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?

The Best value of Lambda we got in Lasso and Ridge are as below:

Ridge: 2.0

Lasso: 0.0001

Different values with Ridge at 2.0 and Lasso at 0.0001 are as below

Metric	Ridge	Lasso
R2SCORE(TRAIN)	0.93006	0.92737
R2 SCORE(TEST)	0.89665	0.90253
RSS(TRAIN)	1.42912	1.48401
RSS(TEST)	1.15425	1.08856
MSE(TRAIN)	0.00158	0.00165
MSE(TEST)	0.00298	0.00281

and after rebuilding the model with doubling the value of Lambda for Ridge and Lasso we get below Metrics

Metric	Ridge	Lasso
R2SCORE(TRAIN)	0.92671	0.92278
R2 SCORE(TEST)	0.89451	0.88561
RSS(TRAIN)	1.49746	1.57782
RSS(TEST)	1.17815	1.27761
MSE(TRAIN)	0.00166	0.00175
MSE(TEST)	0.00304	0.00329

Observations:

Ridge:

- R2 Remains almost the same for Test and Train
- MSE increases

Lasso:

- MSE increases a bit
- Huge fall in value of R2 in Test making prediction worse

Also Higher number of coefficients of variables shrink towards zero

Predictor Variables are as below

Factor	Inference
OverallQual	Increase in Quality increases the price
GrLivArea	Higher the size of living area higher is the price
Total_sqr_foot	With increase in Total Square foot of the House Price increases
YearBuilt	Newer the House the higher is the price, with age the price decreases
Neighborhood_StoneBr	Stone Brook location also is a factor in increasing the price

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?

The Best value of Lambda we got in Lasso and Ridge are as below:

Ridge: 2.0 Lasso: 0.0001

and the r2 values are as below Ridge Train-0.930 Test-0.897 Lasso Train-0.923 Test-0.903

The Mean Squared error in case of Ridge and Lasso is:

Ridge - 0.002951 Lasso - 0.002785

Mean Sqaured Error of Lasso is less than that of Ridge also Since Lasso helps in feature reduction and helps to increase model interpretation, hence i would choose to apply Lasso

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?

I have now created another model, details are in the workbook, and the the five most important predictor variables are as below

Factor	Inference
TotalBsmtSF	Higher the size of Basement area higher is the price
TotRmsAbvGrd	Higher the number of rooms higher is the cost
OverallCond	As the condition increases the Price also increases
Total_bathrooms	Higher the number of Bathrooms higher is the price
LotArea	Price also increases with increase in Lot 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?

To make the model robust and generalizable, we try to keep it simple, Simple models are reliable but may not be perfect

So there is a balance to strike, this is called as Bias- Variance Trade-off one side we do have simple model with a bit of fixed way of doing things called “BIAS”, which makes it predictable
On the other hand we have complex model which is Unpredictable (Variance)

There are chances that both models would fail

Underfitting

A Simple solution may not handle some conditions

Overfitting

A Complex solution might fit too closely to training data but struggle with unseen data

Regularization is something which helps us to find the sweet-spot between Bias and Variance And penalizes coefficients and shrinks the coefficients to zero so that the model don't become too complex

This sweet-spot where your model is simple enough to work well in different situations but not so simple that its unreliable, and achieving this sweet-spot is how we make model more reliable and generalizable

To summarize, regularization is a tool to improvise the accuracy of the model especially while dealing with complex or limited data, by striking balance between Variance and Bias , these techniques can lead to more accurate prediction on unseen data

Below is Graphical representation of the trade-off between Bias and Variance