17	0.00
OverallCond_7	4241.52	2055.78
OverallCond_8	5308.53	3487.99
MasVnrType_BrkFace	0.00	0.00
MasVnrType_None	-0.00	-0.00
MasVnrType_Stone	2508.10	2096.21

Overall, there are no major changes in the metrics for both Ridge and Lasso models and the values are comparable as seen in the above screenshots from the 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:

In case of Ridge with alpha = 60, We have obtained a train R2 score of **0.870**, and a test R2 score of **0.864**.

In case of Lasso with alpha = 1000, We have obtained a train R2 score of **0.868**, and a test R2 score of **0.865**.

As we can see that the results from both ridge and lasso regression are almost same. Moreover, the R2 test score for Lasso is slightly better than Ridge and Lasso provides us with feature selection as well without impacting the model accuracy, hence we will use the Lasso regression.

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:

In case of Lasso, the earlier top 5 predictors were:

```
GrLivArea
prop_age
OverallQual_9
OverallQual_8
OverallQual_10
```

However, when we remove these and create another model, the results obtained are:

	Lasso
GarageArea	26131.18
TotalBsmtSF	22362.14
MasVnrType_Stone	10996.09
LotArea	9735.42
BsmtFinSF1	7001.72
MasVnrType_BrkFace	6324.84
OverallCond_8	1150.46
OverallCond_7	-0.00
MasVnrType_None	-0.00
OverallCond_6	-3650.26

So, now the top 5 predictors are:

GarageArea, TotalBsmtSF, MasVnrType_Stone, LotArea, BsmtFinSF1

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:

A model is said to be robust when in general it is stable i.e it does not change significantly with minor changes in the training data points

A model is considered to be generalisable if it doesn't overfit or if it is able to perform well on unseen data.

Implications on accuracy:

We tend to make the model both robust and generalisable so that it is able to perform equally well on training as well as test data but at the same time, we also need to consider the bias and variance of the model. The bias will increase slightly

with the model becoming more generalizable but that is to be maintained such that it does not impact the model too much. This is also known as the bias variance trade-off.