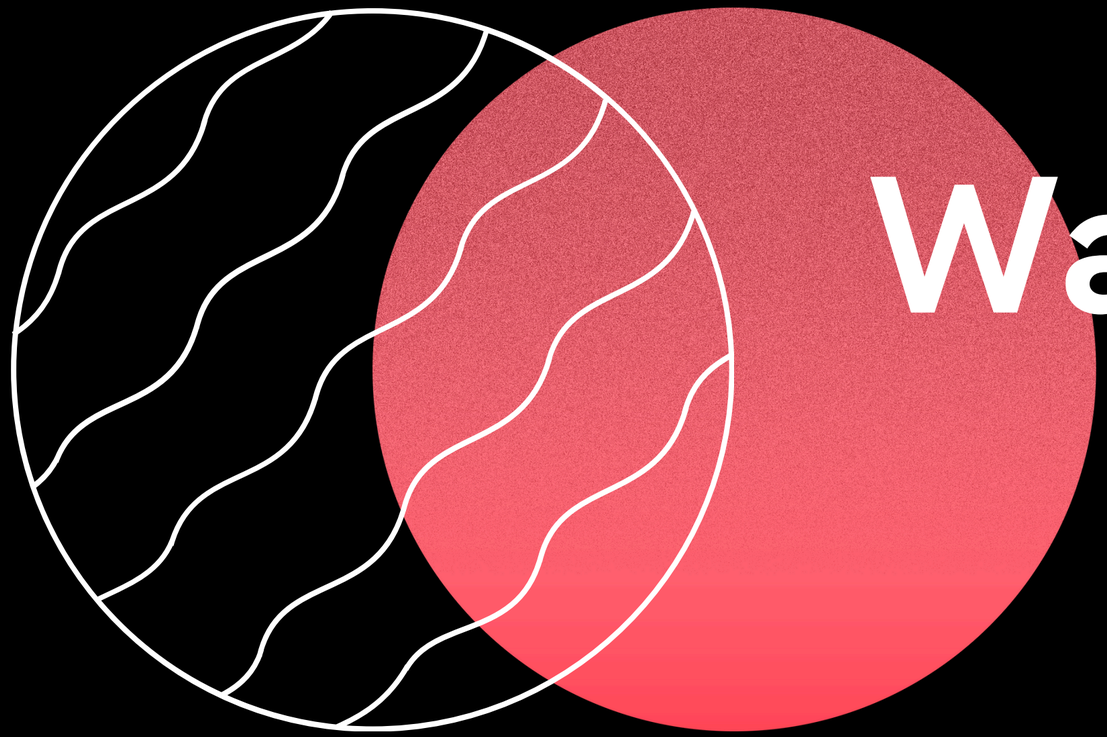


Watt's Up, Doc?

Forecasting Electricity Prices.

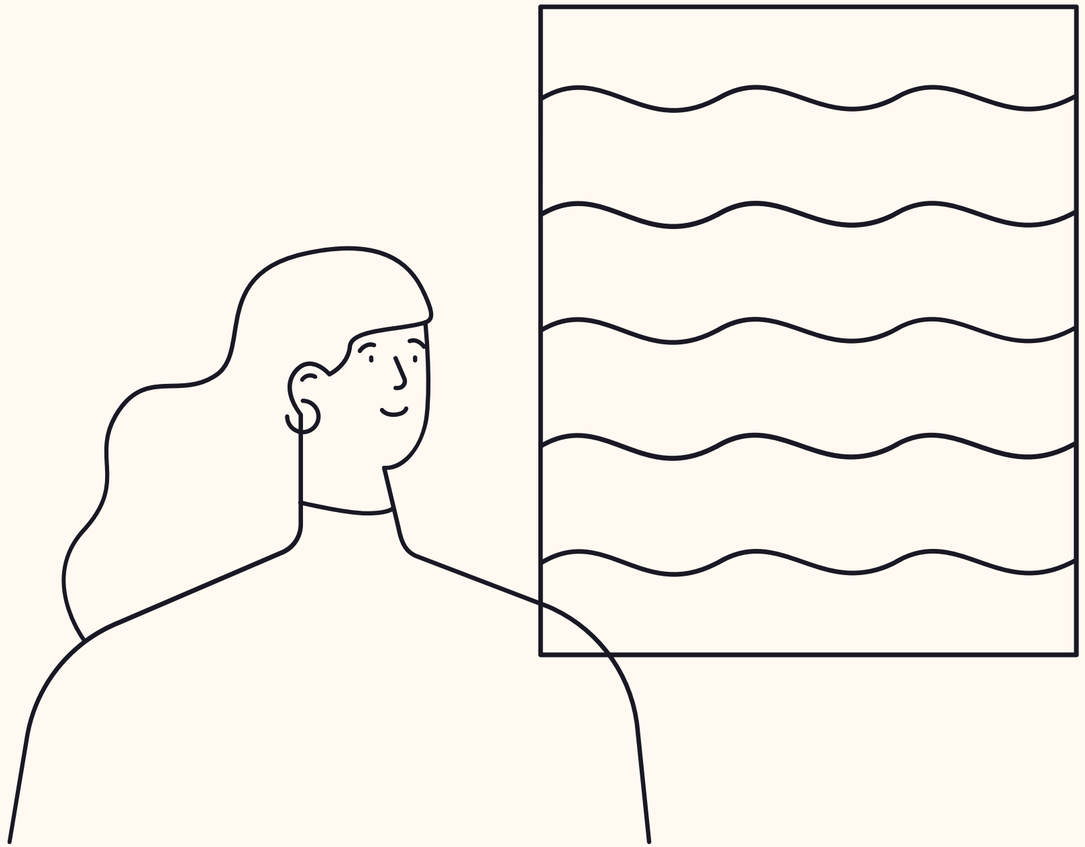
A Machine Learning Project by Gem Mercado

https://tinyurl.com/ML1FPGMZM

Problem: Extrapolation.

We're going to delve into a time series problem to ultimately predict the hourly prices of electricity for 21 days.

Our aim is to equip energy traders with a robust forecast model.



THE DATASET

'time', 'generation biomass', 'generation fossil brown coal/lignite', 'generation fossil coal-derived gas', 'generation fossil gas', 'generation fossil hard coal', 'generation fossil oil', 'generation fossil oil shale', 'generation fossil peat', 'generation geothermal', 'generation hydro pumped storage aggregated', 'generation hydro pumped storage consumption', 'generation hydro run-of-river and poundage', 'generation hydro water reservoir', 'generation marine', 'generation nuclear', 'generation other', 'generation other renewable', 'generation solar', 'generation waste', 'generation wind offshore', 'generation wind onshore', 'forecast solar day ahead', 'forecast wind offshore eday ahead', 'forecast wind onshore day ahead', 'total load forecast', 'total load actual', 'price day ahead', 'price actual'

Specific aspects of **electricity generation data**, and **electricity prices** for **each hour**.

Initial dataset consists of **35,064 rows** and **29 columns**

Data Preprocessing

Missing Values

Categorical Values

Projection Variables

DateTime Variables

Feature Engineering!

Handle NaN Rows and Variables, Categorical, and Projection Variables.

Extract and concatenate DateTime Variables

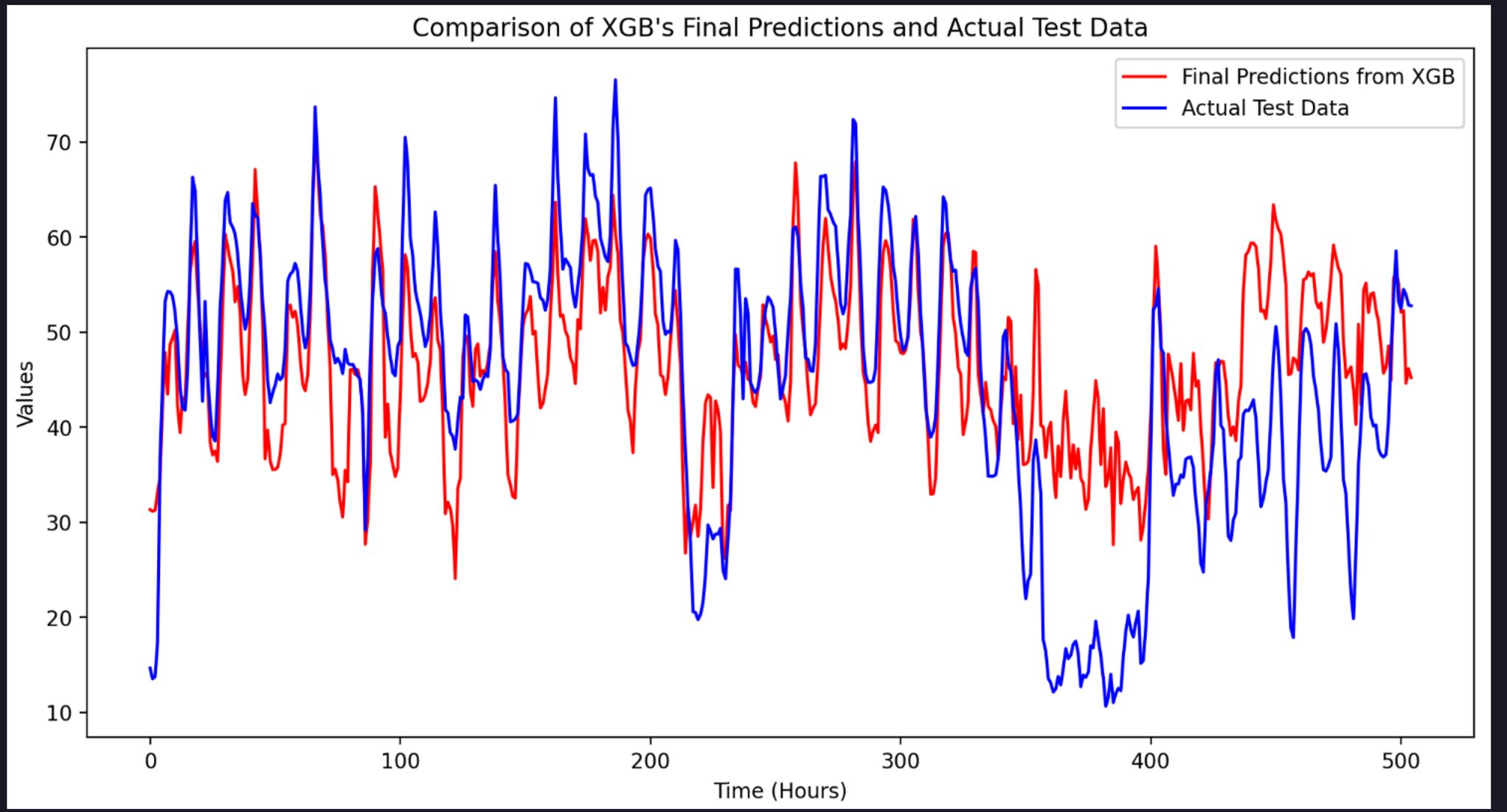
'time', 'year', 'month', 'day', 'hour', 'weekday', 'generation biomass', 'generation fossil brown coal/lignite', 'generation fossil coal-derived gas', 'generation fossil gas', 'generation fossil hard coal', 'generation fossil oil', 'generation fossil oil shale', 'generation fossil peat', 'generation geothermal', 'generation hydro pumped storage aggregated', 'generation hydro pumped storage consumption', 'generation hydro run-of-river and poundage', 'generation hydro water reservoir', 'generation marine', 'generation nuclear', 'generation other', 'generation other renewable', 'generation solar', 'generation waste', 'generation wind offshore', 'generation wind onshore', 'forecast solar day ahead', 'forecast wind offshore eday ahead', 'forecast wind onshore day ahead', 'total load forecast', 'total load actual', 'price day ahead', 'price actual'

AutoML Assessment: Mean Absolute Error

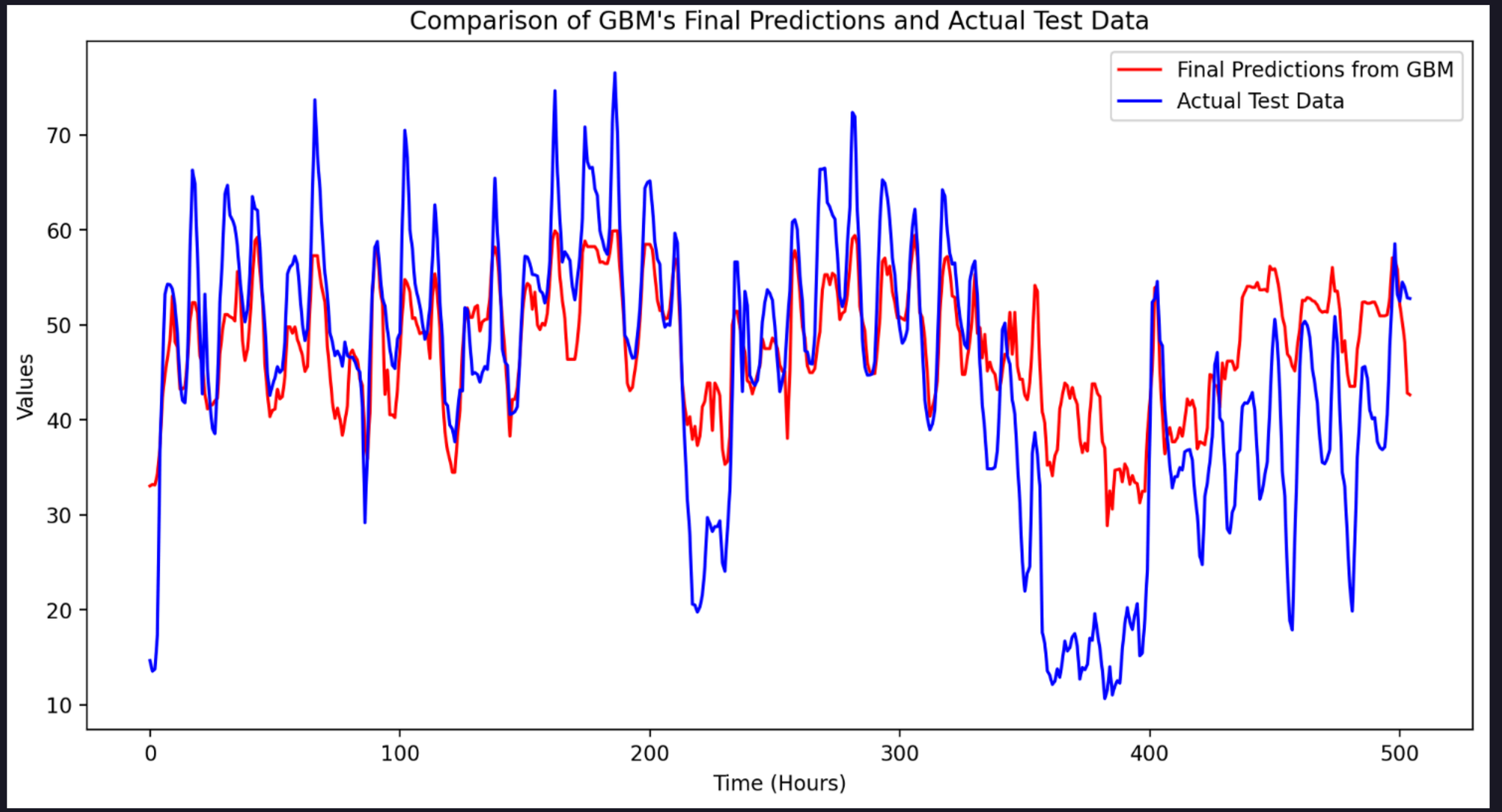
Baseline MAE of 2.52

	Machine Learning Regression Method	MAE Validation	MAE Test	Most Important Feature
0	Linear Regression	6.958866	9.929467	weekday
1	Linear Regression + Lasso	6.949586	9.713221	month
2	Linear Regression + Ridge	6.958866	9.929465	weekday
3	kNN	9.292658	10.857208	NA
4	Decision Tree	8.531879	11.031030	generation fossil hard coal
5	Random Forest	6.478660	8.566274	generation fossil hard coal
6	Gradient Boosting Method	5.668675	7.798040	generation fossil hard coal
7	XGBoost	6.136421	8.072980	year

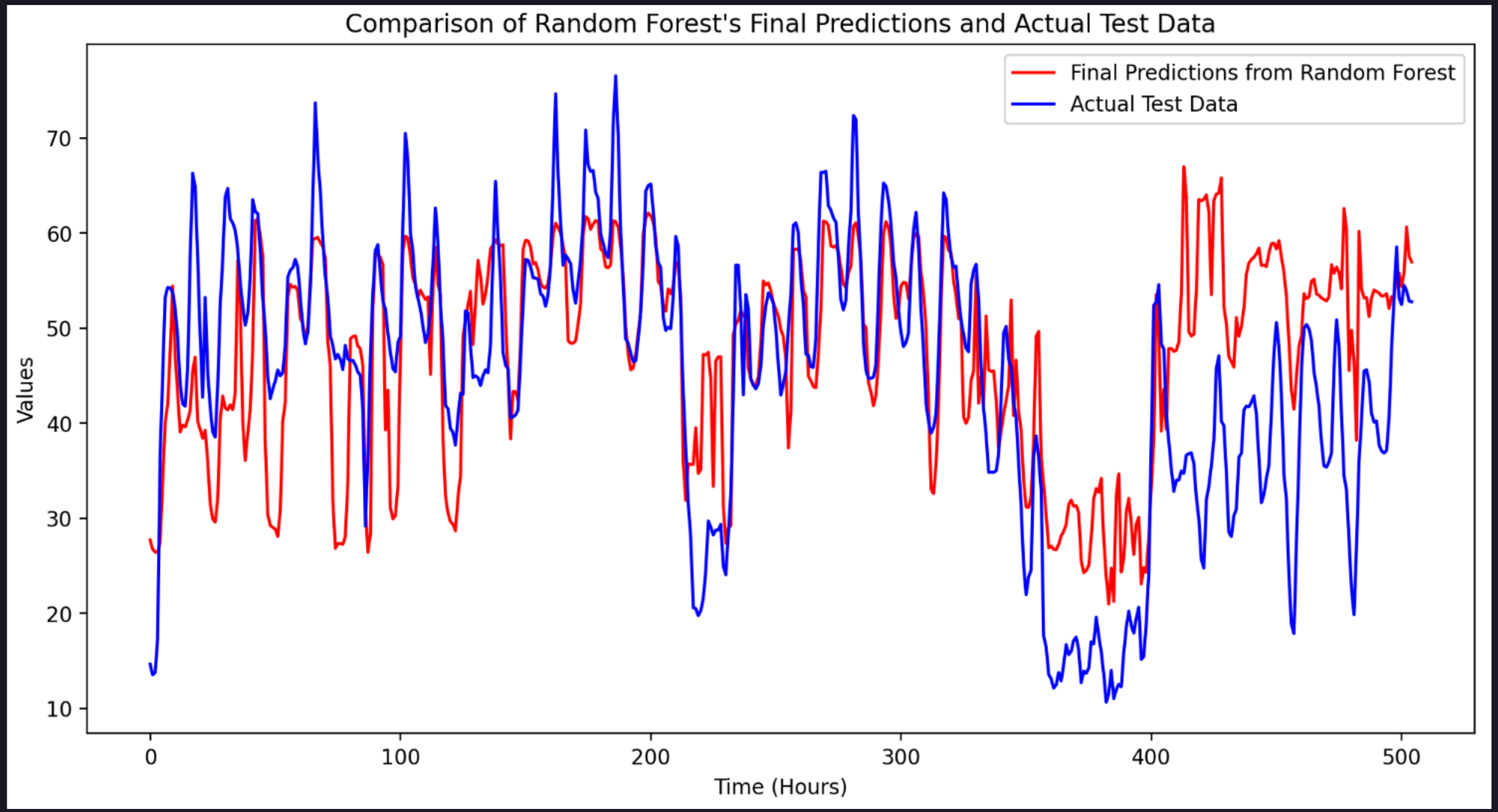
07 A.


Comparison of XGB's Final Predictions and Actual Test Data
Final Predictions from XGB
Actual Test Data
Values
Time (Hours)
10
20
30
40
50
60
70
0
100
200
300
400
500

07 B.


Comparison of GBM's Final Predictions and Actual Test Data
Final Predictions from GBM
Actual Test Data
Values
Time (Hours)
10
20
30
40
50
60
70
0
100
200
300
400
500

07 C.


Comparison of Random Forest's Final Predictions and Actual Test Data
Final Predictions from Random Forest
Actual Test Data
Values
Time (Hours)
10
20
30
40
50
60
70
0
100
200
300
400
500

What if we use history to predict the future?

Enter Lag Values.

lag_1	lag_2	lag_3	lag_4	lag_5	lag_6	lag_7	price actual
51.73	53.63	56.04	59.32	64.48	64.92	65.41	51.43
51.43	51.73	53.63	56.04	59.32	64.48	64.92	48.98
48.98	51.43	51.73	53.63	56.04	59.32	64.48	54.20
54.20	48.98	51.43	51.73	53.63	56.04	59.32	58.94
58.94	54.20	48.98	51.43	51.73	53.63	56.04	59.86

Feature Engineering: Lag Values

AutoML with Lag Values

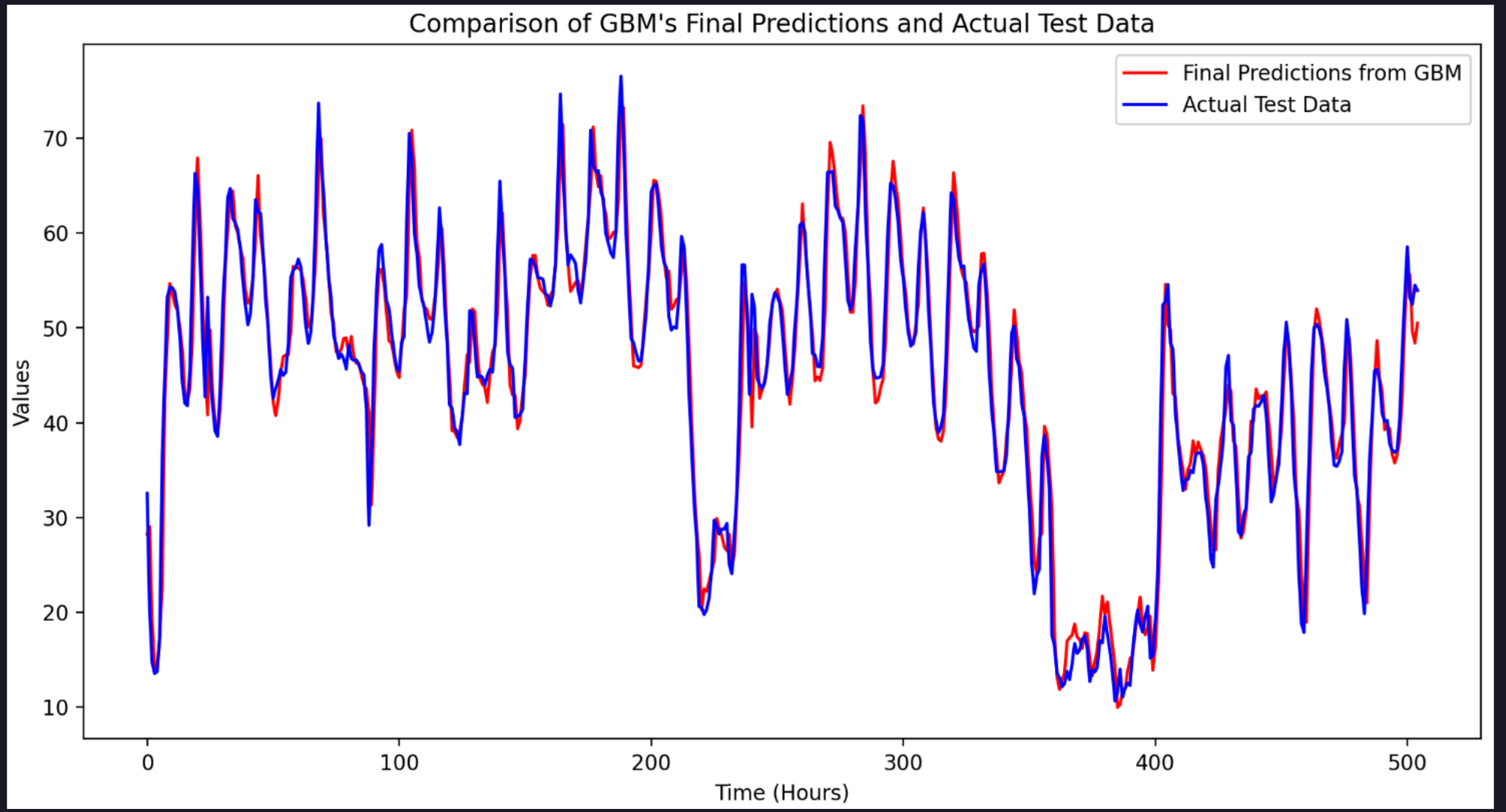
	Machine Learning Regression Method	MAE Validation	MAE Test	Most Important Feature
0	Linear Regression	1.973514	2.587625	lag_1
1	Linear Regression + Lasso	1.993349	2.589428	lag_1
2	Linear Regression + Ridge	1.973507	2.587614	lag_1
3	kNN	9.263018	10.812737	NA
4	Decision Tree	2.732453	3.551366	lag_1
5	Random Forest	1.797893	2.345040	lag_1
6	Gradient Boosting Method	1.814165	2.251013	lag_1
7	XGBoost	1.836604	2.509573	lag_1

AutoML with Lag Values

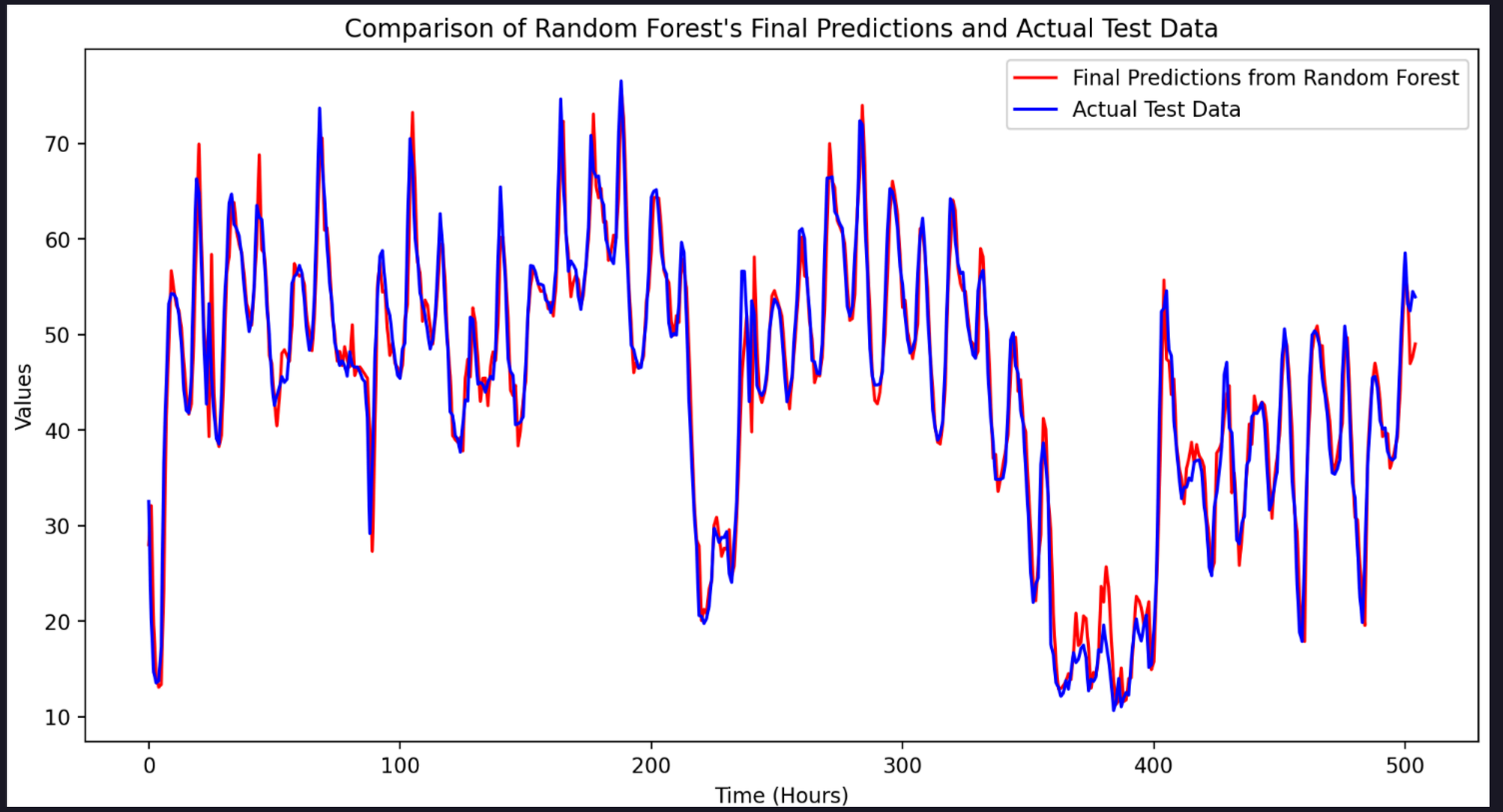
Baseline MAE of 2.52

	Machine Learning Regression Method	MAE Validation	MAE Test	Most Important Feature
0	Linear Regression	1.973514	2.587625	lag_1
1	Linear Regression + Lasso	1.993349	2.589428	lag_1
2	Linear Regression + Ridge	1.973507	2.587614	lag_1
3	kNN	9.263018	10.812737	NA
4	Decision Tree	2.732453	3.551366	lag_1
5	Random Forest	1.797893	2.345040	lag_1
6	Gradient Boosting Method	1.814165	2.251013	lag_1
7	XGBoost	1.836604	2.509573	lag_1

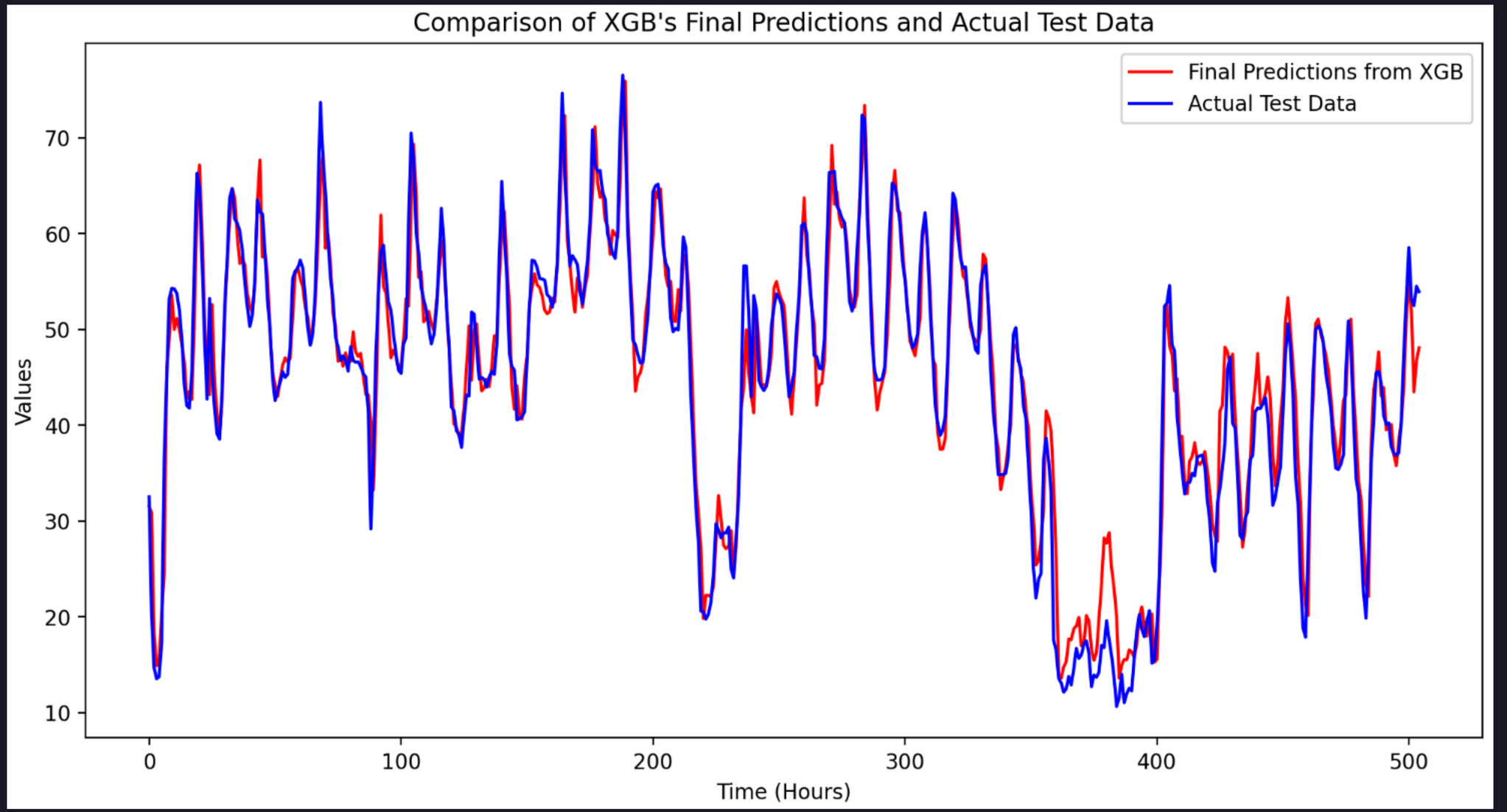
11 A.

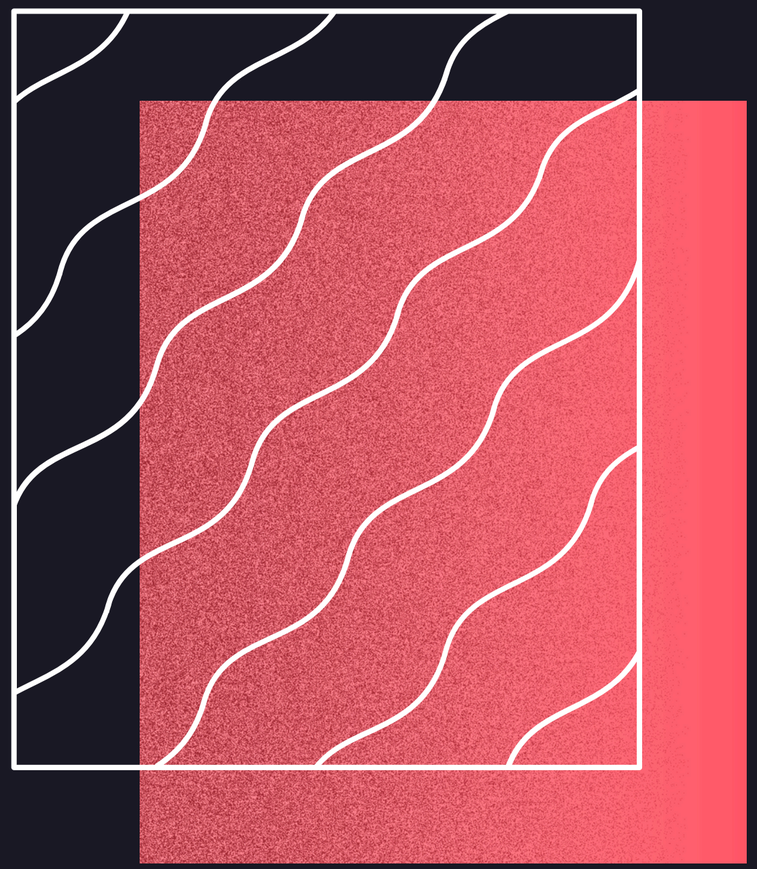

Comparison of GBM's Final Predictions and Actual Test Data
Final Predictions from GBM
Actual Test Data
Values
Time (Hours)
10
20
30
40
50
60
70
0
100
200
300
400
500

11 B.


Comparison of Random Forest's Final Predictions and Actual Test Data
Final Predictions from Random Forest
Actual Test Data
Values
Time (Hours)
10
20
30
40
50
60
70
0
100
200
300
400
500

11 C.


Comparison of XGB's Final Predictions and Actual Test Data
Final Predictions from XGB
Actual Test Data
Values
Time (Hours)
10
20
30
40
50
60
70
0
100
200
300
400
500



HIGHLIGHTS

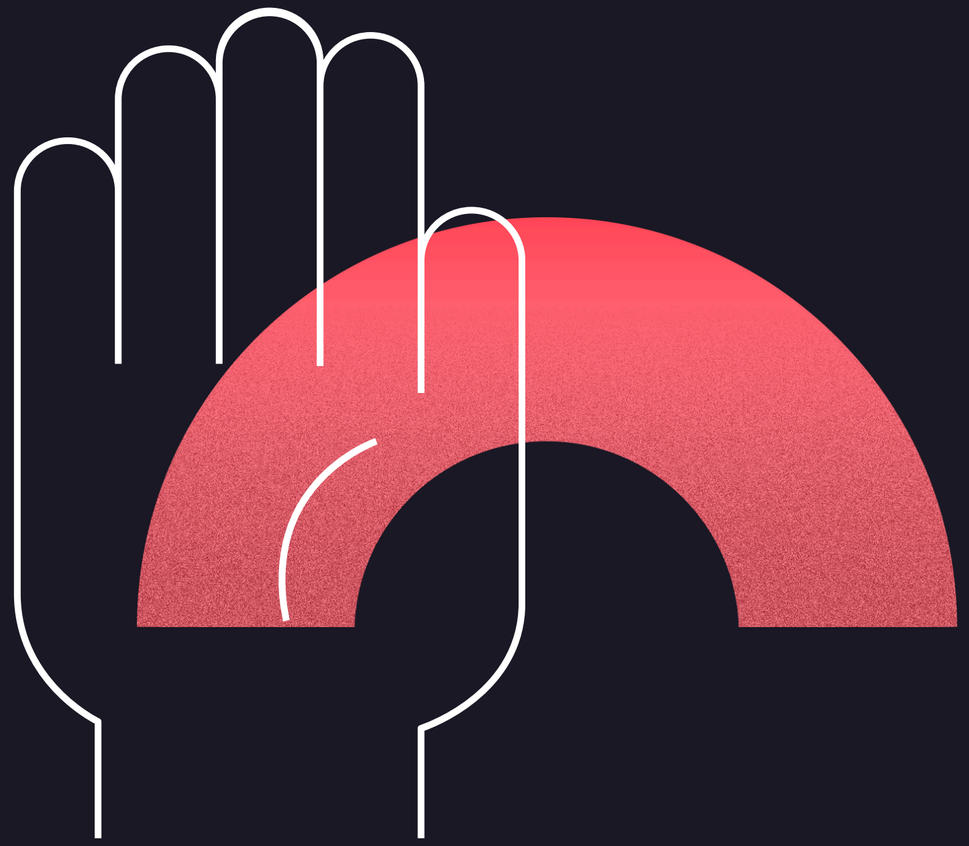
Feature Engineering for Time Series
Lag Values and **DateTime Variables**

Assessment Metric for Time Series
Mean Absolute Error

Emphasis on
Thorough Data Preprocessing

THANK YOU!

https://tinyurl.com/ML1FPGMZM



APPENDICES

THE DATASET


```
# First half of Dataset's Variables
df_energy.head().iloc[:,:14]
```

	time	generation biomass	generation fossil brown coal/lignite	generation fossil coal-derived gas	generation fossil gas	generation fossil hard coal	generation fossil oil	generation fossil oil shale	generation fossil peat	generation geothermal	generation hydro pumped storage aggregated	generation hydro pumped storage consumption	generation hydro run-of-river and poundage	generation hydro water reservoir
0	2015-01-01 00:00:00+01:00	447.0	329.0	0.0	4844.0	4821.0	162.0	0.0	0.0	0.0	NaN	863.0	1051.0	1899.0
1	2015-01-01 01:00:00+01:00	449.0	328.0	0.0	5196.0	4755.0	158.0	0.0	0.0	0.0	NaN	920.0	1009.0	1658.0
2	2015-01-01 02:00:00+01:00	448.0	323.0	0.0	4857.0	4581.0	157.0	0.0	0.0	0.0	NaN	1164.0	973.0	1371.0
3	2015-01-01 03:00:00+01:00	438.0	254.0	0.0	4314.0	4131.0	160.0	0.0	0.0	0.0	NaN	1503.0	949.0	779.0
4	2015-01-01 04:00:00+01:00	428.0	187.0	0.0	4130.0	3840.0	156.0	0.0	0.0	0.0	NaN	1826.0	953.0	720.0

```
# Second half of Dataset's Variables
df_energy.head().iloc[:,14:]
```

	generation marine	generation nuclear	generation other	generation other renewable	generation solar	generation waste	generation wind offshore	generation wind onshore	forecast solar day ahead	forecast wind offshore eday ahead	forecast wind onshore day ahead	total load forecast	total load actual	price day ahead	price actual
0	0.0	7096.0	43.0	73.0	49.0	196.0	0.0	6378.0	17.0	NaN	6436.0	26118.0	25385.0	50.10	65.41
1	0.0	7096.0	43.0	71.0	50.0	195.0	0.0	5890.0	16.0	NaN	5856.0	24934.0	24382.0	48.10	64.92
2	0.0	7099.0	43.0	73.0	50.0	196.0	0.0	5461.0	8.0	NaN	5454.0	23515.0	22734.0	47.33	64.48
3	0.0	7098.0	43.0	75.0	50.0	191.0	0.0	5238.0	2.0	NaN	5151.0	22642.0	21286.0	42.27	59.32
4	0.0	7097.0	43.0	74.0	42.0	189.0	0.0	4935.0	9.0	NaN	4861.0	21785.0	20264.0	38.41	56.04

R^2

No Lag Values

	Machine Learning Regression Method	Train R2	Validation R2	Test R2	Most Important Feature
4	Decision Tree	1.000000	-0.157278	-0.012063	generation fossil hard coal
3	kNN	0.824613	-0.298729	-0.000091	NA
0	Linear Regression	0.441096	0.249850	0.317017	weekday
2	Linear Regression + Ridge	0.441096	0.249850	0.317018	weekday
1	Linear Regression + Lasso	0.438977	0.237974	0.340629	month
5	Random Forest	0.993216	0.276521	0.383007	generation fossil hard coal
6	Gradient Boosting Method	0.867054	0.410452	0.470242	generation fossil hard coal
7	XGBoost	0.978744	0.345710	0.477416	year

With Lag Values

	Machine Learning Regression Method	Train R2	Validation R2	Test R2	Most Important Feature
3	kNN	0.825467	-0.296286	0.021042	NA
4	Decision Tree	1.000000	0.873916	0.883413	lag_1
1	Linear Regression + Lasso	0.956212	0.924360	0.934434	lag_1
0	Linear Regression	0.958229	0.927994	0.935772	lag_1
2	Linear Regression + Ridge	0.958229	0.927994	0.935772	lag_1
7	XGBoost	0.990149	0.941326	0.941319	lag_1
5	Random Forest	0.995869	0.940943	0.946885	lag_1
6	Gradient Boosting Method	0.971434	0.940320	0.952597	lag_1