DSI205

Predicting Bangkok Housing prices

Least-Squares Problem Group Project

Members

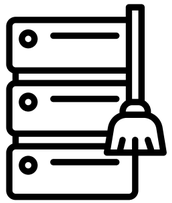
- Thanarak Leenanon 6524650030
- Sirapop Chullapamorn 6524650089
- Watcharanan phanmool 6524650071
- Wasan arumsakul 6524651400
- Warit dongphrachan 6309659180

Overview

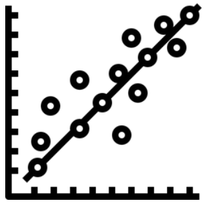
Data source

kaggle

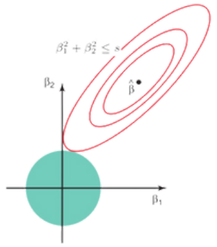
Data cleansing



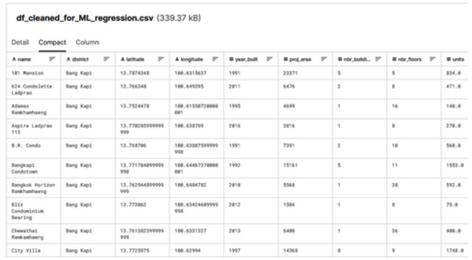
Multiple Linear regression



Ridge regression

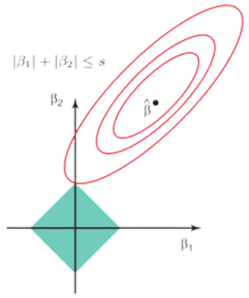


Conclusion

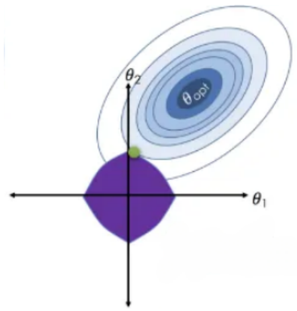


Data set

Data preparation



Lasso regression



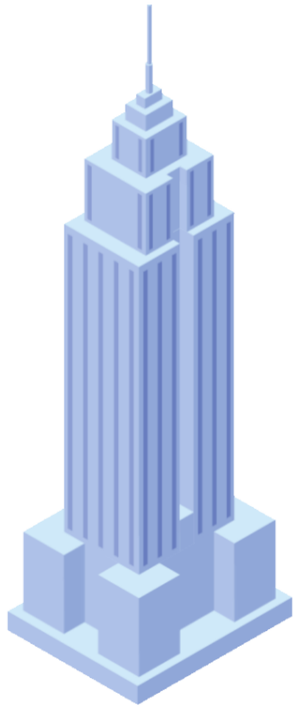
Elasticnet regression

Kaggle Dataset

df_cleaned_for_ML_regression.csv (339.37 kB)

Detail | Compact | Column

name	district	latitude	longitude	year_built	proj_area	nbr_buildi...	nbr_floors	units
101 Mansion	Bang Kapi	13.7874348	100.6315637	1991	23371	5	5	834.0
624 Condolette Ladprao	Bang Kapi	13.766348	100.649395	2011	6476	2	8	471.0
Adamas Ramkhamhaeng	Bang Kapi	13.7524478	100.61550720000001	1995	4699	1	16	140.0
Aspire Ladprao 113	Bang Kapi	13.770285999999999	100.638799	2016	3816	1	8	270.0
B.R. Condo	Bang Kapi	13.768706	100.63887599999998	1991	7391	2	10	560.0
Bangkapi Condotown	Bang Kapi	13.771784099999998	100.64467370000001	1992	15161	5	11	1553.0
Bangkok Horizon Ramkhamhaeng	Bang Kapi	13.762944899999999	100.6484782	2010	5568	1	38	592.0
Bliz Condominium Bearing	Bang Kapi	13.773062	100.63424609999998	2012	1504	1	8	79.0
Chewathai Ramkamhaeng	Bang Kapi	13.761382399999999	100.6331327	2013	6400	1	36	400.0
City Villa	Bang Kapi	13.7725975	100.62994	1997	14368	8	9	1748.0



1019 rows x 55 Columns

from : https://www.kaggle.com/datasets/thedevastator/predicting-bangkok-condominium-prices-using-web

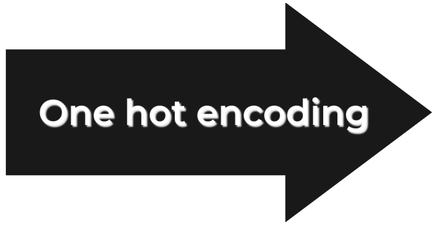
Data Preparation

Add Column price_sqm * proj_area = price AKA : total price of the project.

Data Transformation

no	tran_type1
1	expressway
2	bts
3	mrt

55 columns

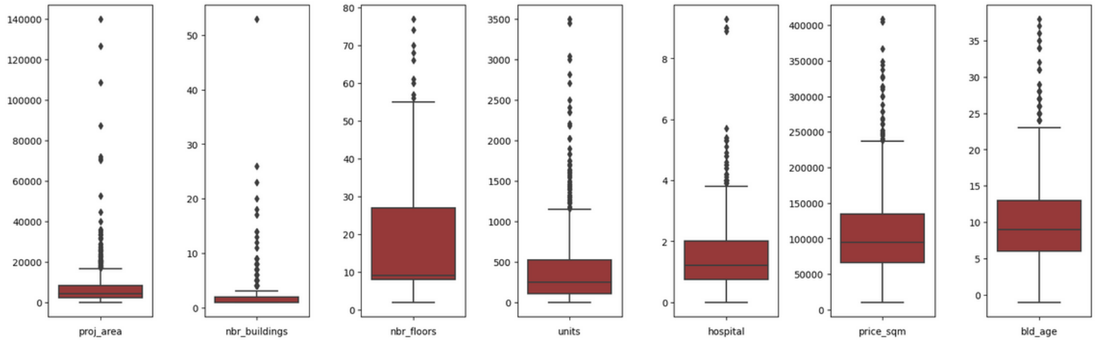


no	tran_type1_expressway	tran_type1_bts	tran_type1_mrt
1	1	0	0
2	0	1	0
3	0	0	1

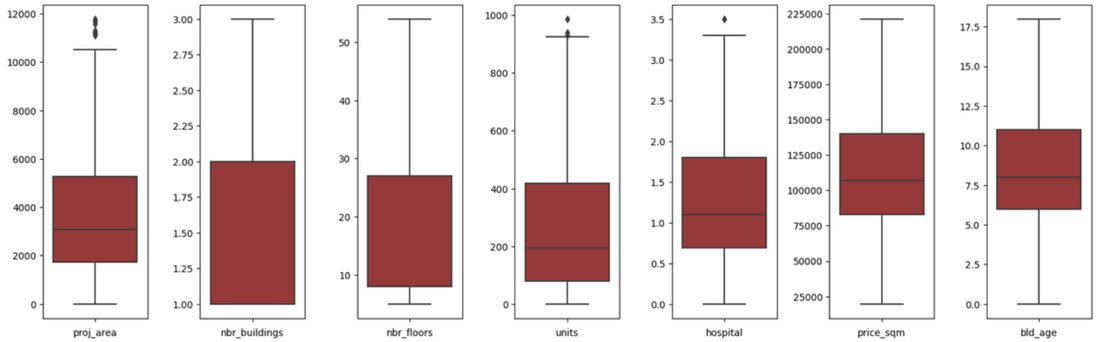
95 columns

Data Cleansing

Before



After



Features Selection

Before deleting features : 95

df_cleaned_for_ML_regression.csv (339.37 kB)

Detail | Compact | Column

name	district	latitude	longitude	year_built	proj_area
101 Mansion	Bang Kapi	13.7874348	100.6315637	1991	23371
624 Condolette Ladprao	Bang Kapi	13.766348	100.649395	2011	6476
Adamas Ramkhamhaeng	Bang Kapi	13.7524478	100.61550720000001	1995	4699
Aspire Ladprao 113	Bang Kapi	13.770285999999999	100.638799	2016	3816
B.R. Condo	Bang Kapi	13.768706	100.63887599999998	1991	7391
Bangkapi Condotown	Bang Kapi	13.771784099999998	100.64467370000001	1992	15161
Bangkok Horizon Ramkhamhaeng	Bang Kapi	13.762944899999999	100.6484782	2010	5568
Bliz Condominium Bearing	Bang Kapi	13.773062	100.63424609999998	2012	1504
Chewathai Ramkamhaeng	Bang Kapi	13.761382399999999	100.6331327	2013	6400
City Villa	Bang Kapi	13.7725975	100.62994	1997	14368
Condo U @ Huamak Station	Bang Kapi	13.751934900000002	100.6365273	2013	5300
D Condo Ramkhamhaeng	Bang Kapi	13.7496291	100.6076671	2012	17194
D Condo Ramkhamhaeng 64	Bang Kapi	13.764691899999999	100.65364699999999	2013	14047

After deleting features : 86

index	price
24 CCTV	NaN
42 district_Bang Khun Thian	NaN
54 district_Dusit	NaN
59 district_Lak Si	NaN
60 district_Lat Krabang	NaN
62 district_Min Buri	NaN
67 district_Pom Prap Sattru Phai	NaN
71 district_Sai Mai	NaN
75 district_Taling Chan	NaN

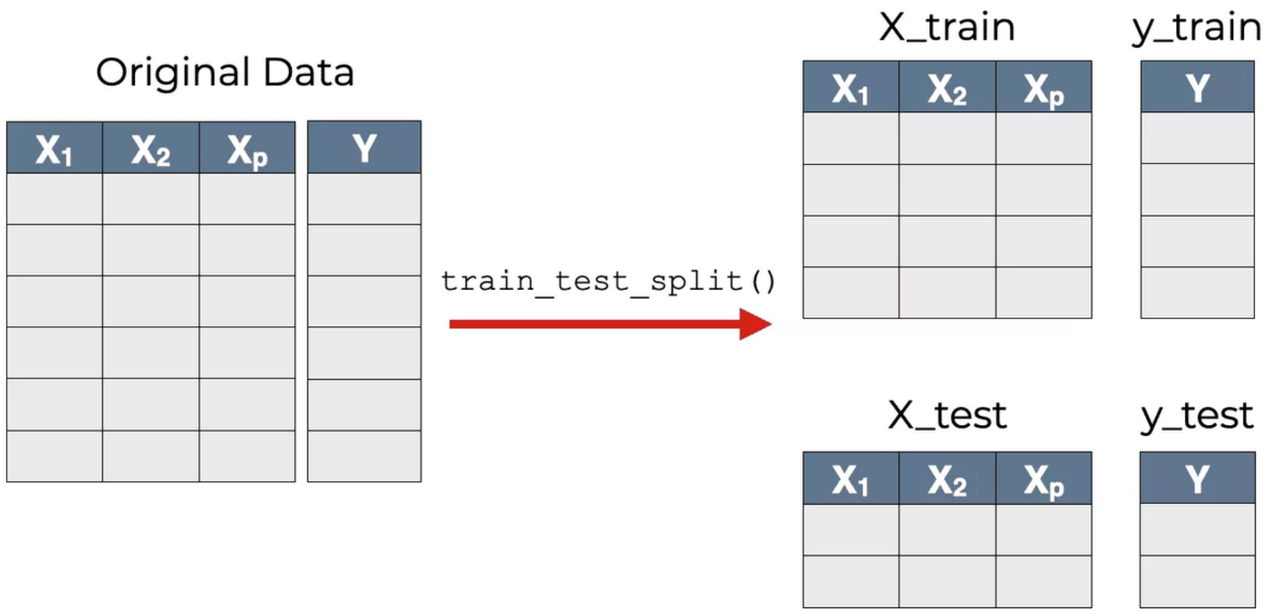
```
คอลัมน์ CCTV มีสมาชิก 1 ตัวนั่นคือ '1'
คอลัมน์ district_Bang Khun Thian มีสมาชิก 1 ตัวนั่นคือ '0'
คอลัมน์ district_Dusit มีสมาชิก 1 ตัวนั่นคือ '0'
คอลัมน์ district_Lak Si มีสมาชิก 1 ตัวนั่นคือ '0'
คอลัมน์ district_Lat Krabang มีสมาชิก 1 ตัวนั่นคือ '0'
คอลัมน์ district_Min Buri มีสมาชิก 1 ตัวนั่นคือ '0'
คอลัมน์ district_Pom Prap Sattru Phai มีสมาชิก 1 ตัวนั่นคือ '0'
คอลัมน์ district_Sai Mai มีสมาชิก 1 ตัวนั่นคือ '0'
คอลัมน์ district_Taling Chan มีสมาชิก 1 ตัวนั่นคือ '0'
```

Dataset for Model

proj_area	nbr_buildings	nbr_floors	units	hospital	bld_age	dist_shop_1	dist_shop_2	dist_shop_3	dist_shop_4	dist_shop_5	dist_school_1	dist_school_2	dist
6476.0	2.0	8.0	471.0	1.80	8.0	0.400	0.79	0.83	0.85	1.10	0.56	0.72	
3816.0	1.0	8.0	270.0	0.68	3.0	0.002	0.53	0.64	0.64	0.65	0.71	0.72	
5568.0	1.0	38.0	592.0	1.40	9.0	0.480	0.64	1.00	1.10	1.30	0.28	0.37	
1504.0	1.0	8.0	79.0	1.80	7.0	1.500	1.50	2.40	2.80	4.90	1.10	1.30	
6400.0	1.0	36.0	400.0	0.46	6.0	0.600	1.00	1.30	1.40	1.50	0.39	0.39	
...	...	...	...	...	...	...	...	...	...	...	...	...	
2400.0	1.0	27.0	621.0	3.30	9.0	0.003	3.00	1.60	1.70	1.70	0.61	1.10	
2768.0	1.0	31.0	185.0	2.60	9.0	0.003	1.20	1.20	1.20	3.00	0.67	0.69	
2268.0	1.0	8.0	34.0	2.40	13.0	1.100	1.20	1.30	3.00	3.00	0.47	0.53	
1432.0	1.0	7.0	176.0	2.90	12.0	0.610	0.75	0.82	3.00	3.00	0.33	0.92	
5372.0	1.0	19.0	277.0	1.00	11.0	2.500	2.80	3.00	8.80	9.90	0.74	0.80	

638 rows x 86 Columns

Train Test Split & Scaling Data



Split the DataFrame
80% for training
20% for testing.

$$x_{i(scaled)} = \frac{x_i - \mu}{\sigma}$$

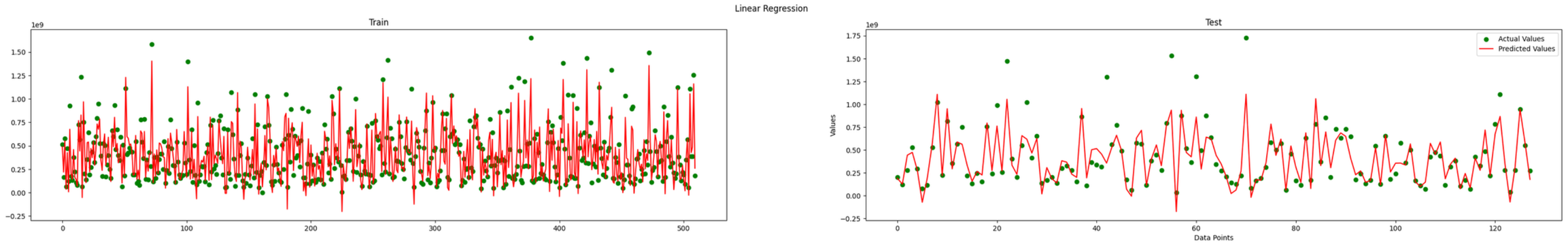
Scaling by Standardization function of scipy library

Multiple Linear regressions

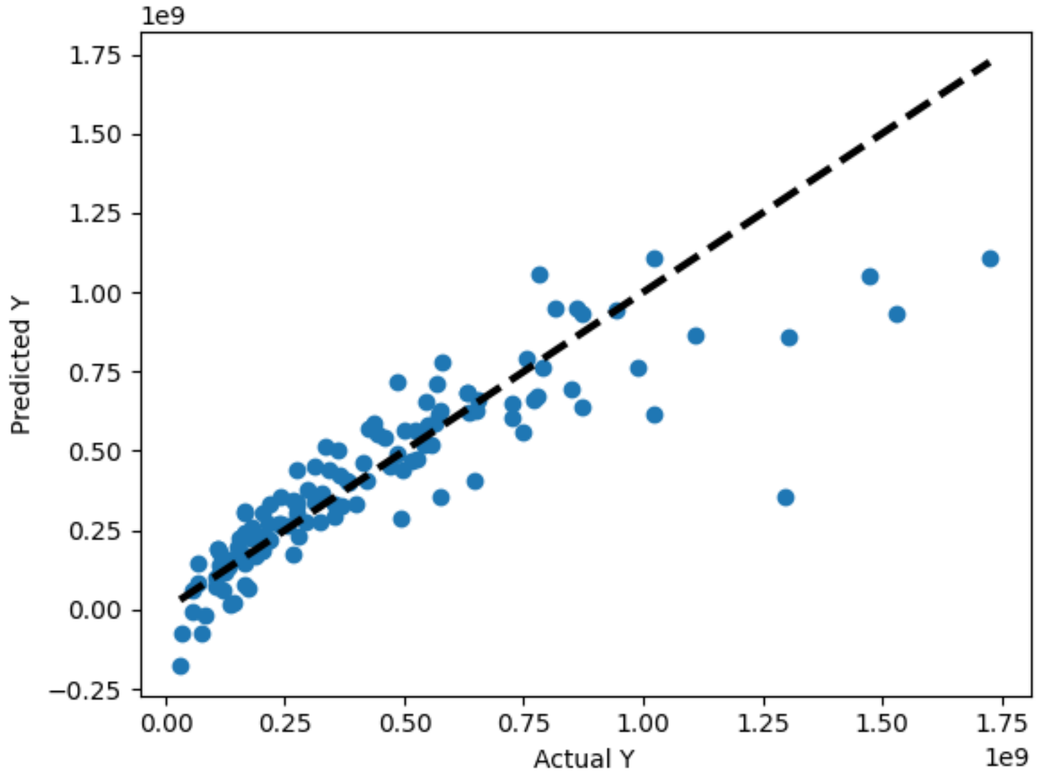
The general form of a Multiple Linear regression equation is :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_n x_n$$

Split the DataFrame into 80% for training and 20% for testing.



Graph showing the relationship of Linear association with Actual values of the train-test split data set of Multiple linear regression model.



The graph shows the relationship between predicted values and actual values of the Multiple linear regression model as follows.

Model	Matrices	Test Score	Train Score
Linear Regression	R2 Score	0.758	0.867
Linear Regression	MAE	98,280,098.99	79,726,924.94
Linear Regression	MSE	2.62E+16	1.18E+16
Linear Regression	RMSE	161,882,797.26	108,798,160.66

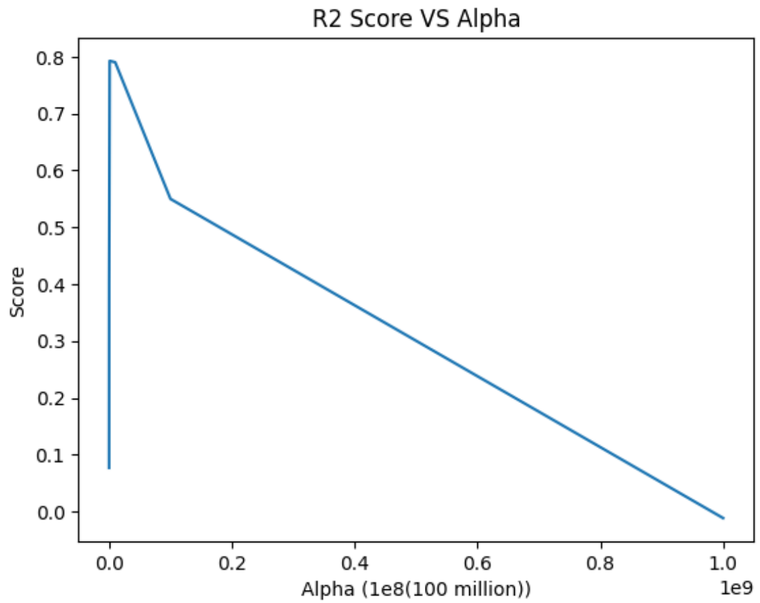
Evaluation metrics used to measure model performance

Lasso regression

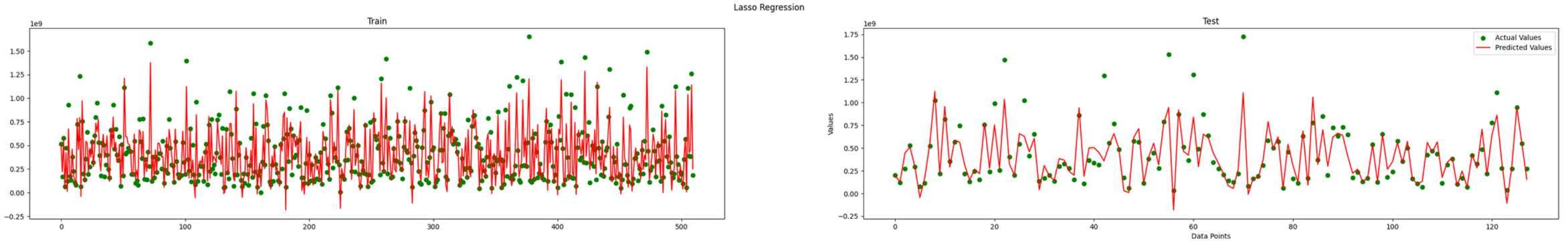
The general form of a Lasso regression equation is :

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 + \underbrace{\lambda \sum_{j=1}^{m}|\hat{\beta}_j|}_{\text{Penalty}}$$

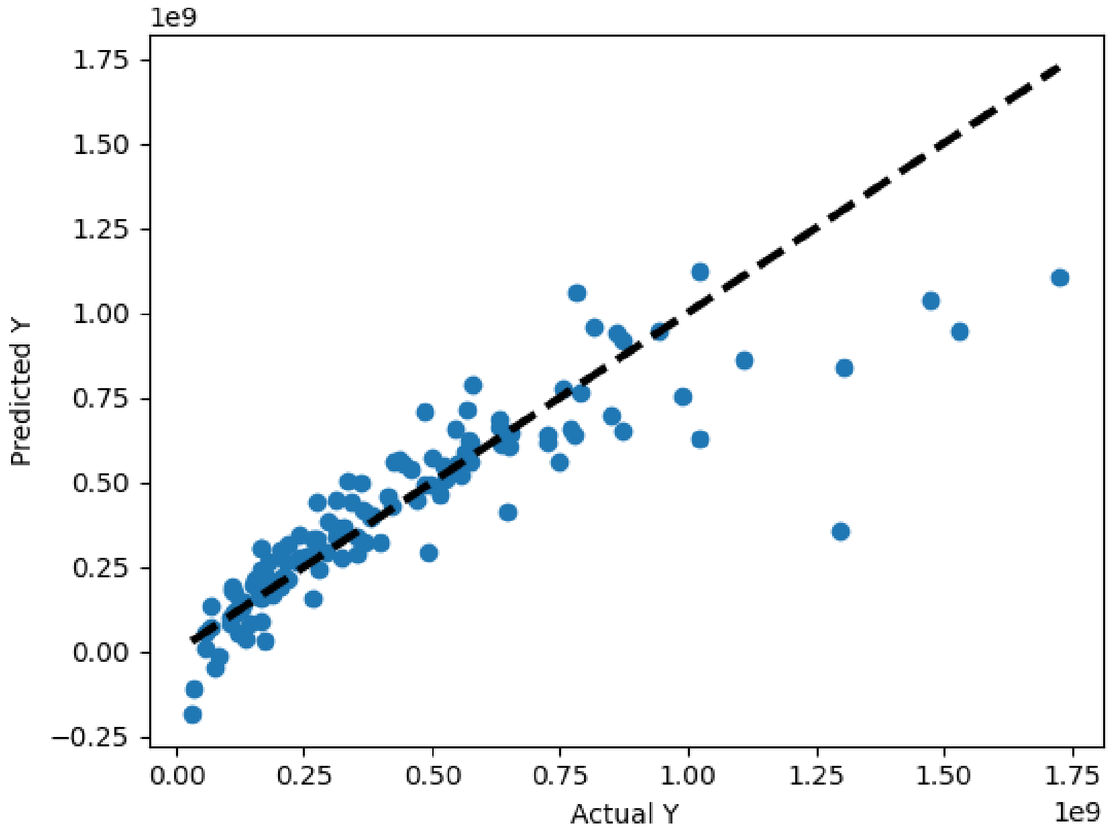
GridSearchCV Result : alpha ~1,000,000



Graph showing Penalty (alpha ~ 1,000,000) makes R2 the highest value.



Graph showing the relationship of Linear associations to Actual values of the train-test split data set of the Lasso regression model.



Graph showing the relationship between predicted values and actual values of the Lasso regression model.

Model	Matrices	Test Score	Train Score
Lasso Regression	R2 Score	0.764	0.864
Lasso Regression	MAE	93,431,447.48	79,392,819.68
Lasso Regression	MSE	2.56E+16	1.20E+16
Lasso Regression	RMSE	159,875,593.11	109,739,739.78

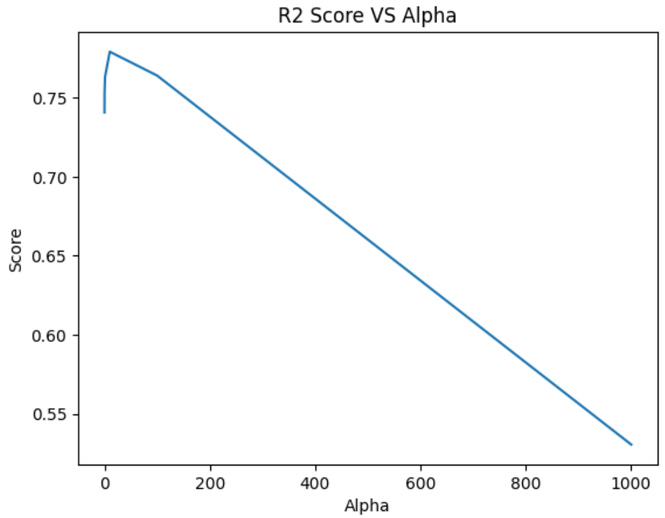
Evaluation metrics used to measure model performance

Ridge regression

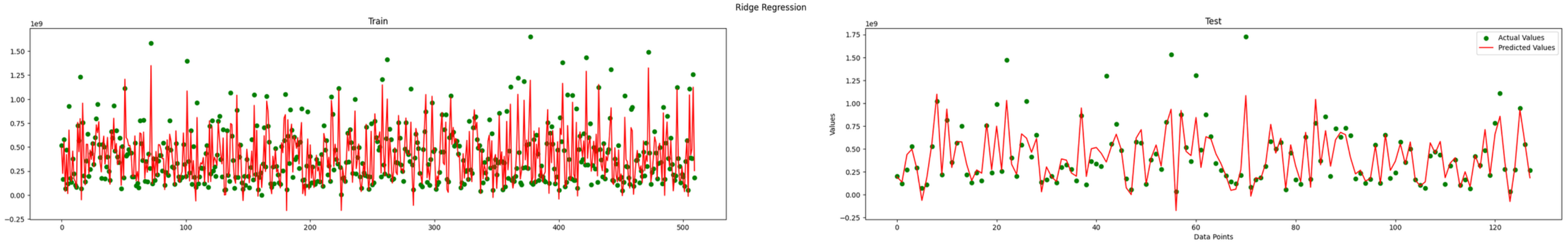
The general form of a Ridge regression equation is :

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - \hat{y})^2 + \lambda \underbrace{\sum_{j=1}^{p}\hat{\beta}_j{}^2}_{\text{Penalty}}$$

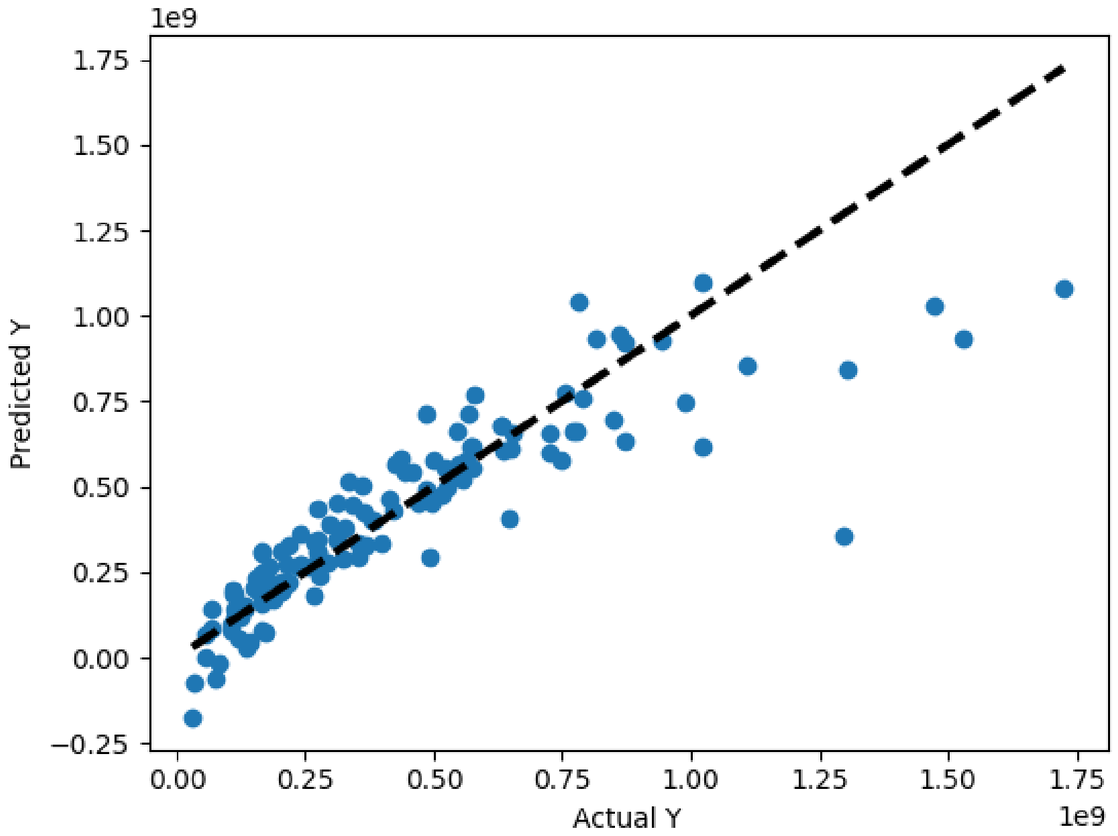
GridSearchCV Result : alpha ~10



Graph showing Penalty(alpha ~ 10) gives highest R2 score.



Graph showing the relationship of Linear associations to Actual values of the train-test split data set of the Ridge regression model.



Graph showing the relationship between predicted values and actual values of the Ridge regression model.

Model	Matrices	Test Score	Train Score
Ridge Regression	R2 Score	0.760	0.866
Ridge Regression	MAE	95,746,212.15	79,495,600.97
Ridge Regression	MSE	2.60E+16	1.19E+16
Ridge Regression	RMSE	161,389,724.16	109,249,479.68

Evaluation metrics used to measure model performance

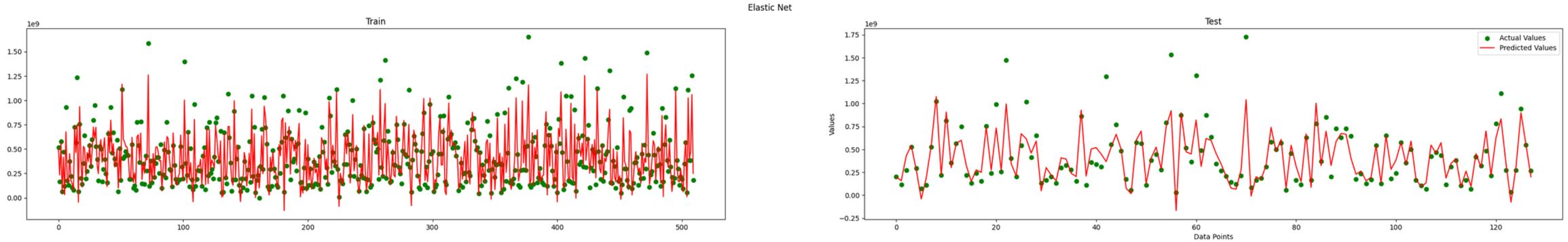
Elastic net regression

The general form of a Elastic net regression equation is :

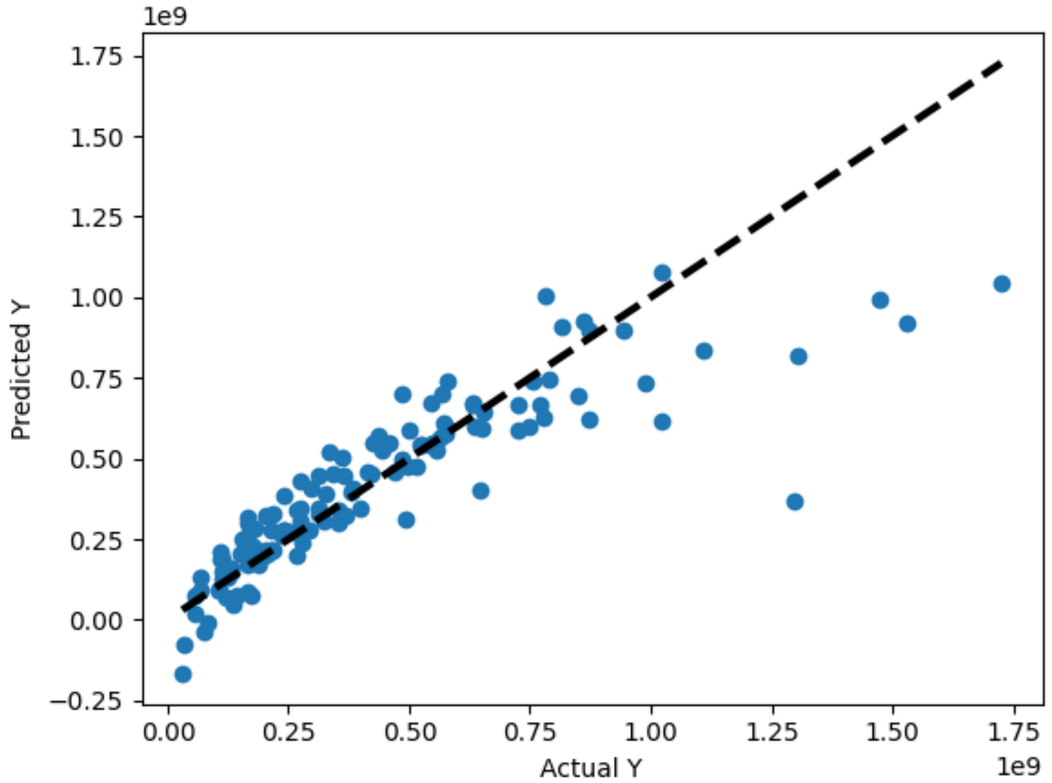
$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i{}^T\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j{}^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|\right)$$

Penalty

GridSearchCV Result : alpha ~0.119, L1 ratio ~0.456



Graph showing the relationship of Linear associations to Actual values of the train-test split data set of the Elastic net regression model.

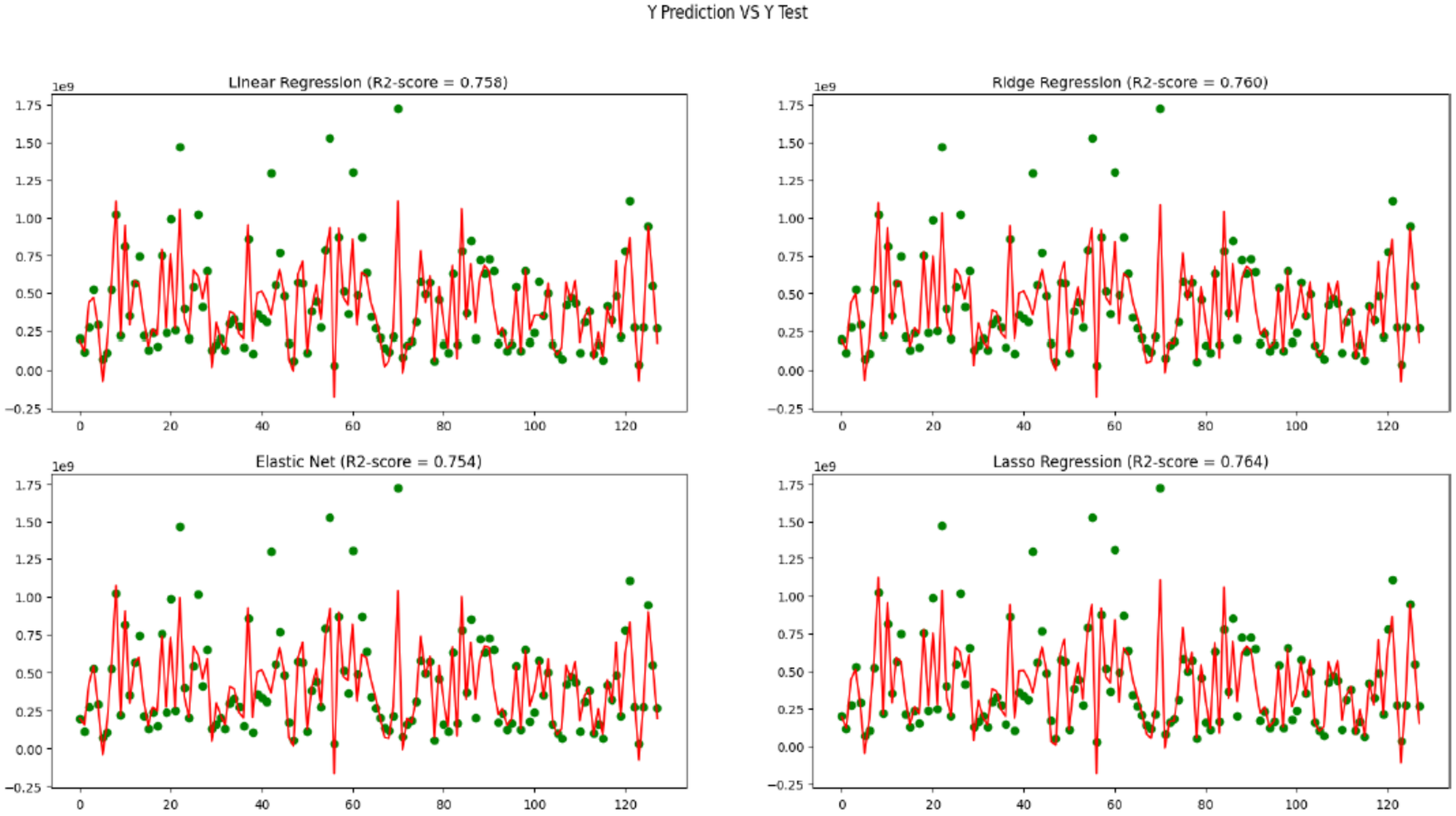


The graph shows the relationship between predicted values and actual values of the Elastic net regression model as follows.

Model	Matrices	Test Score	Train Score
Elastic Net	R2 Score	0.754	0.859
Elastic Net	MAE	95,314,140.23	80,089,091.76
Elastic Net	MSE	2.67E+16	1.25E+16
Elastic Net	RMSE	163,316,555.54	111,973,426.63

Evaluation metrics used to measure model performance

Conclusion



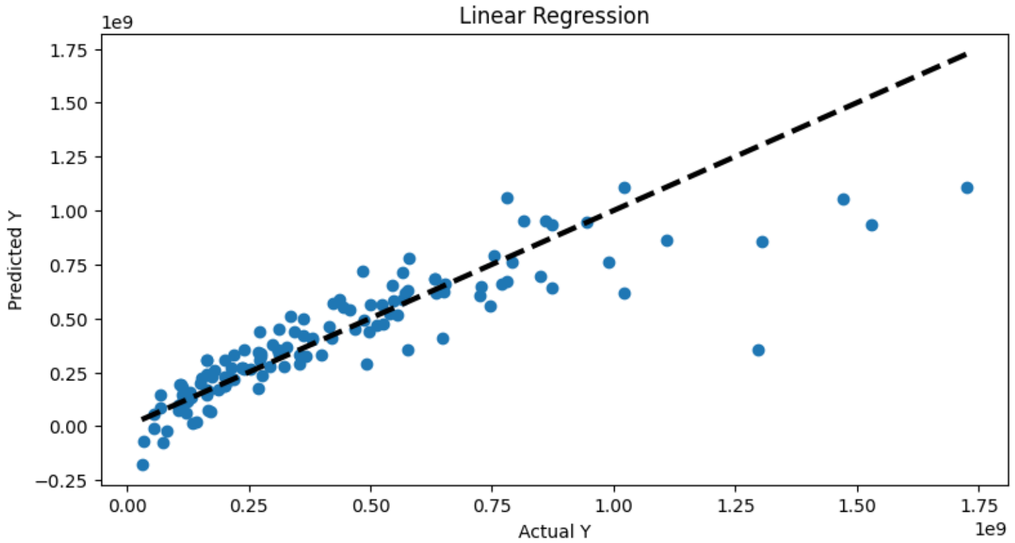
Graph showing the relationship of Actual values (green dots) with predicted values (red lines) of all 4 model types by measuring their performance with R-squared values.

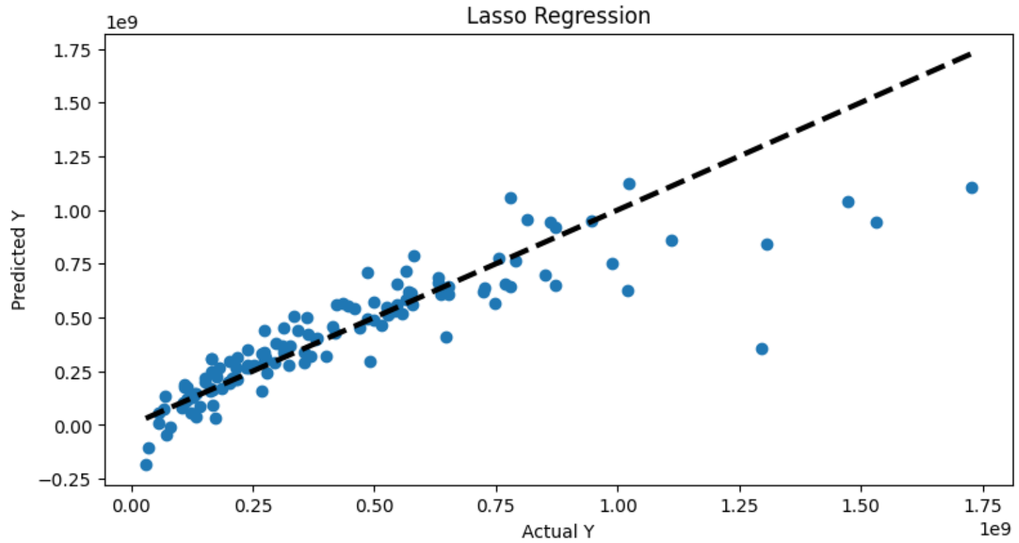
Model	MAE	MSE	RMSE	R2 Score
Lasso Regression	93,431,447.48	2.56E+16	159,875,593.11	0.764
Ridge Regression	95,746,212.15	2.60E+16	161,389,724.16	0.760
Linear Regression	98,280,098.99	2.62E+16	161,882,797.26	0.758
Elastic Net	95,314,140.23	2.67E+16	163,316,555.55	0.754

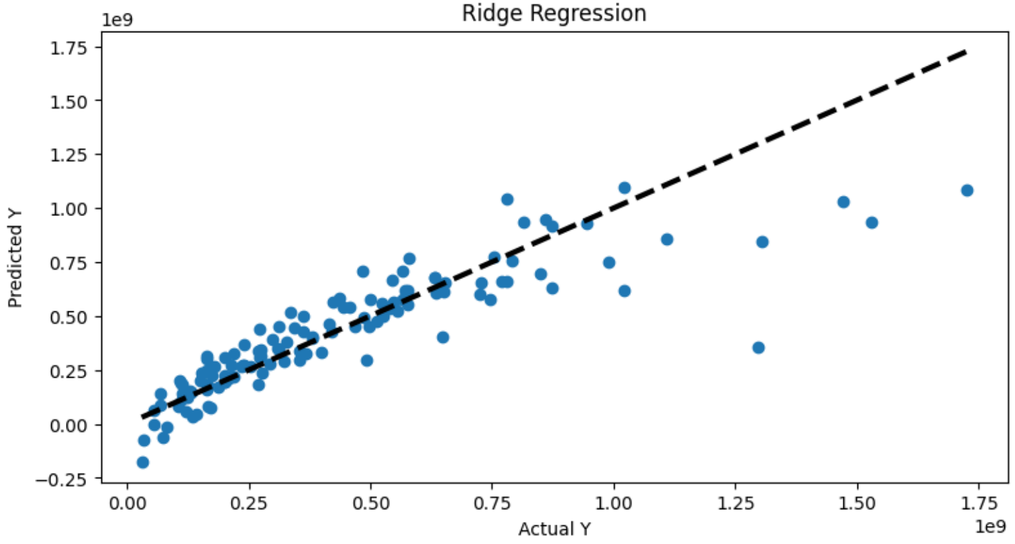
Lasso regression was the most accurate on the Test data set because RMSE value was the lowest R-squared was the highest

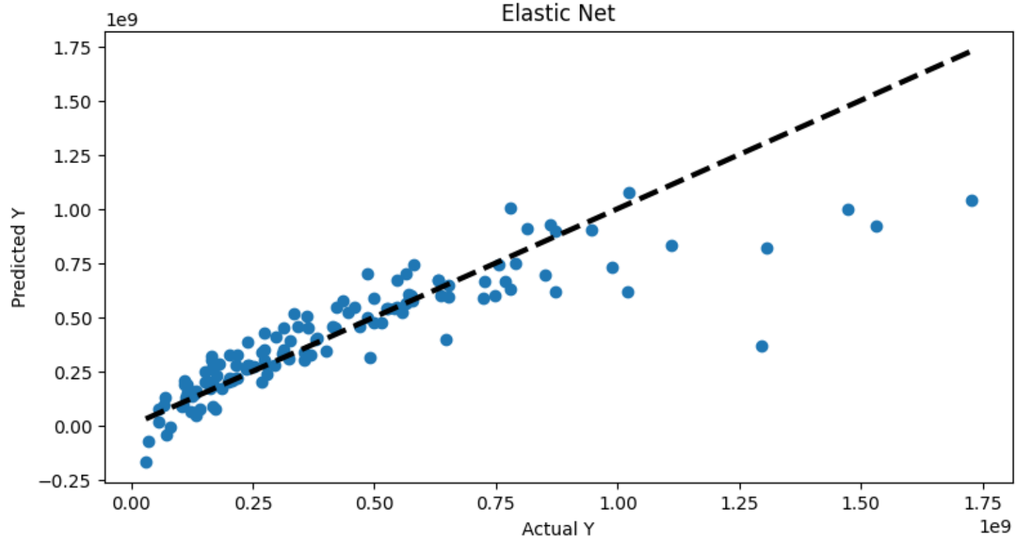
The table summarizes the Evaluation matrices of all 4 types of models by using the train data set (Y-train) and the test data set (Y- train prediction).

Linear association of Actual Y and Predicted Y of the **test scores** dataset









How can adding penalty improve accuracy?

Reduced Multicollinearity Problem

High Correlated Features

Column 1	Column 2	Correlation
dist_tran_5	dist_tran_4	0.986
dist_tran_4	dist_tran_5	0.986
dist_school_4	dist_school_5	0.927
dist_school_5	dist_school_4	0.927
dist_tran_1	dist_tran_2	0.927
dist_tran_2	dist_tran_1	0.927
dist_tran_2	dist_tran_3	0.92
dist_tran_3	dist_tran_2	0.92
dist_school_3	dist_school_2	0.918
dist_school_2	dist_school_3	0.918
dist_school_3	dist_school_4	0.905
dist_school_4	dist_school_3	0.905
dist_tran_3	dist_tran_4	0.889
dist_tran_4	dist_tran_3	0.889
dist_tran_5	dist_tran_3	0.882
dist_tran_3	dist_tran_5	0.882
dist_school_5	dist_school_3	0.86
dist_school_3	dist_school_5	0.86
dist_tran_2	dist_tran_4	0.849
dist_tran_4	dist_tran_2	0.849
dist_school_2	dist_school_4	0.841
dist_school_4	dist_school_2	0.841
dist_tran_1	dist_tran_3	0.837
dist_tran_3	dist_tran_1	0.837
dist_tran_5	dist_tran_2	0.837
dist_tran_2	dist_tran_5	0.837
dist_school_2	dist_school_1	0.828
dist_school_1	dist_school_2	0.828

Unique Feature
dist_school_1
dist_school_2
dist_school_3
dist_school_4
dist_school_5
dist_tran_1
dist_tran_2
dist_tran_3
dist_tran_4
dist_tran_5

Coefficient

Multiple Linear Regression

index	Feature	Coefficients
11	dist_school_1	-8,460,350.47
12	dist_school_2	-5,094,847.35
13	dist_school_3	8,669,949.18
14	dist_school_4	10,382,233.43
15	dist_school_5	-26,913,404.71
32	dist_tran_1	-40,083,683.82
33	dist_tran_2	-12,446,916.01
34	dist_tran_3	-26,122,980.12
35	dist_tran_4	-14,793,880.33
36	dist_tran_5	48,884,855.11

Lasso Regression

index	Feature	Coefficients
13	dist_school_3	2,364,308.43
14	dist_school_4	1,152,451.39
15	dist_school_5	-14,715,091.65
32	dist_tran_1	-43,806,901.03
33	dist_tran_2	-10,629,155.13
34	dist_tran_3	-8,481,859.92
35	dist_tran_4	1,245,921.85
36	dist_tran_5	4,976,637.01

Ridge Regression

index	Feature	Coefficients
11	dist_school_1	-8,117,067.01
12	dist_school_2	-5,219,897.14
13	dist_school_3	6,543,268.73
14	dist_school_4	7,289,386.37
15	dist_school_5	-21,481,352.96
32	dist_tran_1	-35,876,412.85
33	dist_tran_2	-16,995,868.80
34	dist_tran_3	-22,212,971.38
35	dist_tran_4	3,520,985.59
36	dist_tran_5	24,262,193.35

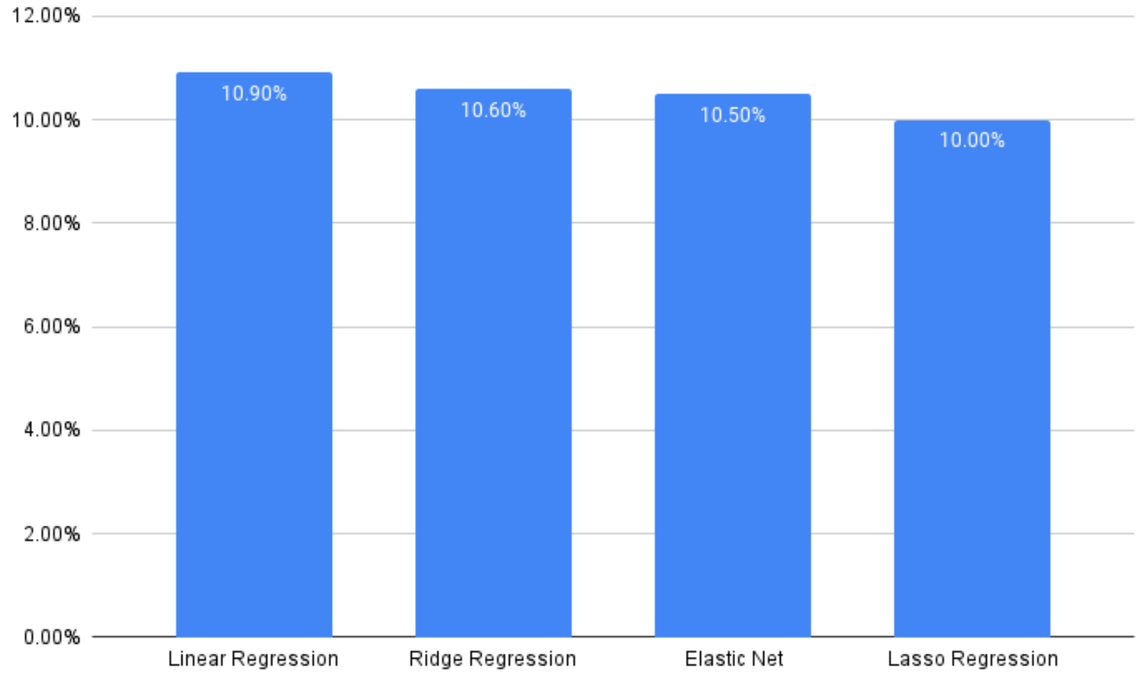
Elastic Net

index	Feature	Coefficients
11	dist_school_1	-7,916,051.33
12	dist_school_2	-4,467,110.85
13	dist_school_3	2,835,066.40
14	dist_school_4	4,364,484.58
15	dist_school_5	-15,332,426.96
32	dist_tran_1	-29,109,111.06
33	dist_tran_2	-17,629,634.70
34	dist_tran_3	-18,814,471.60
35	dist_tran_4	4,134,328.68
36	dist_tran_5	12,436,944.73

Reduced Overfitting

Difference Of Performance

Model	Matrices	Test Score	Train Score	Difference (Train Score - Test Score)
Linear Regression	R2 Score	0.758	0.867	0.109
Ridge Regression	R2 Score	0.760	0.866	0.106
Elastic Net	R2 Score	0.754	0.859	0.105
Lasso Regression	R2 Score	0.764	0.864	0.1



THANK YOU

Q&A