SC1015
Mini Project:

B125 - Group 7

Presented by:
Choo Wei Boon
Wong Kang Yi, Rhyan
Truong Thi Hai Yen

Table of Content

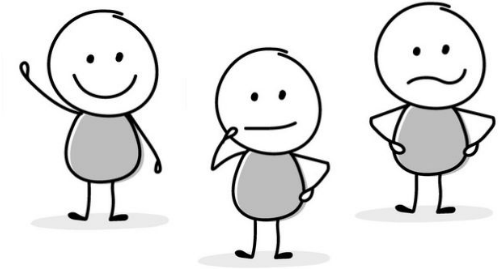
1

Dataset and Motivation

Dataset

Time Series

UCI Machine Learning Repository



Dataset:

Tetuan City
power consumption

Motivation

Background of Tétouan City:

- Reliance on imported energy
- Increasing population size

Importance of energy consumption forecast:

- Reduce production costs
- Avoid power shortages
- Ensure energy demands are met

Motivation

Problem Definitions:

1. Which of the models implemented will be better at predicting the energy consumption?

   XGBoost, LSTM and Random Forest

2. Are we able to predict the energy consumption based on a given week?

   Accuracy metrics: RMSE and MAPE

2

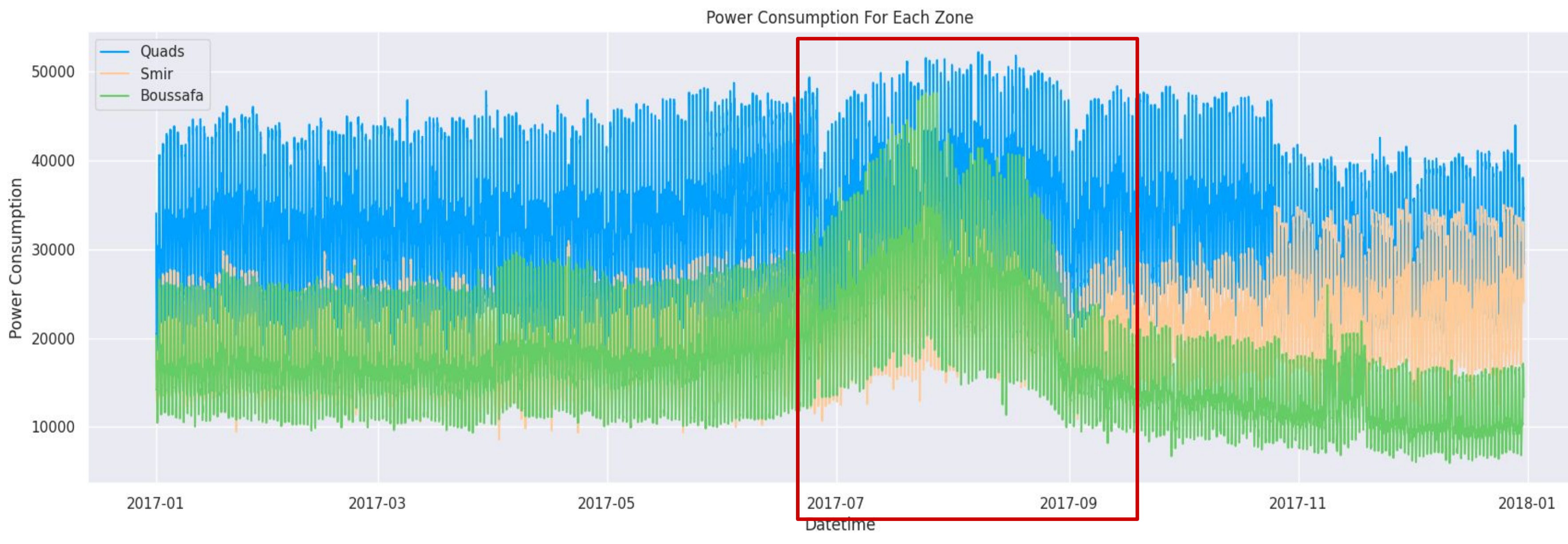
Exploratory Data Analysis

Visualisation of dataset

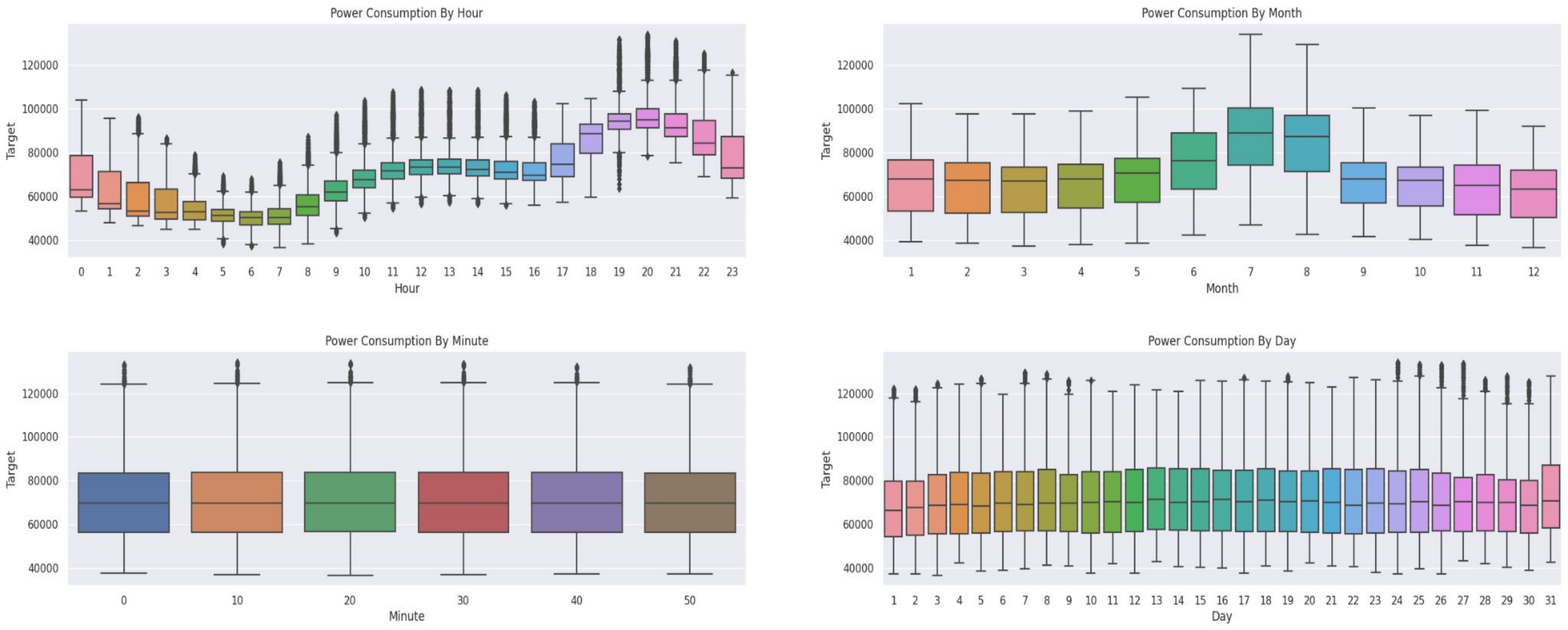
	DateTime	Temperature	Humidity	Wind Speed	general diffuse flows	diffuse flows	Zone 1 Power Consumption	Zone 2 Power Consumption	Zone 3 Power Consumption
0	1/1/2017 0:00	6.559	73.8	0.083	0.051	0.119	34055.69620	16128.87538	20240.96386
1	1/1/2017 0:10	6.414	74.5	0.083	0.070	0.085	29814.68354	19375.07599	20131.08434
2	1/1/2017 0:20	6.313	74.5	0.080	0.062	0.100	29128.10127	19006.68693	19668.43373
3	1/1/2017 0:30	6.121	75.0	0.083	0.091	0.096	28228.86076	18361.09422	18899.27711
4	1/1/2017 0:40	5.921	75.7	0.081	0.048	0.085	27335.69620	17872.34043	18442.40964
...	...	...	...	...	...	...	...	...	...
52411	12/30/2017 23:10	7.010	72.4	0.080	0.040	0.096	31160.45627	26857.31820	14780.31212
52412	12/30/2017 23:20	6.947	72.6	0.082	0.051	0.093	30430.41825	26124.57809	14428.81152
52413	12/30/2017 23:30	6.900	72.8	0.086	0.084	0.074	29590.87452	25277.69254	13806.48259
52414	12/30/2017 23:40	6.758	73.0	0.080	0.066	0.089	28958.17490	24692.23688	13512.60504
52415	12/30/2017 23:50	6.580	74.1	0.081	0.062	0.111	28349.80989	24055.23167	13345.49820

52416 rows × 9 columns

Visualisation of dataset



Visualisation of dataset



Stationary Test

```
adf_test(data['Target'])
```

```
Results of Dickey-Fuller Test:
Test Statistic                     -26.567630
p-value                              0.000000
Lags Used                           58.000000
Number of Observations Used      52357.000000
Critical Value (1%)                 -3.430475
Critical Value (5%)                 -2.861595
Critical Value (10%)                -2.566799
```

p-value < 0.05

Data is <u>stationary</u>

Data Cleaning & Preparation (XGBoost & Random Forest)

1. Merge zones into one feature

2. Remove other features

3. Split 'DateTime' into features

DateTime	Total Power Consumption	date	hour	minute	dayofweek	quarter	month	year	dayofyear	dayofmonth	weekofyear
2017-11-01 00:10:00	61314.21096	2017-11-01 00:10:00	0	10	2	4	11	2017	305	1	44
2017-11-01 00:20:00	60426.11348	2017-11-01 00:20:00	0	20	2	4	11	2017	305	1	44
2017-11-01 00:30:00	59244.04276	2017-11-01 00:30:00	0	30	2	4	11	2017	305	1	44
2017-11-01 00:40:00	58129.05375	2017-11-01 00:40:00	0	40	2	4	11	2017	305	1	44
2017-11-01 00:50:00	56399.66413	2017-11-01 00:50:00	0	50	2	4	11	2017	305	1	44
...	...	...	...	...	...	...	...	...	...	...	...
2017-10-31 23:20:00	68505.52458	2017-10-31 23:20:00	23	20	1	4	10	2017	304	31	44
2017-10-31 23:30:00	66873.99145	2017-10-31 23:30:00	23	30	1	4	10	2017	304	31	44
2017-10-31 23:40:00	65574.82177	2017-10-31 23:40:00	23	40	1	4	10	2017	304	31	44
2017-10-31 23:50:00	64571.18017	2017-10-31 23:50:00	23	50	1	4	10	2017	304	31	44
2017-11-01 00:00:00	62004.76199	2017-11-01 00:00:00	0	0	2	4	11	2017	305	1	44

52416 rows × 12 columns

Data Cleaning & Preparation (LSTM)

1. Merge zones into one feature
2. Remove all other features
3. Check for missing data

DateTime	Target
2017-01-01 00:00:00	70425.53544
2017-01-01 00:10:00	69320.84387
2017-01-01 00:20:00	67803.22193
2017-01-01 00:30:00	65489.23209
2017-01-01 00:40:00	63650.44627
...	...
2017-12-30 23:10:00	72798.08659
2017-12-30 23:20:00	70983.80786

3

Model Comparison

Splitting train, test

1. Picked last 15 days
2. Split into train and test
3. Split ratio (70:30)

DateTime	Total Power Consumption
2017-12-16 00:00:00	61346.076660
2017-12-16 00:10:00	59841.731530
2017-12-16 00:20:00	58469.286360
2017-12-16 00:30:00	57239.781040
2017-12-16 00:40:00	56161.606900
...	...
2017-12-26 11:10:00	66621.400991
2017-12-26 11:20:00	67092.556926
2017-12-26 11:30:00	67690.512441
2017-12-26 11:40:00	68467.392432
2017-12-26 11:50:00	68492.554444

1512 rows × 1 columns

data_train

DateTime	Total Power Consumption
2017-12-26 12:00:00	68619.355085
2017-12-26 12:10:00	69654.302581
2017-12-26 12:20:00	70247.423480
2017-12-26 12:30:00	70852.390870
2017-12-26 12:40:00	70748.015480
...	...
2017-12-30 23:10:00	72798.086590
2017-12-30 23:20:00	70983.807860
2017-12-30 23:30:00	68675.049650
2017-12-30 23:40:00	67163.016820
2017-12-30 23:50:00	65750.539760

648 rows × 1 columns

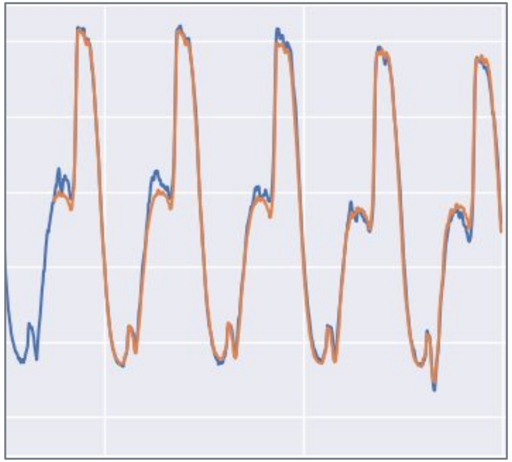
data_test

Prediction results for 15 days

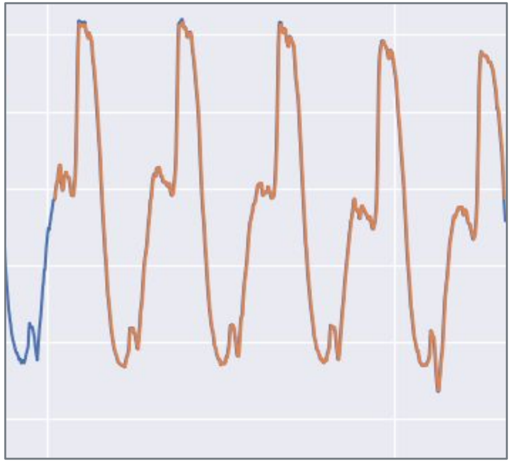
Legend:

15 days dataset

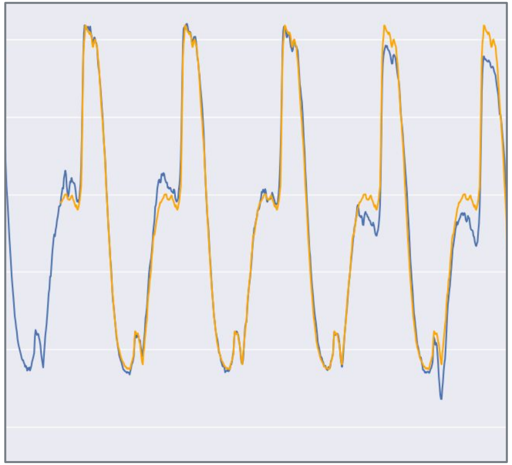
Prediction



XGBoost model



LSTM model

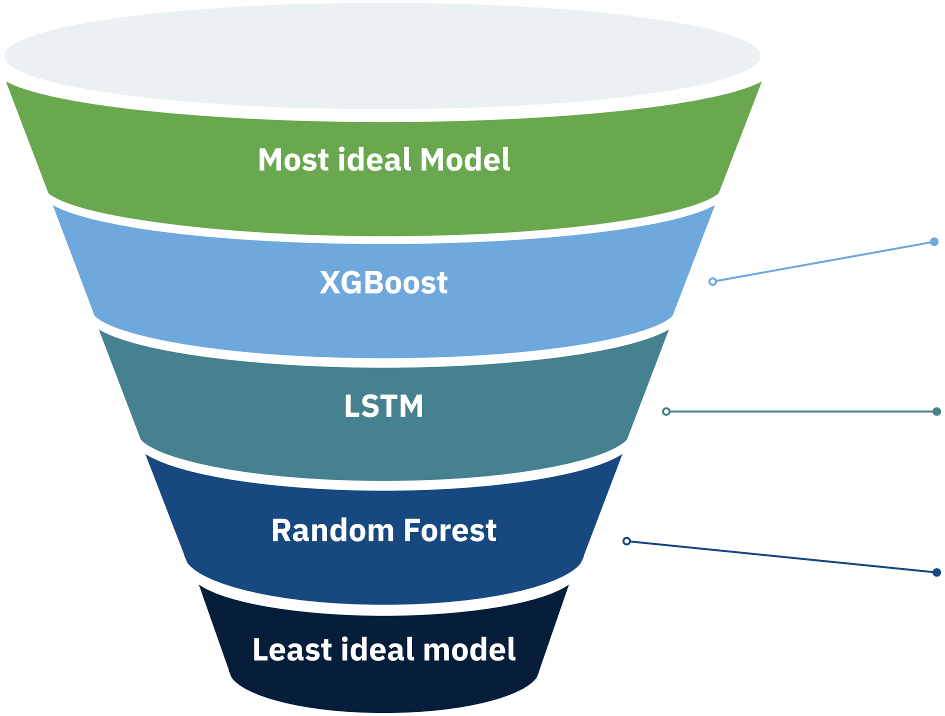


Random Forest model

Accuracy Metrics (15 days)

	Accuracy Metrics / Model	RMSE	MAPE
2	XGBoost	1163.66	1.3604
1	LSTM	1157.74	1.1772
3	Random Forest	1827.26	2.0137

Rationale For Model Chosen



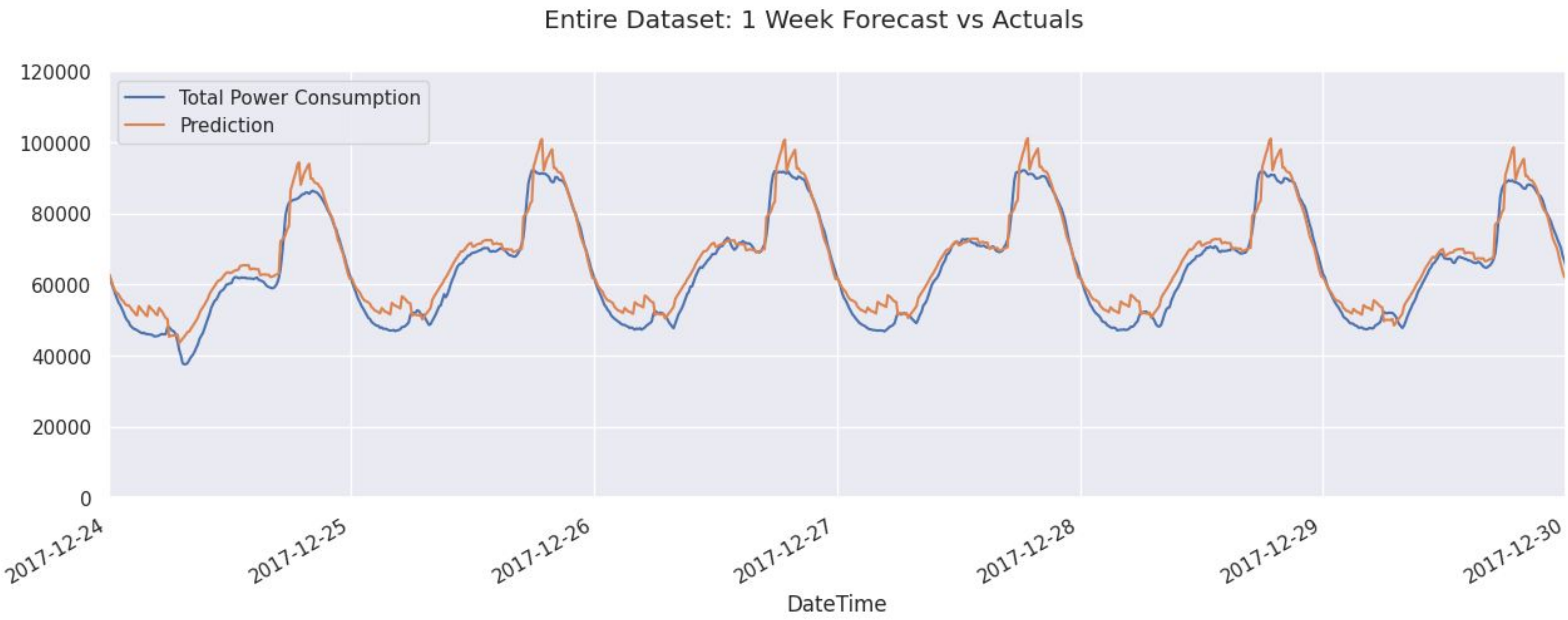
Good accuracy, more straightforward

Best accuracy, but more complex

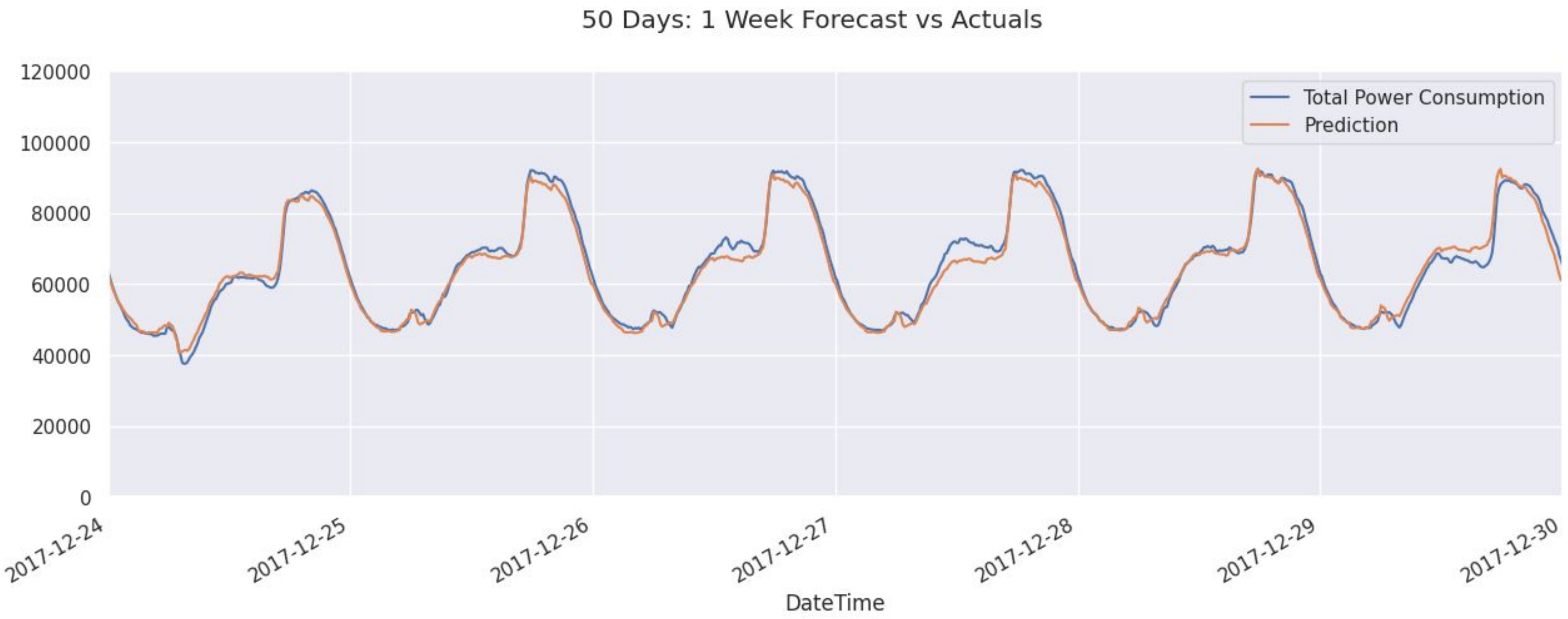
Lowest accuracy metrics in comparison

4

Core Analysis



Prediction results from chosen model (XGBoost) - entire dataset



Prediction results from chosen model (XGBoost) - 50 days dataset

Accuracy Metrics (XGBoost)

Accuracy Metrics / Model	RMSE	MAPE
Entire dataset	4491.2091	6.1412
50 days	**1831.7407**	**2.2841**
15 days	1163.66	1.3604

Time series forecasting

Predicting with sequential data vs data features

Random Forest Model

Not suitable for time series problem

LSTM Model

Requires normalising the data, vanilla and stacked variation

XGBoost Model

Flexible model, can be implemented in either ways

5

Conclusion

Outcome of Mini Project

1. Which of the models implemented will be better at predicting the energy consumption?

   LSTM is more accurate, but we still chose XGBoost.

2. Are we able to predict the energy consumption based on a given week?

   Yes, using able to predict with data from 50 days ahead.

Data-Driven Insights

year	month	dayofmonth	Total Power Consumption	Prediction	error	abs_error
2017	12	1	58574.151543	64762.027344	-6187.872871	6384.171803
	11	6	62288.536668	68277.875000	-5989.338006	6074.192907
	12	2	60104.345342	65827.328125	-5722.979474	5986.959278
		3	59192.378189	64537.035156	-5344.654661	5600.052281
		13	64306.587793	69210.390625	-4903.801123	5150.495580
		16	62336.671485	67069.953125	-4733.282508	4975.200423
		12	64338.518880	68904.492188	-4565.974799	4829.543965
		14	65028.106381	69369.265625	-4341.156260	4706.692615
		7	64821.612249	69106.515625	-4284.898276	4651.329932
	11	7	64588.154859	68880.867188	-4292.710023	4472.850072

Recommendations

Holiday Spike:

1. Prophet’s Birthday (1 & 2 Dec)
2. Green March Day (6 Nov)

Holidays in next 50 days:

-> anticipate potential spike

-> increase energy production (according to prediction)

The End

Thank you for your kind attention