# Advanced Regression Subjective Questions



**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:**

- As per the models built using Cross validation, Optimal values determined for the given dataset are as follows.
    - Ridge - **6**
    - Alpha - **0.0004**
- Below are the key changes we observed when we double the value of alpha for both ridge and Lasso
    - Doubling the value of alpha has shrinked the coefficients further as the penalty imposed has doubled.
    - This caused **slight underfitting of data** as the train and test scores got reduced.
    - In case of Ridge, it was only slight reduction in the scores but where as in case of Lasso, there was a significant decrease in train and test scores after doubling the value of alpha from optimal one.
    - In case of Lasso, more number of variables got their coefficients pushed to zero.
    - Even the order of important predictor variables got slight changes after doubling the value of alpha
    - Below are the key metrics:
        - In case of Ridge, train score got reduced from 0.87 to 0.86
        - In case of Lasso, train score got reduced from 0.88 to 0.84, where as test score got reduced from 0.87 to 0.86
        - Number of predictor variables used in Lasso model got reduced from 65 to 51
    - Important predictor variables after doubling the value of alpha
        - **Ridge** - `OverallQual` , `GrLivArea` , `GarageCars` , `1stFlrSF` , `2ndFlrSF`
        - **Lasso** - `GrLivArea` , `OverallQual` , `GarageCars` , `TotalBsmtSF` , `Fireplaces`

Out[101]:

| | Feature Variable | Ridge Coefficients Double Alpha |
|---|---|---|
| 1 | OverallQual | 0.099463 |
| 11 | GrLivArea | 0.067573 |
| 20 | GarageCars | 0.063899 |
| 8 | 1stFlrSF | 0.062429 |
| 9 | 2ndFlrSF | 0.060980 |

In [102]:
```
# Top five coefficients as per Lasso after doubling the alpha
double_alpha_model_coef_df.sort_values(by='Lasso Coefficients Double Alpha', ascending=False).head()[['Feature Variable','Lasso C
```

Out[102]:

| | Feature Variable | Lasso Coefficients Double Alpha |
|---|---|---|
| 11 | GrLivArea | 0.239859 |
| 1 | OverallQual | 0.215907 |
| 20 | GarageCars | 0.102273 |
| 7 | TotalBsmtSF | 0.088664 |
| 19 | Fireplaces | 0.048575 |

In [103]:
```
print('Number of predictor variables used by Lasso (after excluding ones with zero coefficients): ', len(double_alpha_model_coef_
```

```
Number of predictor variables used by Lasso (after excluding ones with zero coefficients):  51
```

**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?**

**Answer:**

- As per the models built using Cross validation, Optimal values determined for the given dataset are as follows.
  - Ridge - **6**
  - Alpha - **0.0004**
- We decided to go ahead with Lasso regression for the following reasons
  - Our dataset has large number of predictor variables (after performing encoding of categorical variables etc.), this makes the model **complex and overfit the data** as the effect of noisy variables will also get added to the model
  - Unlike ridge, Lasso actually pushes coefficients of some of the predictor variables to actually become zero which have less influence on the target variable essentially performing **feature selection**.
  - Even in our case, lasso model with optimal alpha was using only 65 predictor variables which makes the model little bit simpler.
  - Even though training score of Lasso is little bit less when compared to Ridge, it performed better on the test data.
  - So we decided to go with the Lasso model with optimal value of alpha as 0.0004

**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?**

**Answer:**

- Five most important predictor variables in the Lasso model with optimal value of alpha (0.0004 in our case) are `GrLivArea` , `OverallQual` , `TotalBsmtSF` , `GarageCars` **and** `OverallCond`

- After dropping the above and building the model again, five most important predictor variables are `1stFlrSF`, `2ndFlrSF`, `GarageArea`, `BsmtFinSF1` **and** `Neighborhood_NridgHt`

**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?**

**Answer:**

- As the model becomes more and more complex, it causes overfitting of the data. Because of which it might perform well on the data, but performs poorly on the unseen data.
- Few possible reasons could be adding more and more predictor variables to the model, using more complex models like polynomial regression of higher degrees even when it is not needed etc.
- Because of this, variance in the model is too high such that even for a small change in the data, there might be a significant impact on the performance of the model
- To make the model more robust and generalisable, we need to make the model simple enough such that it performs well and at the same time doesn't underfit the data.
- Implications of making the model simpler could be slight reduction in accuracy of the model. There might be reduction in the training data, but it would perform well on the unseen data.
- This is generally called as Bias Variance trade off.
- We will compromise little bit on the accuracy front to make the model less variant and make it more robust and generalisable