Advance Regression Assignment – Part 2

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:

1.1 Optimal value for alpha for Ridge and Lasso regression is '**2.0**' & '**0.0001**' respectively.

1.2 Changes in the model after using double the alpha:

Out[99]:

	Metric	Linear Regression	Ridge Regression Model-1	Ridge Regression Model-2	Lasso Regression Model-1	Lasso Regression Model-2
0	Predictor Count	2.000000e+02	198.000000	198.000000	124.000000	94.000000
1	R2 (Train)	9.200000e-01	0.920000	0.920000	0.920000	0.910000
2	R2 (Test)	-4.619597e+22	0.880000	0.880000	0.890000	0.880000
3	RSS (Train)	1.351018e+00	1.357138	1.436759	1.416859	1.584495
4	RSS (Test)	3.536647e+23	0.925929	0.935103	0.876722	0.890596
5	MSE (Train)	1.323200e-03	0.001329	0.001407	0.001388	0.001552
6	MSE (Test)	8.056144e+20	0.002109	0.002130	0.001997	0.002029

R2 values sightly decreased and RSS & MSE values increased after the change for both – Ridge and Lasso. Variance is reduced sightly but bias is increased as R2 value decreased.

1.3 The most important predictor variables after implementing the change:

The most important variable across is ‘GrLivArea’ with different coefficient values.

Top 5 predictor variables view before and after the change as below please -

	Ridge M1	Ridge M2	Lasso M1	Lasso M2
GrLivArea	0.096383	0.079792	0.310363	0.304673
OverallQual	0.127016	0.113251	0.16123	0.179266
OverallCond	0.079508	0.070111	0.094591	0.091013
GarageCars			0.078786	0.079679
MSZoning_RL			0.057047	
1stFlrSF	0.084711	0.068035		
2ndFlrSF	0.063604	0.057339		
BsmtFullBath				0.051339
TotRmsAbvGrd				
Neighborhood_Crawfor				
GarageArea				

Note: M1 for before and M2 for after the change

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:

Considering below facts, I will consider Lasso regression as it is simpler than Ridge and metrics are descent:

1) Simplicity: Lasso Model is with 124 features whereas Ridge Regression is with 198. Lasso is simpler than Ridge here as Lasso supports feature elimination

2) Metrics: R2/RSS/MSE scores are descent for both the models; difference between scores of Train and Test is also improved in case of Lasso.

Note: Below screenshots as supporting details from attached Jupyter Notebook

Out[81]:

	Metric	Linear Regression	Ridge Regression Model-1	Ridge Regression Model-2	Lasso Regression Model-1	Lasso Regression Model-2
0	Predictor Count	2.000000e+02	198.000000	198.000000	124.000000	94.000000
1	R2 (Train)	9.204189e-01	0.920059	0.915368	0.916541	0.906666
2	R2 (Test)	-4.619597e+22	0.879054	0.877856	0.885482	0.883670
3	RSS (Train)	1.351018e+00	1.357138	1.436759	1.416859	1.584495
4	RSS (Test)	3.536647e+23	0.925929	0.935103	0.876722	0.890596
5	MSE (Train)	1.323200e-03	0.001329	0.001407	0.001388	0.001552
6	MSE (Test)	8.056144e+20	0.002109	0.002130	0.001997	0.002029

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:

Below are the five most important predictors along with their significance (coefficient value) after removing earlier top five predictors & rebuilding Lasso model

1) 1stFlrSF (0.265459)

2) 2ndFlrSF (0.132028)

3) GarageArea (0.080873)

4) TotRmsAbvGrd (0.058408)

5) Neighborhood_Crawfor (0.049208)

Question-4:

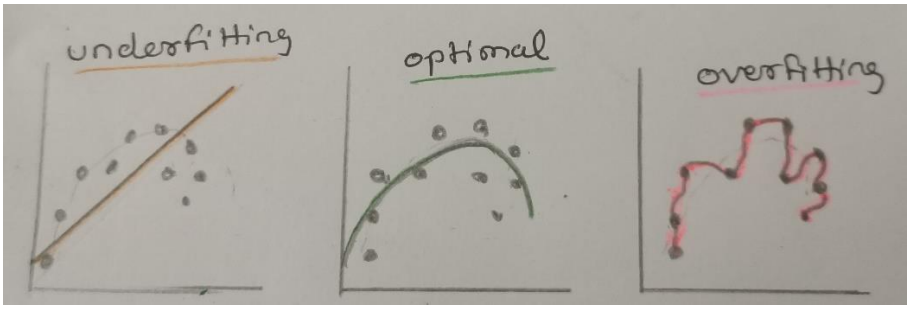
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

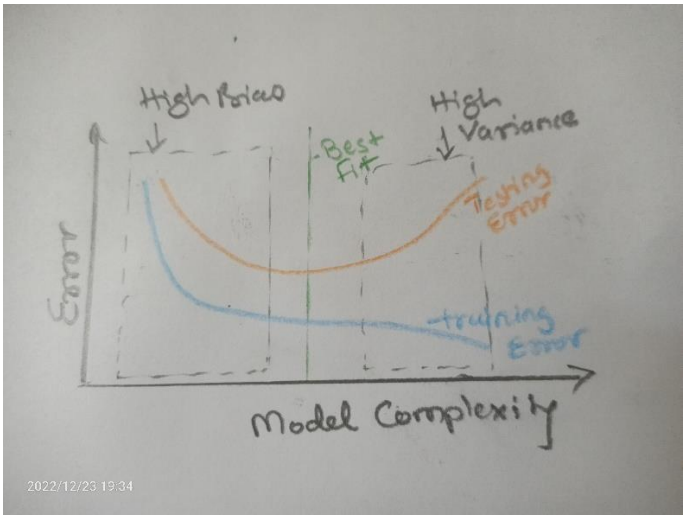
Answer:

We can make a model **robust** by reducing '**Bias'** and make it generalizable by reducing '**Variance'**.

Since bias and variance are inversely proportionate to each other; if we try to get low bias then 'underfitting' problem and if we try to get low variance then 'overfitting' problem as shown in below figures.

To solve these problems, regularization is done to identify 'Best Fit' for optimal accurate model which is robust and generalizable.





During the exercise, focused on getting good Train and Test R2 values along with both values very close. Also identified optimal lambda (alpha) value and addressed overfitting problem by regularization using Ridge and Lasso regression.
