# Problem Statement - Part II

**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 of Alpha
  - Ridge : 3.0
  - Lasso : 100

- Doubling Alpha value
  - Ridge :
    Double Value : 6.0
    Observation :
    - R2 value reduced from 92.69% to 91.89%
    - RSS of Train data increased a little

| | Metric | Aplha=3 | Aplha=6 |
|---|---|---|---|
| 0 | R2 Score (Train) | 9.269326e-01 | 9.185938e-01 |
| 1 | R2 Score (Test) | 8.923161e-01 | 8.874845e-01 |
| 2 | RSS (Train) | 2.748163e+11 | 3.061799e+11 |
| 3 | RSS (Test) | 1.790547e+11 | 1.870887e+11 |
| 4 | MSE (Train) | 1.767177e+04 | 1.865293e+04 |
| 5 | MSE (Test) | 2.176442e+04 | 2.224733e+04 |

- Lasso :
  Double Value : 200
  Observation :
    - R2 value reduced from 91.83% to 90.60%
    - RSS of Train data increased a little

| | Metric | Aplha=100 | Aplha=200 |
|---|---|---|---|
| 0 | Train R2 Score | 9.183758e-01 | 9.060232e-01 |
| 1 | Test R2 Score | 9.055086e-01 | 8.978563e-01 |
| 2 | Train RSS | 3.069996e+11 | 3.534595e+11 |
| 3 | Test RSS | 1.571186e+11 | 1.698426e+11 |
| 4 | Train MSE | 1.867788e+04 | 2.004142e+04 |
| 5 | Test MSE | 2.038768e+04 | 2.119715e+04 |

- Most important predictor variables:

  - Ridge :
    Even after doubling the value of alpha top 5 predictor variables remained the same.

    GrLivArea, OverallQual, BsmtFinSF1, TotalBsmtSF, 2ndFlrSF

| | Ridge |
|---|---|
| GrLivArea | 80065.656814 |
| OverallQual | 51498.783365 |
| BsmtFinSF1 | 45229.489226 |
| TotalBsmtSF | 44475.307495 |
| 2ndFlrSF | 35215.839351 |
| OverallCond | 29311.311572 |
| GarageArea | 28820.050507 |
| Neighborhood_StoneBr | 22649.638400 |
| Fireplaces | 21694.203004 |
| BsmtUnfSF | 19502.618612 |

In [108]:
```
# Ridge with double alpha value
betas = pd.DataFrame(index=X_train.columns)
betas.rows = X_train.columns
betas['Ridge Doubled'] = double_ridge.coef_
betas.sort_values(by=['Ridge Doubled'], ascending=False).head(10)
```

Out[108]:

| | Ridge Doubled |
|---|---|
| GrLivArea | 64276.559169 |
| OverallQual | 47470.779815 |
| BsmtFinSF1 | 40229.743688 |
| TotalBsmtSF | 39102.890546 |
| 2ndFlrSF | 30852.751332 |
| GarageArea | 27428.544643 |
| OverallCond | 25021.589227 |
| Neighborhood_StoneBr | 19532.328532 |
| Fireplaces | 18921.895523 |

- Lasso :
  - Even after doubling the value of alpha top 5 predictor variables remained the same.

    GrLivArea, OverallQual, BsmtFinSF1, TotalBsmtSF, OverallCond

| | Lasso |
|---|---|
| GrLivArea | 147390.030211 |
| OverallQual | 70499.795471 |
| TotalBsmtSF | 58487.851017 |
| OverallCond | 33672.600417 |
| BsmtFinSF1 | 30696.626911 |
| GarageArea | 27109.385078 |
| SaleType_New | 21563.122267 |
| Neighborhood_StoneBr | 19756.002698 |
| Neighborhood_Crawfor | 19205.211785 |
| BsmtExposure_Gd | 15433.762432 |

In [110]:
```
# Lasso with double alpha value
betas = pd.DataFrame(index=X_train.columns)
betas.rows = X_train.columns
betas['Lasso Doubled'] = double_lasso.coef_
betas.sort_values(by=['Lasso Doubled'], ascending=False).head(10)
```

Out[110]:

| | Lasso Doubled |
|---|---|
| GrLivArea | 148501.534617 |
| OverallQual | 80879.639998 |
| TotalBsmtSF | 49654.761389 |
| BsmtFinSF1 | 33342.543369 |
| OverallCond | 30243.585788 |
| GarageArea | 27602.268012 |
| SaleType_New | 20936.057504 |
| Neighborhood_Crawfor | 15410.699555 |
| BsmtExposure_Gd | 14767.665408 |

## 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?

| | Metric | Linear Regression | Ridge Regression | Lasso Regression |
|---|---|---|---|---|
| 0 | Train R2 Score | 9.417435e-01 | 9.269326e-01 | 9.183758e-01 |
| 1 | Test R2 Score | -3.415023e+17 | 8.923161e-01 | 9.055086e-01 |
| 2 | Train RSS | 2.191108e+11 | 2.748163e+11 | 3.069996e+11 |
| 3 | Test RSS | 5.678436e+29 | 1.790547e+11 | 1.571186e+11 |
| 4 | Train MSE | 1.577940e+04 | 1.767177e+04 | 1.867788e+04 |
| 5 | Test MSE | 3.875863e+13 | 2.176442e+04 | 2.038768e+04 |

- Looking at the above Metrics R2 value of Train is slightly more for Ridge but on Test data Lasso performed well and it is more accurate
- MSE Ridge has less value than Lasso but Lasso is more accurate
- Considering Lasso accuracy I prefer using Lasso Regression

**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?

- Dropped top 5 predictor variable in Lasso :

  'GrLivArea','OverallQual','TotalBsmtSF','OverallCond','BsmtFinSF1'

- After dropping 5 top most important predictor variable for Lasso are

  'GarageCond_TA', 'GarageCond_Fa', 'GarageCond_Po', 'GarageCond_Gd', 'LotArea'

**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 generalizable when it's not too complex and avoid overfitting . Simple model should have low variance and high bias. A robust and generalisable model will perform equally well on both training and testing data . i.e accuracy doesn't change much for training and test data.

- When the model is simpler because it has low variance, the variance in its output on test data with respect to training data will also be less.

- Model with high bias will have high accuracy on future test data