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?

The optimal value of alpha for Ridge Regression: 0.4

The optimal value of alpha for Lasso Regression: 0.0001

Doubling alpha for Ridge and Lasso Regression: i.e. alpha = 0.8 and 0.0002 respectively

1. **Effect on Metrics:** As seen in the below table, when we double the alpha value, r2score decreases slightly for train data and increases slightly for test data. This is observed in both Ridge and Lasso regressions.
   RSS and MSE also reduce for train and test data for both these regressions.

Metric (Train/Test)		Ridge		Lasso	
		Alpha 0.4	Alpha 0.8	Alpha 0.0001	Alpha 0.0002
r2score	Train	0.9242	0.92261	0.9216	0.9175
	Test	0.9041	0.9046	0.9075	0.9079
RSS	Train	2.3268	2.3767	2.408	2.5332
	Test	1.2863	1.2792	1.2404	1.2348
MSE	Train	6.61E-06	6.90E-06	7.08E-06	7.84E-06
	Test	1.10E-05	1.09E-05	1.02E-05	1.01E-05

Note: *Kindly refer python notebook* ***'Regularisaion_and_Question_Based_Model.ipynb'*** *for code written to get these above metrics. Cell number "In [4]" has model for the same.*

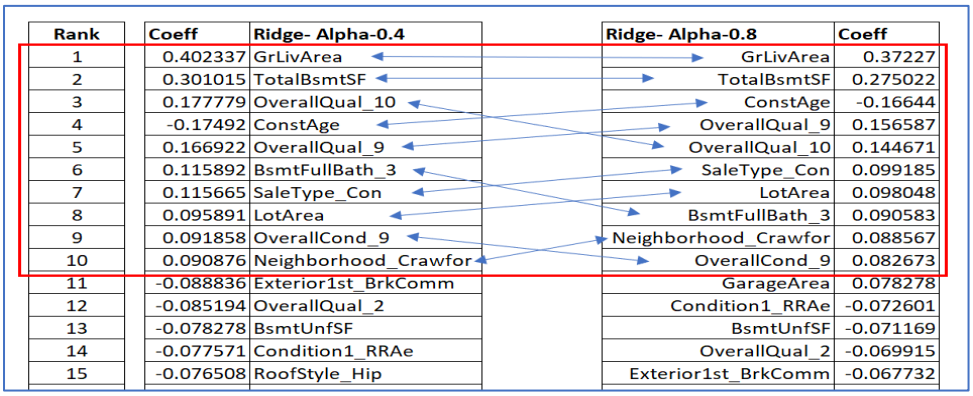
Note: *Also for below co-efficient specific analysis, kindly refer excel sheet,* ***'Ridge_Lasso_Linear_Regression_Comparison.xlsx'*** *available in the repository. There is a worksheet with name "double_best_alpha_effect" which has these details*

2. Effect on Co-efficient and eventually on top predictors:

For Ridge:

The top 10 important predictor variables remain the same, however their ranks change.

Also the co-efficient values are reduced for most of the predictor variables when alpha is doubled.



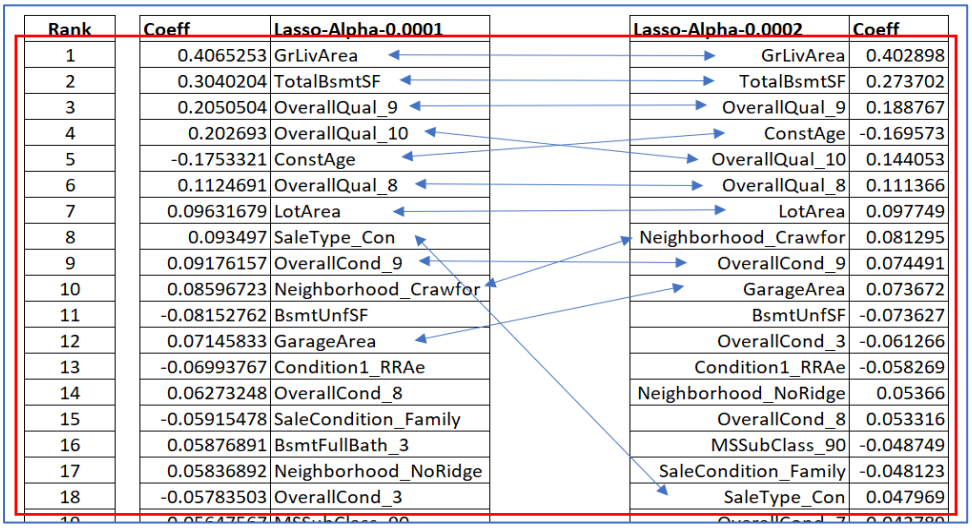
Rank	Coeff	Ridge- Alpha-0.4	Ridge- Alpha-0.8	Coeff
1	0.402337	GrLivArea	GrLivArea	0.37227
2	0.301015	TotalBsmtSF	TotalBsmtSF	0.275022
3	0.177779	OverallQual_10	ConstAge	-0.16644
4	-0.17492	ConstAge	OverallQual_9	0.156587
5	0.166922	OverallQual_9	OverallQual_10	0.144671
6	0.115892	BsmtFullBath_3	SaleType_Con	0.099185
7	0.115665	SaleType_Con	LotArea	0.098048
8	0.095891	LotArea	BsmtFullBath_3	0.090583
9	0.091858	OverallCond_9	Neighborhood_Crawfor	0.088567
10	0.090876	Neighborhood_Crawfor	OverallCond_9	0.082673
11	-0.088836	Exterior1st_BrkComm	GarageArea	0.078278
12	-0.085194	OverallQual_2	Condition1_RRAe	-0.072601
13	-0.078278	BsmtUnfSF	BsmtUnfSF	-0.071169
14	-0.077571	Condition1_RRAe	OverallQual_2	-0.069915
15	-0.076508	RoofStyle_Hip	Exterior1st_BrkComm	-0.067732

For Lasso:

9 out of the top 10 important predictor variables remain the same, 5 of which retain their ranks as well.

Also the co-efficient values are reduced for most of the predictor variables when alpha is doubled.

'GarageArea' made its way to top 10 and **'SaleType_Con'** lost its place from top 10.



Rank	Coeff	Lasso-Alpha-0.0001	Lasso-Alpha-0.0002	Coeff
1	0.4065253	GrLivArea	GrLivArea	0.402898
2	0.3040204	TotalBsmtSF	TotalBsmtSF	0.273702
3	0.2050504	OverallQual_9	OverallQual_9	0.188767
4	0.202693	OverallQual_10	ConstAge	-0.169573
5	-0.1753321	ConstAge	OverallQual_10	0.144053
6	0.1124691	OverallQual_8	OverallQual_8	0.111366
7	0.09631679	LotArea	LotArea	0.097749
8	0.093497	SaleType_Con	Neighborhood_Crawfor	0.081295
9	0.09176157	OverallCond_9	OverallCond_9	0.074491
10	0.08596723	Neighborhood_Crawfor	GarageArea	0.073672
11	-0.08152762	BsmtUnfSF	BsmtUnfSF	-0.073627
12	0.07145833	GarageArea	OverallCond_3	-0.061266
13	-0.06993767	Condition1_RRAe	Condition1_RRAe	-0.058269
14	0.06273248	OverallCond_8	Neighborhood_NoRidge	0.05366
15	-0.05915478	SaleCondition_Family	OverallCond_8	0.053316
16	0.05876891	BsmtFullBath_3	MSSubClass_90	-0.048749
17	0.05836892	Neighborhood_NoRidge	SaleCondition_Family	-0.048123
18	-0.05783503	OverallCond_3	SaleType_Con	0.047969

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:

With 100 predictor variables in the model,

For Ridge: optimal value of lambda is 0.4 and

For Lasso: optimal value of lambda is 0.0001

```
############################################################################
Number Of Columns in Data:100
############################################################################
Fitting 5 folds for each of 28 candidates, totalling 140 fits
Best Alpha/Lambda for RIDGE=0.4
Fitting 5 folds for each of 28 candidates, totalling 140 fits
Best Alpha/Lambda for LASSO=0.0001
############################################################################
######################        FINAL METRIC         ########################
             Metric  Linear Regression  Ridge Regression  Lasso Regression
0  R2 Score (Train)       9.252276e-01          0.924240          0.921597
1   R2 Score (Test)      -1.816230e+19          0.904103          0.907526
2       RSS (Train)       2.296505e+00          2.326841          2.408020
3        RSS (Test)       2.436247e+20          1.286339          1.240425
4       MSE (Train)       5.037435e-02          0.050706          0.051583
5        MSE (Test)       7.924007e+08          0.057579          0.056542
############################################################################
################################ CONCLUSION ################################
With 100 columns, for alpha= 0.0001, Lasso Regression is better
############################################################################
```

As per the above metric table, train R2 score for Ridge regression is better than train R2 score for Lasso regression. However, test R2 score for Lasso is better than test R2 score for Ridge regression.

This tells us that, the variables that we considered in the model to build ridge regression, they all sort of less related to this response variable. So there are more noisy variables in the model. When Lasso regression did feature selection and dropped some variables, it actually removed some of these noisy variables and the remaining variables are more related with the response variable.

Question 3

After building the model, you realized 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:

In Lasso model, five most important predictor variables are:

GrLivArea, TotalBsmtSF, OverallQual_9, OverallQual_10, ConstAge

Once we remove these five most important variables from the incoming test data, below 5 variables become very important in the new model:

1stFlrSF, 2ndFlrSF, OverallQual_4, OverallQual_3, OverallQual_5

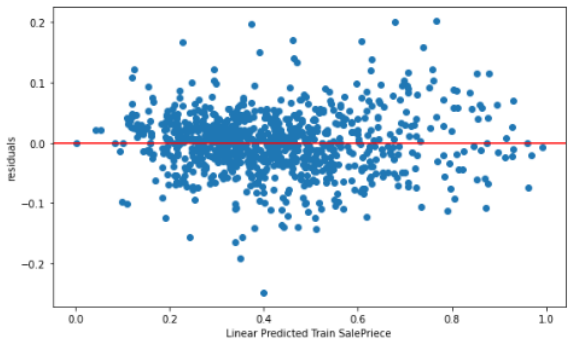
Note: *Kindly refer python notebook* ***'Regularisaion_and_Question_Based_Model.ipynb'*** *for code written to get the model for this analysis. Cell number "In [5]" has model for the same. Further analysis is done in excel sheet,* ***'Ridge_Lasso_Linear_Regression_Comparison.xlsx'*** *available in the repository. There is a worksheet with name "remove_top_5_vars" which has these details*

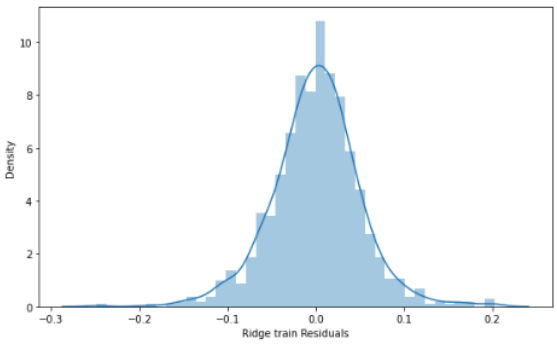
Question 4

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?

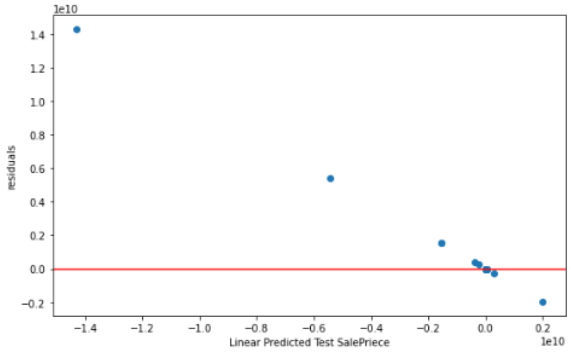
Answer:

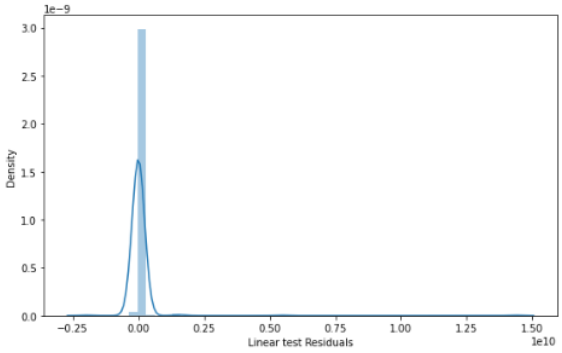
1. **Based upon the R-square metric we can say that**
   a. **With linear model good variation was explained on training set but with test data set it completely failed.**
   b. **After doing regularization with the ridge and lasso regressions, we found improvement in this where good performance was seen in train and test set in terms of R2 score. Ridge plots are given below. Lasso plots are also similar**
2. **Based upon Residual analysis also we can get similar understandings as given below.**
   a. **With linear regression:**



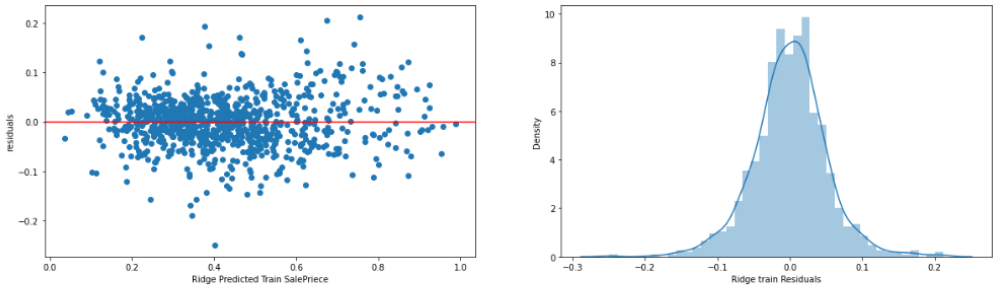


<Figure size 432x288 with 0 Axes>

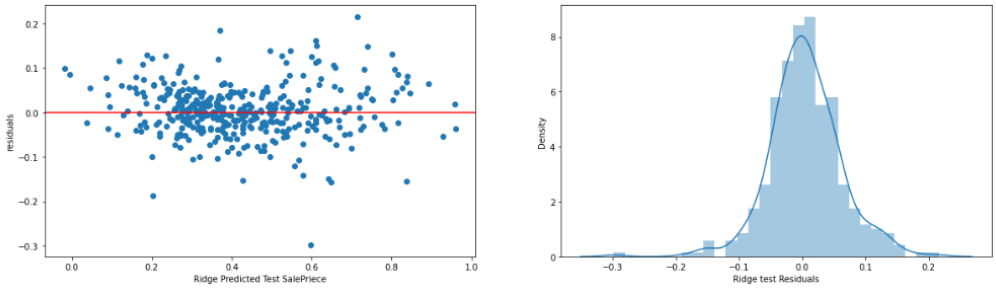




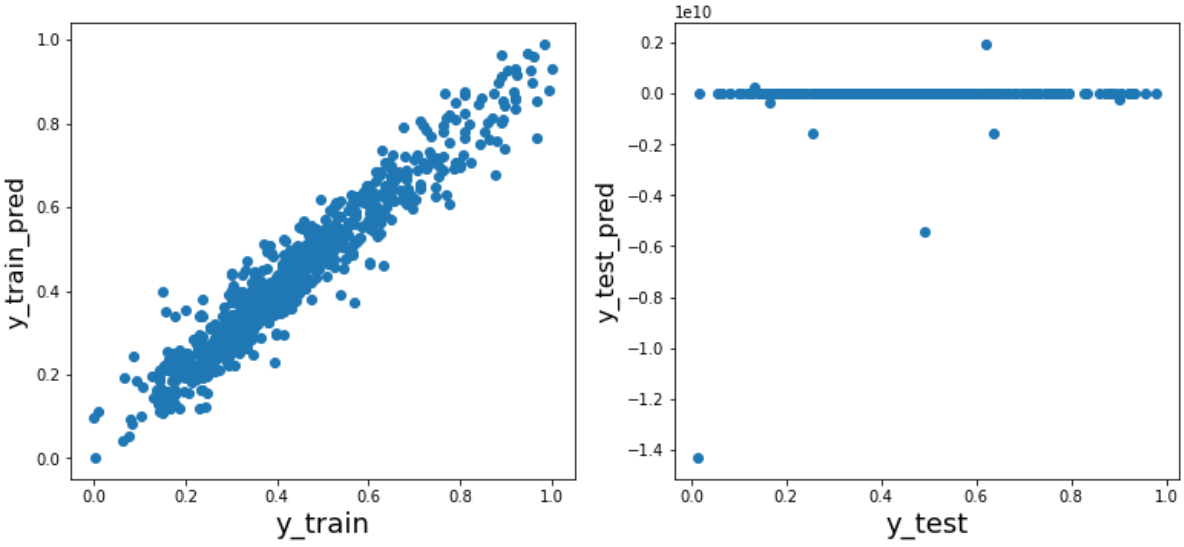
b. **Post regularization: Provided for Ridge, For Lasso it is similar**



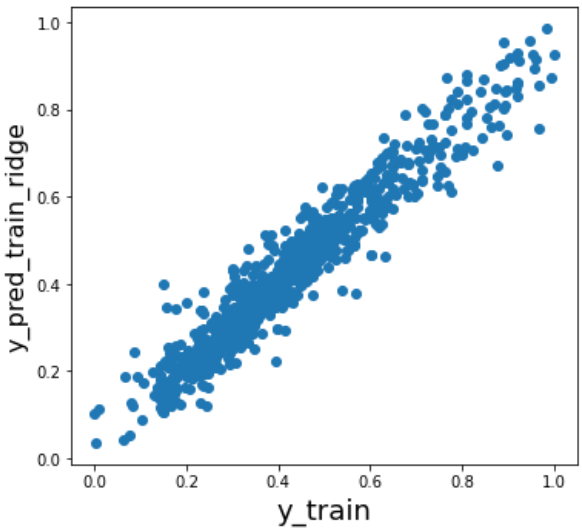
<Figure size 432x288 with 0 Axes>

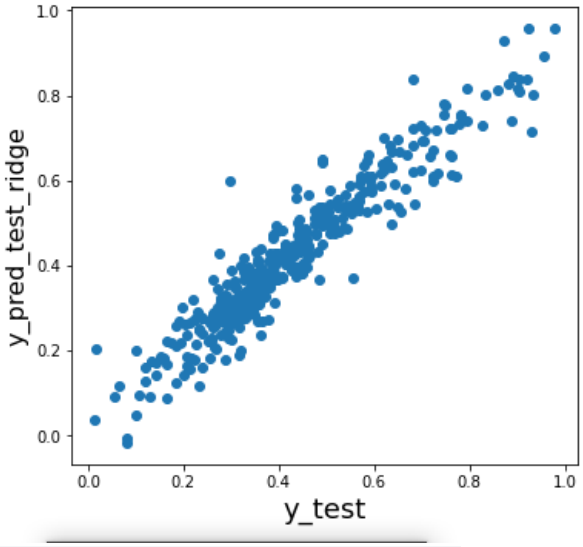


3. **Eventually predicted and actual values from these models also give confirmation that the regularization did its job.**
   a. **Before regularization:**



b. **Post regularization: Provided for Ridge, for Lasso it is similar**





4. **With the help of all this analysis we can safely say that this model is robust and generalizable. I have done RFE for 100 columns. And all this analysis is based upon these 100 columns. Further we can use stats model and VIF on this model to get more accurate model.**