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 1

Optimal alpha for Ridge = 50 gives the below R2 scores

```
Training R2 score 0.904606822679251
Test R2 score 0.8602647918494458
```

Optimal alpha for Lasso = 0.001gives the below R2 scores

```
Training R2 score 0.9070933064894305
Test R2 score 0.8637041005532651
```

R2 scores for doubled alpha in Ridge

```
Training R2 score 0.8984213694918839
Test R2 score 0.8535877918289478
```

R2 scores for doubled alpha in Lasso

```
Training R2 score 0.8950359248012991
Test R2 score 0.8513246756360294
```

The R2 scores seem to decrease with the increase in alpha and also the feature coefficients got changed.

Features
Neighborhood__Crawfor
OverallQual
Neighborhood__NridgHt
Condition1__Norm
Neighborhood__Somerst
OverallCond
Exterior1st__BrkFace
TotalBath
GarageCars
MSZoning__RL

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

Ridge model and lasso both perform similar in terms of R2 score but lasso model is less complex as it uses only 80 features while ridge uses 190 features so Lasso seems better to be used.

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 10 predictor variables were

Features
Neighborhood__Crawfor
Neighborhood__NridgHt
Exterior1st__BrkFace
Neighborhood__Somerst
OverallQual
Neighborhood__ClearCr
Condition1__Norm
Neighborhood__NoRidge
MSZoning__RL
Neighborhood__StoneBr

If the top-5 predictor variables are not there in the data, we can use the next-5 predictor variables i.e. Neighborhood_ClearCr, Condition1_Norm, Neighborhood_NoRidge, MSZOning_RL, Neighborhood_StoneBr.

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

In order to ensure that the model is robust, we can use the techniques such as feature engineering, train-test split, cross validation, regularization

To ensure that the model is generalizable, we ensure that the train r2 score and test r2 score doesn't have much difference. A robust and generalized model should be able to work well on both seen and un-seen data.

The model shouldn't give much importance to the outlier data. Hence, any outliers should be removed. Removing outliers can increase the accuracy of the model.