Assignment Part-II Subjective Questions

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 :

As per the model Built Optimal value of Alpha for

Ridge : 0.6 and Lasso : 0.0002

These alpha value given r-square 83.44 for Ridge and 84.00 for lasso

For alpha value Ridge - 0.6 and Lasso - 0.0002 different metrics parameter

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.904228	0.901216	0.898664
1	R2 Score (Test)	0.810512	0.834496	0.840091
2	RSS (Train)	13.023002	13.432522	13.779596
3	RSS (Test)	10.507662	9.177665	8.867393
4	MSE (Train)	0.119694	0.121562	0.123122
5	MSE (Test)	0.152978	0.142969	0.140532

After doubling the alpha value of ridge model for o.6 to 1.2 and lasso 0.0002 to 0.0004

The value of R2 slightly changes for the both the model. And other metrics also slightly changes.

	Metric	Linear Regression	Ridge Regression	Ridge alpha double	Lasso Regression	Lasso alpha double
0	R2 Score (Train)	0.904228	0.901216	0.897720	0.898664	0.891861
1	R2 Score (Test)	0.810512	0.834496	0.840709	0.840091	0.845336
2	RSS (Train)	13.023002	13.432522	13.907926	13.779596	14.704597
3	RSS (Test)	10.507662	9.177665	8.833133	8.867393	8.576586
4	MSE (Train)	0.119694	0.121562	0.123694	0.123122	0.127188
5	MSE (Test)	0.152978	0.142969	0.140260	0.140532	0.138208

Coefficient after doubling alpha

For Ridge Regression

	Variable	Ridge Coeff
6	TotalBsmtSF	0.502
14	FullBath	0.484
16	BedroomAbvGr	0.397
4	BsmtFinSF2	0.392
15	HalfBath	0.375
17	KitchenAbvGr	0.353
5	BsmtUnfSF	0.293
3	BsmtFinSF1	0.276
8	1stFlrSF	0.245
25	HouseAge	0.241

	Variable	Ridge double Coeff
6	TotalBsmtSF	0.520
14	FullBath	0.397
4	BsmtFinSF2	0.360
16	BedroomAbvGr	0.315
15	HalfBath	0.282
5	BsmtUnfSF	0.277
3	BsmtFinSF1	0.270
17	KitchenAbvGr	0.267
8	1stFlrSF	0.250
25	HouseAge	0.222

For Lasso Regression

	Variable	Coeff
6	TotalBsmtSF	0.604
14	FullBath	0.517
3	BsmtFinSF1	0.467
16	BedroomAbvGr	0.429
15	HalfBath	0.411
17	KitchenAbvGr	0.385
4	BsmtFinSF2	0.307
8	1stFlrSF	0.246
25	HouseAge	0.236
5	BsmtUnfSF	0.225

	Variable	Lasso double Coeff
6	TotalBsmtSF	0.862
3	BsmtFinSF1	0.467
14	FullBath	0.391
16	BedroomAbvGr	0.310
15	HalfBath	0.281
17	KitchenAbvGr	0.263
8	1stFlrSF	0.250
25	HouseAge	0.207
27	GarageAge	0.163
24	MiscVal	0.156

As we can see from above Tables there are Slightly Changes in the predicted features and their values also there order slightly changes due to change in coefficient value . these is not much significant the change in alpha value after doubling of alpha is very small

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-2:-

Model metrics at optimum value

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.904228	0.901216	0.898664
1	R2 Score (Test)	0.810512	0.834496	0.840091
2	RSS (Train)	13.023002	13.432522	13.779596
3	RSS (Test)	10.507662	9.177665	8.867393
4	MSE (Train)	0.119694	0.121562	0.123122
5	MSE (Test)	0.152978	0.142969	0.140532

Both model has given decent value of R-squarred on test data but although Ridge regression is better in r-squarred value of train set,

- But because of features selection property of Lasso its bring and assign a zero value to insignificant features we can choose Lasso Regression to predict Sale Price of Housing Model.
- we should always use simple yet robust model.

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:- Top five most predictors for lasso model are

Variable	Coeff
TotalBsmtSF	0.604
FullBath	0.517
BsmtFinSF1	0.467
BedroomAbvGr	0.429
HalfBath	0.411

After dropping these top five features from dataset R2 value increases and RSS and MSE value for decreases for both train and test set.

The new Top 5 predictors are now:

Variable	Lasso Q3 Coeff
GrLivArea	0.499
1stFlrSF	0.381
MSZoning_FV	0.287
MSZoning_RH	0.261
MSZoning_RL	0.257

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:

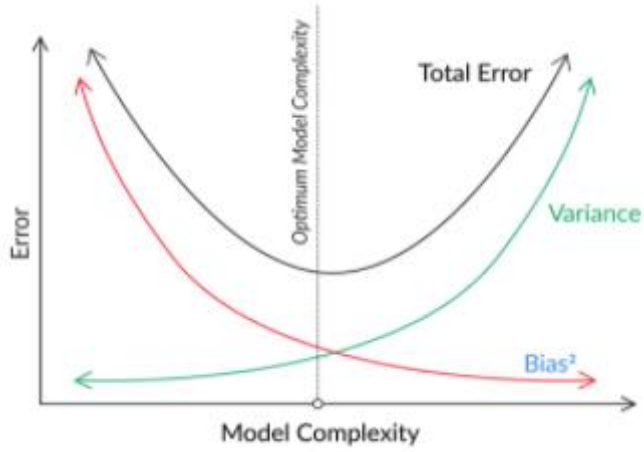
According to Occam Razor – Given two model which show similar in performance in any finite training or test data we should always pick one which makes fewer on the test data due to following reasons.

- ✓ Simpler model are more robust and generic than complex models. and more widely applicable. This becomes important because generic model perform better on unseen dataset.
- ✓ Simpler model requires fewer training point for effective training than that of complex model, which makes simple model easier to train
- ✓ Complex models tends to change widely when any upcoming changes in training dataset.
- ✓ Simple models are more robust and does not change significantly if training datapoint get small changes

Bias Variance Trade off :-

- ✓ Simple model have low variance high bias while complex models have low bias and high variance.

- ✓ Simpler model makes more error in the training set while complex models lead towards overfitting. It works very well for training set but fail miserly when applied to test sample.
- ✓ A perfect model should not have either high bias or high variance, we should find a point where both bias and variance will be minimum. Such model will be more robust and generalised



For achieving this

- ✓ **Transform your data**: if you have data with skewed means some noise present it can be transform using log transformation to remove noise.
- ✓ **Use robust error metric:** switching from mean squared error to mean absolute error difference reduce the influence of outliers
- ✓ Remove the outliers this works if there are few of them and you are fairly certain they are anomalies and not worth predicting.