Advanced Regression – Subjective Question Answers.

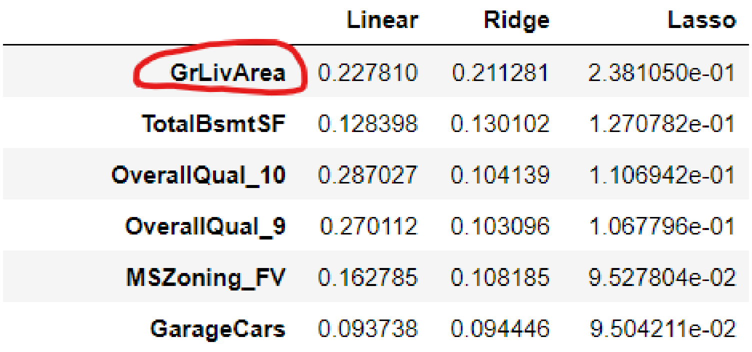
Submitted By – Shikhar Kushwaha

Q.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 Most Optimal Value from my model prediction for Ridge Regression = 1.0 & Lasso Regression = 0.0004
- If we try to change the model by doubling the value of Alpha for both Ridge and Lasso we observe that there is a very slight chance in R2 Score for test and train. Almost the value remains the same.

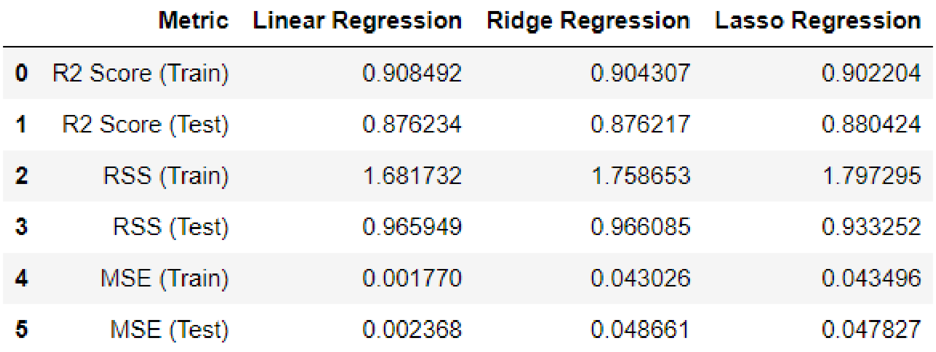
Most important predictor:
GrLivArea, TotalBsmtSF, OverallQual, GarageCars, MSZoning_FV, BsmtFinSF1



	Linear	Ridge	Lasso
GrLivArea	0.227810	0.211281	2.381050e-01
TotalBsmtSF	0.128398	0.130102	1.270782e-01
OverallQual_10	0.287027	0.104139	1.106942e-01
OverallQual_9	0.270112	0.103096	1.067796e-01
MSZoning_FV	0.162785	0.108185	9.527804e-02
GarageCars	0.093738	0.094446	9.504211e-02

Q2. 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?

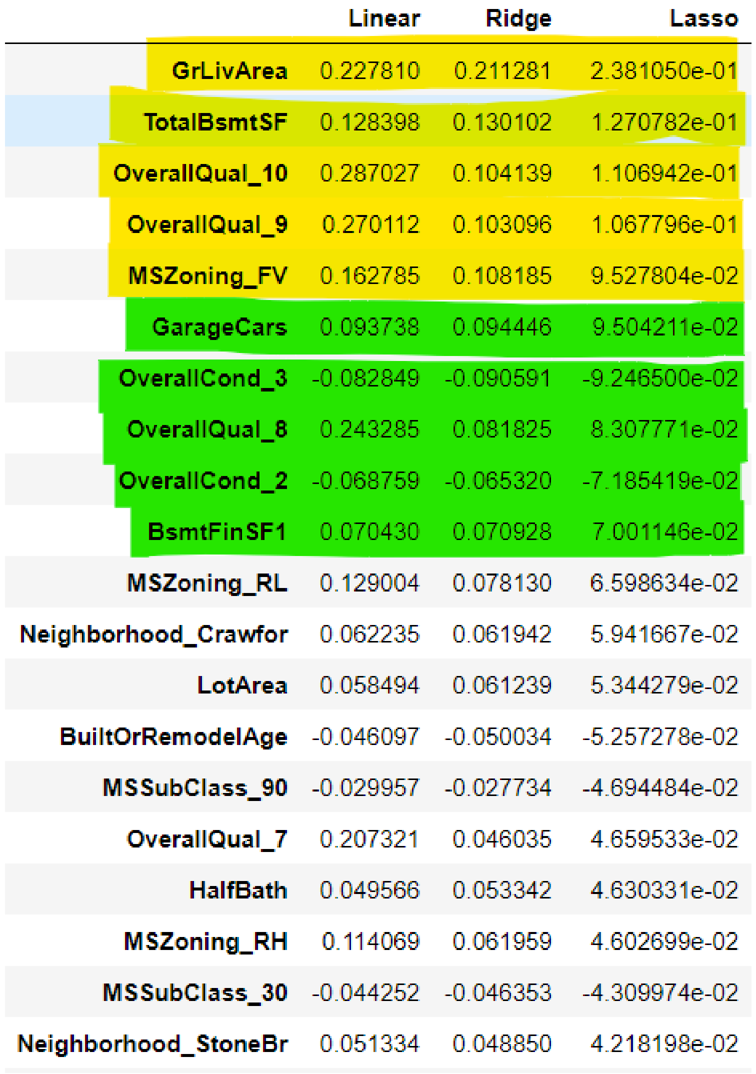
Answer:



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.908492	0.904307	0.902204
1	R2 Score (Test)	0.876234	0.876217	0.880424
2	RSS (Train)	1.681732	1.758653	1.797295
3	RSS (Test)	0.965949	0.966085	0.933252
4	MSE (Train)	0.001770	0.043026	0.043496
5	MSE (Test)	0.002368	0.048661	0.047827

- From the above results we conclude that Lasso Performs slightly better the Ridge.(On Test set ie. Unseen data)
- Lasso Regression remove unnecessary features. Which is help in better feature detection.

Q3. 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?



	Linear	Ridge	Lasso
GrLivArea	0.227810	0.211281	2.381050e-01
TotalBsmtSF	0.128398	0.130102	1.270782e-01
OverallQual_10	0.287027	0.104139	1.106942e-01
OverallQual_9	0.270112	0.103096	1.067796e-01
MSZoning_FV	0.162785	0.108185	9.527804e-02
GarageCars	0.093738	0.094446	9.504211e-02
OverallCond_3	-0.082849	-0.090591	-9.246500e-02
OverallQual_8	0.243285	0.081825	8.307771e-02
OverallCond_2	-0.068759	-0.065320	-7.185419e-02
BsmtFinSF1	0.070430	0.070928	7.001146e-02
MSZoning_RL	0.129004	0.078130	6.598634e-02
Neighborhood_Crawfor	0.062235	0.061942	5.941667e-02
LotArea	0.058494	0.061239	5.344279e-02
BuiltOrRemodelAge	-0.046097	-0.050034	-5.257278e-02
MSSubClass_90	-0.029957	-0.027734	-4.694484e-02
OverallQual_7	0.207321	0.046035	4.659533e-02
HalfBath	0.049566	0.053342	4.630331e-02
MSZoning_RH	0.114069	0.061959	4.602699e-02
MSSubClass_30	-0.044252	-0.046353	-4.309974e-02
Neighborhood_StoneBr	0.051334	0.048850	4.218198e-02

From the Picture above:

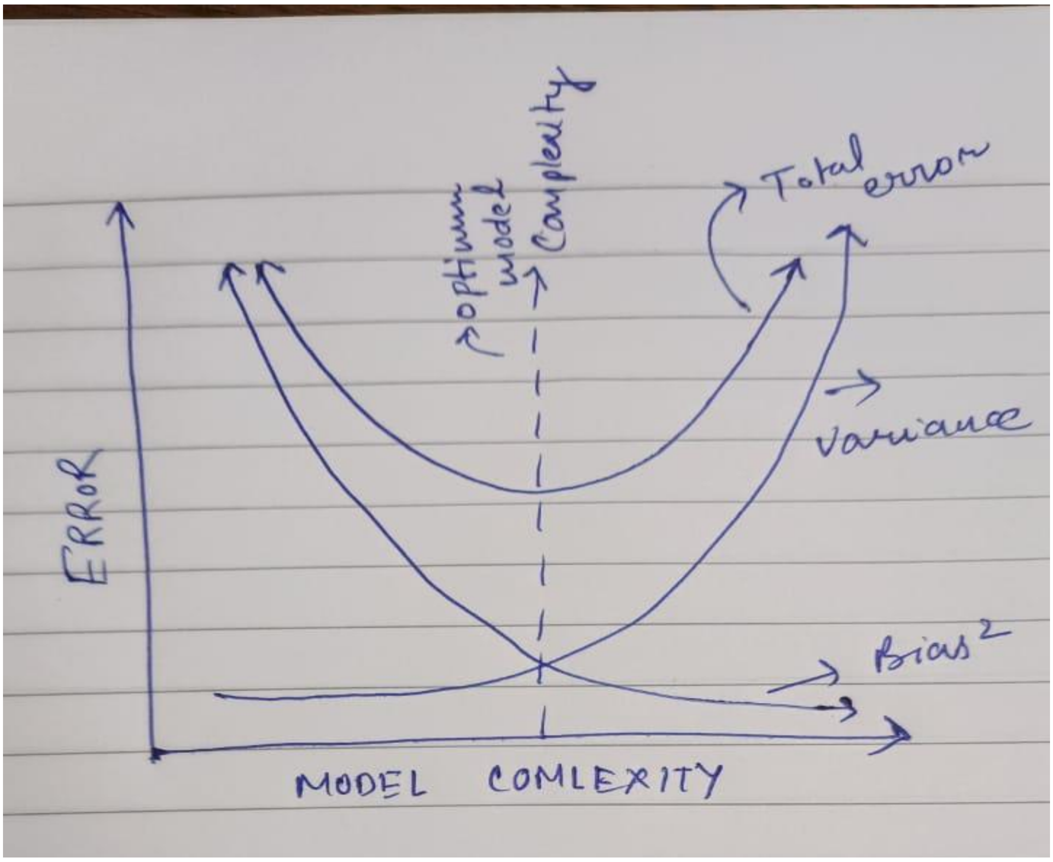
Yellow color is the current five most important predictor variables.

Green color is the another five most important predictor variables.

Q4. 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?

By minimising the Total error which is created between variance and bias trade off, we can make sure our model is robust and generalised.

To make sure our model is accurate we need to do a proper Exploratory Data Analysis.


optimum model Complexity
Total error
Variance
Bias^2
ERROR
MODEL COMLEXITY