Problem Statement - 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?

Answer1:

As per the coding solution solved by me

The Optimal value of alpha for ridge: 4

Doubling the value of alpha for ridge: 8

The Optimal value of alpha for lasso: 0.0006

Doubling the value of alpha for lasso:0.0012

In case of Ridge regression there is no significant change in the R2 and MSE values remains nearly the same.

After doubling:

GrLivArea	0.558779
OverallQual	0.451126
OverallCond	0.172301
GarageArea	0.137731
BsmtFinSF1	0.132628

Before doubling:

OverallQual	0.288297
GrLivArea	0.271558
1stFlrSF	0.219355
OverallCond	0.174195
BsmtFinSF1	0.143023

As observed in Ridge regression after doubling the top 5 features order changes slightly, as well as Garage Area is replaced by 1stFlrSF

In case of Lasso regression there is no significant change in the R2 and MSE values remains nearly the same.

After Doubling: (top 5 features) Before Doubling:

GrLivArea	0.558779	**GrLivArea**	0.540342
OverallQual	0.451126	**OverallQual**	0.422573
OverallCond	0.172301	**OverallCond**	0.195187
GarageArea	0.137731	**BsmtFinSF1**	0.146311
BsmtFinSF1	0.132628	**GarageArea**	0.133949

Only the Garage area takes precedence over BsmtFinF1 after doubling

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:

Based on the results:

r2_score in test dataset:
r2_score for ridge: 0.88
r2_score for lasso: 0.89

Both ridge and lasso give nearly the same r2_score. But I will opt for lasso Regression as it provides a feature selection option also. It has helped in removing unnecessary features from model without affecting the model accuracy.

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 3:

The top most important features in Lasso are as follows:

GrLivArea	0.540342
OverallQual	0.422573
OverallCond	0.195187
BsmtFinSF1	0.146311
GarageArea	0.133949

On dropping the important features, the R2 score has reduced from 88.5 to 85.5.

Now top most features are:

1stFlrSF	0.604992
2ndFlrSF	0.439335
MSZoning_FV	0.413685
MSZoning_RL	0.412696
MSZoning_RH	0.405729

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 4:

We make a model more robust and generalized by applying regularization techniques to control the overfitting to the linear model.

Overfitting occurs when the model perform accurately on the train data but fails badly on the test data. This phenomenon is called overfitting.

Regularization is a process of applying penalty to the coefficients of the linear model in order to prevent overfitting.

A model is said to be accurate if there is fairly explains the variance and there is less bias.

There is a trade-off between bias and variance with respect to model complexity. What we need is total error to be lowest , i.e., low bias and low variance, such that the model identifies all the patterns that it should and is also able to perform well with unseen data.

Regularization helps with managing model complexity by essentially shrinking the model coefficient estimates towards 0. This discourages the model from becoming too complex, thus avoiding the risk of overfitting.

Regularization helps to make simpler models hence more accurate models as per the bias-variance tradeoff.