**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?**

## Solution 1

Optimal Value of alpha:

- Ridge : 1.0
- Lasso : 0.0001
- If we double the value of alpha, the following impacts it will have on our model:
  - If we double the value of alpha it will penalize the curve more and model will be more complex.
  - In case of Lasso more coefficients will be 0.
  - There is a little reduction in r2 value and hence error increased in training and test data.

## Ridge

Before (Alpha = 1.0)

| Features | rfe_support | rfe_ranking | Coefficient |
|---|---|---|---|
| OverallQual | True | 1 | 0.2031 |
| 1stFlrSF | True | 1 | 0.1341 |
| OverallCond | True | 1 | 0.1173 |
| 2ndFlrSF | True | 1 | 0.1132 |
| TotalBsmtSF | True | 1 | 0.0879 |
| GarageArea | True | 1 | 0.0642 |
| MSZoning_FV | True | 1 | 0.0610 |
| BsmtQual | True | 1 | 0.0583 |
| MSZoning_RL | True | 1 | 0.0459 |
| Foundation_Slab | True | 1 | 0.0450 |

After (Alpha = 2.0)

| Features | rfe_support | rfe_ranking | Coefficient |
|---|---|---|---|
| OverallQual | True | 1 | 0.1873 |
| 1stFlrSF | True | 1 | 0.1292 |
| 2ndFlrSF | True | 1 | 0.1121 |
| OverallCond | True | 1 | 0.1116 |
| TotalBsmtSF | True | 1 | 0.0815 |
| GarageArea | True | 1 | 0.0657 |
| BsmtQual | True | 1 | 0.0548 |
| MSZoning_FV | True | 1 | 0.0491 |
| Foundation_Slab | True | 1 | 0.0392 |
| HeatingQC | True | 1 | 0.0386 |

## Lasso

Before (Alpha = 0.0001)

| Features | rfe_support | rfe_ranking | Coefficient |
|---|---|---|---|
| OverallQual | True | 1 | 0.231731 |
| 1stFlrSF | True | 1 | 0.145573 |
| OverallCond | True | 1 | 0.119130 |
| 2ndFlrSF | True | 1 | 0.115552 |
| TotalBsmtSF | True | 1 | 0.081355 |
| GarageArea | True | 1 | 0.064252 |
| BsmtQual | True | 1 | 0.051300 |
| LotArea | True | 1 | 0.036174 |
| KitchenQual | True | 1 | 0.035929 |
| Foundation_Slab | True | 1 | 0.035371 |

After (Alpha = 0.0002)

| Features | rfe_support | rfe_ranking | Coefficient |
|---|---|---|---|
| OverallQual | True | 1 | 0.235686 |
| 1stFlrSF | True | 1 | 0.127186 |
| OverallCond | True | 1 | 0.114835 |
| 2ndFlrSF | True | 1 | 0.097156 |
| GarageArea | True | 1 | 0.067440 |
| TotalBsmtSF | True | 1 | 0.060114 |
| BsmtQual | True | 1 | 0.041579 |
| KitchenQual | True | 1 | 0.038506 |
| HeatingQC | True | 1 | 0.033807 |
| LotArea | True | 1 | 0.032942 |

**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?**

**Solution 2**

- Hyperparameter Lambda:
  - Ridge : 1.0
  - Lasso : 0.0001
- R2 value on train data & test data for Ridge & Lasso
  - Train : Ridge(0.92), Lasso(0.91)
  - Test : Ridge(0.87), Lasso(0.87)
- MSE
  - Ridge : 0.002728
  - Lasso : 0.002730

We can see that in terms of accuracy, best performance is given by the Ridge Regression model (alpha = 1), but Lasso (alpha = 0.0001) is extremely close with an added advantage that it can reduce the number of features if required by fine tuning the alpha value which is not possible with Ridge. Hence, I will be using Lasso (with alpha 0.0001) in this case.

Another advantage of using Lasso is that we can bring down the number of predictor variables by increasing the alpha value without compromising too much on the error in the model. Using lower number of predictor variable without compromising too much on the errors is a great because it helps to keep the model simple.

**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?**

**Solution 3**

If the most 5 important variables are not present in the model in that case following 5 most important predictor variables are:

1) GarageArea
2) BsmtQual
3) LotArea
4) KitchenQual

5) Foundation Slab

**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?**

**Solution 4**

- Model should not be impacted by outliers. Model can be impacted while using test data and model and can fail on that, if too much focus is given to outliers during model building.
- The accuracy of training data and test data should be comparable. If the model performs well on training data but does not perform well on test data then it is a sign of overfitting and hence the model is not robust.
- Regularization is used to ensure that the model is resilient and generalizable. It penalizes the model if it becomes more complex.
- Bias – Variance is achieved with the help of Regularization. It increases the bias to a optimum position. Both variance and bias should be low for a good-fit model.