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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

Note: Please refer to the Jupiter notebook for the coding.

1) Optimal values of Alpha for Ridge and Lasso Regression:

```
Optimal alpha in Ridge:  50
Optimal alpha in Lasso:  0.001
```

2) changes in the model if you choose to double the value of alpha for both Ridge and Lasso:

Ridge Regression:

Ridge Regression with original alpha	**Ridge Regression with doubled alpha**
Ridge Regression with alpha=50:	**Ridge Regression with alpha=100:**
Scores for Train Set:	**Scores for Train Set:**
R2 Score: 0.9197662757946696	R2 Score: 0.9171772754307903
MSE Score: 0.012612037517869377	MSE Score: 0.013019005660585274
MAE Score: 0.0765276854045055	MAE Score: 0.07714061780987087
RMSE Score: 0.11230332816915702	RMSE Score: 0.11410085740512765
Scores for Test Set:	**Scores for Test Set:**
R2 Score: 0.885161713208492	R2 Score: 0.8844972687493229
MSE Score: 0.018895505090551764	MSE Score: 0.01900483285928985
MAE Score: 0.09122978477124648	MAE Score: 0.09093041201805453
RMSE Score: 0.13746092204896548	RMSE Score: 0.13785801702944175

Observations:

i) R2 score slightly decreased with the double alpha in train and test set.
ii) On the other hand, MSE score have increased when doubling the alpha value.
iii) As a conclusion, alpha=50 is performing better than alpha=100 in Ridge Regression.

Lasso Regression:

Lasso Regression with original alpha	Lasso Regression with doubled alpha
Lasso Regression with alpha=0.001:	**Lasso Regression with alpha=0.002:**
Scores for Train Set:	**Scores for Train Set:**
R2 Score: 0.9220783927747972	R2 Score: 0.9169093432670837
MSE Score: 0.012248593013854506	MSE Score: 0.013061122246028371
MAE Score: 0.0761527631613061	MAE Score: 0.07726698418637133
RMSE Score: 0.11230332816915702	RMSE Score: 0.11428526696835586
Scores for Test Set:	**Scores for Test Set:**
R2 Score: 0.8876701333878119	R2 Score: 0.887269307632054
MSE Score: 0.018482769341945256	MSE Score: 0.018548721258505086
MAE Score: 0.09092234954831672	MAE Score: 0.0902369190298715
RMSE Score: 0.13746092204896548	RMSE Score: 0.13619369023014644

Observations:

i) When doubling the alpha value in Lasso Regression, we are seeing similar behavior as we saw with Ridge Regression.
ii) R2 score slightly decreased in train and test set when the alpha value was doubled.
iii) MSE score increased with the doubled the alpha in train and test set.
iv) We can conclude that with the doubled alpha value, Lasso Regression performance is decreasing than the original alpha value. Thus, we can say that Lasso doing better with original alpha value 0.001

3) Most important predictor variables after the change is implemented:

i) Most important features in Ridge Regression with doubled alpha=100:

['GrLivArea', 'OverallQual', 'YearBuilt', '1stFlrSF', '2ndFlrSF', 'OverallCond', 'BsmtFinSF1', 'MSZoning_RL', 'LotArea', 'GarageArea']

ii) Most important features in Lasso Regression with doubled alpha=0.002:

```
['GrLivArea', 'OverallQual', 'YearBuilt', 'OverallCond', 'BsmtFinSF1',
       'MSZoning_RL', 'LotArea', '1stFlrSF', 'GarageArea', 'FireplaceQu']
```

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 values for alpha in Ridge and Lasso Regression are 50 and 0.001 respectively. The R2 and MSE scores from both the models are given below:

Ridge Regression with alpha=50:	**Lasso Regression with alpha=0.001:**
Scores for Train Set: R2 Score: 0.9197662757946696 MSE Score: 0.012612037517869377 MAE Score: 0.0765276854045055 RMSE Score: 0.11230332816915702 **Scores for Test Set:** R2 Score: 0.885161713208492 MSE Score: 0.018895505090551764 MAE Score: 0.09122978477124648 RMSE Score: 0.13746092204896548	**Scores for Train Set:** R2 Score: 0.9220783927747972 MSE Score: 0.012248593013854506 MAE Score: 0.0761527631613061 RMSE Score: 0.11230332816915702 **Scores for Test Set:** R2 Score: 0.8876701333878119 MSE Score: 0.018482769341945256 MAE Score: 0.09092234954831672 RMSE Score: 0.13746092204896548

Observations:

i) By comparing the output from both these models, we can see that R2 score is slightly better in test and train set from Lasso Regression.
ii) The MSE scores is decreased in Lasso Regression as compared to Ridge Regression.
iii) We can say that Lasso is performing better on the test and train dataset as compared to Ridge Regression and we will choose Lasso as our final model.

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:

Note: Please refer to the Jupiter notebook for the coding.

Five most important variables on the original dataset before removal were as follows:

```
['MSZoning_RL', 'GrLivArea', 'MSZoning_RM', 'OverallQual', 'YearBuilt']
```

After removing the above variables from the dataset, the following five most variables were selected by Lasso Regression:

```
['2ndFlrSF', '1stFlrSF', 'Foundation_PConc', 'BsmtFinSF1', 'YearRemodAdd']
```

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:

A model is considered robust or generalizable when it predicts the target variable consistently accurate even if one or more independent variables change drastically. A generalized model performs better on new unseen data. If your model is overfitted then it is not generalized.

An overfit model performs so well on the training data, but it fails miserably on unseen data which makes it low biased and high variance. To make a model generalize, a balance between bias and variance is required so that both bias and variance should be at their lowest. Overfitting can be managed through Regularization. Regularization techniques help 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.

When a model is overfit, the training accuracy is high while the test accuracy is very low. The accuracy of a model can be increased by keeping following things:

1. The presence of missing and outlier values in the training dataset often reduces the accuracy of a model which can make the model biased. To increase the accuracy of a model, missing values and outliers should be treated accordingly which makes the model more generalizable.
2. The scaling and transformations of the independent and dependent features help improve the accuracy of a model as it brings all the different features to same level which eliminate non-linearity between predictors and response variables.