Advanced Regression Assignment

Subjective Answers

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:
Based on the Model execution completed, my model gave the below Best params.

Best alpha value for Lasso: {'alpha': 0.01}

Best alpha value for Ridge: {'alpha': 5.0}

Based on the Doubling factor of the alpha value the score remains same in both Lasso and Ridge whereas, there is some change in the coefficient values (Refer Jupyter for the model interpretation).

Lasso Regression

	Featuere	Coef
37	Exterior1st_BrkComm	0.749964
20	Neighborhood_OldTown	0.531438
19	Neighborhood_NridgHt	0.519386
23	Condition1_Norm	0.518469
50	Exterior2nd_Plywood	0.513832
10	KitchenQual	0.396116
32	RoofMatl_Metal	0.299870
31	RoofMatl_CompShg	0.266980
1	ExterQual	0.244111
12	LotConfig_CulDSac	0.214542

Lasso Regression

	Featuere	double Coef
19	Neighborhood_NridgHt	0.487921
20	Neighborhood_OldTown	0.477383
23	Condition1_Norm	0.451287
10	KitchenQual	0.388175
37	Exterior1st_BrkComm	0.310899
50	Exterior2nd_Plywood	0.268132
31	RoofMatl_CompShg	0.250940
1	ExterQual	0.249108
12	LotConfig_CulDSac	0.201480
13	LotConfig_FR3	0.165744

Ridge Regression

	Feaure	Coef
20	Neighborhood_OldTown	0.475481
19	Neighborhood_NridgHt	0.466185
23	Condition1_Norm	0.438101
37	Exterior1st_BrkComm	0.417249
50	Exterior2nd_Plywood	0.362958
31	RoofMatl_CompShg	0.276984
1	ExterQual	0.241923
10	KitchenQual	0.232892
32	RoofMatl_Metal	0.230210
12	LotConfig_CulDSac	0.209013

Ridge Regression

	Feaure	Double Coef
20	Neighborhood_OldTown	0.400603
19	Neighborhood_NridgHt	0.396169
23	Condition1_Norm	0.342825
31	RoofMatl_CompShg	0.260969
1	ExterQual	0.247354
37	Exterior1st_BrkComm	0.238829
10	KitchenQual	0.231171
50	Exterior2nd_Plywood	0.226202
12	LotConfig_CulDSac	0.190747
13	LotConfig_FR3	0.176700

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:

Best Parameters are listed below

Best alpha value for Lasso: {'alpha': 0.01}

Best alpha value for Ridge: {'alpha': 5.0}

R2Score for Lasso Training: 0.8540079972624912

R2Score for Lasso Test: 0.8687526160111434

R2Score for Ridge Training: 0.894436501914521

R2Score for Ridge Test: 0.7517090642428316

Since Lasso has a better R2 score and considers Feature reduction we can see that Lasso can give a better performance than Ridge.

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:

As per the model (Lasso Regression with best Parameter of 0.001) we have the below best predictors (Features)

1. 'Exterior1st_BrkComm',
2. 'Neighborhood_OldTown',
3. 'Neighborhood_NridgHt',
4. 'Condition1_Norm'
5. 'Exterior2nd_Plywood'

As we build a Lasso model in the Jupyter notebook after removing these attributes from the dataset.

R2 of the new model without the top 5 predictors drops to 0.83

The other different top 5 features are listed below.

```
Lasso Regression
```

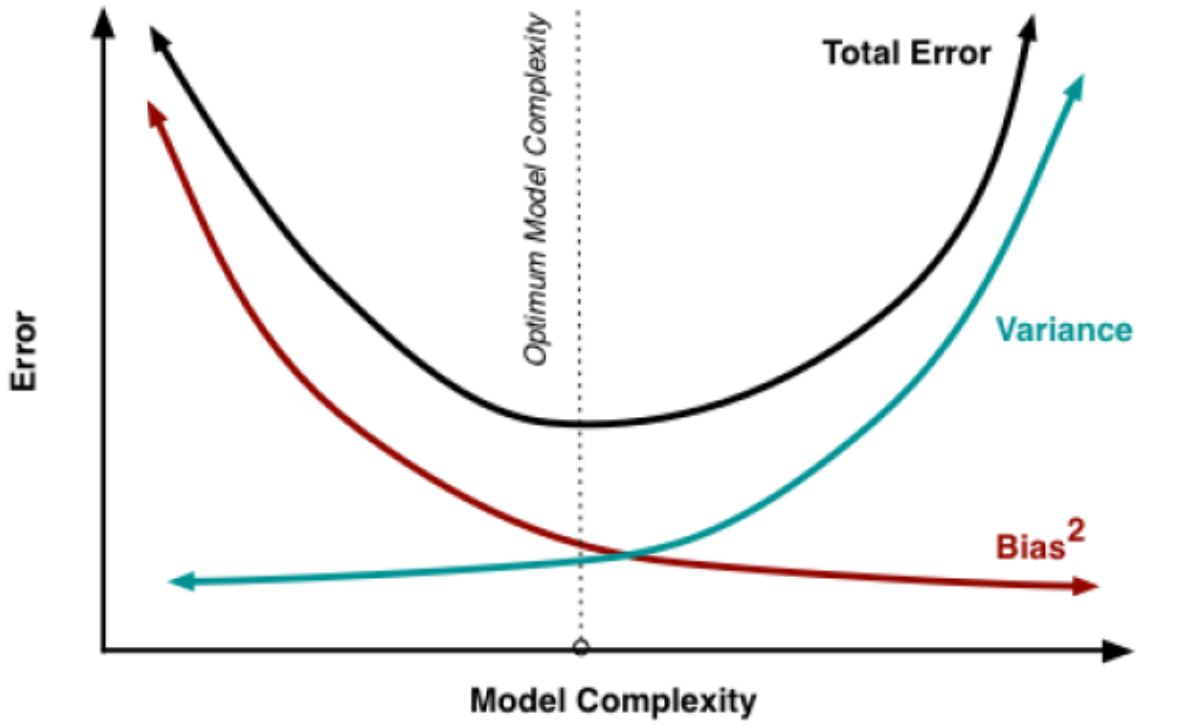
	Feaure	Coef
20	Neighborhood_StoneBr	0.400603
19	Neighborhood_SWISU	0.396169
23	BldgType_Duplex	0.342825
31	RoofMatl_Tar&Grv	0.260969
1	ExterQual	0.247354

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:

Simple models are more generalizable as they can be scalable and can be widely applicable. Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.

Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple and not making it too naive to be of any use.



Above model complexity vs Error diagram shows the exact utilization of Regularization. So, the Bias and Variance tend to move opposite each other when there is some over training or under training, So the regularization can help us keep this in minimum point.

Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve, for e.g., one that gives same answer to all test inputs and makes no discrimination whatsoever has a very large bias as its expected error across all test inputs are very high.

Variance refers to the degree of changes in the model itself with respect to changes in the training data.