Assignment Part-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: -

Optimal value of alpha for ridge and lasso regression:

- Ridge: 3
- Lasso: 0.0001

- Doubling alpha value for ridge and lasso regression:
  - Ridge: 6
  - Lasso: 0.0002

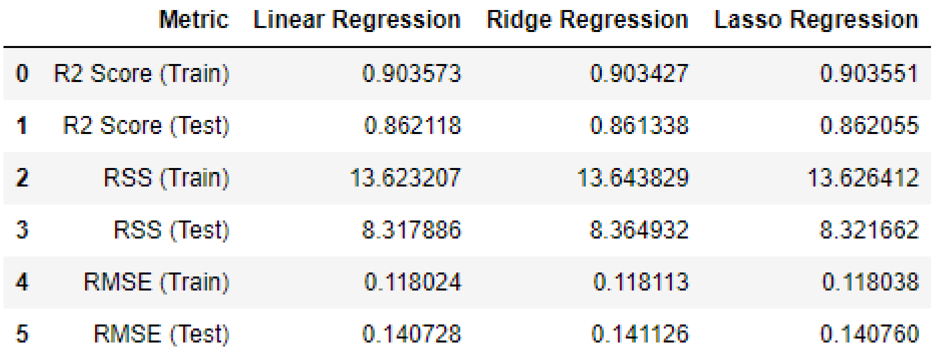
The most important predictor variables after the change is implemented:

1. TotalBsmtSF	:	Total square feet of basement area
2. MSZoning_RL	:	Zoning-Residential Low Density
3. 2ndFlrSF	:	Second floor square feet
4. MSZoning_RM	:	Zoning-Residential Medium Density
5. MSZoning_FV	:	Zoning-Floating Village Residential
6. OverallQual_8	:	Very Good overall material and finish of the house
7. OverallCond_7	:	Good overall condition of the house
8. SaleCondition_Partial	:	Home was not completed when last assessed (associated with New Homes)
9. MSZoning_RH	:	Zoning-Residential High Density
10. OverallQual_7	:	Good overall material and finish of the house

Inferences:

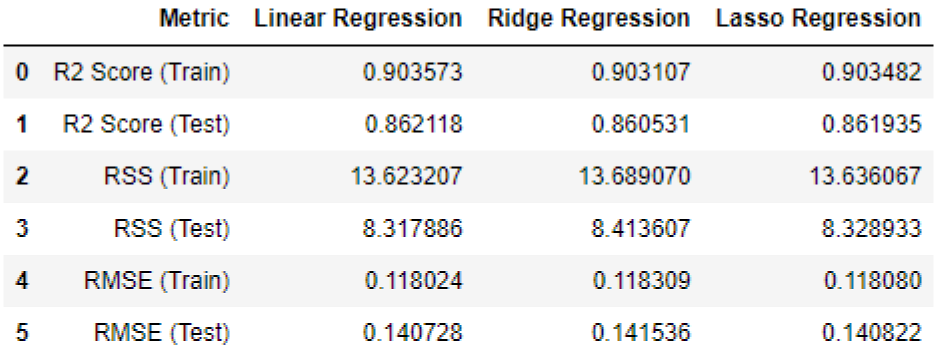
- The ordering of top 10 variable changes a bit.
- R2 score for both train and test data decreases slightly.
- RSS and RMSE for both train and test data increases slightly

Metrics values before doubling alpha/lambda:



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.903573	0.903427	0.903551
1	R2 Score (Test)	0.862118	0.861338	0.862055
2	RSS (Train)	13.623207	13.643829	13.626412
3	RSS (Test)	8.317886	8.364932	8.321662
4	RMSE (Train)	0.118024	0.118113	0.118038
5	RMSE (Test)	0.140728	0.141126	0.140760

Metrics values after doubling alpha/lambda:



	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.903573	0.903107	0.903482
1	R2 Score (Test)	0.862118	0.860531	0.861935
2	RSS (Train)	13.623207	13.689070	13.636067
3	RSS (Test)	8.317886	8.413607	8.328933
4	RMSE (Train)	0.118024	0.118309	0.118080
5	RMSE (Test)	0.140728	0.141536	0.140822

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: -

- Both feature coefficient and other matrices (R2 score, RSS, RMSE) are almost similar in both Ridge and Lasso Regression.
- Broadly we know that Lasso Regression penalizes the co-efficient to zero (helps in feature elimination) and avoids overfitting of the model. By keeping this scenario in mind, its better I go with Lasso Regression for final model building.

Question 3

After building the model, you realized 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: -

Removing the top 5 contributors

- Top 5 features identified when optimal values of alpha or lambda (0.0001) for Lasso Regression is taken:
- 2ndFlrSF, TotalBsmtSF, MSZoning_RL, MSZoning_RM, MSZoning_FV

After dropping previous top 5 features and rebuilding the model.
New top five features identified as below:

FullBath_2	: Full bathrooms above grade 2
LotFrontage	: Linear feet of street connected to property
GarageArea	: Size of garage in square feet
FullBath_3	: Full bathrooms above grade 3
OverallQual_8	: Very Good overall material and finish of the house.

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: -

The conditions which makes sure that model is robust and generalizable are:

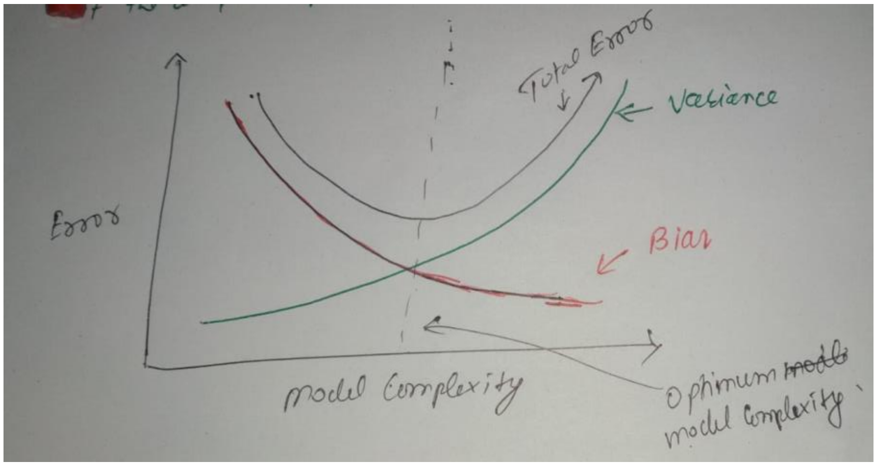
- Model should be immune to outliers.
- Model shouldn’t change significantly if training data points undergo small change.
- Model should be simple because simpler models though makes many errors during training but it is bound to outperform complex models when given new data to process.
- Need to have good balance between bias and variance.
- Difference between r2 score of training and test data shouldn’t be very high (~<5%).
- Remove irrelevant outliers.

Regularization is one of the technique to make model robust and generalizable.

Example: - Tree based models like Random Forest is immune to outliers.

Implications of the same for the accuracy of the model:

- r2 score during training will be very close to 1 which may lead to overfitting. If this happens model will memorize the data instead of intelligently learning. This may lead to bad performance over test data (r2 score will drop drastically over test data) and makes the model less robust and ungeneralizable.


Total Error
Variance
Error
Bias
model Complexity
Optimum ~~model~~
model Complexity