Assignment Part II - Subjective Questions and Answers

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

- Ridge – 3.0.
- Lasso – 0.0002

The changes in the model if we choose to double the value of alpha for both ridge and lasso are:

- When we double the alpha value in case of ridge there is a slight increase in the mean squared error whereas the r2 value of train and test remains almost same.
- When we double the alpha value in case of lasso there is a slight increase in the mean squared error, the r2 value of train slightly decreases whereas there is a huge fall in the r2 value of test thus making the model and prediction worse.
- It also penalizes the model even more and a greater number of coefficients of a variable shrink towards zero.

The most important predictor variable after implementing the change:
The most important predictor variable across is 'GrLivArea' with different coefficient values.
Ridge:-

```
OverallQual     0.107755
GrLivArea       0.086496
OverallCond     0.066089
1stFlrSF        0.064355
TotalBsmtSF     0.057218
BsmtFinSF1      0.057194
Fireplaces      0.042718
LotArea         0.041339
2ndFlrSF        0.040734
GarageArea      0.040665
```

Lasso:-

```
GrLivArea                0.219931
OverallQual              0.210416
TotalBsmtSF              0.106060
OverallCond              0.084730
BsmtFinSF1               0.048986
GarageArea               0.047040
LotArea                  0.043390
Fireplaces               0.038576
SaleCondition_Partial    0.033885
MSZoning_FV              0.028772
```

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

- Ridge – 3.0.
- Lasso – 0.0002

The r2 value we got in case of Ridge and Lasso is:
• Ridge - Train = 0.928, Test = 0.863, difference – 0.065
• Lasso - Train = 0.920, Test = 0.874, difference – 0.046

The Mean Squared error in case of Ridge and Lasso is:
• Ridge - 0.00253
• Lasso - 0.00234

It can be clearly observed that the Mean Squared Error of Lasso is slightly lower than that of Ridge and difference of r2 between train and test is less in lasso as compared to ridge. Since Lasso helps in feature reduction (as the coefficient value of one of the lasso's features 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?

Answer
After dropping the top 5 most important predictor variables in the lasso model and again created again model and we got the below five most important predictor variables:
1. 1stFlrSF
2. 2ndFlrSF
3. GarageArea
4. Total_ MSZoning_FV
5. MSZoning_RL

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
• 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.

- 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 also complex enough to be robust.
- Making a model simple lead to Bias-Variance trade off.