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?

Ans-

- Optimal value for ridge regression is 1.0
- Optimal value for lasso regression is 0.0001

If we double the value of lambda, lambda increases and hence put larger penalty on the coefficients of the features therefore decreasing the complexity of the model.

R2 scores for optimal values of alpha-

- **Ridge-**
    1. train r2 score 0.9437890140795079
    2. test r2 score 0.8836827417306091
- **Lasso**
    1. r2 score for train 0.9319067291877406
    2. r2 score for test 0.8707955785971517

R2 scores when we double the alpha value-

- Ridge-
    1. train r2 score 0.9405198938080329
    2. test r2 score 0.8814188939636445
- Lasso
    1. r2 score for train 0.9214227100361603
    2. r2 score for test 0.8663874635588943

As we can see, **on increasing(doubling) the alpha value there is a decrease in the r2 score for both the models.**

Also, the number of features with 0 coefficients has increased in Lasso. **The model with double alpha uses 169 features whereas the original model used 102 features which means the model complexity decreases when we double the value of alpha.**

Accepting small decrease in the r2 score, we can create a model with lesser complexity.

After the change is implemented, the most important features are

Ridge-

- BsmtFullBath
- OverallCond
- RemodAge
- LowQualFinSF
- BsmtFinSF2

Lasso-

- BsmtFullBath
- LowQualFinSF
- OverallCond
- RemodAge
- MasVnrArea

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?

R2 score for optimal values of alpha-

- **Ridge-**
    3. train r2 score 0.9437890140795079
    4. test r2 score 0.8836827417306091
- **Lasso**
    3. r2 score for train 0.9319067291877406
    4. r2 score for test 0.8707955785971517

Lambda value 1 and 0.0001 are optimal for ridge and lasso regression model respectively.

As we can see Ridge performs better than Lasso. On the other hand model complexity for Lasso is lesser than Ridge regression model**. Ridge regression uses 146 features whereas lasso uses 102 features. Lasso has an inbuilt feature selection**. The difference in r2 score is very less so **we will choose Lasso as it is lesser in complexity and has similar r2 score.**

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?

After the top 5 features (BsmtFullBath, LowQualFinSF, OverallCond, RemodAge, MasVnrArea) are removed, the most important features with their coefficients are-

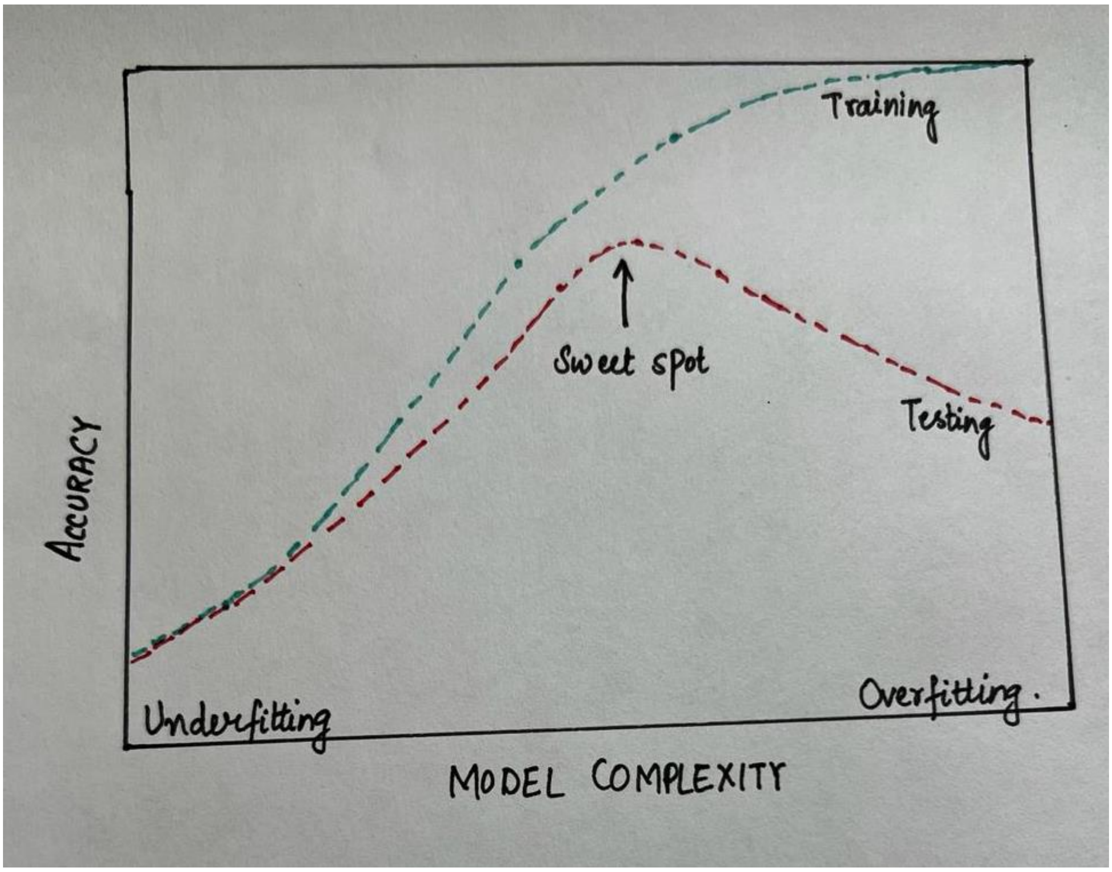
- BsmtHalfBath 0.2837
- BsmtFinSF1 0.1283
- GrLivArea 0.1142
- MSZoning -0.0840
- BsmtFinSF2 0.0695

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?

- A model is robust and generalisable, if it does not demonstrate a significant change in the performance of the training and the test set i.e. it **performs almost equally good on the unseen (test) set as it does on the training dataset**.
- **A more generalized model is one that does not overfit** i.e. performs good on training dataset and the performance significantly reduces for the unseen dataset. In such case, training accuracy is significantly higher than the test accuracy.

- We use regularization to overcome the problem of overfitting. We **use hyperparameter to penalise the features and reduce the complexity of the model**. Choosing an optimal hyperparameter is important as penalising with very large value of hyperparameter can reduce the complexity to a large extent and can leads to underfitting . **The hyperparameter should be chosen so that the model should not overfit or underfit**.
- Therefore, **to make a model robust and generalizable, make the model simple but not simpler to an extent that it loses the accuracy (train and test both) and leads to underfitting.**
- If the model is robust and generalisable, **the accuracy of the model will be steady and will be less vulnerable to outliers or change in the test dataset** i.e. slight changes in the test dataset will not make much changes in the accuracy.



- **If the model is very simple, both the train and test accuracy is low and as we increase the model complexity the accuracy increases to an extent and on increasing the model complexity further, it leads to overfitting, the train accuracy increases and test accuracy decreases**

which loses the generalisation property of the model. The sweet spot is the optimal model complexity where both the accuracies are similar and the model is robust and generalisable.