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 :

An optimal value for Ridge Regression is 3 and for Lasso Regression is 50.

Model performance comparison before and after a change in alpha:

	Metric	Ridge Regression	Lasso Regression	Ridge Regression Double Aplha	Lasso Regression Double Aplha
0	R2 Score (Train)	0.9405	0.9387	0.9405	0.9405
1	R2 Score (Test)	0.9130	0.9148	0.9130	0.9130
2	RSS (Train)	327534556575.3362	337015206088.8757	327534556575.3362	327534556575.3362
3	RSS (Test)	100486421659.6973	98363349075.4278	100486421659.6973	100486421659.6973
4	MSE (Train)	18189.0898	18450.4580	18189.0898	18189.0898
5	MSE (Test)	20129.2619	19915.4816	20129.2619	20129.2619

As shown above, when we double these values, the model performance remains the same in both cases, but there is little change in the position of the importance of variables.

The top 16 Important variables for each model before and after changing alpha are shown below:

	Ridge	Lasso	Rideg Double Alpha	Lasso Double Alpha
0	GrLivArea	GrLivArea	GrLivArea	GrLivArea
1	1stFlrSF	TotalBsmtSF	OverallQual	OverallQual
2	OverallQual	OverallQual	1stFlrSF	TotalBsmtSF
3	TotalBsmtSF	Neighborhood_StoneBr	TotalBsmtSF	OverallCond
4	Neighborhood_StoneBr	OverallCond	Neighborhood_StoneBr	Neighborhood_StoneBr
5	BsmtFinSF1	house_age	BsmtFinSF1	BsmtFinSF1
6	OverallCond	BsmtFinSF1	OverallCond	Neighborhood_NoRidge
7	2ndFlrSF	Neighborhood_NoRidge	FullBath	house_age_when_sold_in_month
8	FullBath	Neighborhood_Crawfor	2ndFlrSF	Neighborhood_Crawfor
9	LotArea	LotArea	LotArea	LotArea
10	Neighborhood_NoRidge	ExterQual_Gd	Neighborhood_NoRidge	ExterQual_TA
11	ExterQual_TA	ExterQual_TA	GarageArea	BsmtExposure_Gd
12	house_age	SaleType_CWD	ExterQual_TA	KitchenQual_TA
13	GarageArea	house_age_when_sold_in_month	KitchenQual_TA	KitchenQual_Fa
14	house_age_when_sold_in_month	HouseStyle_2.5Fin	BsmtExposure_Gd	ExterQual_Gd
15	ExterQual_Gd	BsmtExposure_Gd	KitchenQual_Gd	KitchenQual_Gd

Overall, since the alpha values are small, we do not see a huge change in the model after doubling the alpha.

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 :

- The optimum lambda value in the case of Ridge and Lasso is as follows:-
  - Ridge Regression - 3
  - Lasso Regression - 50
- The R2 Score for Ridge and Lasso on train and test data are as follows:
  - Ridge Regression:
    - Train Data: 94.05 %
    - Test Data: 91.30 %
  - Lasso Regression:
    - Train Data: 93.87 %
    - Test Data: 91.48 %
- Comparing performance matrix R2 score for both is almost the same, but Lasso Regression performs a little better than Ridge Regression
- Since Lasso Regression helps in feature selection (as the coefficient value of some of the features becomes zero), Lasso Regression has a better edge over Ridge. It should be used as the final model as our dataset has over 130+ columns.

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 :

The top 5 important predictor variables in the Lasso in the current model are:

Feature	Description
GrLivArea	Above grade (ground) living area square feet
TotalBsmtSF	Total square feet of basement area
OverallQual	Rates the overall material and finish of the house
Neighborhood	Physical locations within Ames city limits
OverallCond	Rates the overall condition of the house

After the removal of these top 5 variables and rebuilding the Lasso model new model has the following top 5 predictor variables:

Feature	Description
1stFlrSF	First Floor square feet
2ndFlrSF	Second floor square feet
BsmtFinSF1	Type 1 finished square feet
house_age_when_sold_in_month	House Age
BsmtUnfSF	Unfinished square feet of basement area

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:

As Per Occam's Razor— given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes less on the test data due to the following reasons:-

- Simpler models are usually more 'generic' and are more widely applicable
- Simpler models require fewer training samples for effective training than the more complex ones and hence are easier to train.
- Simpler models are more robust.
    - Complex models tend to change wildly with changes in the training data set
    - Simple models have low variance and high bias, and complex models have low bias, high variance
    - Simpler models make more errors in the training set. Complex models lead to overfitting. They work very well for the training samples but fail miserably when applied to other test samples

Therefore, to make the model more robust and generalizable, make the model simple but not Simpler, which will not be of any use.

Regularization can be used to make the model simpler. Regularization helps strike the delicate balance between keeping the model simple and not making it too naive to se. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.

Also, Making a model simple lead to Bias-Variance Trade-off:

- A complex model will need to change for every little change in the dataset and hence is very unstable and extremely sensitive to any changes in the training data.
- A simpler model that abstracts out some pattern followed by the data points given is unlikely to change wildly even if more points are added or removed.

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

bias as its expected error across all test inputs is very high. Variance refers to the degree of changes in the model itself with respect to changes in the training data.

Thus the accuracy of the model can be maintained by keeping the balance between Bias and Variance as it minimizes the total error, as shown in the below graph. The accuracy of the model will go up if we try to fit the model over but that no longer makes it generalizable. When the model is generalized, the accuracy should be pretty good on both the training and the testing dataset making the model robust.