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:

- Initially, the optimal alpha value for **Ridge** regression is **20.0** and **Lasso** regression is **100.0.**
- After doubling the alpha value for **Ridge and Lasso regression**, we can see the **R2 Score decreased,** whereas the **RSS, MSE and RMSE** values are **slightly increased.**
- After doubling the alpha values, the most important predictors are:
    - **OverallQual_8**
    - **OverallQual_9**
    - **GrLivArea**
    - **Neighborhood_Crawfor**
    - **Exterior1st_BrkFace**
    - **Functional_Typ**

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:

- Comparing between the Ridge and Lasso build initially with alpha value 20.0 and 100.0 with the Ridge and Lasso model built by doubling the alpha values, **the ones initially built with Ridge model having 20.0 alpha value and Lasso model having 100.0 will be chosen based the RSS, MSE and RMSE values. Because these metrics were lower for the initial models**

- Comparing between initially build Ridge and Lasso models, **the lasso model will be chosen** because here **feature selection** is performed due to which **all the insignificant predictors equated to exact 0** instead of keeping them in the equation with a value that is close to zero.

- Given our data set has 81 columns, we need **feature selection**. So, for this reason **Lasso regression** will be a better choice here.

Question 3

After building the model, you realized 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:

- These were the top 5 predictors in the initially built model:
  - **OverallQual_8**
  - **OverallQual_9**
  - **GrLivArea**
  - **Neighborhood_Crawfor**
  - **Exterior1st_BrkFace'**

- After removal them and building the ridge and lasso model again these are 5 top predictors:
  - **2ndFlrSF**
  - **Functional_Typ**
  - **1stFlrSF**
  - **Neighborhood_NridgHt**
  - **Neighborhood_Somerst**

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:

- We can make sure a model is **robust** by handling **outliers** properly, make sure the model has **low variance**(**regularization**) such that any future changes made to the data does not affect the model's performance
- We can make sure a model is **generalizable** by making sure the model does not **overfit**. Also can make the data set diverse and split the data set into train and test tests accordingly to make sure the model does well on unseen data.
- In an overfit model, because of **high variance** any change to the data affects the model and will not identify patterns in test data. A **complex** model of such kind will not be robust and generalizable
- In terms of accuracy, we can say that for an overfit model with high variance, high complexity will very **high accuracy**
- With the help of some **regularization techniques** we must strike the **balance** between **complexity** and **accuracy**.