Advanced Regression 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:

Optimal values for alpha are as follows:

- Lasso: 0.0001
- Ridge: 5.0

The alpha parameter is part of the penalty parameter which helps add error to the best fit line so that the line does not overfit. If alpha value is too small or zero, the results will be similar to that of Linear Regression and the model can underfit. If the alpha is too high, model can overfit.

If we double the values of alpha for the two methods, the model fit will start tending to overfitting.

Here, the change in lambda did not make any change in the actual predictor variables; however, the coefficients start moving further towards zero.

RIDGE REGRESSION

Actual alpha

	Features	Coefficients
7	GrLivArea	0.265310
4	TotalBsmtSF	0.244413
6	2ndFlrSF	0.230877
5	1stFlrSF	0.209258
19	Street_Pave	0.137111
14	houseAge	-0.276526

Double alpha

	Features	Coefficients
7	GrLivArea	0.255839
4	TotalBsmtSF	0.222360
6	2ndFlrSF	0.211298
5	1stFlrSF	0.203779
25	OverallQual_8	0.126627
14	houseAge	-0.246457

LASSO REGRESSION

Actual Alpha

	Features	Coefficients
7	GrLivArea	0.265310
4	TotalBsmtSF	0.244413
6	2ndFlrSF	0.230877
5	1stFlrSF	0.209258
14	houseAge	-0.276526

Double Alpha

	Features	Coefficients
7	GrLivArea	0.255839
4	TotalBsmtSF	0.222360
6	2ndFlrSF	0.211298
5	1stFlrSF	0.203779
14	houseAge	-0.246457

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:

	Metric	Lasso Regression	Ridge Regression
0	R2 Score (Train)	0.881727	0.881323
1	R2 Score (Test)	0.854286	0.853947
2	RSS (Train)	19.528685	19.595445
3	RSS (Test)	9.861356	9.884274
4	MSE (Train)	0.138233	0.138469
5	MSE (Test)	0.150048	0.150223

As can be observed, the different metrics are quite similar for both techniques with Lasso Regression showing slightly better results. Additionally, when the predictions are quite close for Ridge and Lasso, it would be better to opt for Lasso Regression since Lasso Regression assists in Feature Selection by pushing some of the coefficients to zero.

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:

As observed, the most important variables were
'houseAge','GrLivArea','TotalBsmtSF','2ndFlrSF','1stFlrSF'

If these features are dropped from the dataset, the new important predictors are:

'BsmtFinSF1','TotRmsAbvGrd','Street_Pave','BsmtUnfSF', 'houseRemodAge'.

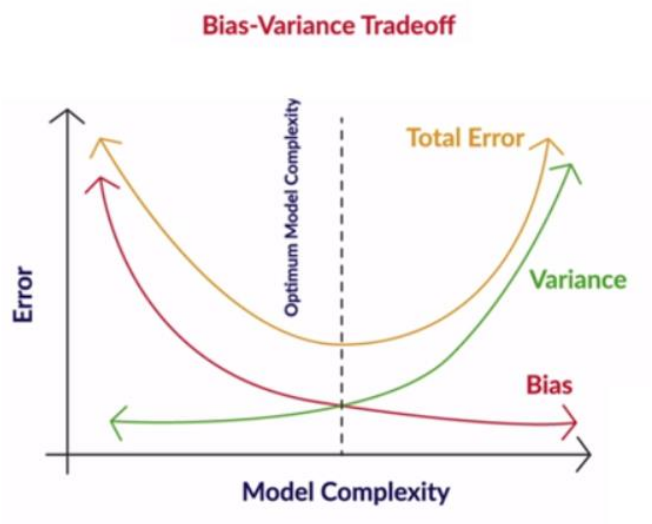
	Features	Coefficients
3	BsmtFinSF1	0.297011
7	TotRmsAbvGrd	0.257599
18	Street_Pave	0.210812
4	BsmtUnfSF	0.191654
13	houseRemodAge	-0.188004

Sale Price is positively correlated with ‘BsmtFinSF1’, ’TotRmsAbvGrd’, ’Street_Pave’, ’BsmtUnfSF’and negatively correlated with ’houseRemodAge’.

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:



Whenever designing a model, it is important to ensure that we look at the Bias-variance tradeoff.

Bias is a lack of signal - When your model misses capturing relationships that actually exist between dependent and independent variables. These relationships are important for accurate prediction. This is also called underfitting.

Variance is when the model takes noise as a signal i.e. captures patterns and relationships in data that do not exist. These patterns will not appear in the test set and the model does not generalize well. This is also called overfitting.

Hence, we look at the Bias-Variance trade-off. We find the best complexity of model such that error can be minimized while both bias and variance can also be maintained as low.

For example, some of the regularization techniques to do this for linear regression are Ridge and Lasso Regression techniques. When it is observed that the Linear Regression model tends to overfit, we can go for these methods.

These techniques add a penalty term to the actual error making the model believe that there is more of an error than actually present. This would cause smoothening of the actual fit line and make the system more generalizable. By giving up a little on Bias, we can achieve a greater reduction in variance – resulting in better generalizability.