Assignment-based Subjective Questions

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?

Ans:

- Optimal value of alpha for Ridge Regression is 2.0
- Optimal value of alpha for Lasso Regression is 0.001

# Changing the alpha values double to from 2.0 to 4.0 for Ridge and 0.001 to 0.002 for Lasso Regression

# Alpha (1.0 and 0.001)
Ridge Regression - R^2 Score on Train Data: 0.949774688762443
Ridge Regression - R^2 Score on Test Data: 0.8685457733694272
Lasso Regression - R^2 Score on Train Data: 0.928937802340879
Lasso Regression - R^2 Score on Test Data: 0.8777015866323229

# Alpha (2.0 and 0.002)
Ridge Regression - R^2 Score on Train Data: 0.9417423761649065
Ridge Regression - R^2 Score on Test Data: 0.8712298647328072
Lasso Regression - R^2 Score on Train Data: 0.9033777884532062
Lasso Regression - R^2 Score on Test Data: 0.8664613307973125

Conclusion:
Changes in Ridge Regression Model:

- R^2 Score on Train Data remain same 0.94(without rounding off)
- R^2 Score on Test Data increase from 0.86 to 0.87

Changes in Lasso Regression Model:

- R^2 Score on Train Data decrease from 0.92 to 0.90
- R^2 Score on Test Data decrease from 0.87 to 0.86

Important predictor features post doubling the alpha:

Ridge positive features:

GrLivArea	0.175040
TotalBsmtSF	0.145325
1stFlrSF	0.142742
BsmtFinSF1	0.128196
2ndFlrSF	0.109206
GarageArea	0.107519
GarageCars	0.099305
Neighborhood_Crawfor	0.096491
FullBath	0.082962
OverallQual_9	0.077880
Exterior1st_BrkFace	0.076314
OverallQual_8	0.074603
TotRmsAbvGrd	0.073154
YearBuilt	0.072690
SaleType_CWD	0.067802

Lasso positive features:

GrLivArea	0.684937
TotalBsmtSF	0.171451
GarageCars	0.136910
YearBuilt	0.130182

BsmtFinSF1	0.111507
GarageArea	0.082926
OverallQual_8	0.070358
Functional_Typ	0.067668
Neighborhood_Crawfor	0.058801
YearRemodAdd	0.045076
Neighborhood_Somerst	0.044713
BsmtCond_TA	0.032834
Foundation_PConc	0.028488
Condition1_Norm	0.025008
GarageType_Attchd	0.023734

Below are the common list of features:

- GrLivArea
- TotalBsmtSF
- GarageCars
- BsmtFinSF1
- GarageArea
- Neighborhood_Crawfor

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?
Ans:

- Lasso should be considered, where the primary objective is feature selection and number of feature are more for huge data set.
- Ridge should be considered, where there are higher multicollinearity where feature are correlated. For Robust model, use all the feature for model rather than neglecting the features.
- Conclusion: Choosing the model should be w.r.t to use cases, which will give the optimal outcome.

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?

Ans:
Lasso Regression Training - MSE: 0.007952975927461655, R^2: 0.9240221465626874, RSS: 3.8969582044562108, RMSE: 0.08917945911173523
Lasso Regression Testing - MSE: 0.01826619188614485, R^2: 0.8687040828448955, RSS: 2.2467416019958164, RMSE: 0.13515247643363718

Lasso Top 15 positive features:

1stFlrSF	0.634122
2ndFlrSF	0.360898
GarageArea	0.183009
Neighborhood_Crawfor	0.125662
Functional_Typ	0.099429
OverallQual_9	0.091172
GarageYrBlt	0.088709
OverallQual_8	0.083695
Neighborhood_Somerst	0.068826
BsmtCond_TA	0.048229
GarageQual_TA	0.045781
WoodDeckSF	0.044579

FullBath	0.044424
BsmtFullBath	0.044376
Condition1_Norm	0.041513

Below are the list of the top 5 feature post changes and removal of top fearure.

1stFlrSF	0.634122
2ndFlrSF	0.360898
GarageArea	0.183009
Neighborhood_Crawfor	0.125662
Functional_Typ	0.099429

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?

Ans:
Model is considered to be robust provided, any changes/Variation in data set doesn't impact much on the outcome of the model prediction. The overall performance of the model is not affected with the changes. On other hand it is generally considered as generalisable where the performance of the model is optimal or better for unseen data set or adapt to new data set without much changes or alteration. Model should perform well not only with train data set but also with the new and unseen data. Overfitting and underfitting of data should be handled before training the model, hence the performance will be optimal and also the complexity of the model should be less. In other word we should try to create model which shouldn't be complex in nature.

Higher Accuracy of model will have trade off, since higher the accuracy the complexity of the model is increase. In same time, we should consider the proper balance between the accuracy and complexity by considering the variance and bias.
There are techniques such L1(Lasso) and L2(Ridge) for regularization for handling the same for better performances.

Balance Bais and variance, same time the complexity of model should be optimal.