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 for Ridge Regression = **10**

The optimal value for Lasso Regression = **0.0002**

Please find the metric for different models below

Metric	RidgeRegression	RidgeRegression_Double_Alpha	LassoRegression	LassoRegression_Double_Alpha
R2 Score (train)	0.922734	0.919564	0.926494	0.921956
R2 Score (test)	0.891708	0.891272	0.894503	0.893740
RMSE (train)	0.110190	0.112427	0.107475	0.110743
RMSE (test)	0.133596	0.133865	0.131860	0.132337

Ridge Regression:

Double of optimal value: 20

If we double the optimal value, R2 score (both train and test) decreases and RMSE (both train and test) slightly increases

Lasso Regression:

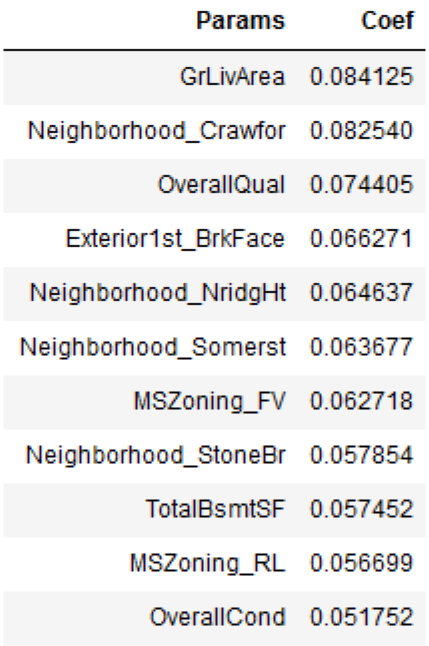
Double of optimal value: 0.0004

If we double the optimal value, R2 score (both train and test) slightly decreases and RMSE (both train and test) slightly increases

Predictor Variables:

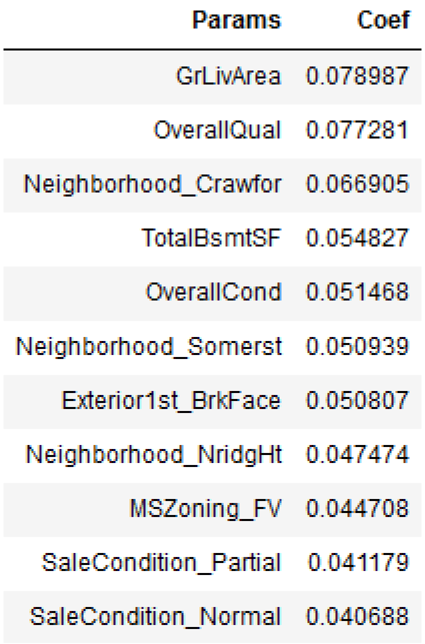
Ridge Regression:

Predictor Variables for optimal value of 10:



Params	Coef
GrLivArea	0.084125
Neighborhood_Crawfor	0.082540
OverallQual	0.074405
Exterior1st_BrkFace	0.066271
Neighborhood_NridgHt	0.064637
Neighborhood_Somerst	0.063677
MSZoning_FV	0.062718
Neighborhood_StoneBr	0.057854
TotalBsmtSF	0.057452
MSZoning_RL	0.056699
OverallCond	0.051752

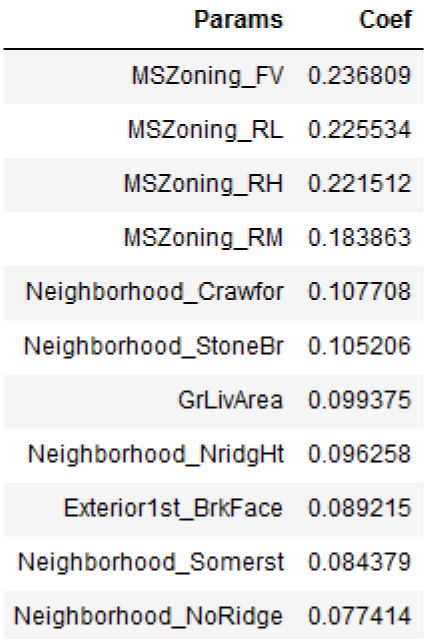
Predictor Variables for doble the optimal value ie 20:



Params	Coef
GrLivArea	0.078987
OverallQual	0.077281
Neighborhood_Crawfor	0.066905
TotalBsmtSF	0.054827
OverallCond	0.051468
Neighborhood_Somerst	0.050939
Exterior1st_BrkFace	0.050807
Neighborhood_NridgHt	0.047474
MSZoning_FV	0.044708
SaleCondition_Partial	0.041179
SaleCondition_Normal	0.040688

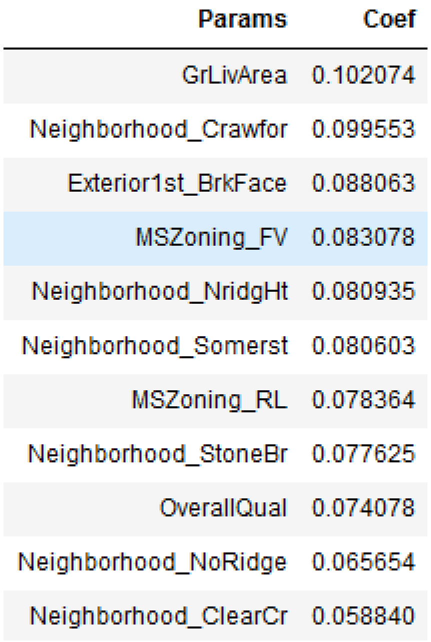
Lasso Regression:

Predictor Variables for optimal value of 0.0002:



Params	Coef
MSZoning_FV	0.236809
MSZoning_RL	0.225534
MSZoning_RH	0.221512
MSZoning_RM	0.183863
Neighborhood_Crawfor	0.107708
Neighborhood_StoneBr	0.105206
GrLivArea	0.099375
Neighborhood_NridgHt	0.096258
Exterior1st_BrkFace	0.089215
Neighborhood_Somerst	0.084379
Neighborhood_NoRidge	0.077414

Predictor Variables for doble the optimal value ie 0.0004:



Params	Coef
GrLivArea	0.102074
Neighborhood_Crawfor	0.099553
Exterior1st_BrkFace	0.088063
MSZoning_FV	0.083078
Neighborhood_NridgHt	0.080935
Neighborhood_Somerst	0.080603
MSZoning_RL	0.078364
Neighborhood_StoneBr	0.077625
OverallQual	0.074078
Neighborhood_NoRidge	0.065654
Neighborhood_ClearCr	0.058840

Result: In both Ridge and Lasso regression models, the list of top 10 features and the co-efficient’s are changed after doubling the values.

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?

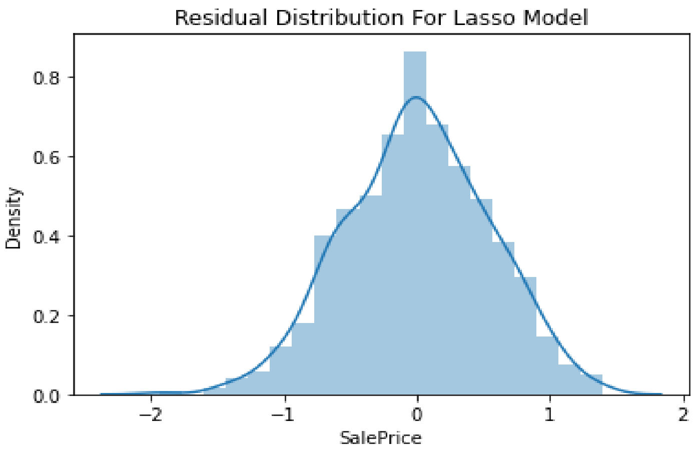
Ans:

The optimal value for Ridge Regression = **10**

The optimal value for Lasso Regression = **0.0002**

Please find the metric for different models below

Metric	LinearRegression	RidgeRegression	LassoRegression
R2 Score (train)	0.852030	0.922734	0.926494
R2 Score (test)	0.813680	0.891708	0.894503
RMSE (train)	0.152487	0.110190	0.107475
RMSE (test)	0.175236	0.133596	0.131860


Residual Distribution For Lasso Model

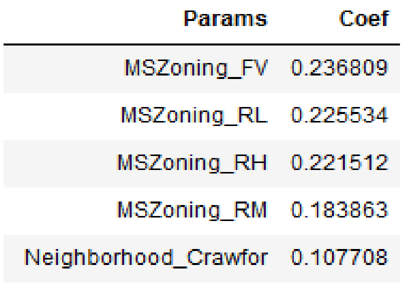
1. Linear Regression (RFE) model has low R2 value in compared to Ridge Regression and Lasso Regression models. So, rejecting the same.
2. In-comparison to Ridge Regression, Lasso Regression model has minimal increase in R2 and can say Lasso Regression model is slightly better in compared to Ridge Regression.
3. Error terms are normally distributed in Lasso model
4. Considering above points, we can consider Lasso Regression Model for housing Sales Prize prediction as it has high R2 value, low RMSE value and normal error term distribution

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?

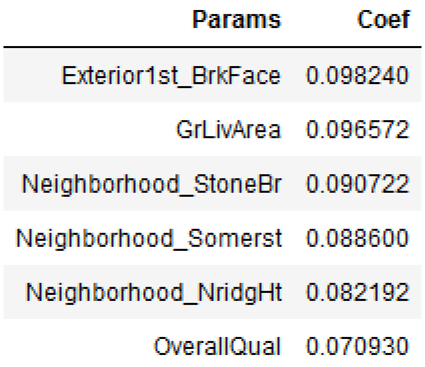
Ans:

Below are the first 5 important variables as per Lasso model with given dataset.



Params	Coef
MSZoning_FV	0.236809
MSZoning_RL	0.225534
MSZoning_RH	0.221512
MSZoning_RM	0.183863
Neighborhood_Crawfor	0.107708

By creating another model after dropping above variables, below are the new 5 important predictor variables.



Params	Coef
Exterior1st_BrkFace	0.098240
GrLivArea	0.096572
Neighborhood_StoneBr	0.090722
Neighborhood_Somerst	0.088600
Neighborhood_NridgHt	0.082192
OverallQual	0.070930

Question 4:

How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

Ans:

To ensure the model is Robust and generalized, model should be resistant to outliers.

We can treat the outlier data either by capping the data to the acceptable level or by removing the data if you feel its not required as per business terms.

In case of data has a very pronounced right tail, we can transform to log/exp/square/square root.

If model is not robust, the accuracy of the model will not be good and it won’t perform well on test data as it may be overfitting. Such that, we observe error in training and test scores.

So, the model we select should be robust and generalized to perform well in both train and test data set.