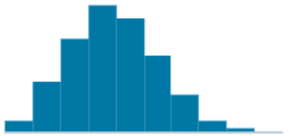
DSI205 PROJECT PRESENTATION

AVERAGE MEDICAL COST ANALYSIS

DATASET FROM KAGGLE

# age age of person	A sex sex	# bmi BMI	A ranking ranking of overall weight	# children number of children	✓ smoker whether or not a person smokes	A region region person lives in
 18 64	male 50% female 50%	 16 53.1	obesity 53% overweight 28% Other (251) 19%	 0 5	 true 274 20% false 1063 80%	southeast 27% southwest 24% Other (648) 48%
19	female	27.90	overweight	0	yes	southwest
18	male	33.77	obesity	1	no	southeast
28	male	33.00	obesity	3	no	southeast
33	male	22.71	healthy weight	0	no	northwest
32	male	28.88	overweight	0	no	northwest
31	female	25.74	overweight	0	no	southeast
46	female	33.44	obesity	1	no	southeast

1337 ROWS × 8 COLUMNS

แหล่งอ้างอิง : HTTPS://WWW.KAGGLE.COM/DATASETS/CANDACEGOSTINSKI/AVERAGE-MEDICAL-COST-ANALYSIS?SELECT=MEDICAL_INSURANCE_DATA_CLEANED_REV.CSV

DATA PREPARATION AND DATA CLEANING

Check Missing value and Duplicate

Check Missing value

	0
age	0
sex	0
bmi	0
ranking	9
children	0
smoker	0
region	0
charges	0

dtype: int64

Check Duplicate

```
age  sex  bmi  ranking  children  smoker  region  charges
```

No Duplicate

DATA PREPARATION AND DATA CLEANING

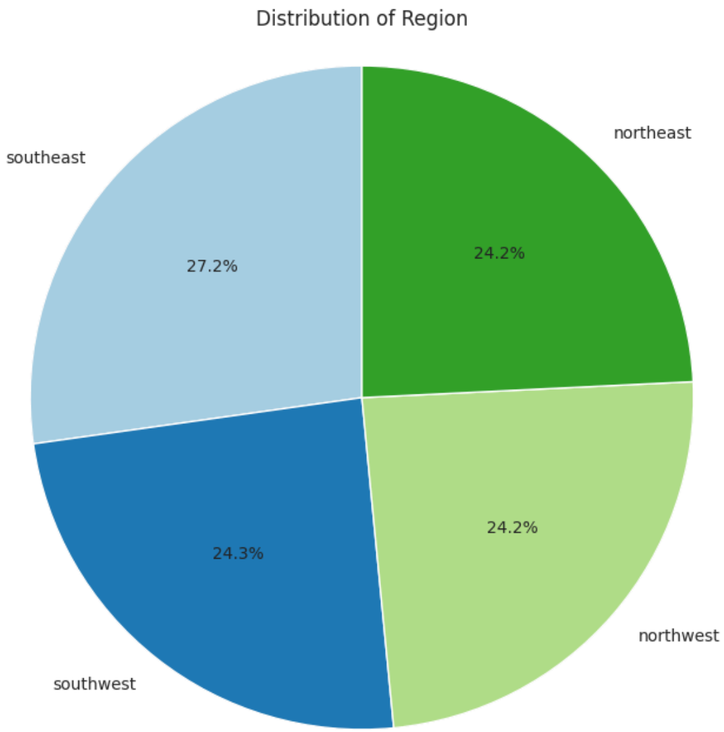
Ordinal encode

ranking
overweight
obesity
obesity
healthy weight
overweight
...
obesity
obesity
obesity
overweight
overweight

ranking
3.0
4.0
4.0
2.0
3.0
...
4.0
4.0
4.0
3.0
3.0

4		3		2		1
obesity	>	overweight	>	healthy weight	>	underweight

DATA PREPARATION AND DATA CLEANING



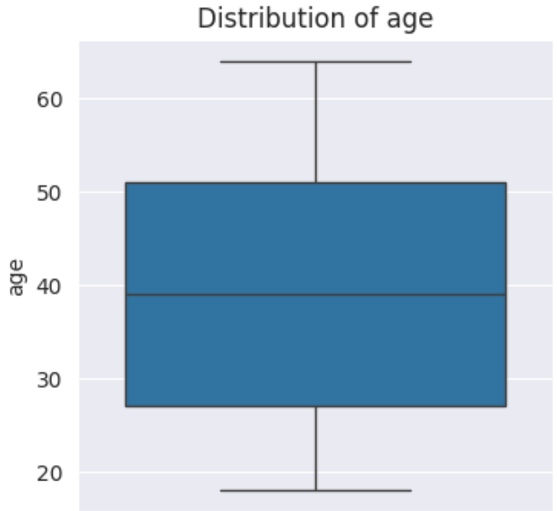
Pie chart showing distribution of region

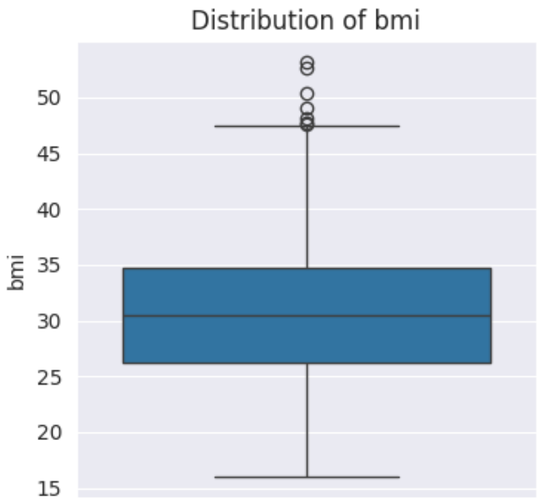
region
southwest
southeast
southeast
northwest
northwest
...
northwest
northeast
southeast
southwest
northwest

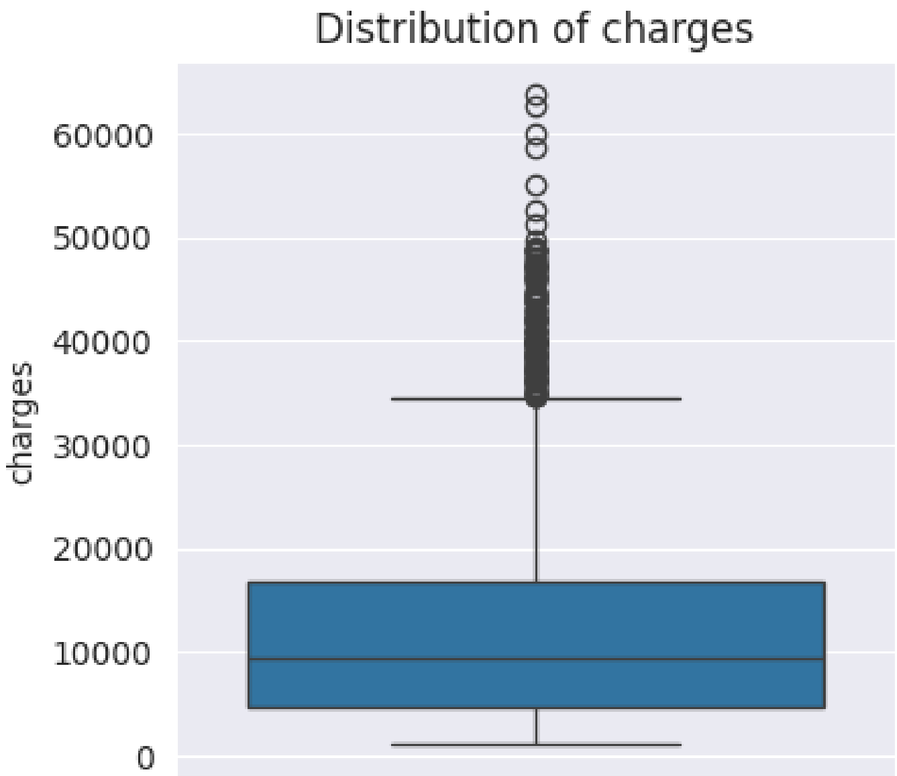
One-hot encode

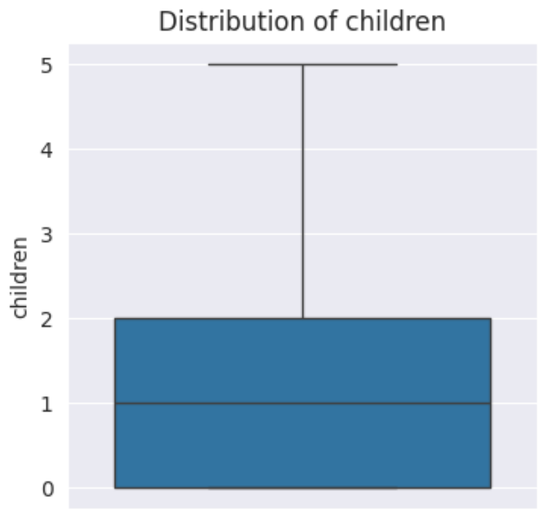
region_northeast	region_northwest	region_southeast	region_southwest
0	0	0	1
0	0	1	0
0	0	1	0
0	1	0	0
0	1	0	0
...	...	...	...
0	1	0	0
1	0	0	0
0	0	1	0
0	0	0	1
0	1	0	0

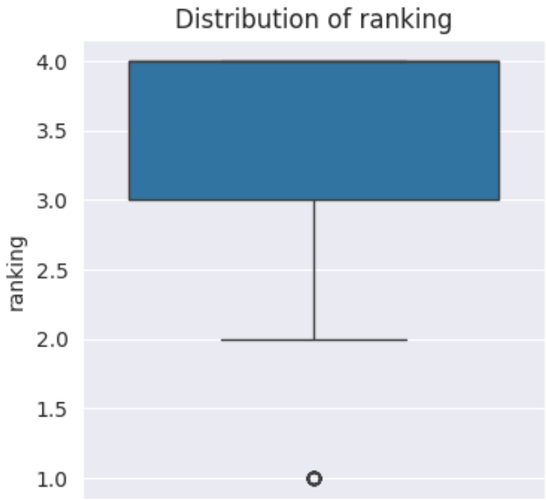
EXPLORATORY DATA ANALYSIS (OUTLIERS)











Target

DATASET FOR MODEL

	age	sex	bmi	ranking	children	smoker	charges	region_northeast	region_northwest	region_southeast	region_southwest
0	0.021739	0	-0.452689	-1.0	-0.5	1	16884.92	0	0	0	1
1	0.000000	1	0.506737	0.0	0.0	0	1725.55	0	0	1	0
2	0.217391	1	0.380884	0.0	1.0	0	4449.46	0	0	1	0
3	0.326087	1	-1.300972	-2.0	-0.5	0	21984.47	0	1	0	0
4	0.304348	1	-0.292512	-1.0	-0.5	0	3866.86	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...
1332	0.695652	1	0.049089	0.0	1.0	0	10600.55	0	1	0	0
1333	0.000000	0	0.204362	0.0	-0.5	0	2205.98	1	0	0	0
1334	0.000000	0	1.010149	0.0	-0.5	0	1629.83	0	0	1	0
1335	0.065217	0	-0.795925	-1.0	-0.5	0	2007.95	0	0	0	1
1336	0.934783	0	-0.261458	-1.0	-0.5	1	29141.36	0	1	0	0

1328 rows × 11 columns

1328 ROWS × 11 COLUMNS

MAKE A MODEL

TRAIN TEST SPLIT

Original Data

x_1	x_2	x_P

y

Train Test split →

X_train

x_1	x_2	x_P

Y_train

y

X_test

x_1	x_2	x_P

Y_test

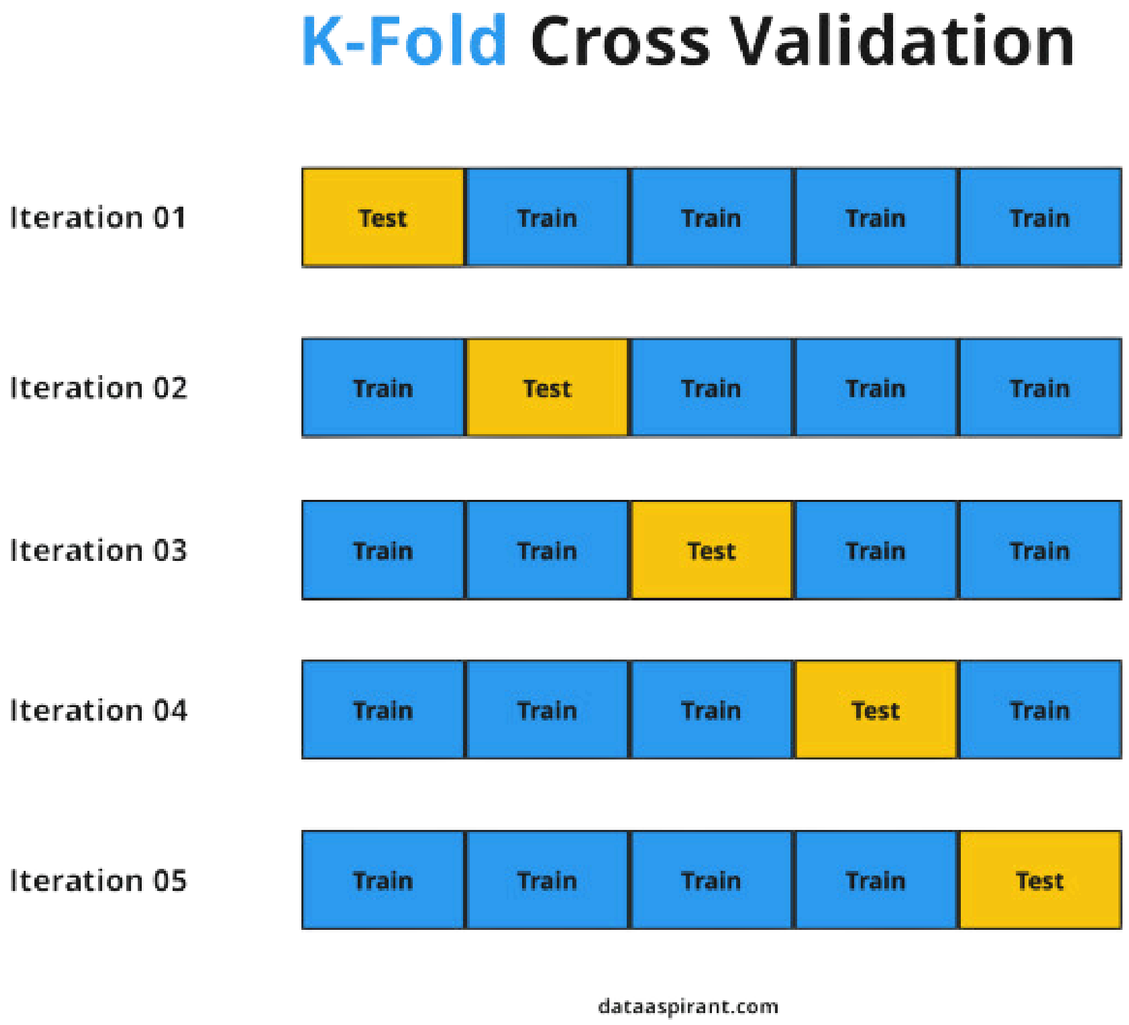
y

Split the DataFrame

80% for training

20% for testing

CROSS-VALIDATION



https://dataaspirant.com/cross-validation/

EVALUATION

R2

$$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

MSE

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

MAE

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

RMSE

$$\sqrt{MSE}$$

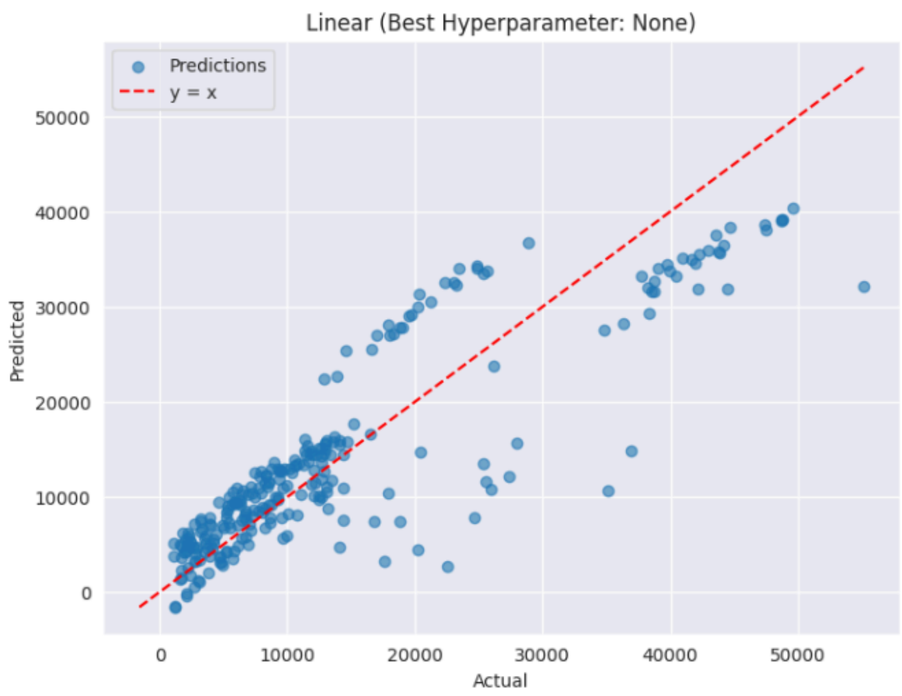
MULTIPLE LINEAR REGRESSIONS

The general form of a Multiple Linear regression equation is :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

$$\beta = \left(X^T X\right)^{-1} X^T y$$

MULTIPLE LINEAR REGRESSIONS



MODEL	MATRICES	TEST SCORE	TRAIN SCORE
Linear Regression	R2 score	0.7729	0.7384
Linear Regression	MSE	3.5277e+07	3.6583e+07
Linear Regression	RMSE	5939.4635	6048.4181
Linear Regression	MAE	4321.2180	4289.0964

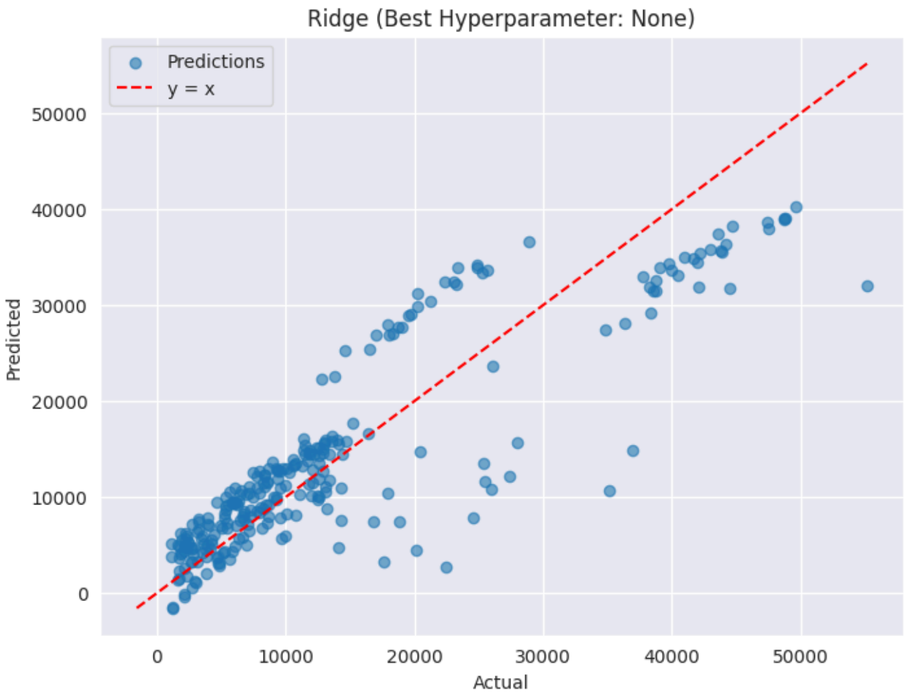
Evaluation metrics to measure model performance

RIDGE REGRESSION

The general form of a Ridge regression equation is :

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \beta \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

RIDGE REGRESSION



MODEL	MATRICES	TEST SCORE	TRAIN SCORE
Ridge Regression	R2 score	0.7727	0.7384
Ridge Regression	MSE	3.5310e+07	3.6586e+07
Ridge Regression	RMSE	5942.2248	6048.6755
Ridge Regression	MAE	4331.7117	4298.5781

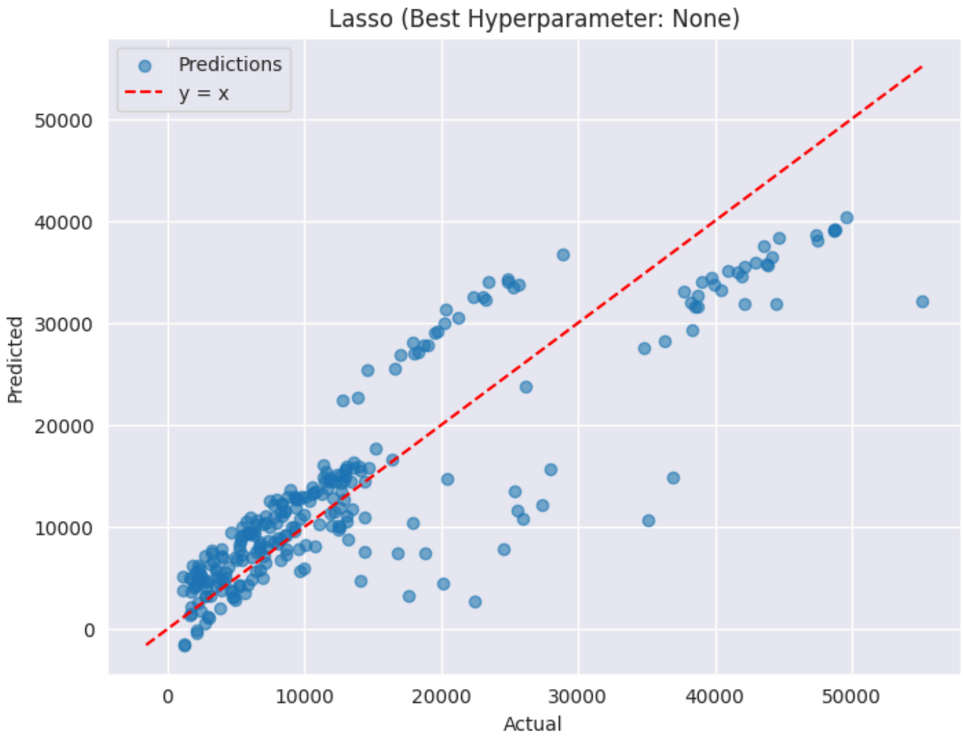
Evaluation metrics to measure model performance

LASSO REGRESSION

The general form of a Lasso regression equation is :

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \beta \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

LASSO REGRESSION



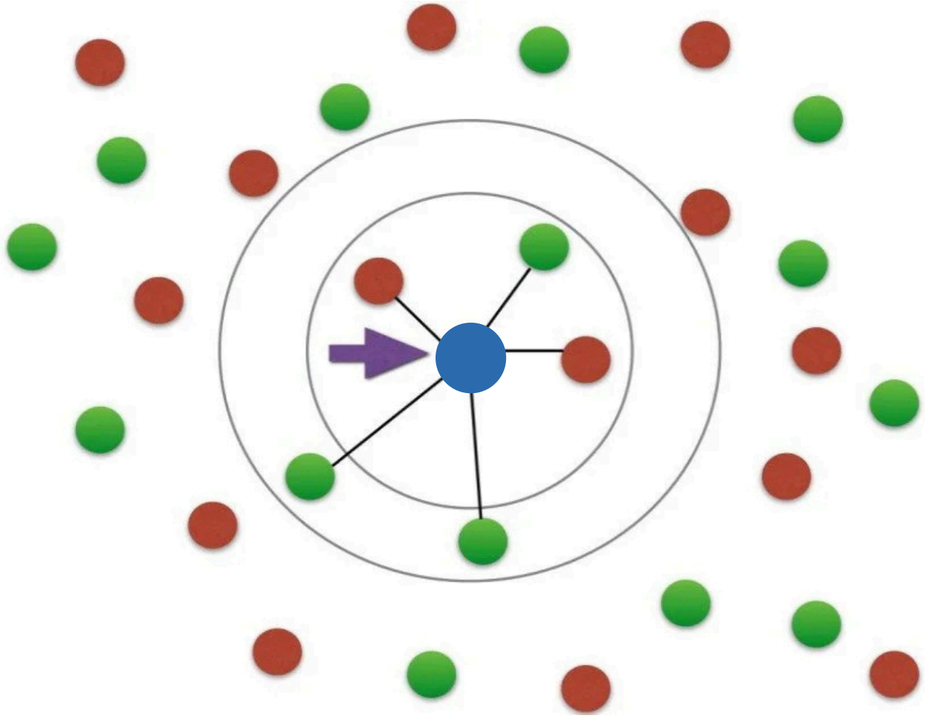
MODEL	MATRICES	TEST SCORE	TRAIN SCORE
Lasso Regression	R2 score	0.7729	0.7384
Lasso Regression	MAE	4320.6320	4288.9472
Lasso Regression	MSE	3.5272e+07	3.6583e+07
Lasso Regression	RMSE	5939.0945	6048.4209

Evaluation metrics to measure model performance

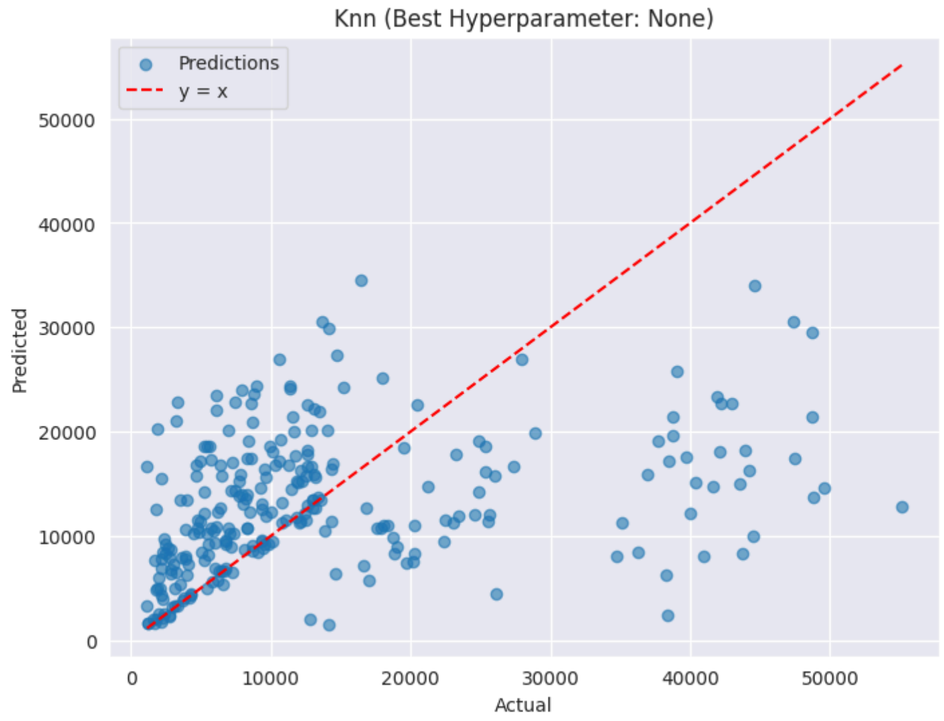
KNN REGRESSION

The general form of a KNN regression equation is :

$$\hat{y} = \frac{1}{k} \sum_{i \in S} y_i$$



KNN REGRESSION



MODEL	MATRICES	TEST SCORE	TRAIN SCORE
KNN Regression	R2 score	0.1017	0.0980
KNN Regression	MAE	8361.2241	5956.3115
KNN Regression	MSE	1.3956e+08	7.8448e+07
KNN Regression	RMSE	11813.7344	8857.1173

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Evaluation metrics to measure model performance

COMPARISON

Lasso Regression was the most accurate on the Test data set because it had the best R^2, MSE, and MAE values.

MODEL	R2 SCORE	MSE	RMSE	MAE
Lasso Regression	0.7729	3.5272e+07	5939.0945	4320.6320
Ridge Regression	0.7727	3.5310e+07	5942.2248	4331.7117
KNN regression	0.1017	1.3956e+08	11813.7344	8361.2241
Linear Regression	0.7729	3.5277e+07	5939.4635	4321.280

OUR EXPERIENCE

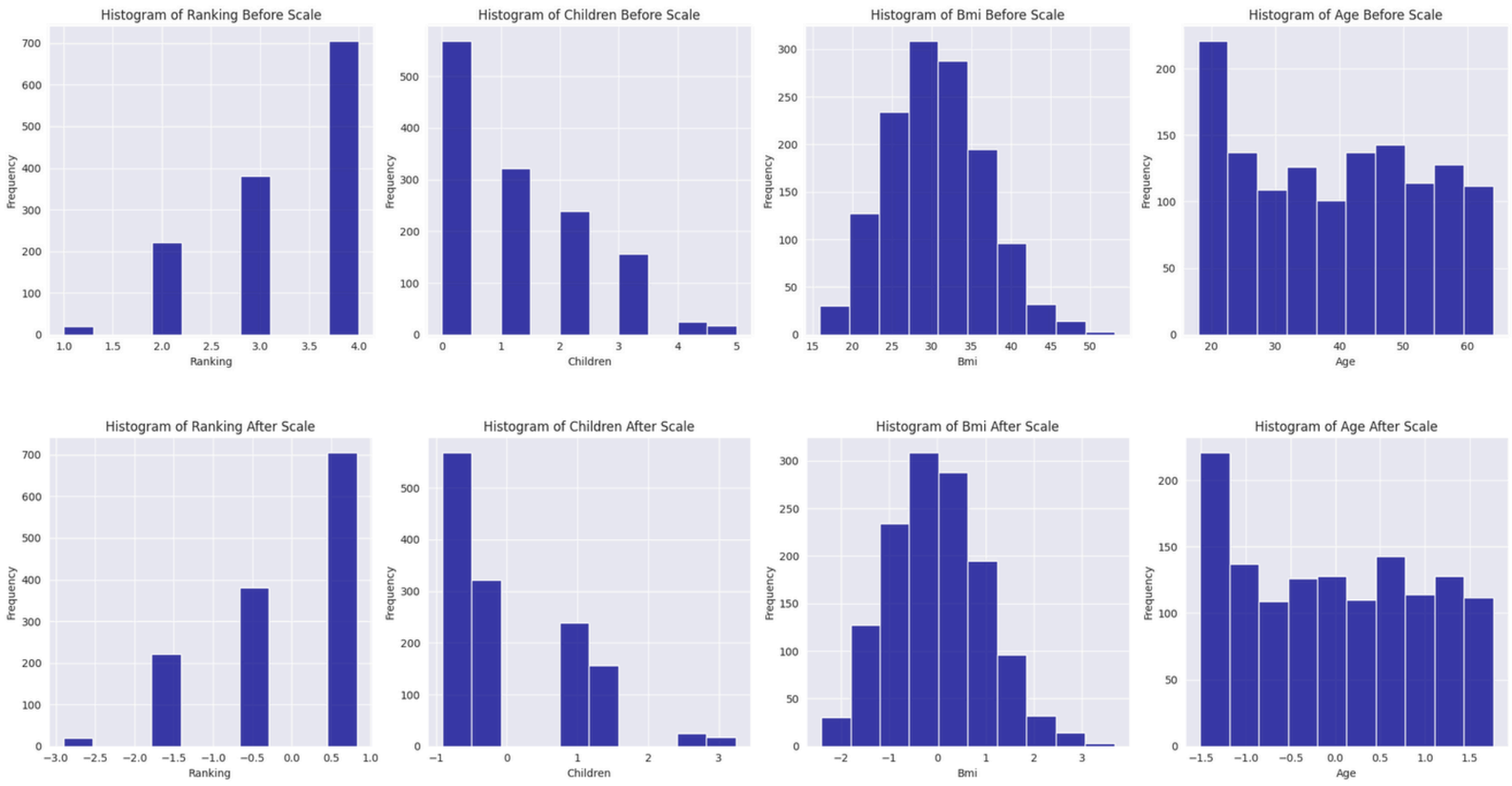
SCALE

STANDARD SCALE

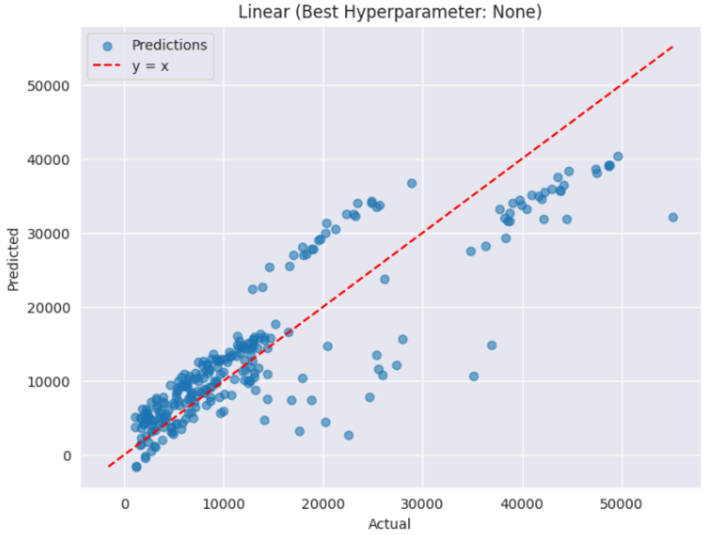
$$X_{scaled} = \frac{X - \mu}{\sigma}$$

Before

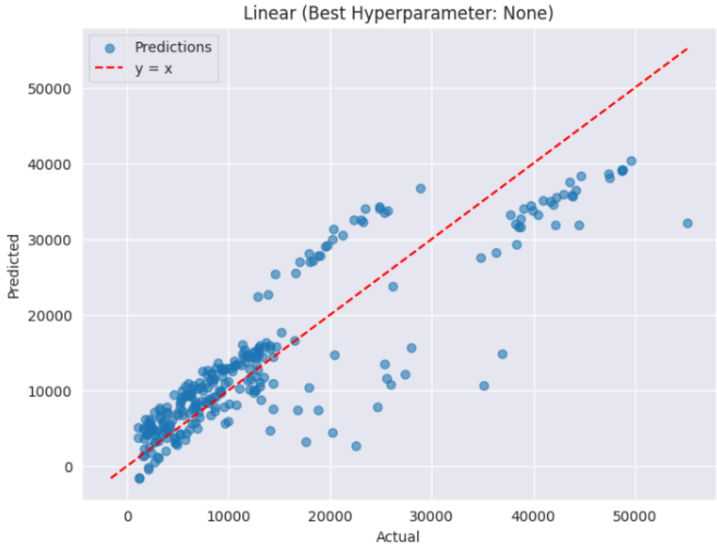
After



LINEAR REGRESSIONS



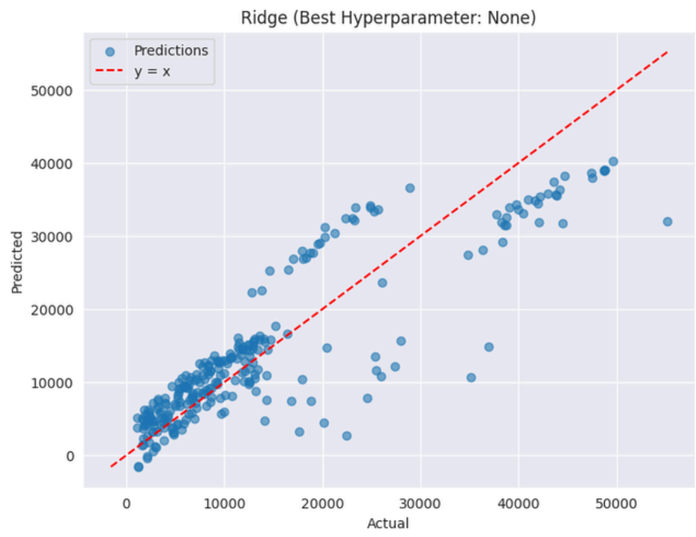
MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Linear Regression	0.77294	3.6583e+07	6048.4181	4289.0964	-

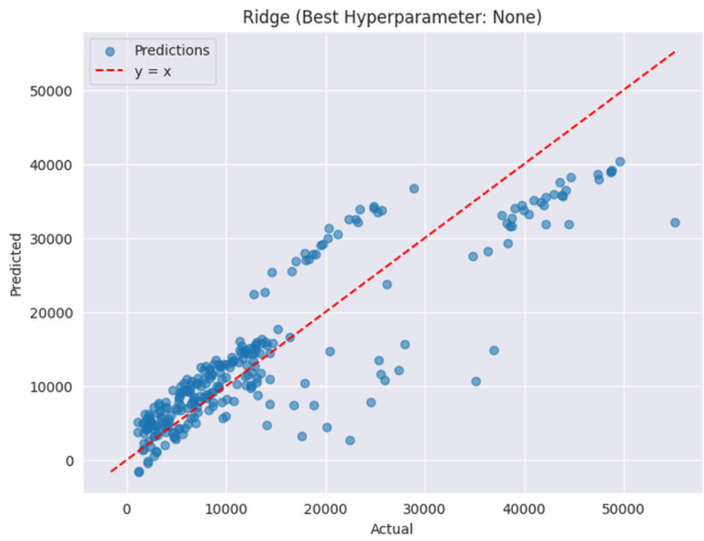


MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Linear Regression	0.77294	3.5277e+07	5939.4635	4321.2180	SCALE

Test score before scale and after scaled

RIDGE REGRESSIONS



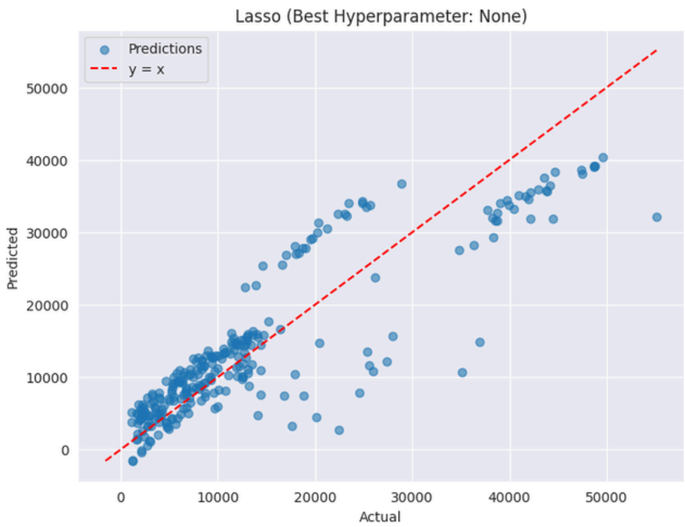


MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Ridge Regression	0.77273	3.6583e+07	6048.4261	4290.4143	-

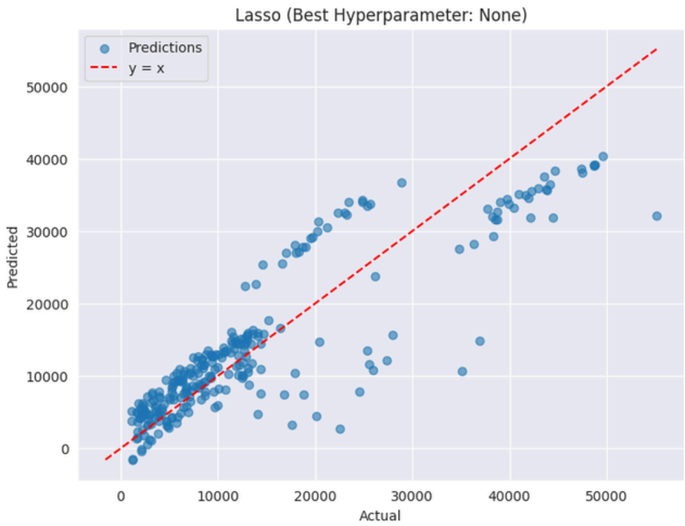
MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Ridge Regression	0.77289	3.5285e+07	5940.1655	4322.8544	SCALE

Test score before scale and after scaled

LASSO REGRESSIONS



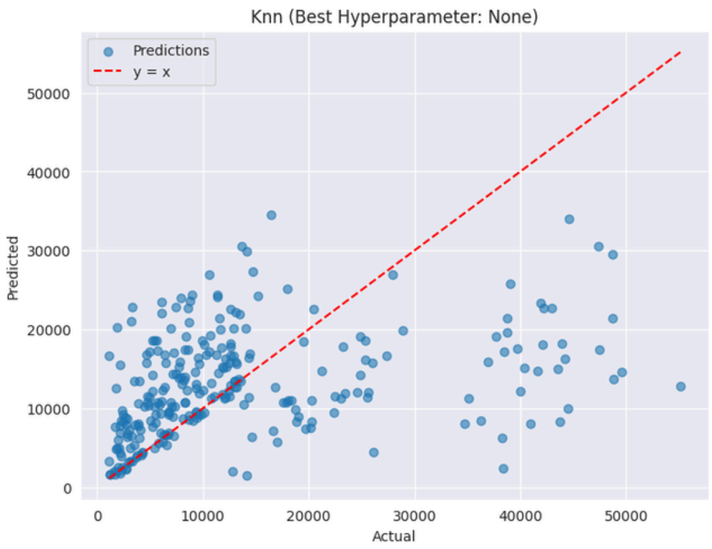
MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Lasso Regression	0.77297	3.6538e+07	6048.4188	4288.9171	-



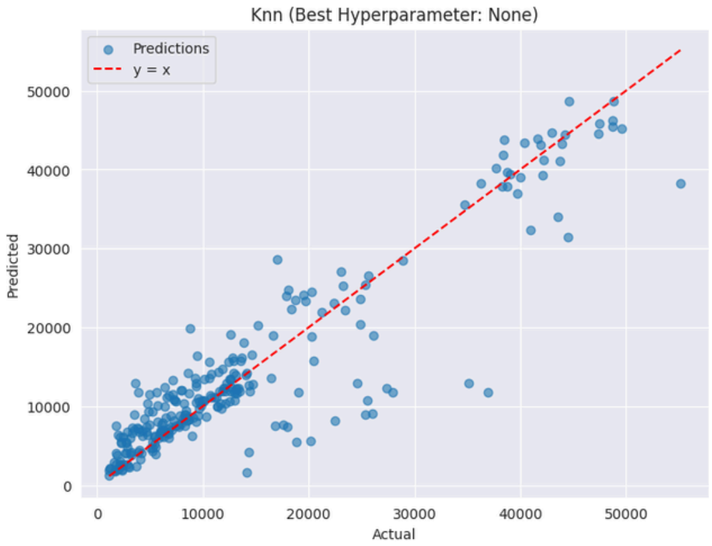
MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Lasso Regression	0.77294	3.5277e+07	5939.4648	4320.9714	SCALE

Test score before scale and after scaled

KNN REGRESSIONS



MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
KNN Regression	0.101717	1.395e+08	4317.0390	2652.9340	-



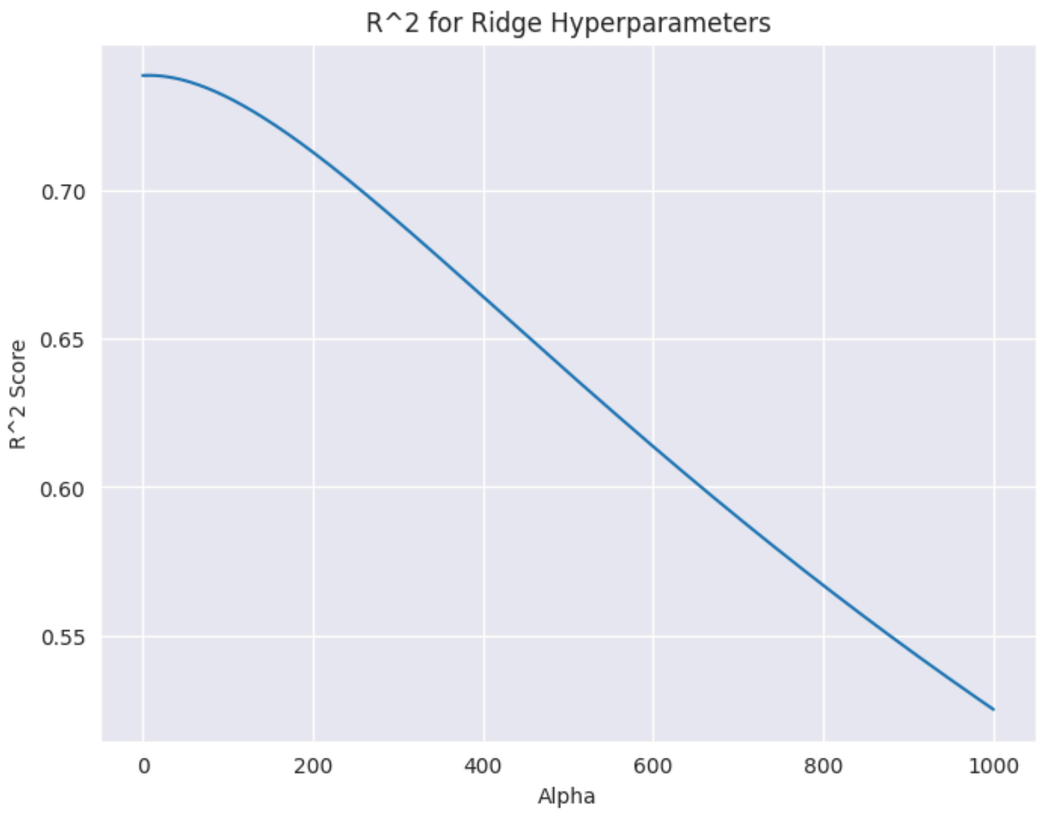
MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
KNN Regression	0.84558	2.394e+07	4893.4685	2983.2659	SCALE

Test score before scale and after scaled

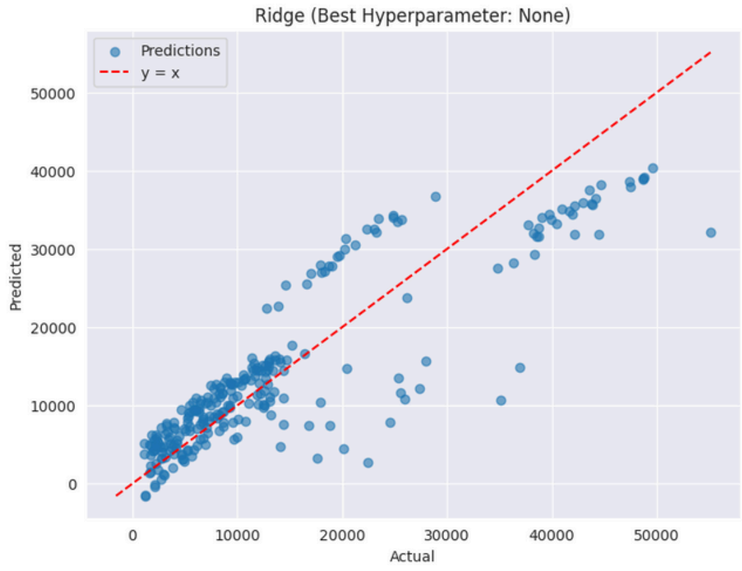
HYPERPARAMETER

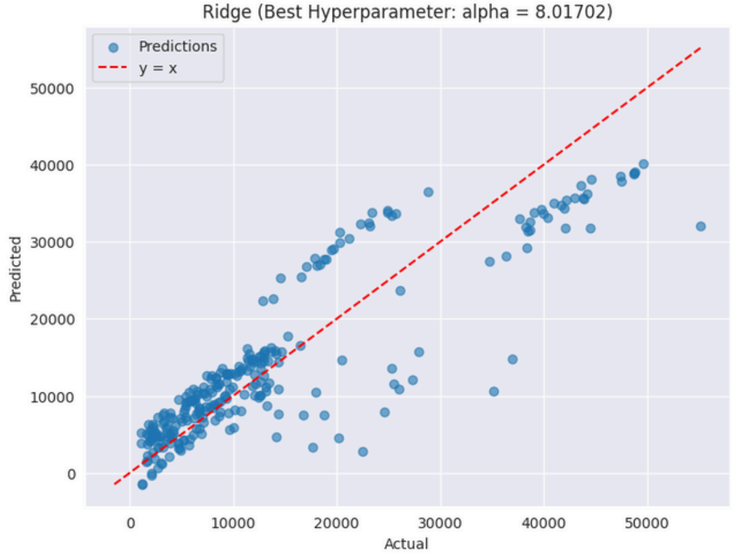
GRID SEARCH

RIDGE REGRESSION



ALPHA = 8.01702



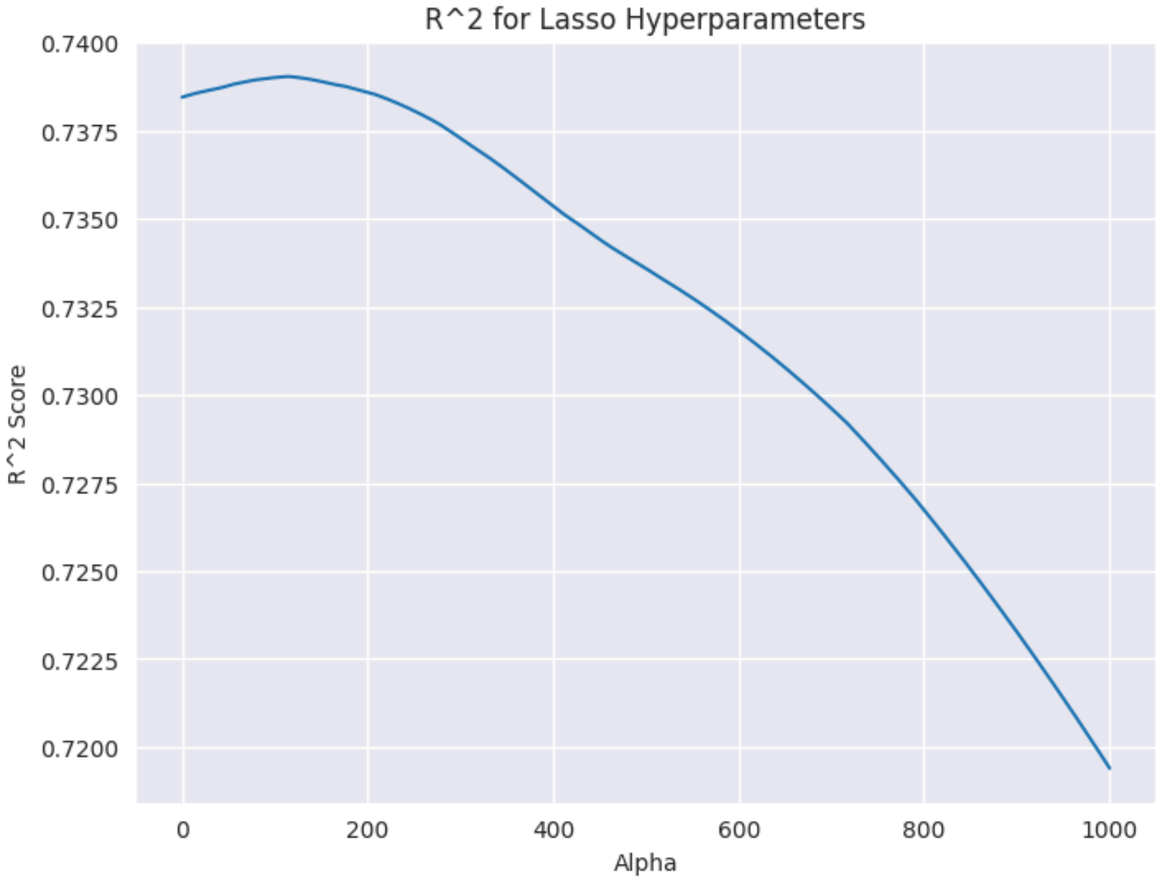


MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Ridge Regression	0.77289	3.528e+07	5940.16	4322.85	SCALE

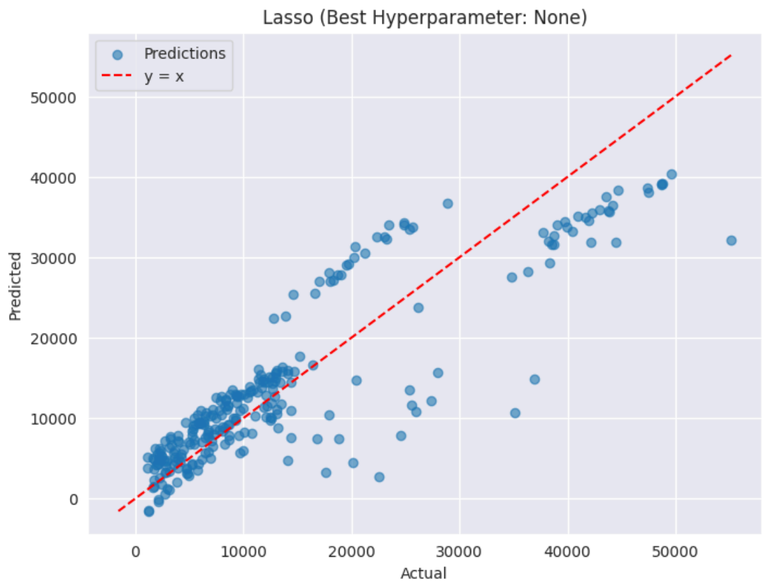
MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Ridge Regression	0.77248	3.534e+07	5945.51	4334.27	ALPHA = 8.01702

Test score before Hyperparameter and after Hyperparameter

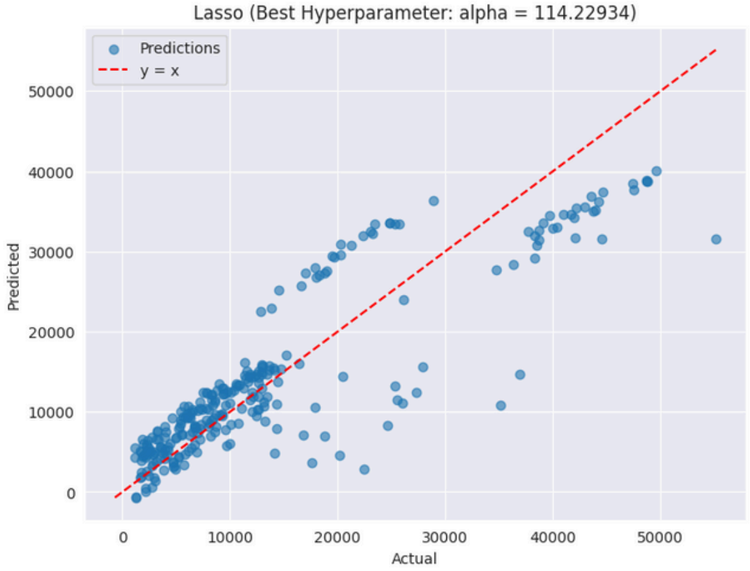
LASSO REGRESSION



ALPHA = 114.229



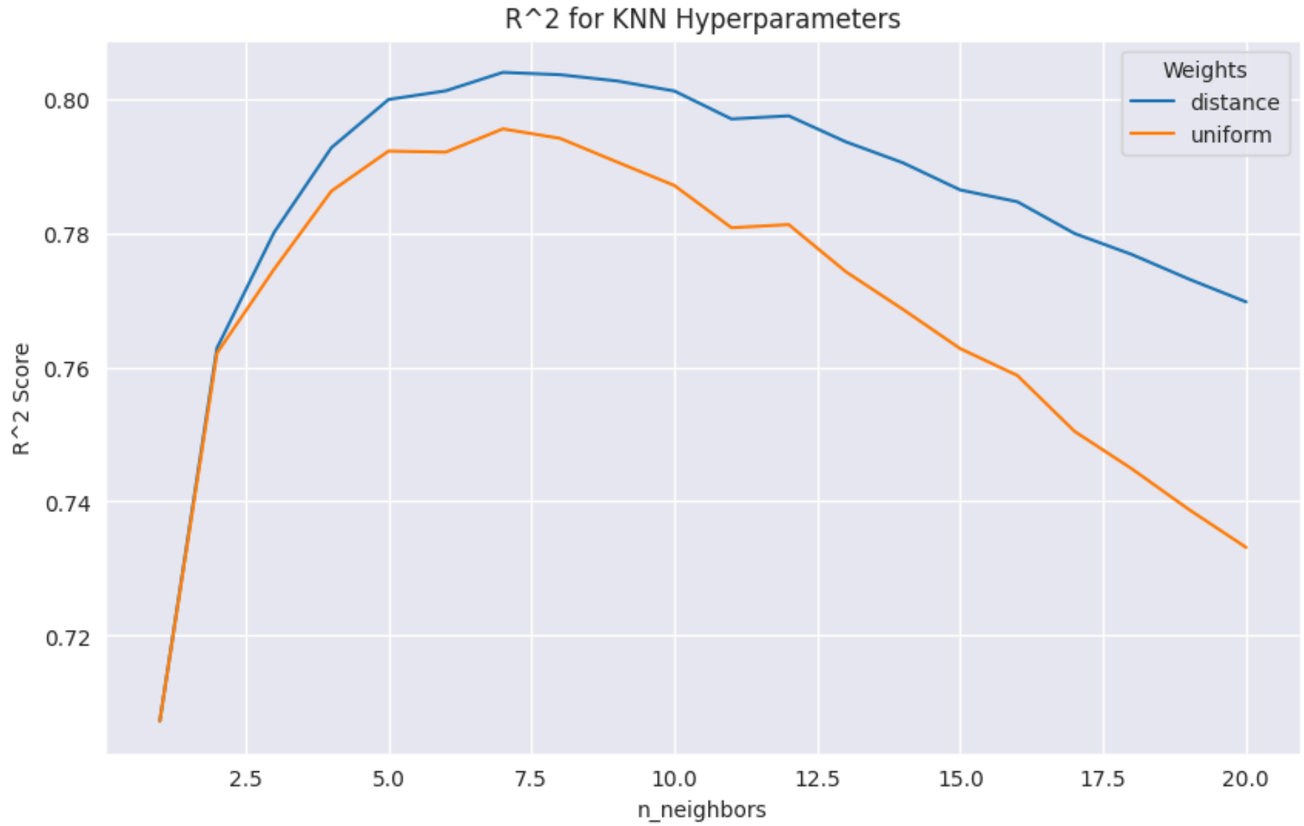
MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Lasso Regression	0.77294	3.527e+07	5939.46	4320.97	SCALE



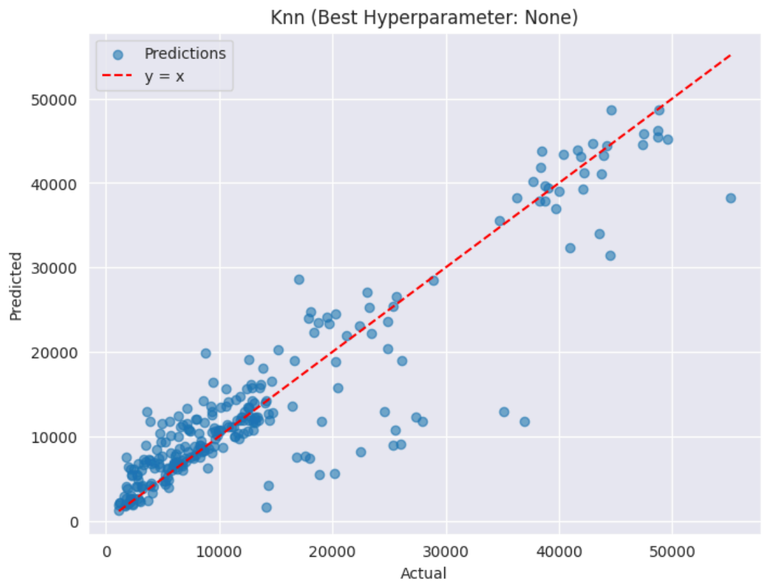
MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
Lasso Regression	0.77240	3.536e+07	5946.52	4302.85	ALPHA = 114.229

Test score before Hyperparameter and after Hyperparameter

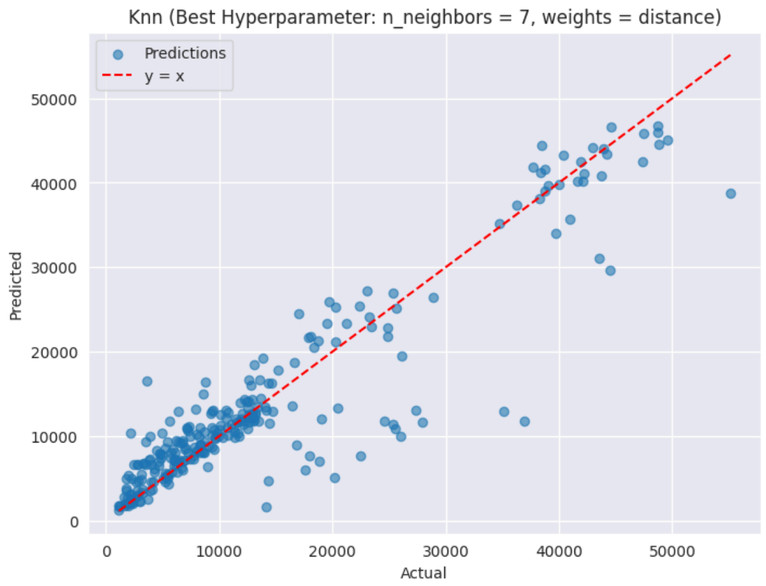
KNN REGRESSION



n_neighbors = 7, weights = distance



MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
KNN Regression	0.845875	2.394e+07	4893.46	293.26	SCALE



MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
KNN Regression	0.85441	2.261e+07	4755.97	2812.45	n_neighbors = 7, weights = distance

Test score before Hyperparameter and after Hyperparameter

CONCLUSION

MODEL	R2 SCORE	MSE	RMSE	MAE	OTHER
KNN Regression	0.85441	2.261e+07	4755.97	2812.45	n_neighbors = 7, weights = distance
KNN Regression	0.84587	2.394e+07	4893.46	2983.26	SCALE
Lasso Regression	0.77297	3.527e+07	5939.09	4320.63	-
Lasso Regression	0.77294	3.527e+07	5939.46	4320.97	SCALE
Linear Regression	0.77294	3.527e+07	5939.46	4321.21	-

KNN Regression When hyperparameter is ‘n_neighbors = 7, weights = distance’ was the most accurate on the Test data set because it had the best R^2, MSE, RMSE and MAE values.

MODEL	MATRICES	TEST SCORE	TRAIN SCORE	OHTER
KNN Regression	R2 score	0.85441	0.80370	n_neighbors = 7 , weights = distance
KNN Regression	MSE	2.261e+07	2.841e+07	n_neighbors = 7 , weights = distance
KNN Regression	RMSE	4755.97	6048.4181	n_neighbors = 7 , weights = distance
KNN Regression	MAE	4321.2286	4289.1058	n_neighbors = 7 , weights = distance

THANK YOU

Q&A