

Uczenie maszynowe dla szeregów czasowych

sages

kodo/amacz

Ros Apostol

Agenda

1. Dlaczego szeregi czasowe są ważne?
2. Czym się różni uczenie maszynowe dla szeregów czasowych od regresji?
3. Czy jesteśmy pewni, że nasze dane to szeregi czasowe?
4. Co robić, jak nasze dane to nie szeregi czasowe, a muszą być?
5. Feature Engineering - jak wzbogacić nasze dane?
6. Naiwna prognoza - czy nasz model uczenia maszynowego w ogóle ma sens?
7. Dlaczego walidacja krzyżowa nie działa?
8. Czy tradycyjne metody statystyczne już do niczego?
9. Kiedy lepiej się już nie da?

Po co nam szeregi czasowe?

- Prognozy gospodarcze
- Prognozy sprzedażowe
- Analizy rynku akcji / surowców / nieruchomości
- Prognozy plonów
- Prognozowanie zużycia gazu
- Monitorowanie działanie sprzętu technologicznego

Podstawa kluczowych i strategicznych decyzji

Po co nam ML dla szeregów czasowych?

Zalety

1. Wyższa dokładność prognoz
2. Nieliniowe zależności
3. Duża ilość danych, wiele wymiarów
4. Wydajność obliczeń

Wady

1. ML nie zawsze ma sens
2. Problem z tłumaczeniem modeli (tzw. modele - czarne skrzynki)

Szeregi czasowe vs. Regresja

Czas jest wszystkim.

Napoleon I

Czas się nie śpieszy - to my nie nadążamy.

Lew Tolstoj

**Z naprawdę wielkich, posiadamy tylko
jednego wroga - czas.**

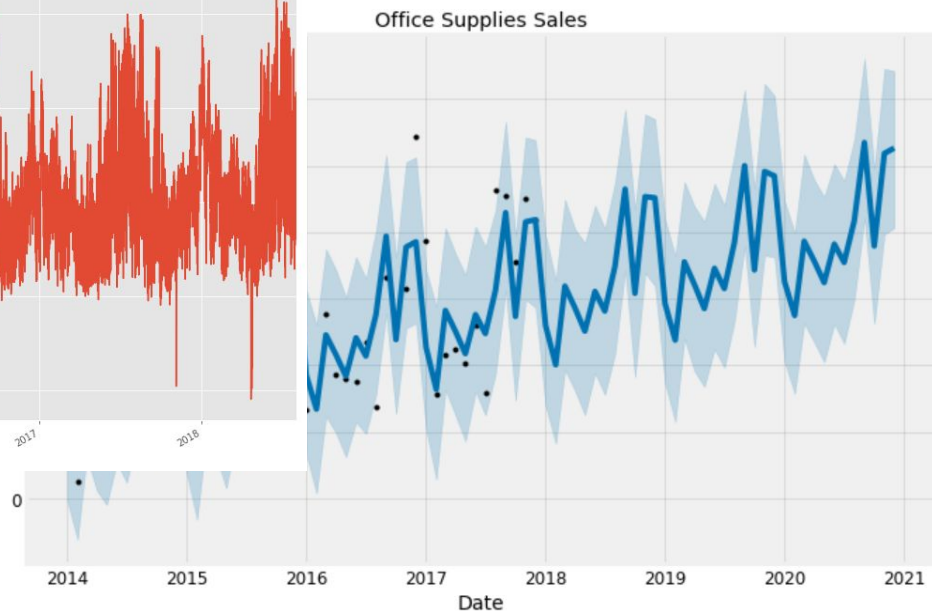
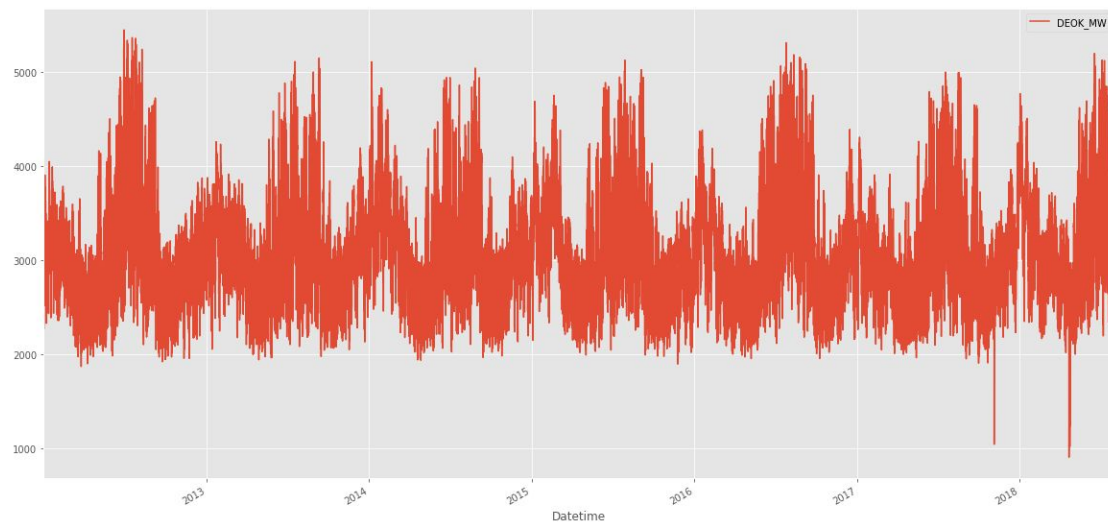
Joseph Conrad

**Znane są tysiące sposobów zabijania czasu,
ale nikt nie wie, jak go wskrzesić.**

Albert Einstein

Szeregi czasowe

Kolejność jest kluczowa



Czy nasze dane to szeregi czasowe?

1. Dane zawierają zmienną odzwierciedlającą czas lub kolejność - indeks czasowy.
2. Dane są posortowane.
3. Indeks czasowy ma stabilną częstotliwość.

Tabela 1

Data	Wartość
2020-01-01	100
2020-01-02	200
2020-01-05	120
2020-01-09	170

Tabela 2

Data	Wartość
2020-01-01	120
2020-01-10	140
2020-01-05	160
2020-01-03	150


Tabela 3

Data	Wartość
2020-01-01	100
2020-01-02	130
2020-01-03	150
2020-01-04	140

Jak zmusić dane do bycia szeregiem czasowym?

1. Wprowadzić sztuczną kolumnę z kolejnością - 1,2,3,4,....
2. Posortować indeks czasowy.
3. Ponowne próbkowanie - resampling z potrzebną częstotliwością.

Resampling - wkładamy brakujące indeksy

Tabela 1			Tabela 2	
Data	Wartość		Data	Wartość
2020-01-01	100		2020-01-01	100
2020-01-02	200		2020-01-02	200
2020-01-04	120		2020-01-03	
2020-01-07	170		2020-01-04	120
			2020-01-05	
			2020-01-06	
			2020-01-07	170

Z pustego i Salomon nie należy - wartości brakujące

1. Fill forward - wartość brakująca taka jak poprzednia
2. Backfilling - wartość brakująca tak jak następna
3. Interpolacja - wartości pośrodku pomiędzy istniejącymi wartościami
4. Wartość średnia / mediana

Data	Wartość	Fill Forward	Backfilling	Interpolacja	Srednia
2020-01-01	100	100	100	100	100
2020-01-02	200	200	200	200	200
2020-01-03		200	120	160	148
2020-01-04	120	120	120	120	120
2020-01-05		120	170	137	148
2020-01-06		120	170	154	148
2020-01-07	170	170	170	170	170

Feature Engineering

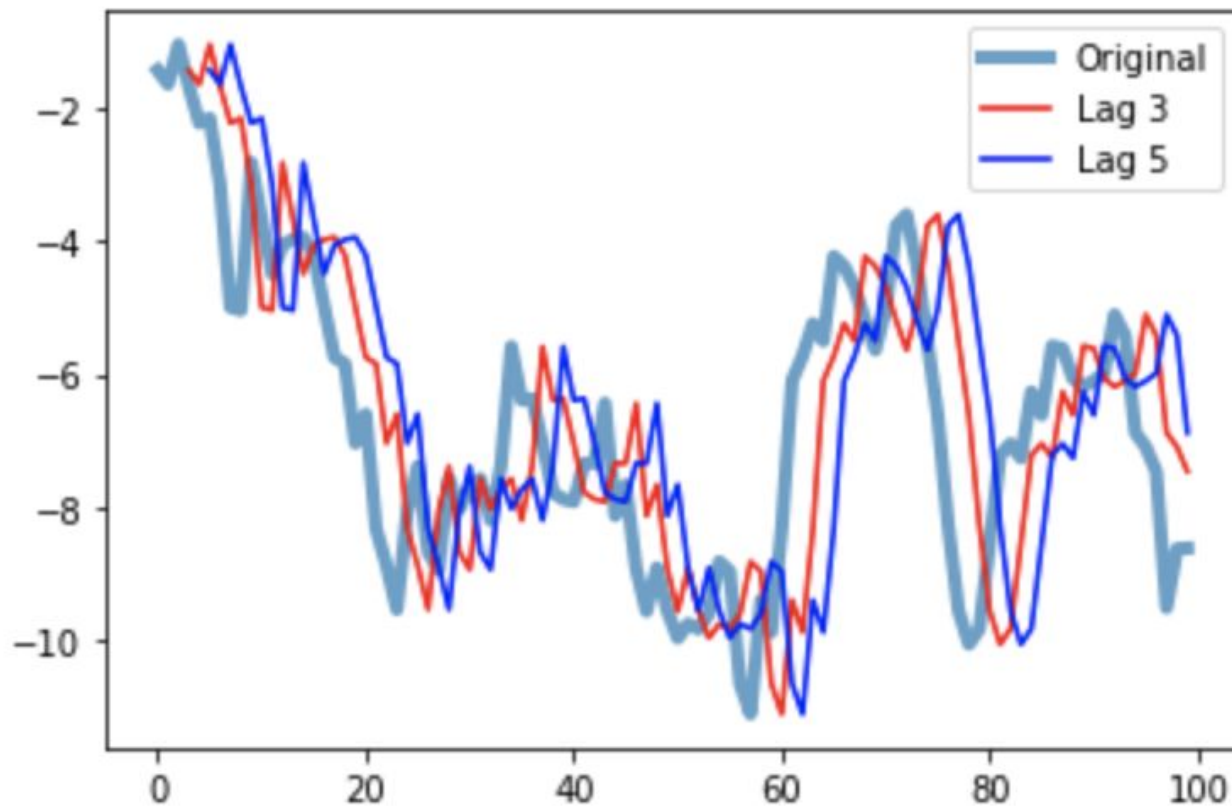
“Stosowane uczenia maszynowe to w zasadzie inżynieria zmiennych”

- *Andrew Ng, profesor*

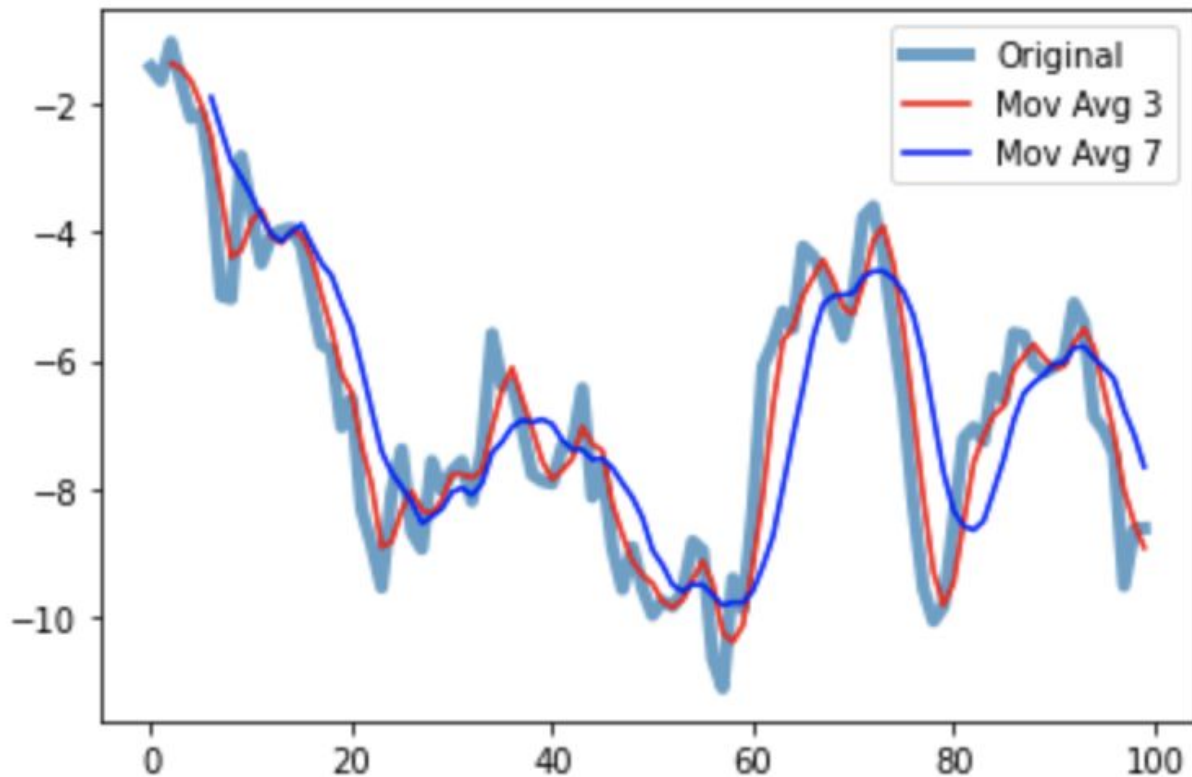
Wzbogacamy nasze dane

1. **Opóźnione wartości** - co było wczoraj, tydzień temu, miesiąc temu?
2. **Kroczące statystyki** - średnia, mediana, odchylenie standardowe, suma itd. o różnych wartościach “rolling window”
3. **Trend i sezonowość**
4. **Zmienne związane z datą**
 - Miesiąc
 - Dzień tygodnia
 - Czy to weekend?
 - Czy to dzień wolny od pracy?
 - Czy to początek miesiąca / kwartału?
5. **Różnicowanie**

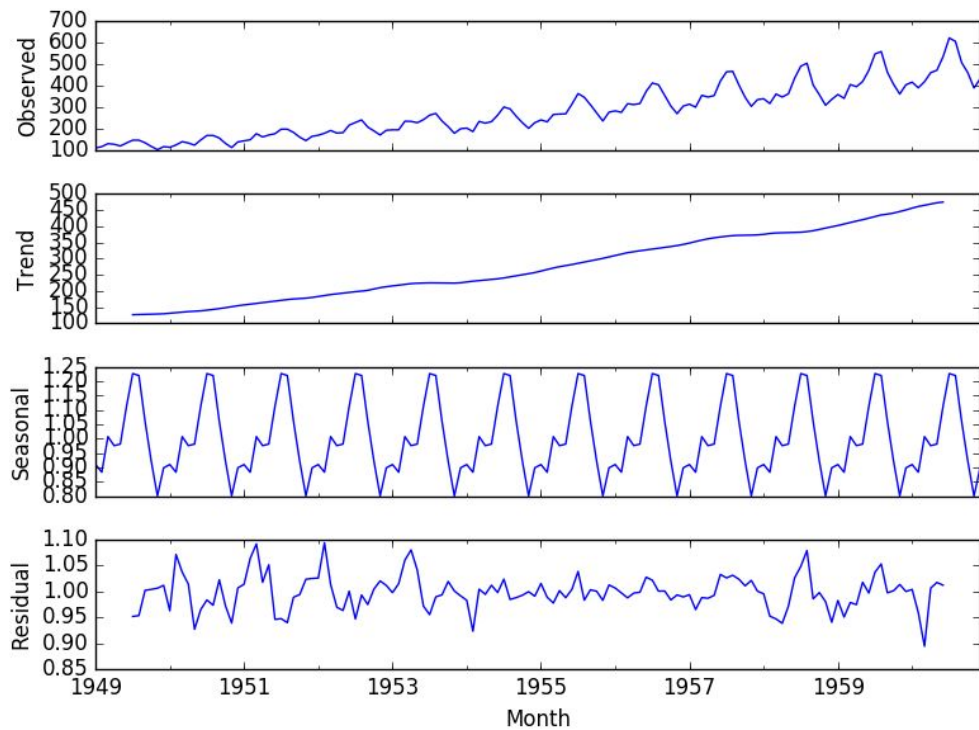
Wartości opóźnione



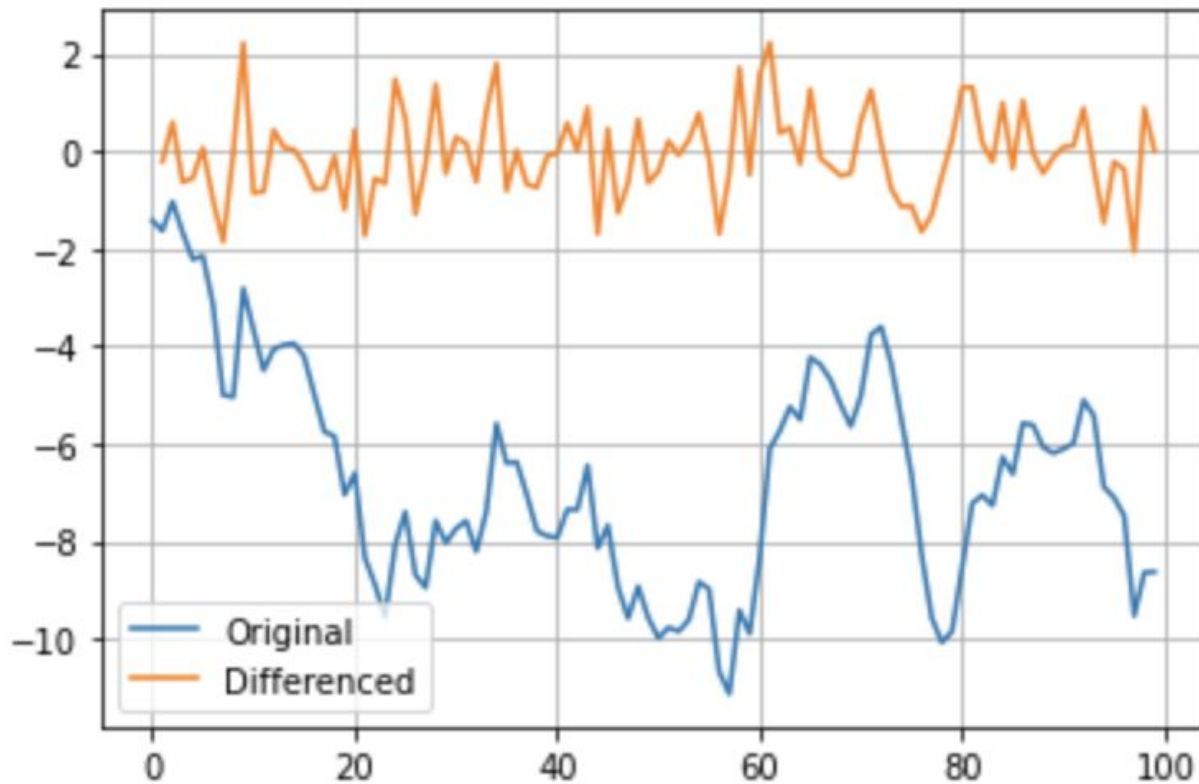
Kroczące statystyki



Rozkład na trend i sezonowość



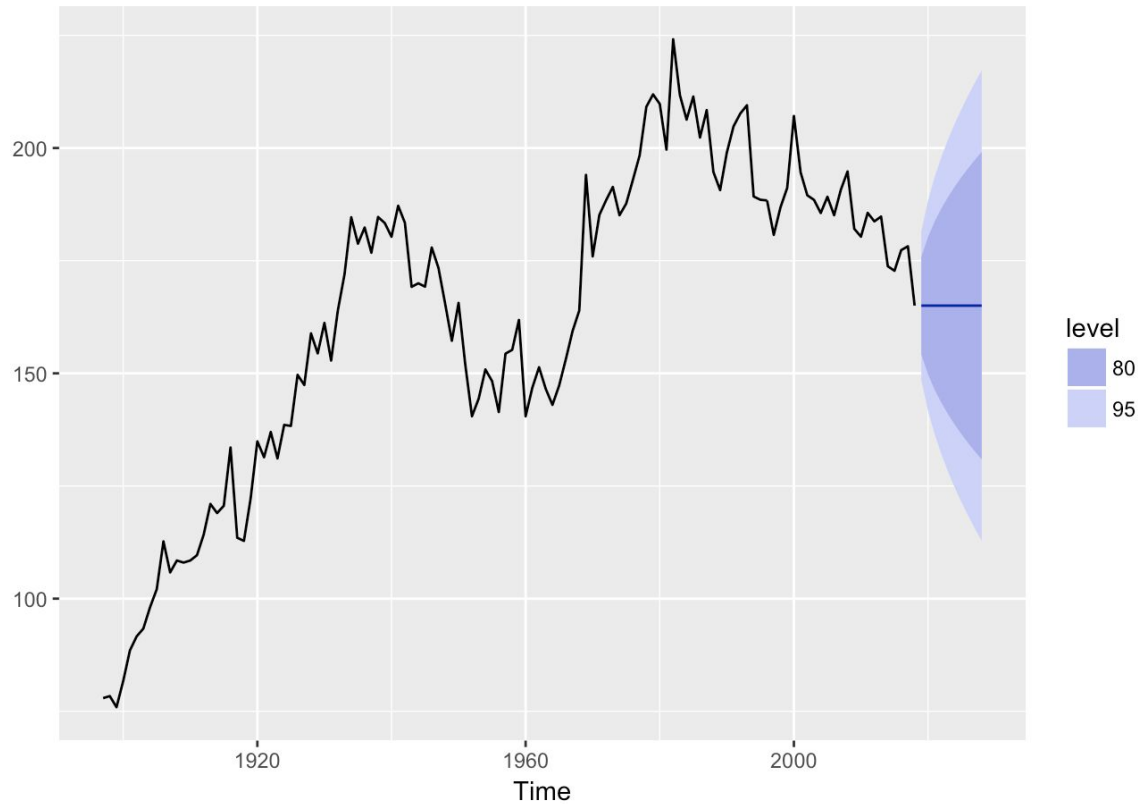
Różnicowanie szeregu czasowego



Feature Engineering - Przykład

Data	Wartość	Lag_1	Lag_3	Srednia_3	Dzień tyg.	Weekend
2020-01-01	100	-	-	-	3	0
2020-01-02	200	100	-	-	4	0
2020-01-03	200	200	-	166.7	5	0
2020-01-04	120	200	100	173.3	6	1
2020-01-05	120	120	200	146.7	0	1
2020-01-06	120	120	200	120	1	0
2020-01-07	170	120	120	136.7	2	0

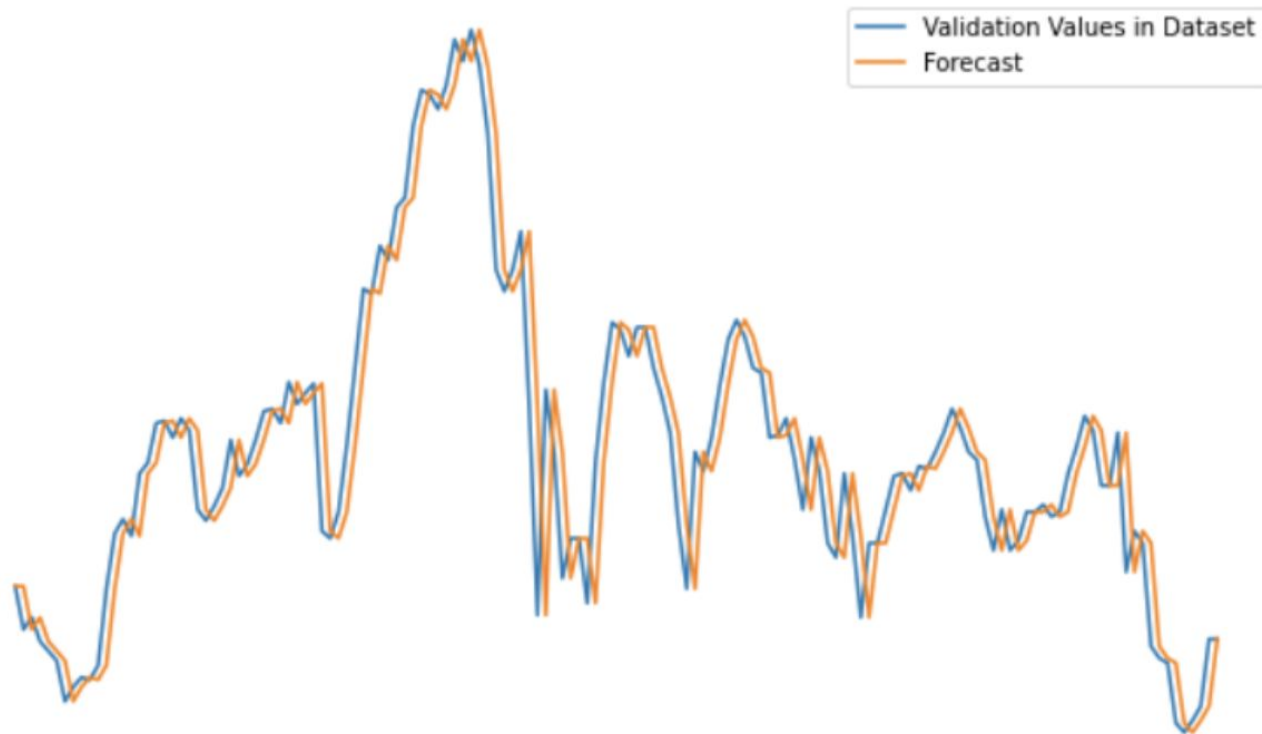
Naiwna Prognoza



Jutro będzie tak jak dzisiaj.

I pojutrze też ...

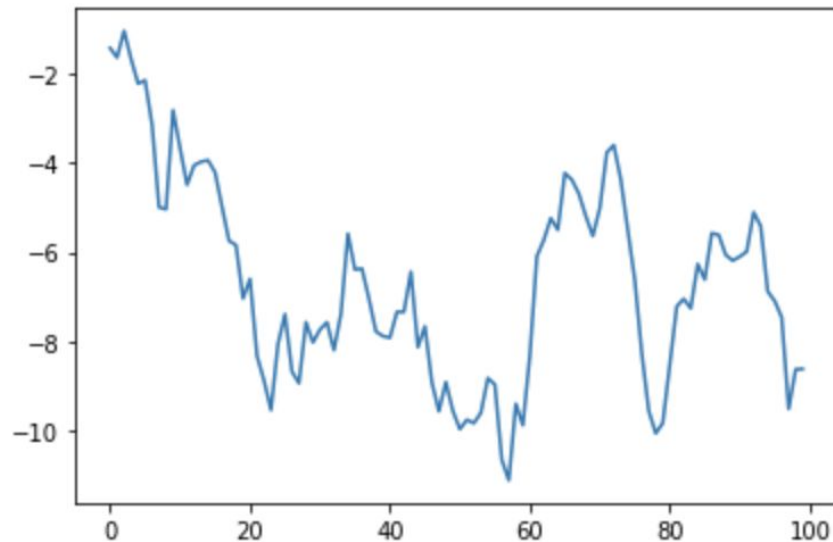
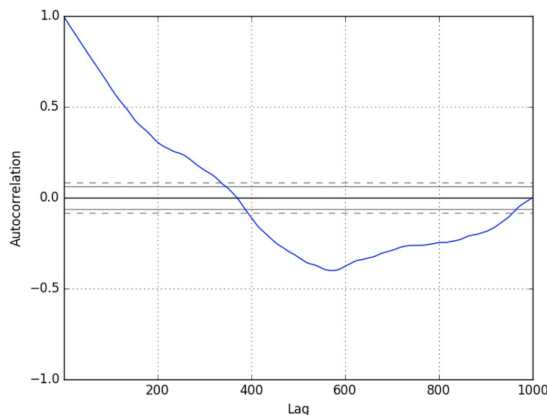
Naiwna prognoza - walidacja



Naiwna prognoza najlepszym rozwiązaniem

Naiwna prognoza będzie najlepszym rozwiązaniem, jeżeli szereg czasowy jest błędzeniem losowym (random walk).

- 1) Stopniowo zapadająca autokorelacja
- 2) Niestacjonarność
- 3) Prognoza naiwna najlepszą prognozą



Jak przygotować dane do trenowania?

Data	Wartość	Lag_1	Lag_3	Srednia_3
2020-01-01	100	100	100	166.7
2020-01-02	200	100	100	166.7
2020-01-03	200	200	100	166.7
2020-01-04	120	200	100	173.3
2020-01-05	120	120	200	146.7
2020-01-06	120	120	200	120
2020-01-07	170	120	120	136.7

Data	Wartość				
2020-01-01	100	Wartość	Lag_1	Lag_3	Srednia_3
2020-01-02	200	100	100	100	166.7
2020-01-03	200	200	100	100	166.7
2020-01-04	120	200	200	100	166.7
2020-01-05	120	120	200	100	173.3
2020-01-06	120	120	120	200	146.7
2020-01-07	170	120	120	200	120
	?	170	120	120	136.7

Data	Wartość	Lag_1	Lag_3	Srednia_3
2020-01-01	100	100	100	166.7
2020-01-02	200	100	100	166.7
2020-01-03	200	200	100	166.7
2020-01-04	120	200	100	173.3
2020-01-05	120	120	200	146.7
2020-01-06	120	120	200	120
2020-01-07	170	120	120	136.7

Data	Wartość				
2020-01-01	100				
2020-01-02	200	Wartość	Lag 1	Lag 3	Srednia 3
2020-01-03	200	100	100	100	166.7
2020-01-04	120	200	100	100	166.7
2020-01-05	120	200	200	100	166.7
2020-01-06	120	120	200	100	173.3
2020-01-07	170	120	120	200	146.7
		120	120	200	120
	?	170	120	120	136.7

Dlaczego walidacja krzyżowa nie działa?

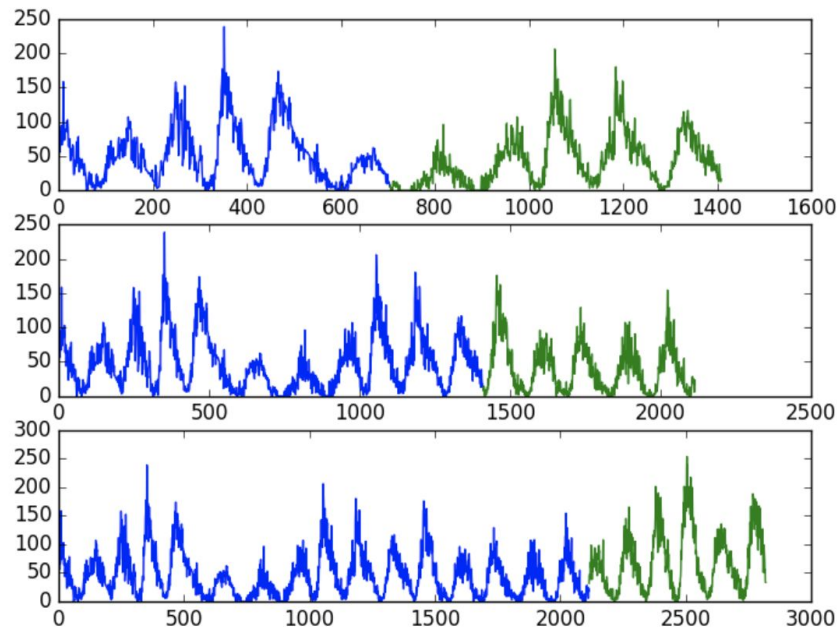
Główny powód - randomizacja



Metody walidacji modeli dla szeregów czasowych

1. Train-test split
2. Multiple train-test split
3. Walkforward validation

Od czego zależy?



Czy tradycyjne metody już do niczego?

Spyros Markidakis - profesor Uniwersytetu w Nikozji, Cypr, 2018

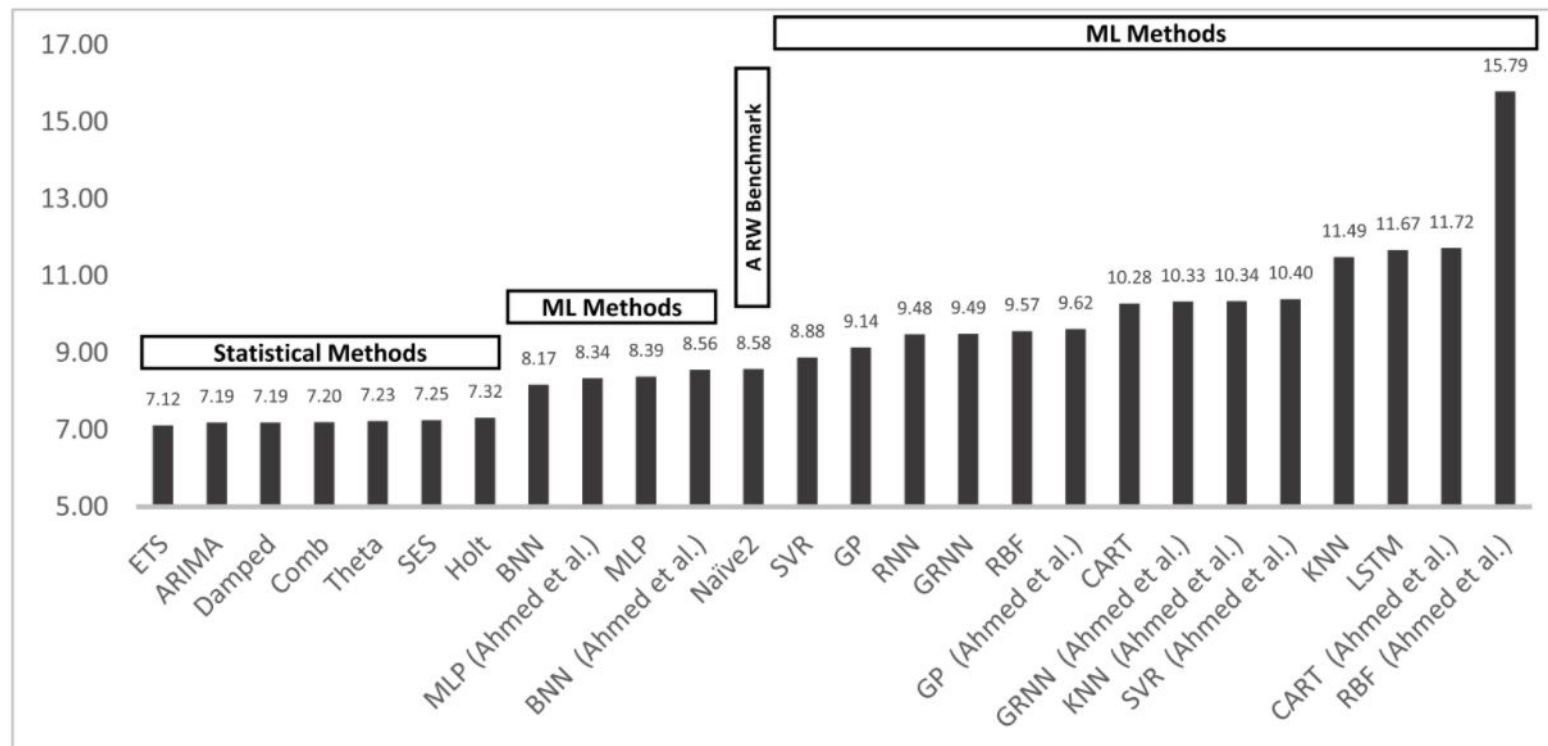
Algorytmy: 8 tradycyjnych + 10 ML

Zestawy danych: 1045 jednowymiarowych szeregów czasowych, częstotliwość - od godzinowych do rocznych.

Przygotowanie danych: różne kombinacje

Walidacja: walk forward, 18 ostatnich obserwacji

Wyniki badań - błędy prognozowania



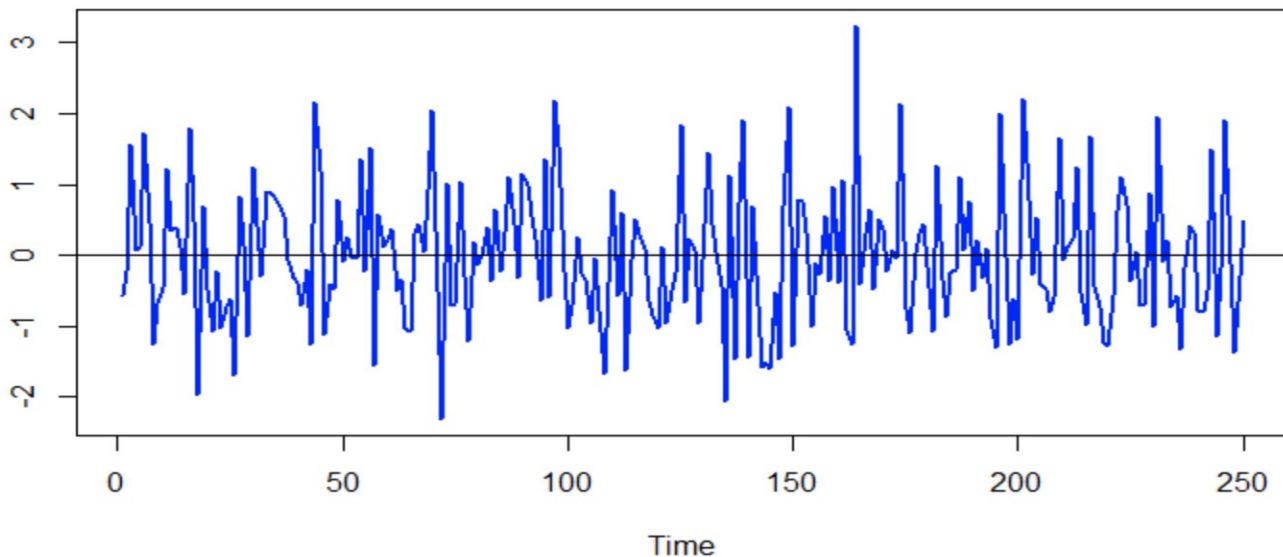
Algorytmy do spróbowania

1. Prognoza Naiwna (Naive, Naive 2)
2. Modele auto regresyjne (AR, ARIMA, SARIMA)
3. Wygładzanie wykładnicze (Holta, Wintersa)
4. ML - modele liniowe (Linear, Ridge, Lasso, ElasticNet)
5. ML - modele nieliniowe (KNN, SVR, drzewa decyzyjne)
6. ML - Ensemble Learning (laso losowe, XGBoost)
7. Głębokie uczenie (MLP, CNN, LSTM, hybrydy)

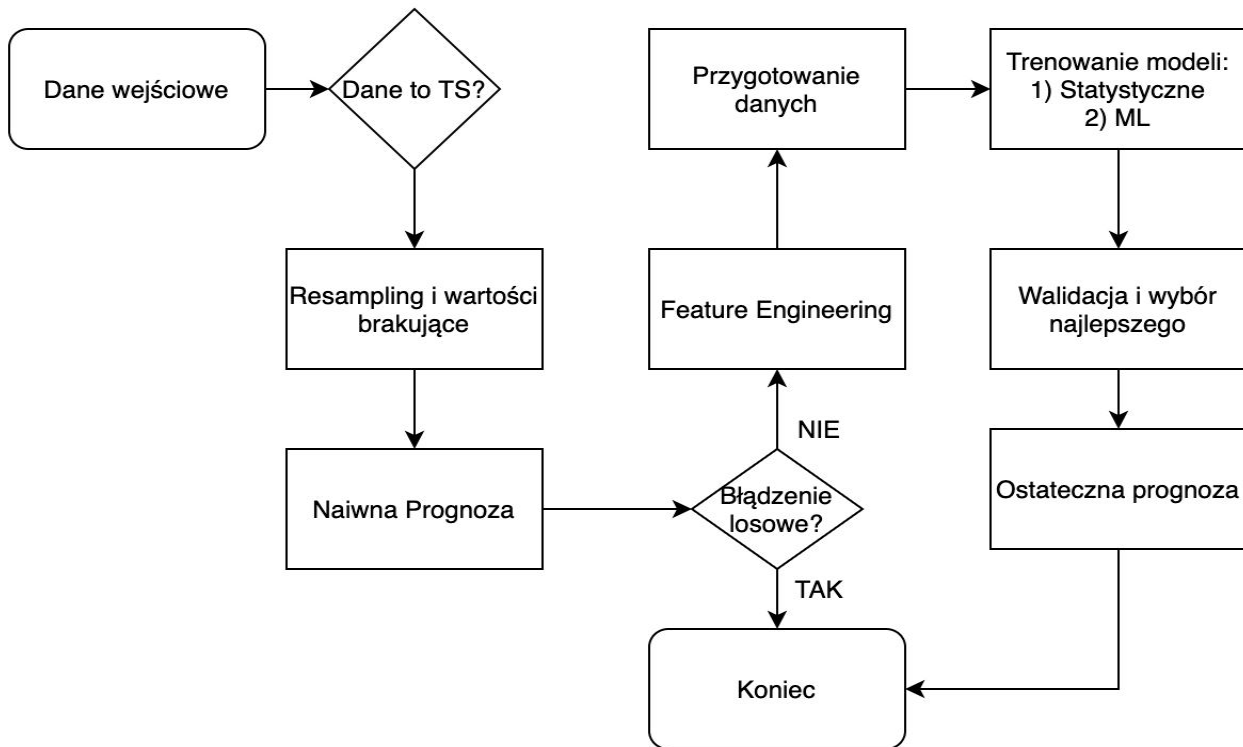


Kiedy lepiej się już nie da?

- Błędy prognozowania są białym szumem.
- Średnia błędów - 0.
- Normalny rozkład



Podsumowanie - Proces ML dla szeregów czasowych



Dziękuję!
Pytania?