

### 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:

The optimal value of alpha for Ridge is 2 and for Lasso it is 0.001.

With these alphas the R2 of the model was approximately 0.83.

After doubling the alpha values in the Ridge and Lasso, the prediction accuracy remains around 0.82 but there is a small change in the co-efficient values. The new model is created and demonstrated in the Jupiter notebook. Below are the changes in the co-efficients.

- Ridge Regression Model

Ridge Co-Efficient		Ridge Doubled Alpha Co-Efficient	
Total_sqr_footage	0.169122	Total_sqr_footage	0.149028
GarageArea	0.101585	GarageArea	0.091803
TotRmsAbvGrd	0.067348	TotRmsAbvGrd	0.068283
OverallCond	0.047652	OverallCond	0.043303
LotArea	0.043941	LotArea	0.038824
CentralAir_Y	0.032034	Total_porch_sf	0.033870
LotFrontage	0.031772	CentralAir_Y	0.031832
Total_porch_sf	0.031639	LotFrontage	0.027526
Neighborhood_StoneBr	0.029093	Neighborhood_StoneBr	0.026581
Alley_Pave	0.024270	OpenPorchSF	0.022713
OpenPorchSF	0.023148	MSSubClass_70	0.022189
MSSubClass_70	0.022995	Alley_Pave	0.021672
RoofMatl_WdShngl	0.022586	Neighborhood_Veenker	0.020098
Neighborhood_Veenker	0.022410	BsmtQual_Ex	0.019949
SaleType_Con	0.022293	KitchenQual_Ex	0.019787
HouseStyle_2.5Unf	0.021873	HouseStyle_2.5Unf	0.018952
PavedDrive_P	0.020160	MasVnrType_Stone	0.018388
KitchenQual_Ex	0.019378	PavedDrive_P	0.017973
LandContour_HLS	0.018595	RoofMatl_WdShngl	0.017856
SaleType_Oth	0.018123	PavedDrive_Y	0.016840

- Lasso Regression Model

Lasso Co-Efficient		Lasso Doubled Alpha Co-Efficient	
Total_sqr_footage	0.202244	Total_sqr_footage	0.204642
GarageArea	0.110863	GarageArea	0.103822
TotRmsAbvGrd	0.063161	TotRmsAbvGrd	0.064902
OverallCond	0.046686	OverallCond	0.042168
LotArea	0.044597	CentralAir_Y	0.033113
CentralAir_Y	0.033294	Total_porch_sf	0.030659
Total_porch_sf	0.028923	LotArea	0.025909
Neighborhood_StoneBr	0.023370	BsmtQual_Ex	0.018128
Alley_Pave	0.020848	Neighborhood_StoneBr	0.017152
OpenPorchSF	0.020776	Alley_Pave	0.016628
MSSubClass_70	0.018898	OpenPorchSF	0.016490
LandContour_HLS	0.017279	KitchenQual_Ex	0.016359
KitchenQual_Ex	0.016795	LandContour_HLS	0.014793
BsmtQual_Ex	0.016710	MSSubClass_70	0.014495
Condition1_Norm	0.015551	MasVnrType_Stone	0.013292
Neighborhood_Veenker	0.014707	Condition1_Norm	0.012674
MasVnrType_Stone	0.014389	BsmtCond_TA	0.011677
PavedDrive_P	0.013578	SaleCondition_Partial	0.011236
LotFrontage	0.013377	LotConfig_CulDSac	0.008776
PavedDrive_Y	0.012363	PavedDrive_Y	0.008685

Overall since the alpha values are small, we do not see a huge change in the model after doubling the alpha

## 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:-

- The optimal lambda value in case of :

Ridge

```

Ridge(alpha=2, random_state=100)

```

Lasso:

```

Lasso(alpha=0.0001, random_state=100)

```

- The MSE Value in case of :  
Ridge: 0.0018396090787924247  
Lasso: 0.0018634152629407766
- The Mean Squared Error of both the models are almost same.
- Since Lasso helps in feature reduction (as the coefficient value of some of the features become zero), Lasso has a better edge over Ridge and should be used as the final 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?

Answer:

The five most important are:

1. Total\_sqr\_footage
2. GarageArea
3. TotalRmsAbvGrd
4. OverallCond
5. LotArea

After building Lasso Model in Jupiter notebook after removing these attributes from the data set

We get below result and the new Five Predictor variables

The R2 Score of the model on the test dataset is 0.7330077964268464  
The MSE of the model on the test dataset is 0.0028575670906482546  
The most important predictor variables are as follows:

:

	Lasso Co-Efficient
LotFrontage	0.146535
Total_porch_sf	0.072445
HouseStyle_2.5Unf	0.062900
HouseStyle_2.5Fin	0.050487
Neighborhood_Veenker	0.042532

**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:

The model should be generalized so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much importance should not be given to the outliers so that the accuracy predicted by the model is high. To ensure that this is not the case, the outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. If the model is not robust, it cannot be trusted for predictive analysis.