House Price Assignment - Part II

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: Optimal Value of alpha for Ridge Regression is 0.6 and for Lasso Regression it is 0.0001 .

When we double the value of alpha as Ridge Regression = 1.2 and Lasso Regression = 0.0002 , which means we are doubling the penalty terms and penalizing more on the coefficients .

Observations as per below table with Ridge Regression = 1.2 and Lasso Regression = 0.0002,

- Observed lesser Coefficient values with Higher alpha value. This means coefficients are penalized thus their magnitude is less.
- More number of coefficients are reduced to absolute 0 in Lasso Regression model

Below is the list of Predictors after doubling the alpha for Ridge Regression = 1.2 and Lasso Regression = 0.0002

	Ridge (alpha = 0.6)	Lasso (alpha = 0.0001)
MSSubClass	-0.053894	-0.051807
LotFrontage	-0.025400	-0.007462
LotArea	0.117785	0.091775
OverallQual	0.148582	0.156275
OverallCond	0.065891	0.058091
YearBuilt	0.057359	0.059882
MasVnrArea	0.030416	0.024733
ExterQual	0.046119	0.048320
BsmtQual	0.054613	0.041814
BsmtCond	-0.040763	-0.027131
BsmtExposure	0.046161	0.049292
BsmtFinSF1	0.089013	0.072495
BsmtFinSF2	0.024981	0.012800
TotalBsmtSF	0.009063	-0.000000
1stFlrSF	0.180878	0.008185
2ndFlrSF	0.090097	0.000000
LowQualFinSF	-0.010113	-0.017907
GrLivArea	0.184903	0.402643
BedroomAbvGr	-0.026070	-0.023151
KitchenAbvGr	-0.011422	-0.001951
GarageArea	0.063896	0.051908
PoolArea	-0.023422	-0.013028
MiscVal	0.003042	-0.000000
Street_Pave	0.026843	0.001078
Neighborhood_Crawfor	0.039774	0.037128
Neighborhood_NoRidge	0.068666	0.068086
Neighborhood_NridgHt	0.052955	0.052931
Neighborhood_StoneBr	0.042383	0.035959
Condition2_PosN	-0.299413	-0.370036
RoofMatl_CompShg	0.132491	0.046188
RoofMatl_Membran	0.105452	0.000000
RoofMatl_Metal	0.101014	0.000000
RoofMatl_Roll	0.075776	0.000000
RoofMatl_Tar&Grv	0.116422	0.013851
RoofMatl_WdShake	0.095367	0.000000
RoofMatl_WdShngl	0.214986	0.141227
Exterior1st_CemntBd	-0.021777	0.000000
Exterior2nd_CmentBd	0.034290	0.015238
Heating_OthW	-0.051424	-0.000000
Heating_Wall	0.036627	0.008121
Functional_Mod	-0.016429	-0.000000
Functional_Sev	-0.070616	-0.000000
GarageType_No_Garage	0.016327	0.007630

	Ridge (alpha = 1.2)	Lasso (alpha = 0.0002)
MSSubClass	-0.053551	-0.048251
LotFrontage	-0.020855	-0.000000
LotArea	0.097262	0.046853
OverallQual	0.148515	0.165604
OverallCond	0.062673	0.045377
YearBuilt	0.056844	0.057587
MasVnrArea	0.030235	0.018146
ExterQual	0.049202	0.051245
BsmtQual	0.053465	0.032550
BsmtCond	-0.039982	-0.013806
BsmtExposure	0.048021	0.051643
BsmtFinSF1	0.077327	0.058988
BsmtFinSF2	0.024309	0.003489
TotalBsmtSF	0.021238	0.000000
1stFlrSF	0.159466	0.013844
2ndFlrSF	0.087306	0.000000
LowQualFinSF	-0.008004	-0.002846
GrLivArea	0.166230	0.363530
BedroomAbvGr	-0.015176	-0.000000
KitchenAbvGr	-0.008391	-0.000000
GarageArea	0.067638	0.051681
PoolArea	-0.021542	-0.000000
MiscVal	0.001045	0.000000
Street_Pave	0.021434	0.000000
Neighborhood_Crawfor	0.041708	0.036180
Neighborhood_NoRidge	0.071957	0.070573
Neighborhood_NridgHt	0.054273	0.053654
Neighborhood_StoneBr	0.043019	0.030439
Condition2_PosN	-0.207127	-0.234448
RoofMatl_CompShg	0.072728	0.018340
RoofMatl_Membran	0.049326	0.000000
RoofMatl_Metal	0.046581	0.000000
RoofMatl_Roll	0.027165	-0.000000
RoofMatl_Tar&Grv	0.055957	0.000000
RoofMatl_WdShake	0.044001	0.000000
RoofMatl_WdShngl	0.149469	0.099556
Exterior1st_CemntBd	-0.014057	0.000000
Exterior2nd_CmentBd	0.025481	0.011259
Heating_OthW	-0.035597	-0.000000
Heating_Wall	0.030900	0.000000
Functional_Mod	-0.012400	-0.000000
Functional_Sev	-0.048162	-0.000000
GarageType_No_Garage	0.014893	0.002065

Most important predictor variables after doubling the alpha value :

	Ridge (alpha = 1.2)
GrLivArea	0.166230
1stFlrSF	0.159466
RoofMatl_WdShngl	0.149469
OverallQual	0.148515
LotArea	0.097262
2ndFlrSF	0.087306
BsmtFinSF1	0.077327
RoofMatl_CompShg	0.072728
Neighborhood_NoRidge	0.071957
GarageArea	0.067638
OverallCond	0.062673
YearBuilt	0.056844
RoofMatl_Tar&Grv	0.055957
Neighborhood_NridgHt	0.054273
BsmtQual	0.053465
RoofMatl_Membran	0.049326
ExterQual	0.049202
BsmtExposure	0.048021
RoofMatl_Metal	0.046581
SaleType_Con	0.045245

	Lasso (alpha = 0.0002)
GrLivArea	0.363530
OverallQual	0.165604
RoofMatl_WdShngl	0.099556
Neighborhood_NoRidge	0.070573
BsmtFinSF1	0.058988
YearBuilt	0.057587
Neighborhood_NridgHt	0.053654
GarageArea	0.051681
BsmtExposure	0.051643
ExterQual	0.051245
LotArea	0.046853
OverallCond	0.045377
Neighborhood_Crawfor	0.036180
BsmtQual	0.032550
Neighborhood_StoneBr	0.030439
RoofMatl_CompShg	0.018340
MasVnrArea	0.018146
SaleType_New	0.016926
1stFlrSF	0.013844
Exterior2nd_CmentBd	0.011259

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: Decision of which regression model to use was done based on R2 Score and RSME value. Below is the R2 Score and RSME values for Ridge and Lasso Regressions obtained with their respective optimal alpha values.

```
===== Ridge Regression (alpha = 0.6) =====

[('R2 (Train)', 0.8717663857389981),
 ('R2 (Test)', 0.8547549145119415),
 ('RMSE (Train)', 0.03931595025937328),
 ('RMSE (Test)', 0.04246103832050107)]
```

```
===== Lasso Regression (alpha = 0.0001) =====

[('R2 (Train)', 0.8609481670945432),
 ('R2 (Test)', 0.8461708658416498),
 ('RMSE (Train)', 0.04094078771408868),
 ('RMSE (Test)', 0.04369776094806514)]
```

Summary of Ridge and Lasso Regression –

- Ridge Regression Model - Observed R2 score of 0.871 for train dataset and 0.854 for test dataset which means this model explained about 85.4% of the variation in the Test Data . Also, Train and Test R2 Scores are decent enough for this model.
- Lasso Regression Model - Observed R2 score of 0.860 for train dataset and 0.846 for test dataset which means this model explained about 84.6% of the variation in the Test Data. Also Train and Test R2 Scores are decent enough for this model.

Now, as we observe for both the models, R2 Scores of Train and Test dataset are decent.

To decide on the Final Model we will consider "BETTER" R2 Scores on the test dataset and as we see, R2 Score of Test dataset for Ridge Regression model (0.8547) is better than Lasso Regression Model (0.8461)

So, we choose Ridge Regression model as our final model.

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: Current Model's 5 most important predictor variables in Lasso Model are → **GrLivArea, OverallQual, RoofMatl_WdShngl, LotArea, BsmtFinSF1.**

Now, as mentioned in the question, above predictors are not available in the incoming data and we need to create a new model without them.

In this case, we remove above 5 predictors from the dataset and recreate the Lasso Model to get new 5 most important predictor variables → **1stFlrSF**, **2ndFlrSF**, **Neighborhood_NoRidge**, **BsmtQual**, **Neighborhood_NridgHt**

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: For a model to be robust and generalisable,

- It should not be complex.
- Model should perform well on train dataset as well as test dataset. Metrics can be R2 scores and RMSE.
- Find a Bias-Variance trade-off to avoid model overfitting and underfitting.

Above requirements can be met using Regularization methods along with existing data cleaning methods like outlier treatments ..etc

With Regularization, complexity of model is reduced and methods like Hyperparameter tuning and Cross Validation helps to find the optimal value of lambda needed to find the Bais-Variance -trade-off best fit values. With the small compromise in the bias there is a significant reduction in variance.