# Housing Price Prediction-Subjective Assignment

**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:**

a) Optimal value of alpha for ridge regression is 2 with below metric values:

```
R2 Score (Train) --> 0.9535576912085453
R2 Score (Test)  --> 0.916283384936825
RSS (Train)      --> 0.4794511490030525
RSS (Test)       --> 0.3079957427696738
MSE (Train)      --> 0.00047706581990353486
MSE (Test)       --> 0.0007129531082631338
```

Top 5 important predictor variables with coeff. values are as below.

| Variable | Coeff |
|---|---|
| constant | 0.101 |
| GrLivArea | 0.090 |
| 1stFlrSF | 0.081 |
| BsmtFinSF1 | 0.073 |
| LotArea | 0.051 |
| OverallQual_9 | 0.040 |

b) Optimal value of alpha for lasso regression is 0.0001 with below metric values:

```
R2 Score (Train) --> 0.9463610016510096
R2 Score (Test)  --> 0.9228034224070627
RSS (Train)      --> 0.5537467894905397
RSS (Test)       --> 0.28400834454512125
MSE (Train)      --> 0.0005509918303388454
MSE (Test)       --> 0.0006574267234840769
```

Top 5 important predictor variables with coeff. values are as below.

| Variable | Coeff |
|---|---|
| GrLivArea | 0.215 |
| BsmtFinSF1 | 0.074 |
| OverallQual_10 | 0.067 |
| LotArea | 0.067 |
| OverallQual_9 | 0.065 |
| constant | 0.042 |

c) If alpha is doubled from optimal value of 2 to 4 for ridge regression then below is changed metric values. R2 score of train dataset is reduced and RSS of train is increased.

```
R2 Score (Train) --> 0.9493674492892745
R2 Score (Test)  --> 0.9159292271599025
RSS (Train)      --> 0.5227094700270232
RSS (Test)       --> 0.30929869902845925
MSE (Train)      --> 0.000520108925400023
MSE (Test)       --> 0.000715969210714026
```

Top 5 important predictor variables remain same. Only Coeff values are changed.

| Variable | Coeff |
|---|---|
| constant | 0.135 |
| GrLivArea | 0.077 |
| 1stFlrSF | 0.069 |
| BsmtFinSF1 | 0.061 |
| OverallQual_9 | 0.039 |
| LotArea | 0.039 |

d) If alpha is doubled from optimal value of 0.0001 to 0.0002 for lasso regression then below is changed metric values. R2 score of train dataset is reduced and RSS of train is increased.

```
R2 Score (Train) --> 0.9386236397966589
R2 Score (Test)  --> 0.9206141186137901
RSS (Train)      --> 0.6336241067009916
RSS (Test)       --> 0.2920628537658852
MSE (Train)      --> 0.0006304717479611857
MSE (Test)       --> 0.0006760714207543638
```

Top 5 important predictor variables remain same. Only Coeff values are changed.

| Variable | Coeff |
|---|---|
| GrLivArea | 0.217 |
| OverallQual_9 | 0.065 |
| OverallQual_10 | 0.063 |
| BsmtFinSF1 | 0.058 |
| constant | 0.052 |
| LotArea | 0.046 |

RSS= residual sum of squared error

MSE= mean squared error

**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?

**Ans:** Below is comparison of metric of ridge and lasso regression:

| | Metric | Linear Regression | Ridge Regression | Lasso Regression |
|---|---|---|---|---|
| 0 | R2 Score (Train) | 9.601487e-01 | 0.953558 | 0.946361 |
| 1 | R2 Score (Test) | -1.034406e+18 | 0.916283 | 0.922803 |
| 2 | RSS (Train) | 4.114088e-01 | 0.479451 | 0.553747 |
| 3 | RSS (Test) | 3.805609e+18 | 0.307996 | 0.284008 |
| 4 | MSE (Train) | 2.023270e-02 | 0.021842 | 0.023473 |
| 5 | MSE (Test) | 9.385777e+07 | 0.026701 | 0.025640 |

The model performance by Lasso Regression is better than Ridge Regression in terms of R2 values of Test.

Also, it is better to use Lasso, since it assigns a zero value to insignificant features, enabling us to choose only important predictive variables.

It is always advisable to use simple but robust model.

**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?

**Ans:** Creating new lasso model after dropping top 5 important predictor variables i.e. 'GrLivArea', 'BsmtFinSF1', 'OverallQual_10', 'LotArea' and 'OverallQual_9'.

R2 score of test dataset is reduced and RSS of train & test is increased.

```
R2 Score (Train) --> 0.9410267567033125
R2 Score (Test)  --> 0.9076354495478678
RSS (Train)      --> 0.6088153236739144
RSS (Test)       --> 0.3398117363037656
MSE (Train)      --> 0.0006057863917153378
MSE (Test)       --> 0.0007866012414439018
```

Below are new top 5 important predictor variables with coeff values:

| Variable | Coeff |
|---|---|
| BsmtFinSF2 | 0.264 |
| BsmtUnfSF | 0.110 |
| constant | 0.089 |
| Neighborhood_Sawyer | 0.040 |
| RoofMatl_Membran | 0.034 |

**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?

**Ans:** To make sure that a model is robust and generalizable, model should be simple. Its accuracy will decrease. We can understand using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in the terms of accuracy that model will perform equally well on the both train and test data i.e. very less change in accuracy for train and test data.