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?

The optimal value of alpha for Ridge = 300

The optimal value of alpha for Lasso = 0.001

Model Accuracy Metrics with optimal Lambda:

[3389]:

	Metric	Linear RegressionRFE	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.876013	0.912543	0.920007
1	R2 Score (Test)	0.868990	0.885811	0.889359
2	RSS (Train)	23.387896	16.497107	15.089313
3	RSS (Test)	5.711987	4.978566	4.823901
4	MSE (Train)	0.141627	0.118947	0.113759
5	MSE (Test)	0.139863	0.130575	0.128531

Comparison with alpha and doubled alpha

	Metric	Linear RegressionRFE	Ridge Regression	Lasso Regression	Ridge Regression Alpha Double	Lasso Regression Alpha Double
0	R2 Score (Train)	0.876013	0.912543	0.920007	0.903510	0.915417
1	R2 Score (Test)	0.868990	0.885811	0.889359	0.882105	0.893029
2	RSS (Train)	23.387896	16.497107	15.089313	18.201041	15.955064
3	RSS (Test)	5.711987	4.978566	4.823901	5.140182	4.663865
4	MSE (Train)	0.141627	0.118947	0.113759	0.124939	0.116977
5	MSE (Test)	0.139863	0.130575	0.128531	0.132678	0.126381

In [ ]:

Observations:

R2 score and test accuracy on Alpha is slightly higher than 2*Alpha.

MSE increases with 2*Alpha as compared to Alpha

Both Ridge and Lasso regression seem to perform better on Alpha than 2*Alpha

The most important variables with Alpha for Ridge regression as compared to 2* Alpha:

ut[3380]:

	Features	Coefficient	Abs_Coefficient_Ridge(Desc_Sort)
0	GrLivArea	0.0330	0.0330
1	OverallQual_Excellent	0.0291	0.0291
2	GarageCars	0.0274	0.0274
3	FullBath	0.0257	0.0257
4	TotRmsAbvGrd	0.0252	0.0252
5	1stFlrSF	0.0250	0.0250
6	Neighborhood_NridgHt	0.0238	0.0238
7	OverallQual_Very Good	0.0233	0.0233
8	HalfBath	0.0198	0.0198
9	Neighborhood_Crawfor	0.0186	0.0186

Out[3392]:

	Features	Coefficient	Abs_Coefficient_Ridge(Desc_Sort)
0	GrLivArea	0.0291	0.0291
1	OverallQual_Excellent	0.0259	0.0259
2	GarageCars	0.0242	0.0242
3	TotRmsAbvGrd	0.0239	0.0239
4	1stFlrSF	0.0232	0.0232
5	FullBath	0.0219	0.0219
6	Neighborhood_NridgHt	0.0205	0.0205
7	OverallQual_Very Good	0.0196	0.0196
8	GarageArea	0.0193	0.0193
9	TotalBsmtSF	0.0167	0.0167

The most important variables with Alpha for Lasso regression as compared to 2* Alpha:

Out[3388]:

	Features	Coefficient	Abs_Coefficient_Lasso(Desc_Sort)
0	GrLivArea	0.0913	0.0913
1	GarageCars	0.0461	0.0461
2	OverallQual_Excellent	0.0353	0.0353
3	OverallQual_Very Good	0.0330	0.0330
4	SaleType_New	0.0326	0.0326
5	FullBath	0.0316	0.0316
6	Neighborhood_NridgHt	0.0315	0.0315
7	MSZoning_RL	0.0308	0.0308
8	SaleCondition_Normal	0.0306	0.0306
9	AgeofProperty	-0.0285	0.0285

Out[3396]:

	Features	Coefficient	Abs_Coefficient_Lasso(Desc_Sort)
0	GrLivArea	0.0872	0.0872
1	GarageCars	0.0466	0.0466
2	OverallQual_Excellent	0.0385	0.0385
3	OverallQual_Very Good	0.0343	0.0343
4	Neighborhood_NridgHt	0.0311	0.0311
5	FullBath	0.0298	0.0298
6	SaleType_New	0.0288	0.0288
7	AgeofProperty	-0.0256	0.0256
8	SaleCondition_Normal	0.0255	0.0255
9	Neighborhood_Crawfor	0.0255	0.0255

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?

[3389]:

	Metric	Linear RegressionRFE	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.876013	0.912543	0.920007
1	R2 Score (Test)	0.868990	0.885811	0.889359
2	RSS (Train)	23.387896	16.497107	15.089313
3	RSS (Test)	5.711987	4.978566	4.823901
4	MSE (Train)	0.141627	0.118947	0.113759
5	MSE (Test)	0.139863	0.130575	0.128531

As per the above table R2 score on both train and test data is slightly better in Lasso Regression than Ridge regression.

The MSE is also slight lower in Lasso regression than Ridge regression.

Given the number of features is huge, it is better to reduce coefficients to zero and opt for Lasso regression.

Choosing Lasso may be less prone to risk as it reduces the features and complexity of 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?

Out[3388]:

	Features	Coefficient	Abs_Coefficient_Lasso(Desc_Sort)
0	GrLivArea	0.0913	0.0913
1	GarageCars	0.0461	0.0461
2	OverallQual_Excellent	0.0353	0.0353
3	OverallQual_Very Good	0.0330	0.0330
4	SaleType_New	0.0326	0.0326
5	FullBath	0.0316	0.0316
6	Neighborhood_NridgHt	0.0315	0.0315
7	MSZoning_RL	0.0308	0.0308
8	SaleCondition_Normal	0.0306	0.0306
9	AgeofProperty	-0.0285	0.0285

Top predictors before were: GrLivArea, GarageCars, OverallQual_Excellent, OverallQual_Very Good and SaleType_New

If these are not available, we will need to perform Lasso regression again on remaining data and get the top predictors.

New predictors:

1stFlrSF

2ndFlrSF

KitchenQual_TA

GarageArea

Neighborhood_NridgHt

3410]:

	Features	Coefficient	Abs_Coefficient_Lasso(Desc_Sort)
0	1stFlrSF	0.0721	0.0721
1	2ndFlrSF	0.0630	0.0630
2	KitchenQual_TA	-0.0469	0.0469
3	GarageArea	0.0384	0.0384
4	Neighborhood_NridgHt	0.0374	0.0374
5	FullBath	0.0361	0.0361
6	TotRmsAbvGrd	0.0351	0.0351
7	MSZoning_RL	0.0327	0.0327
8	AgeofProperty	-0.0323	0.0323
9	KitchenQual_Gd	-0.0314	0.0314

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?

If the testing error is consistent with the training error and there isn't a huge difference, it means the model is robust. Thus, if a model can be applied on unseen data and can perform well, it is said to be robust.

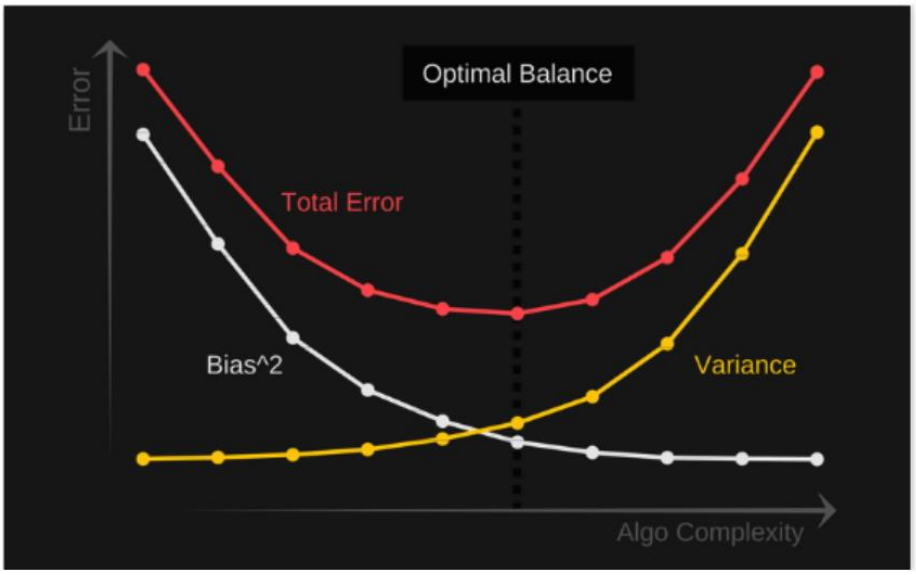
One of the most common problems every Data Science practitioner face is Overfitting- the situation where machine learning model performs exceptionally well on the train data but was not able to predict on the unseen data.

Avoiding overfitting can single-handedly improve our model's performance.

Regularization helps in overcoming the problem of overfitting and increases the model interpretability.

Regularization works by adding a penalty or complexity term or shrinkage term with Residual Sum of Squares (RSS) to the complex model.

Ideally, one should choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data.



TO get good predictions, we need to find a balance of Bias and Variance.

In other words, bias in a model is high when it does not perform well on the training data itself, and variance is high when the model does not perform well on the test data

We can use Ridge and Lasso regression, which both allow some bias to get a significant decrease in variance, thereby pushing the model coefficients towards 0.