## 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 :

Optimal value for alpha for Ridge - 20

Optimal value for alpha for Ridge - 0.0001

After doubling the alpha for both , we see that metrics for both haven't changed much. While looking at the coefficients, we can see that it has reduced slightly for all the predictors for both Ridge and Lasso

For Lasso, with alpha 0.0001, the model had made 57 out of 230 features to 0. However, after doubling the alpha, 91(34 more features) were 0 , making the model more simpler

The most important predictor variables are as follow

For Ridge:

| **Features** | **Coefficients** |
|---|---|
| TotalFlrSF | 0.11925 |
| OverallQualCond | 0.072715 |
| Neighborhood_Edwards | -0.05526 |
| property_age | -0.0532 |
| Condition1_Norm | 0.047118 |
| Neighborhood_NridgHt | 0.042218 |
| Neighborhood_Crawfor | 0.042016 |
| LotArea | 0.039342 |
| TotalBsmtSF | 0.037556 |
| Neighborhood_NoRidge | 0.037493 |

For Lasso:

| **Features** | **Coefficients** |
|---|---|
| Condition2_PosN | -0.7583 |
| MSZoning_RH | 0.2246 |
| MSZoning_RL | 0.2235 |
| MSZoning_FV | 0.2129 |
| TotalFlrSF | 0.205 |
| MSZoning_RM | 0.1963 |
| RoofMatl_WdShngl | 0.1913 |
| FullBath_3 | 0.1258 |
| Neighborhood_NoRidge | 0.1169 |
| SaleType_ConLD | 0.1154 |

## 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 metrics for both are quite similar

| **Metric** | **Ridge Regression** | **Lasso Regression** |
|---|---|---|
| R2 Score (Train) | 0.925845 | 0.947999 |
| R2 Score (Test) | 0.896888 | 0.891321 |
| RSS (Train) | 11.901279 | 8.345779 |
| RSS (Test) | 7.431114 | 7.832346 |
| MSE (Train) | 0.107965 | 0.090411 |
| MSE (Test) | 0.130254 | 0.133724 |

But since Lasso puts more penalty and has made around 57 predictors coefficient to 0, I would choose the Lasso model for prediction

## 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 removing the 5 most important features which were –

Condition2_PosN, RoofMatl_WdShngl, RoofMatl_CompShg, RoofMatl_Roll, RoofMatl_Tar&Grv

The five most important features now are

| **Features** | **Coefficients** |
|---|---|
| TotalFlrSF | 0.2045 |
| Neighborhood_NoRidge | 0.0916 |
| Neighborhood_Somerst | 0.089 |
| OverallQualCond | 0.0767 |
| Neighborhood_NridgHt | 0.0761 |

## Question 4

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

## Answer :

The model can be made more robust and generalizable by handling the outliers, getting more data to train on, by feature engineering and transforming and deriving new features/information from the existing features and selecting  the best features that explains the relationship of the predictors with the features.

When the model is more robust, it will be able to predict correctly even if the dataset is changed. There will be no overfitting problem. In general there will be less bias and less variance