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: The optimal value of alpha for RIDGE and LASSO regression is the value obtained after tuning the hyper parameter for minimum total error duly penalizing the Regression Model coefficients for any over fitting in the Model. In the present assignment I could find the most **optimal alpha value** for LASSO and RIDGE as under:

RIDGE- 20
LASSO- 100

In case we choose the double the value of the alpha, we can see reduction in the r2 score values for RIDGE and LASSO as given under:

With alpha=20 (RIDGE) and 100 (LASSO)

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.747831e-01	8.743218e-01	8.744435e-01
1	R2 Score (Test)	8.589506e-01	8.596567e-01	8.622748e-01
2	RSS (Train)	2.547692e+11	2.557078e+11	2.554603e+11
3	RSS (Test)	1.259535e+11	1.253229e+11	1.229850e+11
4	MSE (Train)	1.725189e+04	1.728364e+04	1.727527e+04
5	MSE (Test)	1.850040e+04	1.845403e+04	1.828109e+04

With alpha=40 (RIDGE) and 200 (LASSO)

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	8.747831e-01	8.732476e-01	8.734432e-01
1	R2 Score (Test)	8.589506e-01	8.596577e-01	8.647893e-01
2	RSS (Train)	2.547692e+11	2.578934e+11	2.574955e+11
3	RSS (Test)	1.259535e+11	1.253220e+11	1.207397e+11
4	MSE (Train)	1.725189e+04	1.735734e+04	1.734395e+04
5	MSE (Test)	1.850040e+04	1.845397e+04	1.811344e+04

The most important predictor variables have been as under and there is no change in the major predictor variables even after doubling the alpha values:
GrLivArea, GarageArea, OverallQual, TotalBsmtSF, GarageCars

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: Once Regularisation Regression analysis was carried out on Multiple Linear regression Model output, it was done by applying various lambda values using Grid Search function with Cross Validation for evaluating the model fit and output performance for such lambda values. An optimum or best lambda value is recommended by the function; theoretically speaking gives the value at which total error is minimum for the output and variance is optimum resulting is efficient model interpretation of the input variations.

See following example for RIDGE as implemented in the assignment for better understanding.

Ridge Regression

In [128]:
```
# list of alphas has been selected here to tune the hyperparameters- if value too high it will lead to underfitting,
# if it is too low,it will not handle the overfitting
params = {'alpha': [0.00001, 0.0001, 0.001, 0.01, 0.05, 0.1,
 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0,
 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000 ]}

ridge = Ridge()

# cross validation
folds = 5
model_cv = GridSearchCV(estimator = ridge,
                        param_grid = params,
                        scoring= 'neg_mean_absolute_error',
                        cv = folds,
                        return_train_score=True,
                        verbose = 1)
model_cv.fit(X_train_new, y_train)
```

```
Fitting 5 folds for each of 29 candidates, totalling 145 fits
```

Out[128]:
```
GridSearchCV(cv=5, estimator=Ridge(),
             param_grid={'alpha': [1e-05, 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2,
                                   0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0,
                                   3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20,
                                   50, 100, 500, 1000]},
             return_train_score=True, scoring='neg_mean_absolute_error',
             verbose=1)
```

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: Considering the LASSO methodology and concept, model coefficient values can become zero at points LASSO penalty output on beta axis touches the output error or residual value contours. In case we find that most important predictor variable beta coefficients have become zero, then it would be better to analyze the feature engineering carried out at the Regression Model building stage for ensuring that all significant predictor variables are part of the final model for good accuracy of prediction. In the present assignment, all major predictor variables are part of final LASSO regularized model as well, so this possibility has not arises, however five most important predictor variables in the assignment are as under:

GrLivArea, GarageArea, OverallQual, TotalBsmtSF, GarageCars

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: We can ensure that Model is generisable and robust by ensuring that feature engineering has been done carefully to ensure all predictor variables with significant impact on the response variable are part of the modeling process. This will definitely help us in creating a simple and robust model with good performance on train as well test data sets having very good model accuracy to be evaluated for performance using metrics like R2 score, RMSE etc.

Once a Regression model is trained and tested for good accuracy, we can always use Regularisation techniques of RIDGE or LASSO (as applicable) for improving the robustness and performance with simplification of model complexity. Hyperparameter tuning needs to be carried out for optimal value of alpha during the regularization process, which results in shrinkage of model parameters or simplification of the model, which would always be more generelisable.

Model accuracy on test data increases with the application of Regularisation techniques, as over fitting effect is reduced on the model otherwise severely affecting the Model accuracy on test data. Genralisation of the model is not carried out py paying high cost of the reduction in model accuracy, however some accuracy can always be compromised for the sake of better interpretability by the model with better variance characterstics.