Eksploracja danych

Zastosowanie algorytmów Gradient Boosted Decision Trees do prognozowania szeregów czasowych

Sprawozdanie z projektu - 18.01.2019

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Bartłomiej Bukowski
Ryszard Sikora

Na ile dni do przodu mogę przewidzieć kurs euro

- Jak dużo danych historycznych potrzebuję, żeby przewidzieć przyszłe wartości?
- Na ile dni do przodu mogę przewidywać przyszłe wartości?
- Z jaką dokładnością jestem w stanie przewidzieć kurs za X dni mając Y dni danych historycznych?
- Jak częstotliwość danych wpływa na te przewidywania?

Image taken from: https://www.research.ibm.com/ibm-q/

Gradient Boosted Decision Trees

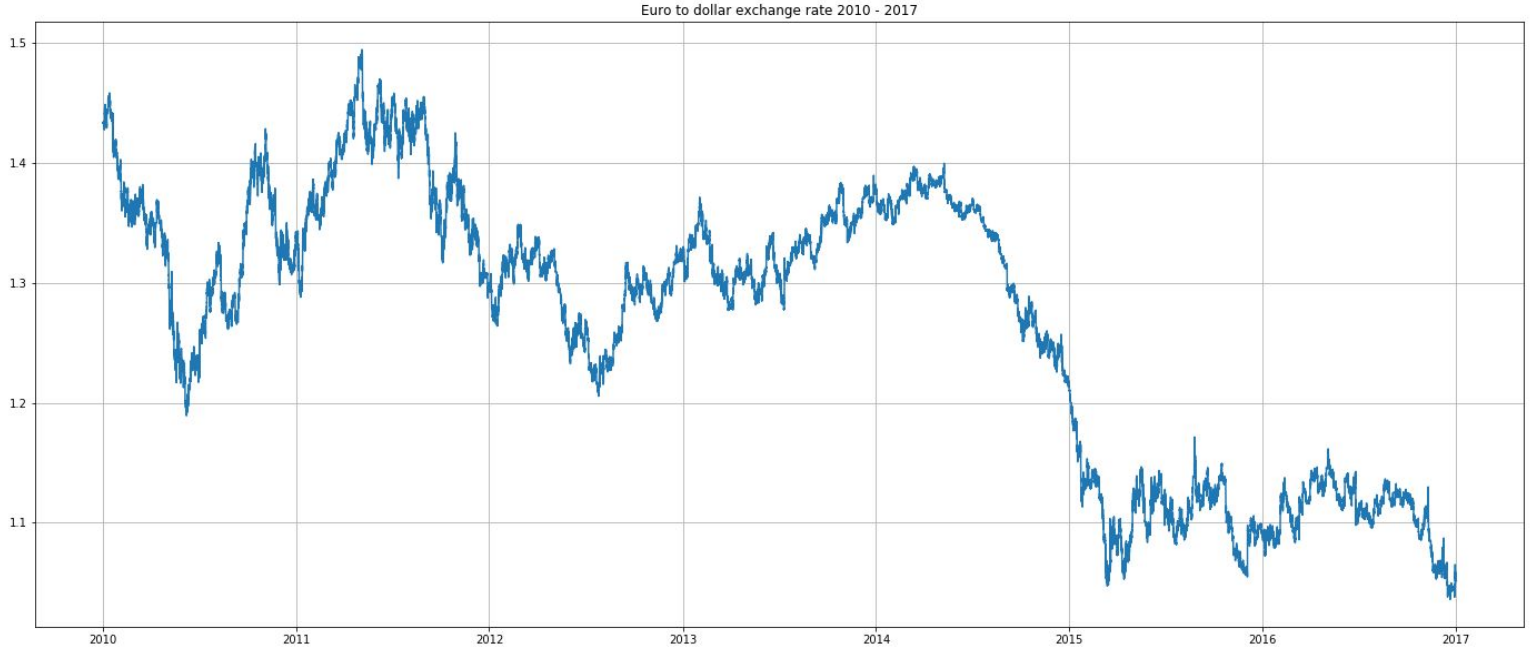
Gradient Boosting to technika uczenia zespołowego, w której klasyfikatory uczone są sekwencyjnie - kolejny klasyfikator uczy się na błędach poprzednich klasyfikatorów

- Zmniejsza stronniczość (bias) i wariancję
- Może przeuczać (overfitting)
- Najczęściej używany z drzewami decyzyjnymi

XGBoost - eXtreme Gradient Boosting

AGH

Kurs EURO do USD 2010-2017

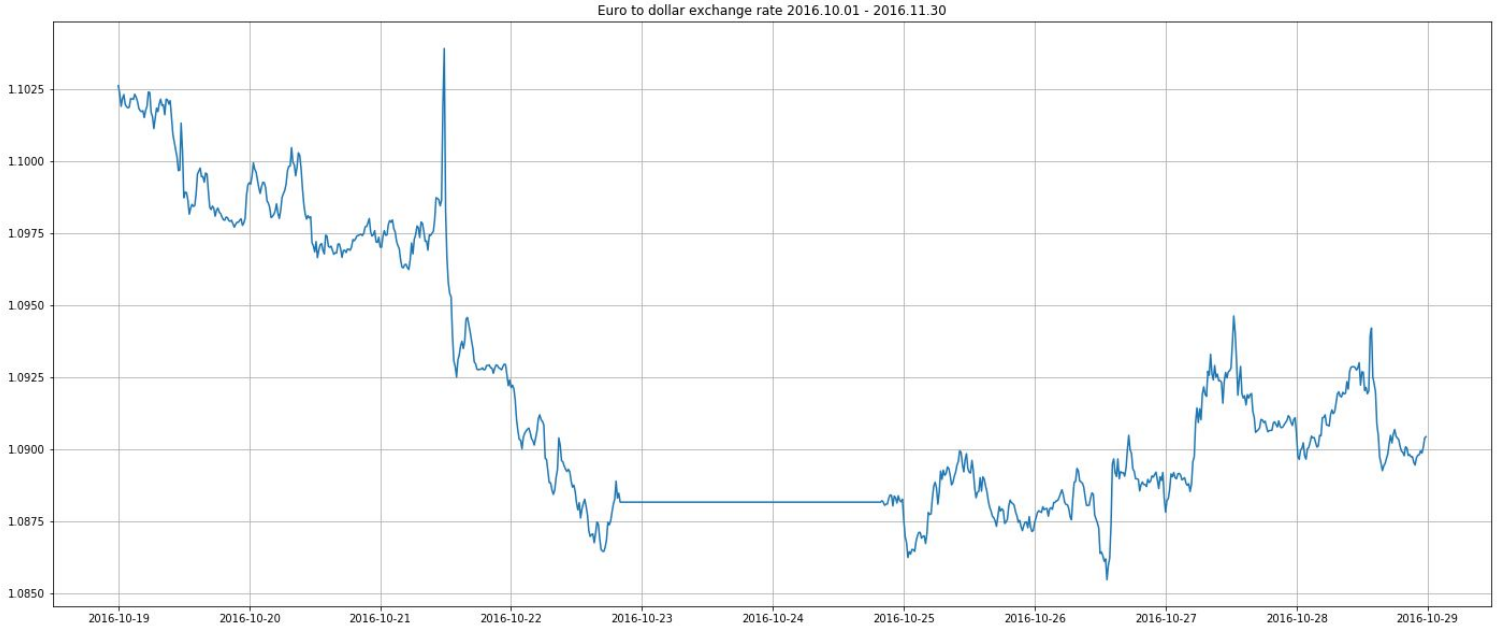


XGBoost - wykorzystanie

- Algorytm XGBoost parametryzować można dwoma zmiennymi:
  - długością wektorów uczących (w seriach czasowych rozumianych jako ilość próbek z puntów czasowych wstecz)
  - ilością wektorów uczących (dla każdego punktu czasowego w serii, konstruujemy wektor złożony z n-1 wartości próbek czasowych wstecz + wartości danego punktu czasowego)
- Z zaprezentowanej serii czasowej wybrano podserię z zakresu dat: 2016.10.01 - 2016.11.30
  - granularność próbek to 15 minut (96 próbek na dzień)
- Długość wektorów uczących była zmienna
- Przewidywano różne okresy czasowe wprzód

AGH

Kurs EURO do USD 2016.10.01 - 2016.11.30 (wybrana podseria czasowa)



Opis wykresów

Każdy wykres jest opatrzony dwoma parametrami:

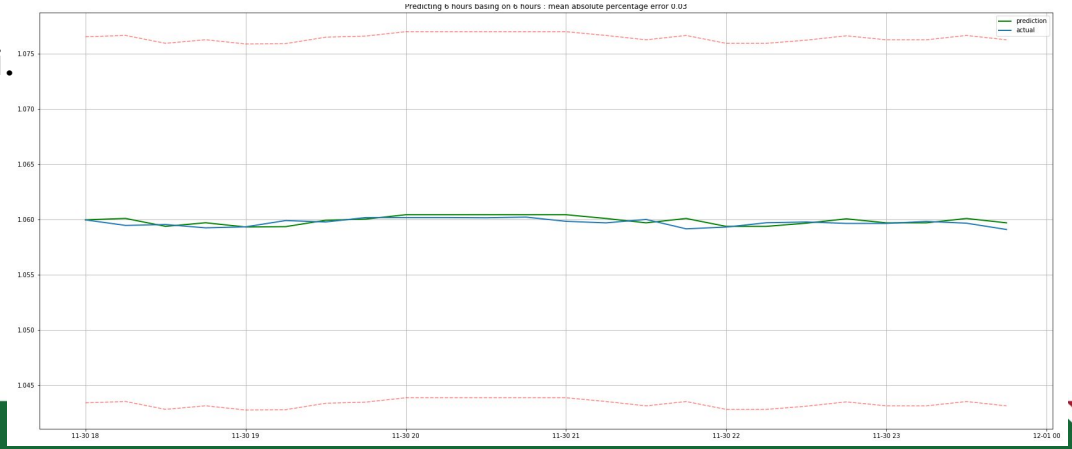
długość wektorów uczących → przewidywany okres

Przewidywany okres jest wizualizowany na wykresie w postaci:

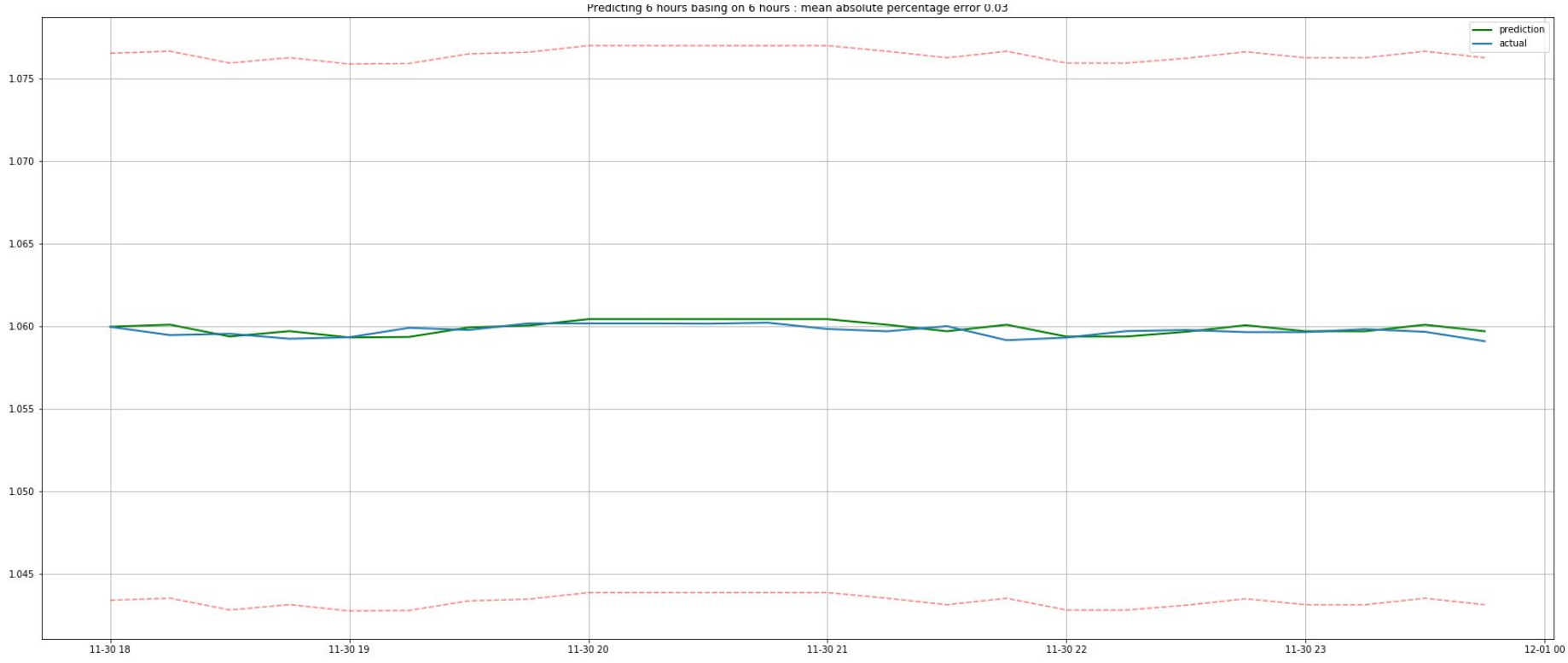
- faktycznych wartości dla danego czasu
- przewidywanych wartości
- dodatkowo górnej i dolnej granicy błędu

Każdy wykres wizualizuje uczenie na podstawie wcześniej wskazanej subserii.

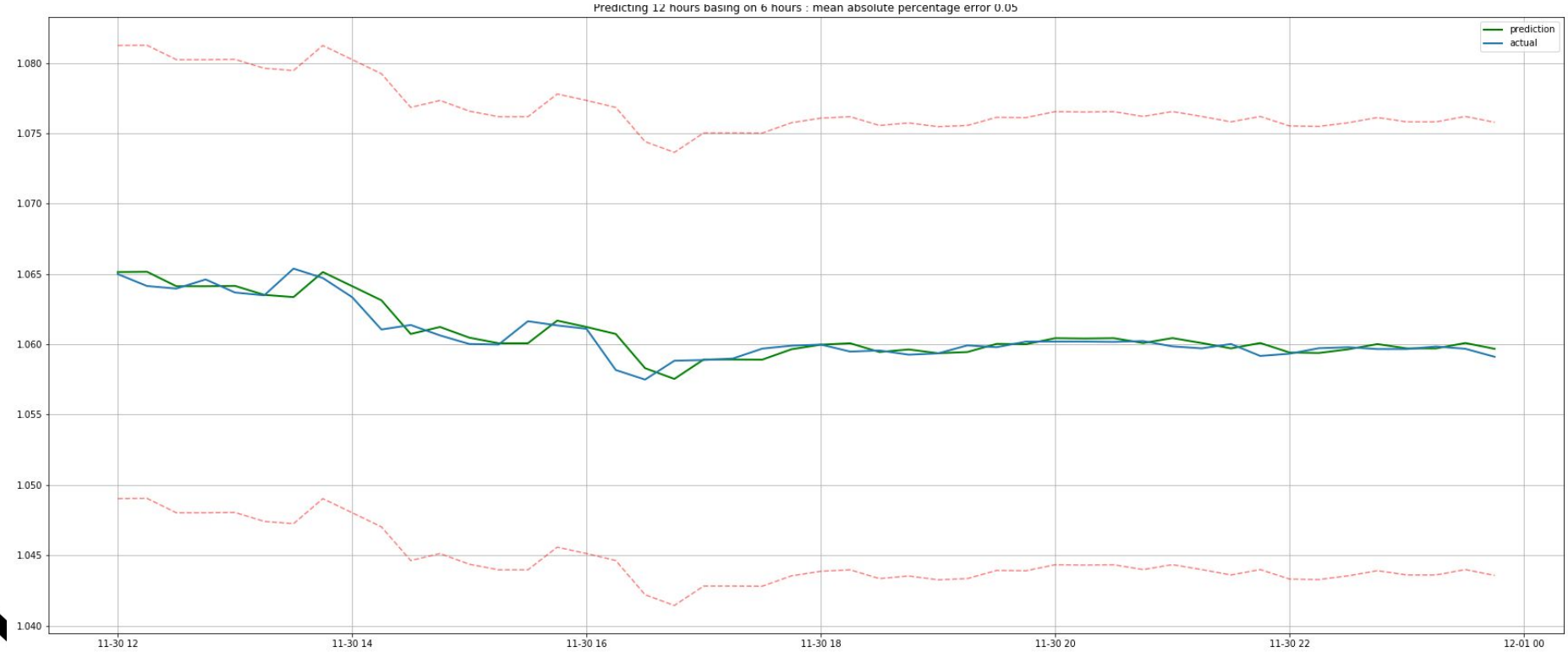
6 godzin → 6 godzin



6 godzin → 6 godzin



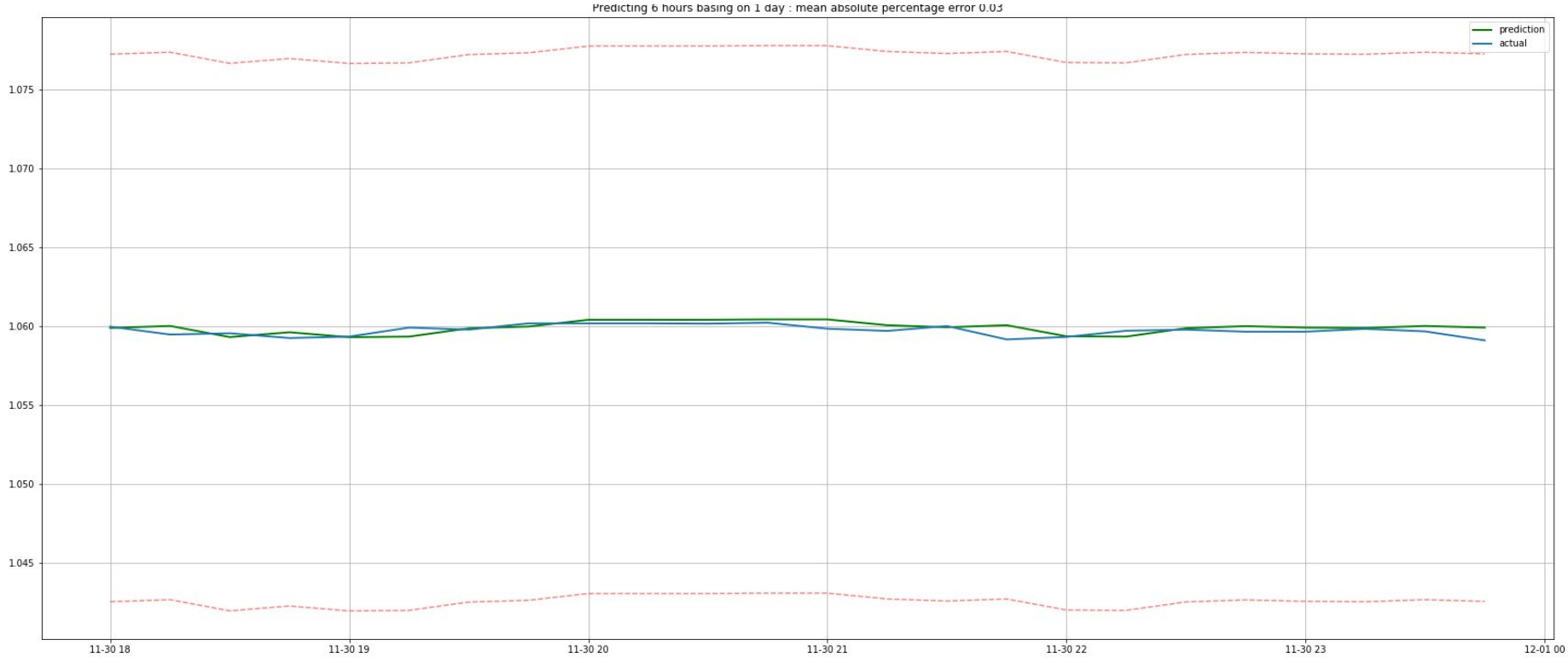
6 godzin → 12 godzin



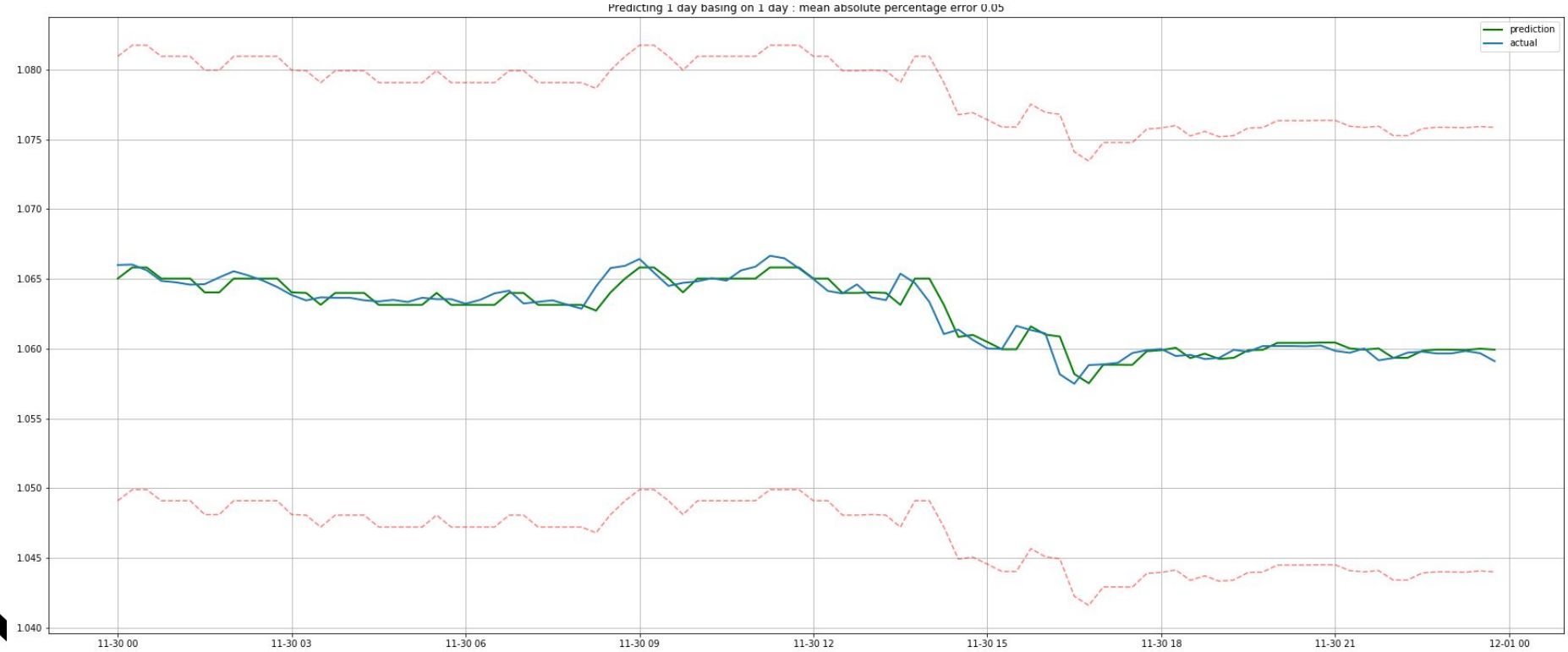
6 godzin → 1 dzień

6 godzin → 2 dni

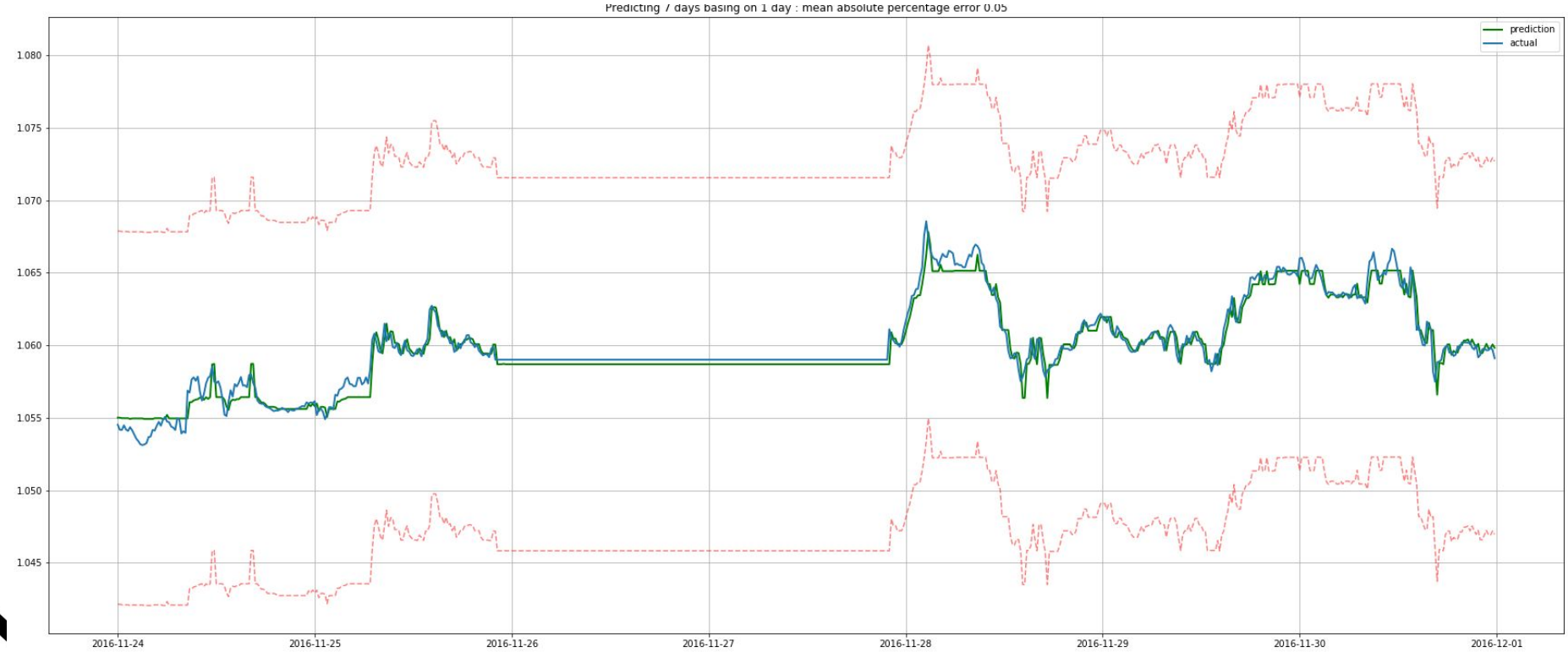
1 dzień → 6 godzin



Przewidywanie 1 dzień → 1 dzień

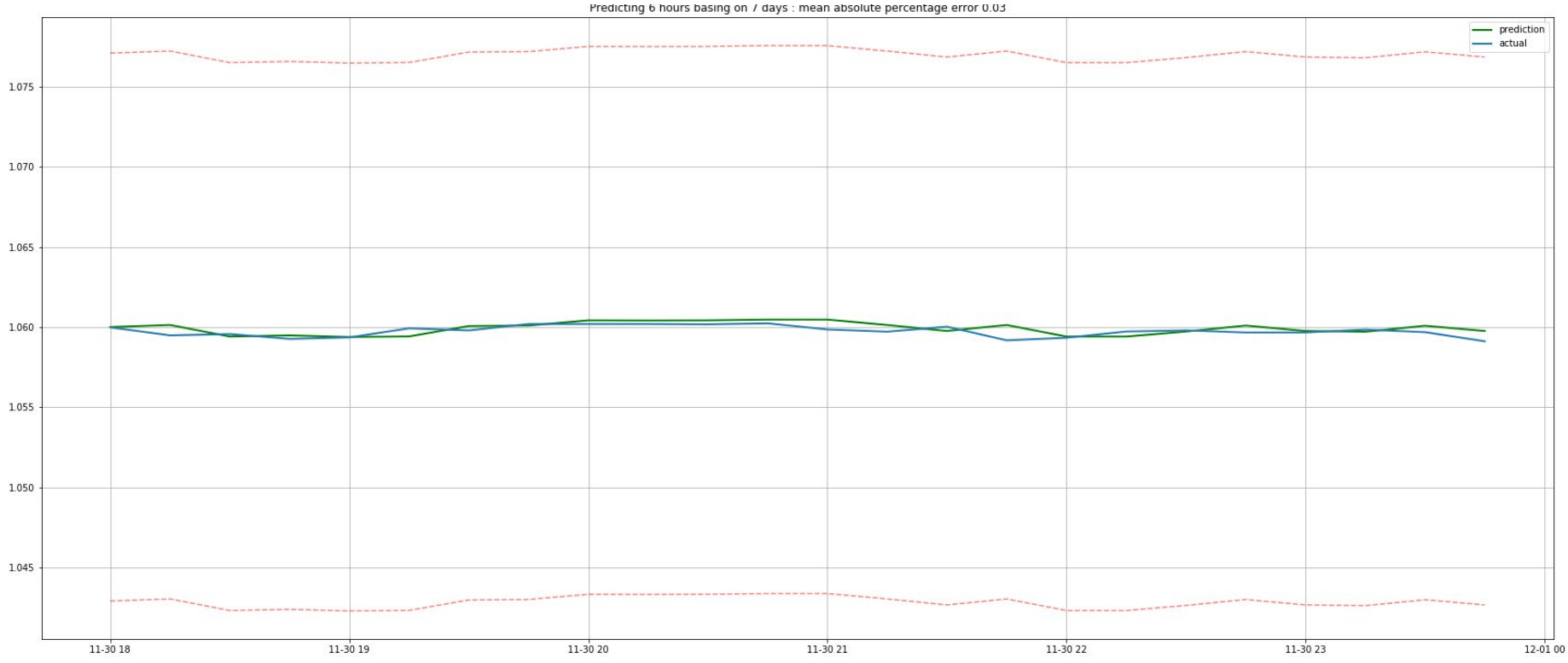


1 dzień → 7 dni

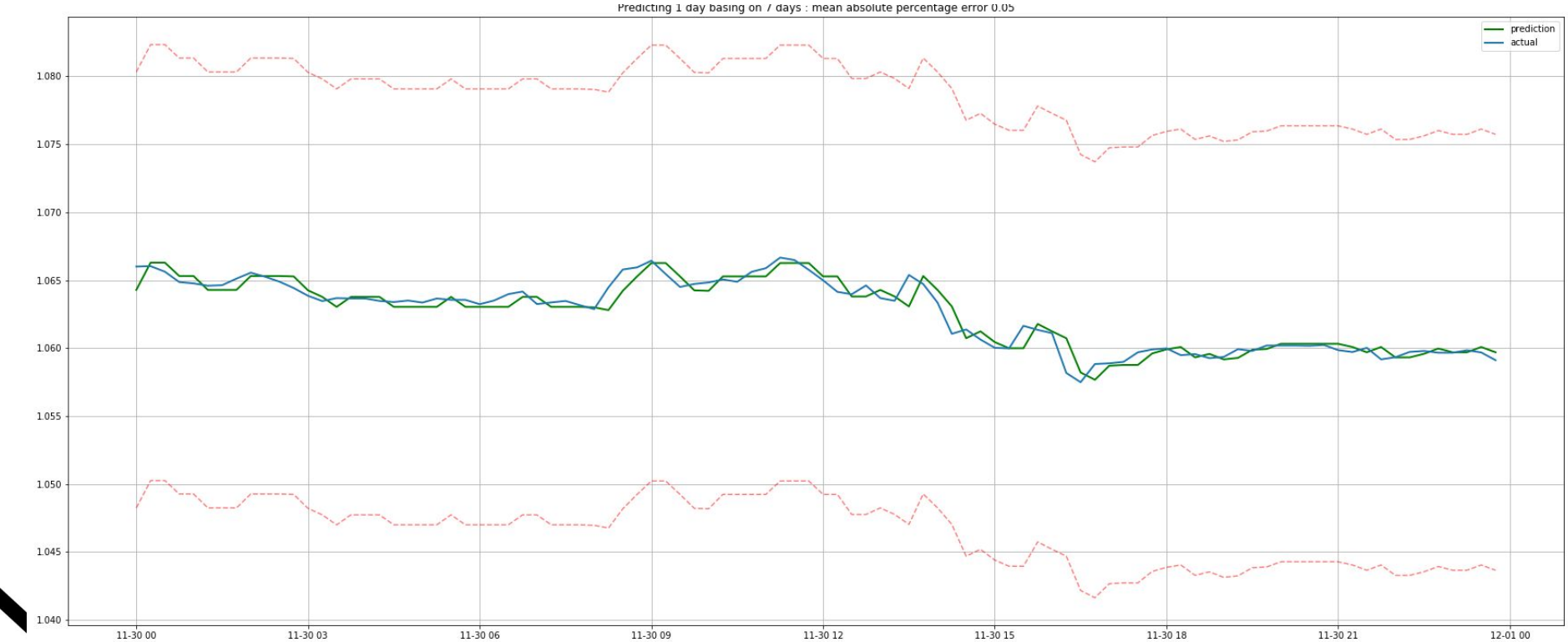

Predicting 7 days basing on 1 day : mean absolute percentage error 0.05
prediction
actual
1.080
1.075
1.070
1.065
1.060
1.055
1.050
1.045
2016-11-24
2016-11-25
2016-11-26
2016-11-27
2016-11-28
2016-11-29
2016-11-30
2016-12-01

1 dzień → 12 dni

7 dni → 6 godzin



7 dni → 1 dzień


Predicting 1 day basing on 7 days : mean absolute percentage error 0.05
prediction
actual
1.080
1.075
1.070
1.065
1.060
1.055
1.050
1.045
1.040
11-30 00
11-30 03
11-30 06
11-30 09
11-30 12
11-30 15
11-30 18
11-30 21
12-01 00

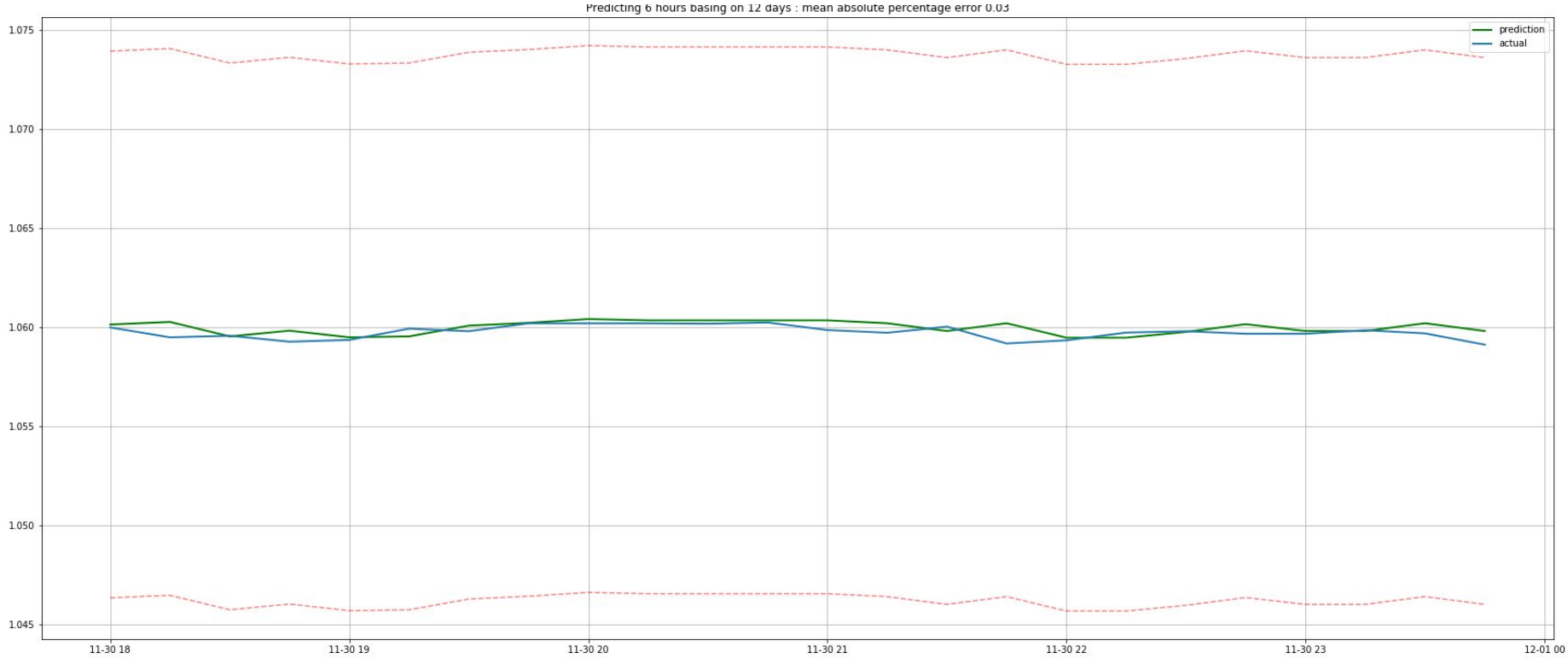
7 dni → 7 dni


Predicting 7 days basing on 7 days : mean absolute percentage error 0.15
prediction
actual
1.075
1.070
1.065
1.060
1.055
1.050
1.045
2016-11-24
2016-11-25
2016-11-26
2016-11-27
2016-11-28
2016-11-29
2016-11-30
2016-12-01

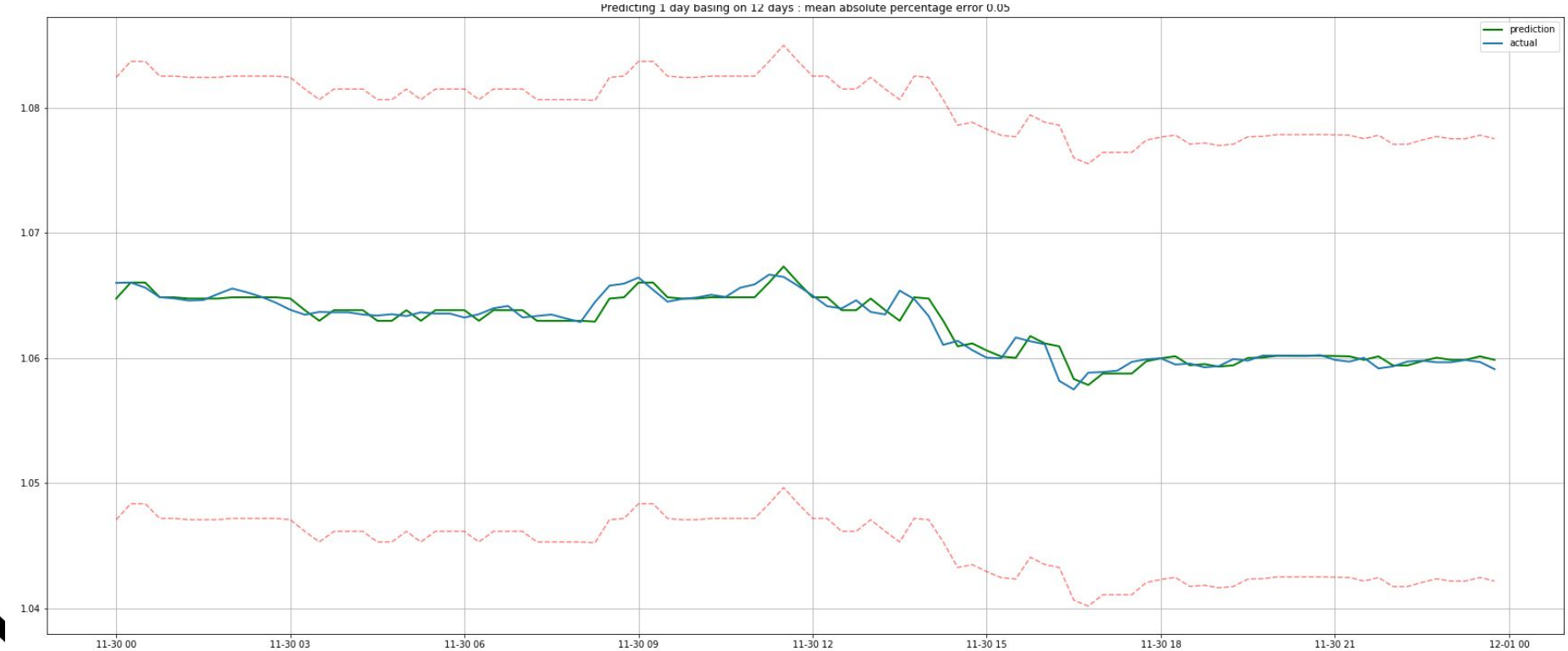
7 dni → 12 dni


Predicting 12 days basing on 7 days : mean absolute percentage error 0.13
prediction
actual
1.075
1.070
1.065
1.060
1.055
1.050
2016-11-19
2016-11-21
2016-11-23
2016-11-25
2016-11-27
2016-11-29
2016-12-01

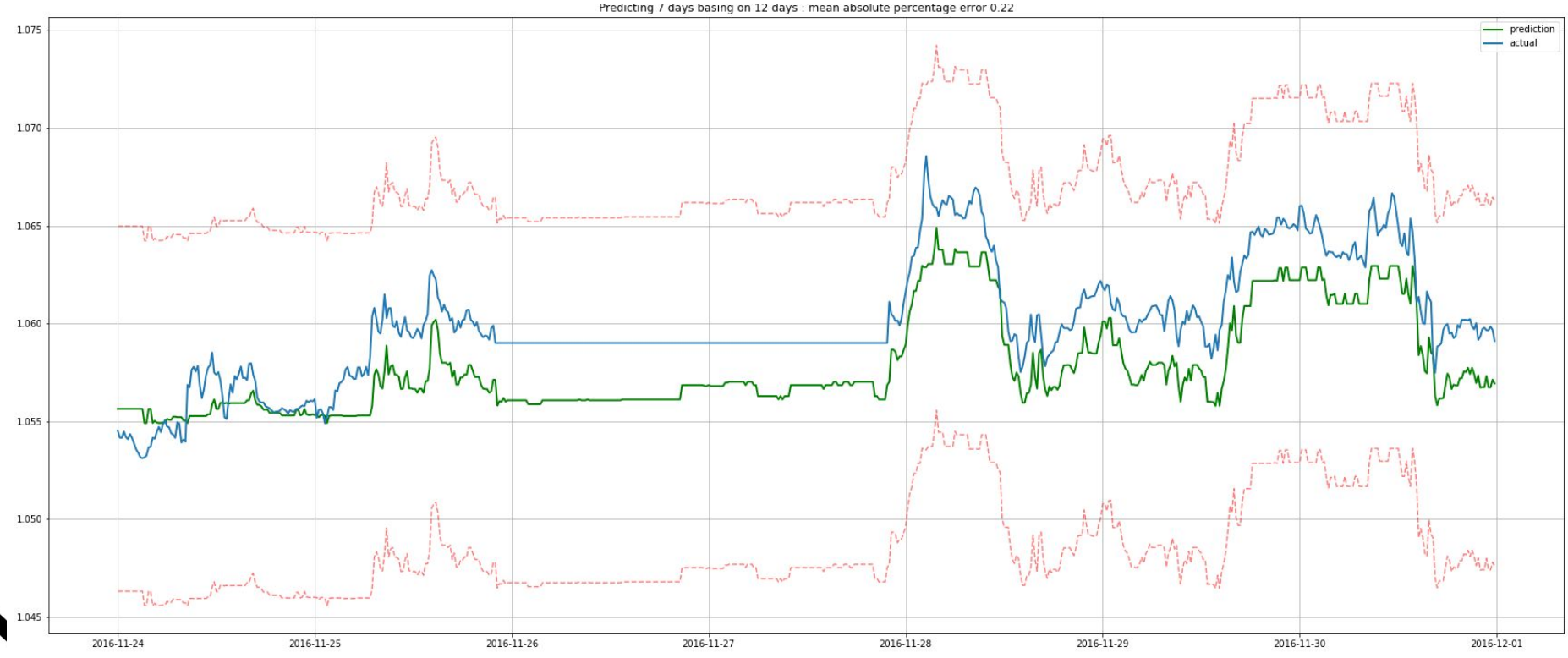
12 dni → 6 godzin



12 dni → 1 dzień



12 dni → 7 dni


Predicting 7 days basing on 12 days : mean absolute percentage error 0.22
prediction
actual
1.075
1.070
1.065
1.060
1.055
1.050
1.045
2016-11-24
2016-11-25
2016-11-26
2016-11-27
2016-11-28
2016-11-29
2016-11-30
2016-12-01

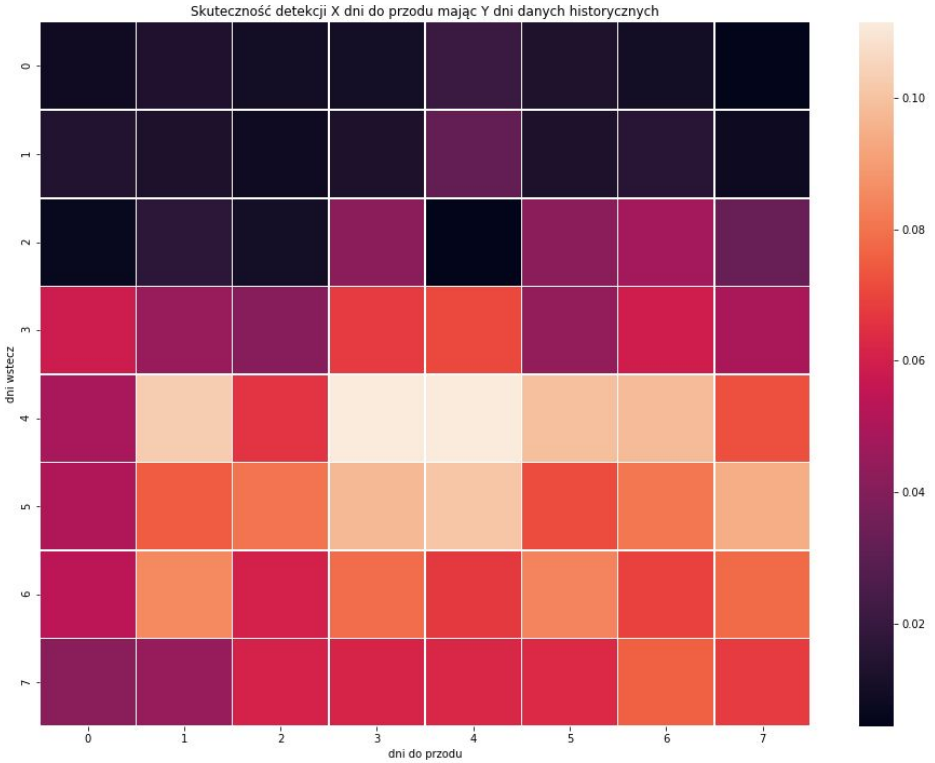
12 dni → 12 dni


Predicting 12 days basing on 12 days : mean absolute percentage error 0.15
prediction
actual
1.075
1.070
1.065
1.060
1.055
1.050
1.045
2016-11-19
2016-11-21
2016-11-23
2016-11-25
2016-11-27
2016-11-29
2016-12-01

heatmapa

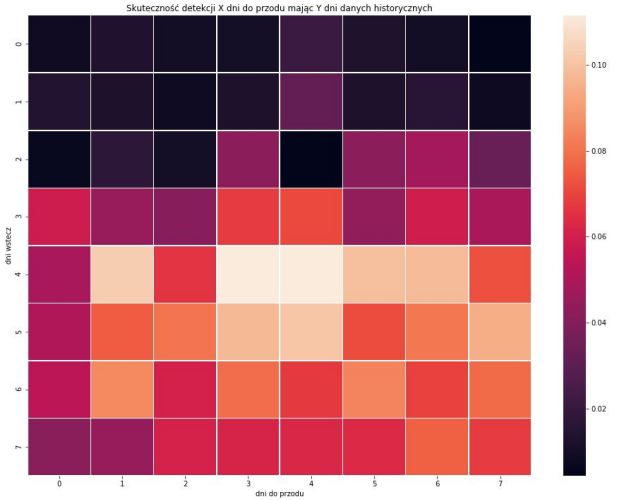
Mapowanie wartości na osiach:

0 → 6 godzin
1 → 12 godzin
2 → 1 dzień
3 → 2 dni
4 → 3 dni
5 → 4 dni
6 → 5 dni
7 → 6 dni

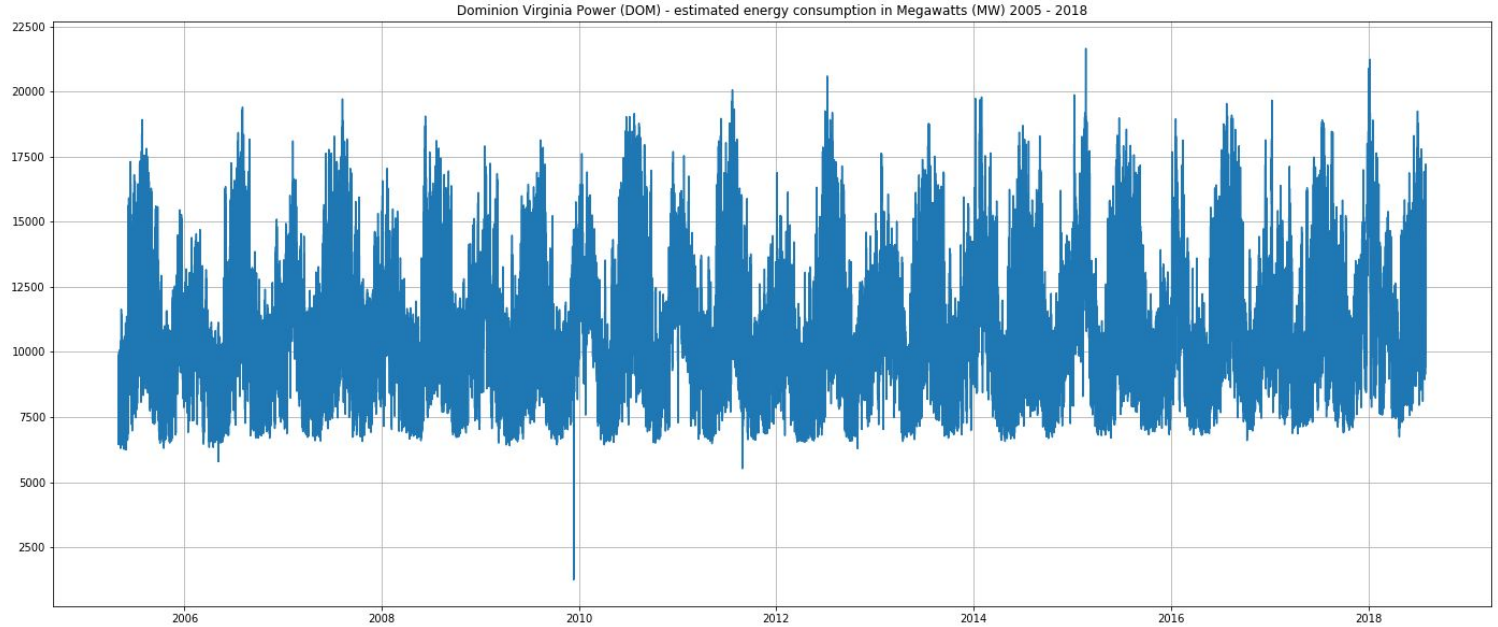


heatmapa - wytłumaczenie

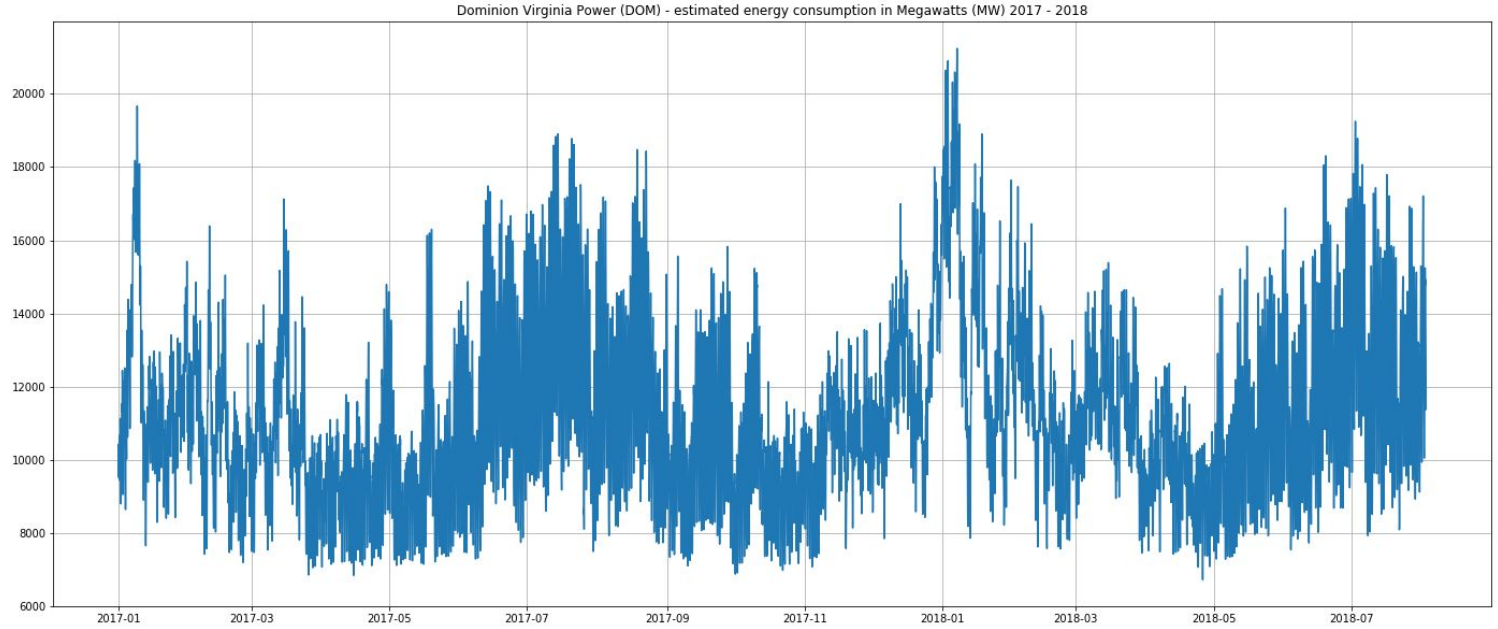
- Wartości heatmapy to błąd średniokwadratowy między przewidywanymi próbkami wprzód a faktycznymi wartościami
- Widzimy, że przewidywanie od 6h do 1 dnia wprzód daje najlepsze wyniki, lecz nie jest to zbyt zaskakujące
- Uczenie na podstawie długich wektorów uczących daje lepsze wyniki oraz przewidywanie małego okresu wprzód
- Świadczy to o braku wykrywalnego trendu, a bardziej lokalnych charakterystyk serii czasowej



Szacunkowe zapotrzebowanie na energię elektryczną DOM - 2005-2018



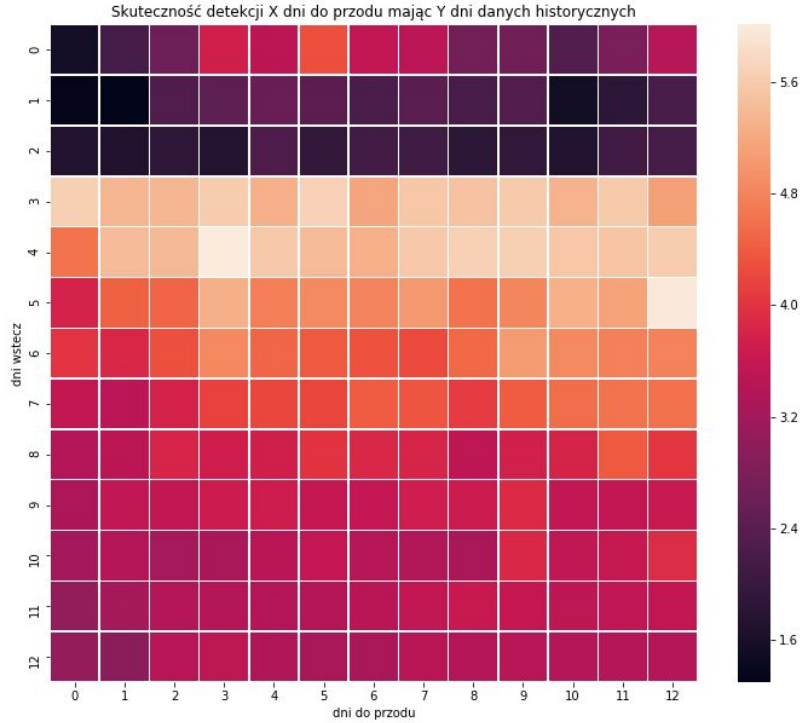
DOM - 2017-2018



heatmapa

Mapowanie wartości na osiach:
0 → 1 dzień
:
:
:
12 → 13 dni

Tutaj widzimy, że ta seria czasowa ma trend, który objawia się, gdy wektory uczące są dłuższe. Pomijając przewidywanie najbliższych 1-3 dni jest najlepsze (wyjąwszy przewidywanie jednego dnia na podstawie wektorów jednodniowych)



Bibliografia

1. https://nbviewer.jupyter.org/github/Yorko/mlcourse_open/blob/master/jupyter_english/topic09_time_series/topic9_part1_time_series_python.ipynb
2. https://www.kaggle.com/thebrownviking20/everything-you-can-do-with-a-time-series
3. https://xgboost.readthedocs.io/en/latest/
4. https://github.com/rychuhardy/eksploracja-danych-timeseries/tree/master/final

AGH

Zastosowanie algorytmów Gradient Boosted Decision Trees do prognozowania szeregów czasowych

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

Bartłomiej Bukowski
Ryszard Sikora