## 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:

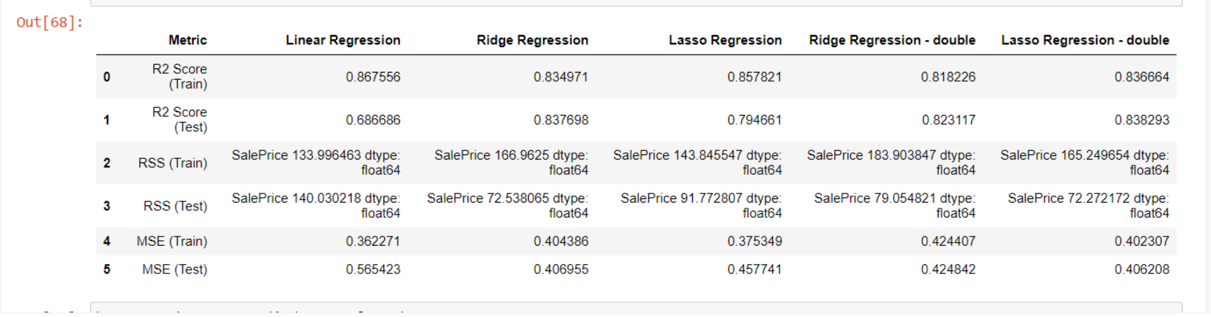
Optimal Value as per model building:

- Ridge: 500
- Lasso: 0.1

Before the alpha value was doubled, there is a difference between R2 values of training and test data indicating the overfitting in case of Lasso.

When the alpha values doubled for both Ridge and Lasso,

1. R2 value suffered a bit for Ridge. However, difference in R2 values of test and training data for Lasso went away and thus indicating a better fit for Lasso with alpha = 0.2.
2. In Lasso regression, more predictors are reduced to zero. Approx 35 with alpha=0.2 as compared to 21 with alpha = 0.1
3. Beta coefficients values also changed.



Out[68]:

| | Metric | Linear Regression | Ridge Regression | Lasso Regression | Ridge Regression - double | Lasso Regression - double |
|---|---|---|---|---|---|---|
| 0 | R2 Score (Train) | 0.867556 | 0.834971 | 0.857821 | 0.818226 | 0.836664 |
| 1 | R2 Score (Test) | 0.686686 | 0.837698 | 0.794661 | 0.823117 | 0.838293 |
| 2 | RSS (Train) | SalePrice 133.996463 dtype: float64 | SalePrice 166.9625 dtype: float64 | SalePrice 143.845547 dtype: float64 | SalePrice 183.903847 dtype: float64 | SalePrice 165.249654 dtype: float64 |
| 3 | RSS (Test) | SalePrice 140.030218 dtype: float64 | SalePrice 72.538065 dtype: float64 | SalePrice 91.772807 dtype: float64 | SalePrice 79.054821 dtype: float64 | SalePrice 72.272172 dtype: float64 |
| 4 | MSE (Train) | 0.362271 | 0.404386 | 0.375349 | 0.424407 | 0.402307 |
| 5 | MSE (Test) | 0.565423 | 0.406955 | 0.457741 | 0.424842 | 0.406208 |

## The important predictor variables having non-zero beta coeff:

```
'LotArea', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1',
       '1stFlrSF', 'LowQualFinSF', 'GrLivArea', 'BsmtFullBath', 'FullBa
th',
       'Fireplaces', 'GarageCars', 'WoodDeckSF', 'PoolArea', 'MSSubClas
s',
       'Street', 'Alley', 'LotShape', 'LandContour', 'Utilities', 'Land
Slope',
       'Neighborhood', 'Condition2', 'OverallQual', 'OverallCond', 'Roo
fMatl',
       'Exterior2nd', 'ExterQual', 'Foundation', 'BsmtQual', 'BsmtCond'
,
       'BsmtExposure', 'BsmtFinType1', 'HeatingQC', 'KitchenQual',
       'Functional', 'FireplaceQu', 'GarageFinish', 'GarageQual', 'Gara
geCond',
       'PoolQC', 'SaleCondition'
```

## The predictor variables having significant beta coeff:

```
'LotArea', 'YearBuilt', 'YearRemodAdd', 'MasVnrArea', 'BsmtFinSF1',
       '1stFlrSF', 'GrLivArea', 'BsmtFullBath', 'FullBath', 'Fireplaces
',
       'GarageCars', 'WoodDeckSF', 'PoolArea', 'MSSubClass', 'LandConto
ur',
       'LandSlope', 'Neighborhood', 'OverallQual', 'OverallCond', 'Roof
Matl',
       'ExterQual', 'BsmtQual', 'BsmtExposure', 'BsmtFinType1', 'Kitche
nQual',
       'Functional', 'FireplaceQu', 'PoolQC'
```

**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 optimal values of lambda, lasso regression should be considered due to:

1. Many variables are reduced to zero and hence have limited predictor variables and thus complexity is reduced.
2. Ridge regression has very high value of lambda suggesting that high degree of regularization happened and thus the high penalty. This high value of lambda also causes high bias.

**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:**

1. **1stFlrSF**
2. **2ndFlrSF**
3. **GarageCars**
4. **ExterQual**
5. **PoolArea**

**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:**

To make sure that a model is robust and generalizable:

1. Pre-process the data like clean, transform, scale etc.
2. Use ridge and lasso techniques to regularize the model by adding the penalty. This penalty makes the model robust by compromising the bias slightly. It also resolves the problem of over-fitting due to reduced variance.
3. The alpha/lambda values should be selected as such so that the model does not become under fit nor it should become overfit.
4. After creating the model, the r2 values of training and test data should not have much difference and should have high values.

By compromising a bit on bias by applying the penalty to the model such as in ridge and lasso, the variance in the model is highly reduced and thus increasing the accuracy of the model. The low variance model makes sure that the model generalizes well and provide more consistent predictions.