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?

The optimal value of Alpha are

Ridge - 9
Lasso - 0.0004

When we double the value of alpha for **Ridge**,

- the mean squared error slightly increases
- the r2_score for both test and training data slightly decreases
- The most important predictor variables will be
    - GrLivArea
    - TotalBsmtSF
    - 1stFlrSF
    - OverallQual_Very Good
    - OverallQual_Excellent
    - 2ndFlrSF
    - YearRemodAdd
    - LotArea
    - Neighborhood_Crawfor
    - OverallQual_Below Average

When we double the value of alpha for **Lasso**,

- the mean squared error slightly increases
- the r2_score for both test and training data decreases
- The most important predictor variables will be
    - GrLivArea
    - OverallQual_Excellent
    - TotalBsmtSF
    - OverallQual_Very Good
    - YearBuilt
    - GarageCars
    - Neighborhood_Crawfor
    - YearRemodAdd
    - LotArea
    - OverallQual_Below Average

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?

The optimal value of Alpha are

Ridge - 9
Lasso - 0.0005

Now let us look at the mean squared error values

Ridge - 0.0051
Lasso - 0.0048

R2_Score for test data
Ridge - 0.924
Lasso - 0.929

So we can clearly see that

- mean squared error value is lesser for lasso
- lasso is giving better r-squared score on test data
- Also lasso will do **feature selection**

-
Hence I will choose Lasso

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?

When we exclude the five most important predictor variables in the lasso model and create another model the five most important predictor variables that we got were

- 1stFlrSF
- 2ndFlrSF
- GarageCars
- BsmtFinSF1
- YearRemodAdd

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?

- Ensuring that a machine learning model is robust and generalizable is crucial for its effectiveness in real-world scenarios. They refer to the capability of models to apply learned patterns to new, unseen data.
- Sometimes underfitting and overfitting are the problems associated with the machine learning models.. Hence, it is important to have balance in **Bias** and **Variance** to avoid such problems. This is possible with “Regularization”
- **High bias** typically leads to underfitting, where the model is too simple to capture the underlying patterns in the data. As a result, it performs poorly on both the training data and unseen data.
- **High variance** can lead to overfitting, where the model fits the training data too closely. While it performs well on the training data, it may generalise poorly to new, unseen data.

We need to work on both

Bias Reduction:

- Increase model complexity using a more sophisticated algorithm
- Add more features that capture relevant information.

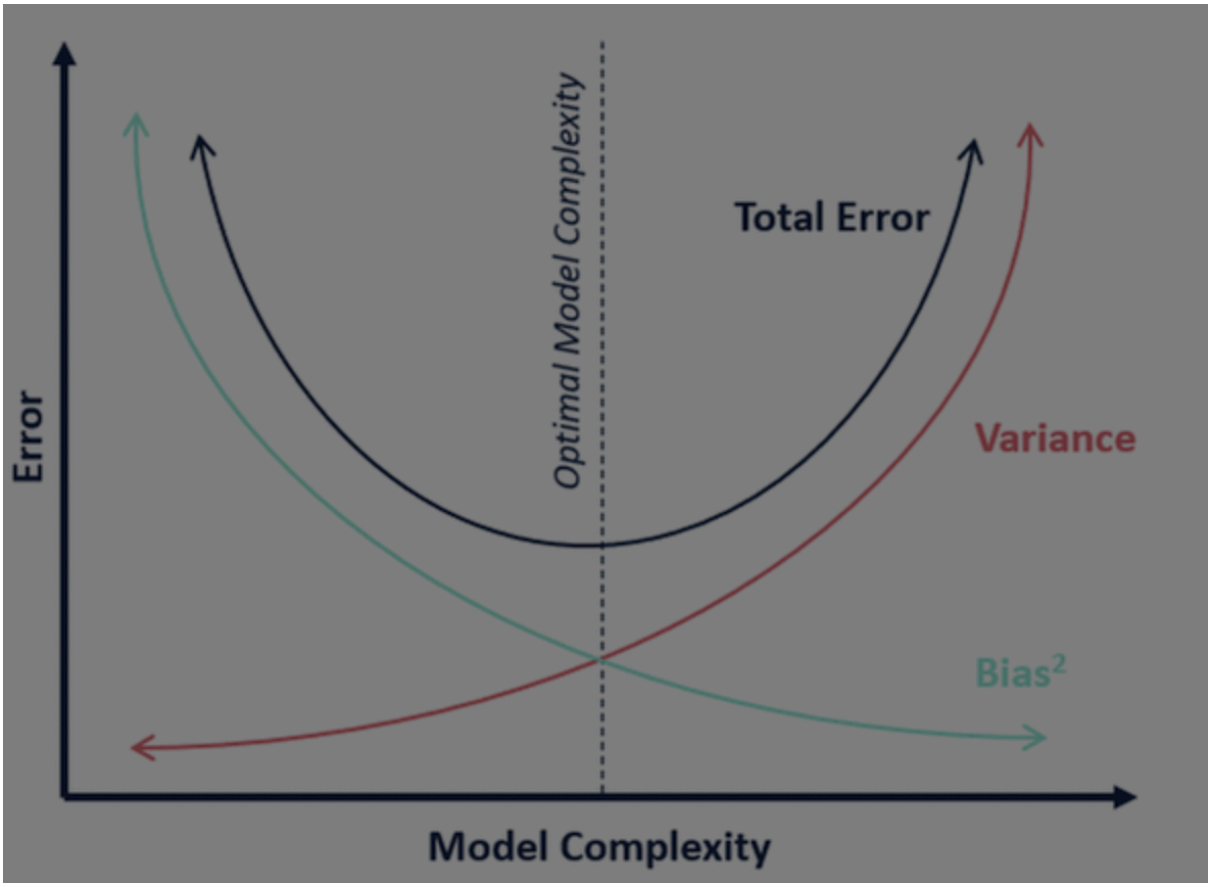
Variance Reduction:

- Simplify the model (e.g., reduce the number of features or use regularisation).
- Increase the amount of training data.

Balancing Bias and Variance:

- Fine-tune model hyperparameters to find the right balance.
- Use techniques like cross-validation to assess model performance on different subsets of the data.

This below diagram explains how we find the optimal value for model complexity which balances both bias and variance


Error
Optimal Model Complexity
Total Error
Variance
Bias^2
Model Complexity