Szymon Baczyński KD-158

Raport z projektu:

Prognozowanie ultra-krótkoterminowe zapotrzebowania na energię systemu elektroenergetycznego z kwantyzacją 15 minutową

1. Prognozowanie ultra-krótkoterminowe i wskaźniki jakości prognozy

Prognozowanie ultra-krótkoterminowe wykorzystywane w elektroenergetyce wykorzystuje się do sterowania produkcją energii w elektrowniach oraz zarządzania siecią elektroenergetyczną. Badania nad prognozowaniem, w tej dziedzinie, istnieją od kilkudziesięciu lat. Podczas prognozowania wykorzystuje się różne dane - głównie dane historyczne o zapotrzebowaniu elektroenergetycznym oraz dane pogodowe (lokalne czy dla całego kraju).

Wskaźniki jakości prognozy - przede wszystkim są to *MAE* oraz *MAPE*.

MAE (*mean absolute error*) - średni błąd bezwzględny, informuje on o ile średnio w okresie prognoz, będzie wynosić odchylenie od wartości rzeczywistej. Krótko mówiąc, jakim błędem miarowym jest obarczona prognoza.

MAPE (*mean absolute percentage error*) - średni bezwzględny błąd procentowy informuje o średniej wielkości błędów prognoz dla okresu testowego, wyrażonych w procentach. Wartość *MAPE* pozwala na porównanie dokładność prognoz różnych modeli.

2. Dane wykorzystane w projekcie

Cały projekt wykonany został w pakiecie R, wykorzystując odpowiednie biblioteki:
readr, MASS, e1071, rpart, randomForest, tidyverse, lubridate, stringr, caret, FNN, plotly

Dane do projektu zostały pobrane z serwisu: *https://data.open-power-system-data.org/*
Są tam dane *Time Series* z 37 krajów Europy, które mają znaczniki czasowe co 15, 30 i 60 minut. Dane są szczegółowo zebrane z ostatnich kilku lat.

Do projektu wybrane zostały dane z Węgier, z kwantyzacją 15 minutową. W pliku znajdowały się dane:

1) Znacznik czasowy UTC
2) Teraźniejsze zapotrzebowanie na energię
3) Prognoza zapotrzebowania
4) Teraźniejsza generacja energii z farm wiatrowych

Podczas wstępnej obróbki danych dokonano podziału i odpowiedniej interpretacji danych. W wyniku wcześniejszych czynności powstała tablica, w której były kolumny z danymi:

1) Time = Czas UTC
2) Load_Now = Teraźniejsze zapotrzebowanie na energię
3) Hour = godziny (wydobyte z kolumny “Time”
4) Load_Min15 = dane cofnięte o 15 minut
5) Load_Min30 = dane cofnięte o 30 minut
6) Load_Min45 = dane cofnięte o 45 minut
7) Load_Day1B = dane cofnięte o 1 dzień
8) Load_Day1B15min = dane cofnięte o 1 dzień i 15 minut
9) Load_Day1Bp15min = dane cofnięte o 23 godziny i 45 minut

Dane do uczenia modeli obejmowały zakres:
od 2017-04-30 & 23:00 do 2018-04-30 & 22:45
Dane do testowania modeli obejmowały zakres:
od 2018-04-30 & 23:00 do 2019-04-30 & 22:45

3. Wyniki i porównanie modeli

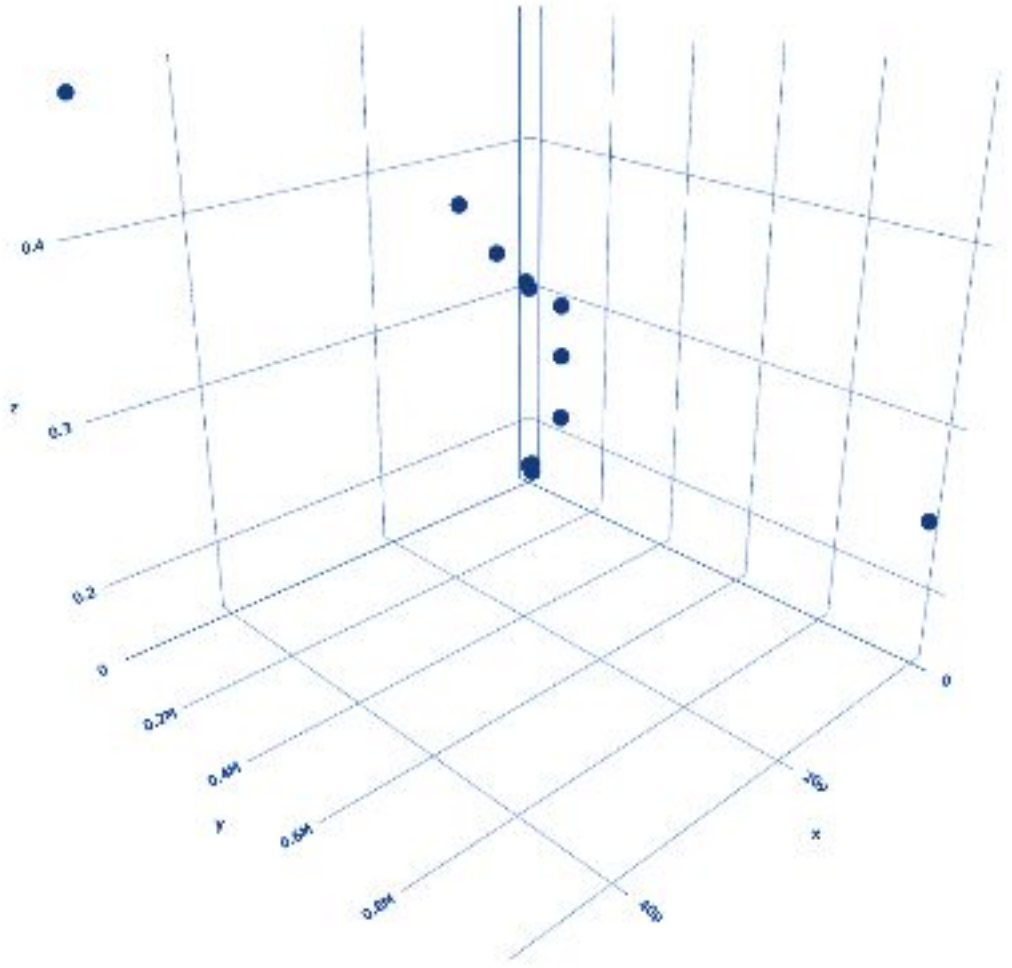
Jako wyznacznik jakości modelu, można każdy z modeli porównać do modelu z metody “naiwnej”, która jako wynik podaje wartości, które były 15 minut wcześniej. Modele, których wyniki, są gorsze od metody “naiwnej” można uznać za nie rokujące do dalszych analiz.
Notka: Predykcja modelu musi być cofnięta o 1 wiersz - lepsze pokrycie wyników.

Metoda Naiwna: MAE = 43,704 | MAPE = 0,907

- Wyniki dla modeli wykorzystujących: **Load_Now ~ Load_Min15 + Load_Day1B**

Model	***MAE***	***MAPE***
kNN (k=43)	21,241	0,443
Drzewo (ANOVA)	24,879	0,526
Lasy Losowe (n=10)	27,747	0,581
Lasy Losowe (n=100)	24,777	0,518
SVM (□=$5 \cdot 10^{-5}$ | C=$2 \cdot 10^{3}$)	22,416	0,474
SVM (□=$1 \cdot 10^{-5}$ | C=$2 \cdot 10^{3}$)	17,422	0,371
SVM (□=$5 \cdot 10^{-6}$ | C=$5 \cdot 10^{3}$)	15,547	0,333
SVM (□=$1 \cdot 10^{-6}$ | C=$1 \cdot 10^{4}$)	14,258	0,308

SVM (□=5 · 10^{-7} | C=1 · 10^{4})	14,047	0,303
SVM (□=1 · 10^{-7} | C=1 · 10^{5})	13,707	0,297
SVM (□=1 · 10^{-8} | C=1 · 10^{5})	12,071	0,261
SVM (□=1 · 10^{-9} | C=1 · 10^{6})	11,579	0,246
SVM (□=1 · 10^{-9} | C=1 · 10^{5})	10,132	0,216
SVM (□=1 · 10^{-9} | C=1 · 10^{4})	8,103	0,171
SVM (□=1 · 10^{-9} | C=1 · 10^{3})	8,184	0,169
SVM (□=1 · 10^{-10} | C=1 · 10^{4})	7,980	0,165
Best SVM (Tuning)	4,325	0,089



Wykres dla SVM ***(Load_Min15 + Load_Day1B)***
Zależność *MAPE* (oś Z) od *Gammy* (oś X) i *Kosztu* (oś Y)

- Wyniki dla modeli wykorzystujących:
  Load_Now ~ Load_Min15 + Load_Day1B + Hour

Model	***MAE***	***MAPE***
Drzewo (ANOVA)	35,173	0,749
Lasy Losowe (n=10)	47,469	1,014
Lasy Losowe (n=100)	44,716	0,951
SVM ($\square$=$1 \cdot 10^{-7}$ | C=$1 \cdot 10^{5}$)	15,388	0,328
SVM ($\square$=$1 \cdot 10^{-7}$ | C=$1 \cdot 10^{4}$)	13,522	0,272
SVM ($\square$=$1 \cdot 10^{-8}$ | C=$1 \cdot 10^{4}$)	11,772	0,257
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{4}$)	7,985	0,165
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{3}$)	7,981	0,165
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{5}$)	7,446	0,155

- Wyniki dla modeli wykorzystujących:
  Load_Now ~ Load_Min15 + Load_Min30 + Load_Day1B + Hour

Model	***MAE***	***MAPE***
Drzewo (ANOVA)	37,275	0,789
Lasy Losowe (n=10)	41,307	0,879
Lasy Losowe (n=100)	36,756	0,777
SVM ($\square$=$1 \cdot 10^{-7}$ | C=$1 \cdot 10^{5}$)	32,393	0,685
SVM ($\square$=$1 \cdot 10^{-7}$ | C=$1 \cdot 10^{4}$)	32,459	0,686
SVM ($\square$=$1 \cdot 10^{-8}$ | C=$1 \cdot 10^{4}$)	32,479	0,685
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{4}$)	17,254	0,356
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{3}$)	31,348	0,649
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{5}$)	24,361	0,496

- Wyniki dla modeli wykorzystujących:
  Load_Now ~ Load_Min15 + Load_Min30 + Load_Day1B

Model	***MAE***	***MAPE***
Drzewo (ANOVA)	34,147	0,724
Lasy Losowe (n=10)	31,613	0,667
Lasy Losowe (n=100)	29,298	0,619
SVM ($\square$=$1 \cdot 10^{-7}$ | C=$1 \cdot 10^{5}$)	32,403	0,685
SVM ($\square$=$1 \cdot 10^{-7}$ | C=$1 \cdot 10^{4}$)	32,287	0,682
SVM ($\square$=$1 \cdot 10^{-8}$ | C=$1 \cdot 10^{4}$)	32,479	0,685
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{3}$)	24,360	0,495
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{4}$)	17,255	0,356
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{5}$)	31,349	0,649

- Wyniki dla modeli wykorzystujących:
  Load_Now ~ Load_Min15 + Load_Min30 + Load_Min45 + Load_Day1B

Model	***MAE***	***MAPE***
Drzewo (ANOVA)	35,296	0,749
Lasy Losowe (n=10)	34,165	0,722
Lasy Losowe (n=100)	31,339	0,665
SVM ($\square$=$1 \cdot 10^{-7}$ | C=$1 \cdot 10^{5}$)	33,275	0,703
SVM ($\square$=$1 \cdot 10^{-7}$ | C=$1 \cdot 10^{4}$)	33,388	0,705
SVM ($\square$=$1 \cdot 10^{-8}$ | C=$1 \cdot 10^{4}$)	32,940	0,697
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{3}$)	26,779	0,545
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{4}$)	25,662	0,531
SVM ($\square$=$1 \cdot 10^{-10}$ | C=$1 \cdot 10^{5}$)	32,474	0,675

- Wyniki dla modeli wykorzystujących:

Load_Now ~ Load_Min15 + Load_Day1B + Load_Day1B15min + Load_Day1Bp15min

Model	***MAE***	***MAPE***
Drzewo (ANOVA)	35,321	0,744
Lasy Losowe (n=10)	51,816	1,073
Lasy Losowe (n=100)	44,728	0,930
SVM ($\gamma=1 \cdot 10^{-7}$ | $C=1 \cdot 10^{5}$)	37,888	0,790
SVM ($\gamma=1 \cdot 10^{-7}$ | $C=1 \cdot 10^{4}$)	37,918	0,791
SVM ($\gamma=1 \cdot 10^{-8}$ | $C=1 \cdot 10^{4}$)	38,277	0,798
SVM ($\gamma=1 \cdot 10^{-10}$ | $C=1 \cdot 10^{3}$)	27,032	0,557
SVM ($\gamma=1 \cdot 10^{-10}$ | $C=1 \cdot 10^{4}$)	34,105	0,715
SVM ($\gamma=1 \cdot 10^{-10}$ | $C=1 \cdot 10^{5}$)	37,592	0,786

4. Podsumowanie i wnioski

Model predykcyjny prognozujący ultra-krótkoterminowe zapotrzebowania na energię systemu elektroenergetycznego musi być bardzo dokładny. Takie modele potrzebują dużej ilości danych, by ich prognoza była warta uwagi. W branży elektroenergetycznej do prognozowania wykorzystuje się skomplikowane sieci neuronowe, które są badane i również rozwijane pod tym kątem.

Wyniki osiągnięte przez Drzewo Decyzyjne, bazujące na regresji ANOVA, osiąga wyniki lepsze od metody Naiwnej, lecz nie wystarczające by mówić o zadowalającym prognozowaniu.

Lasy Losowe - model, który osiągał współczynniki jakości gorsze od Drzewa Decyzyjnego. Wykorzystanie funkcji do szukania najlepszego modelu nie dały lepszych rezultatów.

SVM - Maszyna Wektorów Nośnych - teoretycznie modele, które dostały w danych treningowych więcej zmiennych powinny uzyskiwać lepsze wyniki. Jednakże model SVM nauczony na 2 zmiennych - *Load_Min15* + *Load_Day1B* - uzyskał wynik *MAPE* na poziomie 0,089%. Odchylenia tego rzędu pozwalają na ponowne spojrzenie na klasyczne modele predykcyjne. Możliwe, że dla 3 zmiennych - *Load_Min15* + *Load_Day1B* + *Hour* - najlepszy model uzyskałby lepsze współczynniki jakości niż model uczący się na 2 zmiennych.

Czas uczenia poszczególnych modeli oscylował od 20 minut do 40 minut. Problematyczne, aczkolwiek możliwe, zrównoleglenie obliczeń nie było wykonane przez brak znajomości narzędzi w pakiecie R. Niektóre fragmenty kodu nie zostały wykonane przez czas wykonywania oraz sprzęt.