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 optimal value of alpha is:

1. Ridge – 2.0
2. Lasso – 0.0001

If we choose to double the value of alpha, then the changes are:

1. In case of Ridge, there is a slight increase in the MSE whereas the r2 value of train and test remains almost same.
2. In case of Lasso, there is a slight increase in MSE and there is a huge drop in the r2 value of test data set.

The most important predictor variables after the change are:

1. For Ridge:

	Coefficient
Total_sqr_footage	0.121485
OverallQual	0.110806
GrLivArea	0.103789
Neighborhood_StoneBr	0.080577
OverallCond	0.071055
TotalBsmtSF	0.057590
LotArea	0.052714
YearBuilt	0.047783
Neighborhood_Crawfor	0.046705
Fireplaces	0.041324

2. For Lasso:

	Coefficient
Total_sqr_footage	0.185429
OverallQual	0.172690
YearBuilt	0.119993
GrLivArea	0.105465
Neighborhood_StoneBr	0.093874
OverallCond	0.089137
LotArea	0.054712
Neighborhood_Crawfor	0.053322
Neighborhood_NridgHt	0.047466
GarageCars	0.044402

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 optimal value of LAMBDA we got in case of Ridge and Lasso is:

- Ridge - **2.0**
- Lasso - **0.0001**

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

- Ridge - **0.00297**
- Lasso - **0.00280**

We can clearly observe that the Mean Squared Error of Lasso is slightly lower than that of Ridge.

Also, since Lasso helps in feature reduction (as the coefficient value of one of the lasso's feature to be shrunk toward 0) and helps to increase model interpretation by taking the magnitude of the coefficients thus Lasso has a better edge over Ridge.

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?

After dropping the five most important predictor variables in the lasso model:

```
Top 5 correlated features when alpha is 0.0001 are:

                Coefficient
TotalBsmtSF        0.323329
TotRmsAbvGrd       0.126096
OverallCond        0.094291
Total_Bathrooms    0.086659
LotArea            0.067859
```

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?

- The model is expected to be as simple as possible and simpler models are considered as more **generic**, though its accuracy will be decreased but it will be more **robust**.
- This can be understood from the Bias-Variance trade-off. The simpler the model the more the bias but less variance becoming generalizable. Whereas the complex model will have high variance and low bias.
- Sometimes underfitting and overfitting are the problems associated with the model. Hence, it is important to have balance in Bias and Variance to avoid such problems. This is possible with **Regularization**.
- Regularization helps in managing the model complexity by essentially shrinking the coefficients towards zero. This avoids the model becoming too complex, thus reducing the risk of overfitting.
- Regularization method should be used to keep the model optimum simpler. It penalizes the model if it becomes more complex.
- Regularization method helps to achieve the Bias-Variance trade off. It compromises by increasing bias to an optimum position where Total Error is minimum.
- This point also known as Optimum Model Complexity where Model is sufficient simpler to be generalisable and complex enough to be robust.