Wydział Elektryczny
POLITECHNIKA WARSZAWSKA



# *Analiza ważności danych wejściowych dla różnych technik uczenia maszynowego w zadaniu krótkoterminowej prognozy produkcji energii elektrycznej przez elektrownię słoneczną w zakładzie przemysłowym*

1.Paweł PIOTROWSKI, 2. Marcin KOPYT, 3.Łukasz ROKICKI

Politechnika Warszawska

DIEGO DIGITAL SOLUTIONS

# Plan prezentacji



- Motywacja
- Opis zadania projektu badawczego
- Pytania badawcze
- Charakterystyka dostępnych zmiennych wejściowych do prognoz energii generowanej z PV
- Prognozy generacji z PV z horyzontem 15 minut
- Odpowiedzi na pytania badawcze oraz wnioski

# Motywacja



- Redukcja śladu węglowego zakładu przemysłowego może zostać osiągnięta za pomocą poprawy efektywności energetycznej, częściowe zasilanie obiektu z OZE, ustalanie reżimów pracy urządzeń oraz zarządzanie odpadami przemysłowym
- Niniejsze badania stanowią fragment międzynarodowego Projektu Diego, mającego zapewnić skalowalne i powtarzalne rozwiązania kompleksowe służące redukcji śladu węglowego obiektów przemysłowych
- Przedstawione badania związane są z prognozowaniem energii wytwarzanej przez instalację PV zakładu przemysłowego. Prognozy te, będą w dalszej kolejności stosowane do optymalizacji nadążnej pracy zakładu przemysłowego w sposób umożliwiający redukcję jego śladu węglowego.
- Ze względu na konieczność automatyzacji procesu istotne jest poznanie potencjalnej wrażliwości modeli prognostycznych na zaburzenia mogące wystąpić w dostępnych w danym momencie danych wykorzystywanych przez model prognostyczny.

# Opis zagadnienia



- W badaniach rozpatruje się zakład przemysłowy wykonujący obróbkę stali, posiadający jako dodatkowe źródło energii instalację PV.
- Prognozy generacji energii czynnej były wykonywane nadążnie, 1 krokowo, z horyzontem 15-min.
- Historyczne wartości prognozowanej generacji pochodziły z systemu opomiarowania właściciela zakładu. Dodatkowe zmienne objaśniające pochodziły z układu pomiaru pogody.
- W celu anonimizacji danych zmienną wyjściową znormalizowano wykorzystując wartość mocy znamionowej transformatora.
- Zmienne wejściowe pogodowe przed analizą oraz do zastosowania w modelach prognostycznych poddano normalizacji min-max.

# Pytania badawcze dot. danych prognostycznych



- Czy możliwa jest dalsza, wielokryterialna poprawa dokładności prognoz względem z założenia dokładnej dla krótkiego horyzontu czasu metody naiwnej (modelu odniesienia do jakości innych metod)?
- Jak złożone muszą być modele uczenia maszynowego dla opisanego problemu?
- Czy występują różnice w ważności zmiennych wejściowych, będące przesłanką do tworzenia modeli zastępczych, przełączanych albo łączonych?
- Na dokładność, których zmiennych wejściowych należy zwrócić uwagę w zależności od modelu prognostycznego?

# Charakterystyka zmiennej prognozowanej i nasłonecznienia



| Statystyka | Generacja energii | Nasłonecznienie |
|---|---|---|
| Średnia arytmetyczna | 0,0584 [p.u.] | 259 [W/m$^2$] |
| Odchylenie standardowe | 0,0627 [p.u.] | 284 [W/m$^2$] |
| Minimum | 0 [p.u.] | 0 [W/m$^2$] |
| Maksimum | 0,203 [p.u.] | 1000 [W/m$^2$] |
| Rozstęp | 0,203 [p.u.] | 1000 [W/m$^2$] |
| Współczynnik zmienności | 107,362 [%] | 110 [%] |
| Percentyl 10 | 0 [p.u.] | 0 [W/m$^2$] |
| Percentyl 25 (dolny kwartyl) | 0,00414 [p.u.] | 25 [W/m$^2$] |
| Percentyl 50 (mediana) | 0,03159 [p.u.] | 131 [W/m$^2$] |
| Percentyl 75 (górny kwartyl) | 0,10443 [p.u.] | 447 [W/m$^2$] |
| Percentyl 90 | 0,16449 [p.u.] | 733 [W/m$^2$] |
| Wariancja | 0,003936 | 80738 |
| Skosność | 0,8655 | 0,9828 |
| Kurtoza | -0,6024 | -0,28475 |

# Zmienne wejściowe zastosowane w badaniach z ich wartościami korelacji liniowej Pearsona (R)



| Opis zmiennej wejściowej | Kod | R |
|---|---|---|
| Miesiąc | Month | -0,1155 |
| Godzina | Hour | 0,3800 |
| Okres wzrostu nasłonecznienia | R_SI | -0,2406 |
| Okres spadku nasłonecznienia | D_SI | -0,2406 |
| Wygładzona generacja w okresie T-1 [p.u.] | SEG(T-1) | 0,9274 |
| Generacja w okresie T-1 [p.u.] | EG(T-1) | 0,9245 |
| Generacja w okresie T-2 [p.u.] | EG(T-2) | 0,8820 |
| Generacja w okresie T-3 [p.u.] | EG(T-3) | 0,8460 |
| Generacja w okresie T-4 [p.u.] | EG(T-4) | 0,8103 |
| Generacja w okresie T-5 [p.u.] | EG(T-5) | 0,7729 |
| Generacja w okresie T-6 [p.u.] | EG(T-6) | 0,7373 |
| Generacja w okresie T-96 [p.u.] | EG(T-96) | 0,6719 |
| Generacja w okresie T-192 [p.u.] | EG(T-192) | 0,6030 |

| Opis zmiennej wejściowej | Kod | R |
|---|---|---|
| Nasłonecznienie w okresie T-1 [W/m$^2$] | SI(T-1) | 0,8818 |
| Nasłonecznienie w okresie T-2 [W/m$^2$] | SI(T-2) | 0,8406 |
| Nasłonecznienie w okresie T-3 [W/m$^2$] | SI(T-3) | 0,8054 |
| Nasłonecznienie w okresie T-4 [W/m$^2$] | SI(T-4) | 0,7684 |
| Nasłonecznienie w okresie T-5 [W/m$^2$] | SI(T-5) | 0,7328 |
| Nasłonecznienie w okresie T-6 [W/m$^2$] | SI(T-6) | 0,6960 |
| Nasłonecznienie w okresie T-96 [W/m$^2$] | SI(T-96) | 0,6566 |
| Nasłonecznienie w okresie T-192 [W/m$^2$] | SI(T-192) | 0,5862 |
| Temperatura powietrza w okresie T-1 [$^0$C] | AT(T-1) | 0,3899 |
| Temperatura powietrza w okresie T-2 [$^0$C] | AT(T-2) | 0,3815 |
| Prędkość wiatru w okresie T-1 [m/s] | WS(T-1) | -0,0720 |

# Ranking ważności zmiennych wejściowych



W celu określenia ważności poszczególnych zmiennych wejściowych posłużono się 5 miarami:

- Korelacją liniową Pearsona (R)
- Punktami z analizy wrażliwości dla sztucznej sieci neuronowej (MLP)
- Punktami z analizy ważności dla lasów losowych (RF)
- Punktami z analizy ważności dla drzew decyzyjnych wzmacnianych gradientowo (GBDT)
- Punktami z analizy ważności dla regresji liniowej wielorakiej (MLR)

- Metody analizy ważności były różne dla poszczególnych modeli. Dla przykładu, dla lasu losowego była to częstotliwość zastosowania zmiennej do podziału w drzewie, a w MLP było to usunięcie zmiennej na wejściu i zastosowanie wartości średniej)
- Następnie zmienne posortowano malejąco względem ważności, przyznając punkty według formuły (25 - miejsce na podium ważności dla danej metryki). Następnie punkty posumowano uzyskując wypadkową ważność

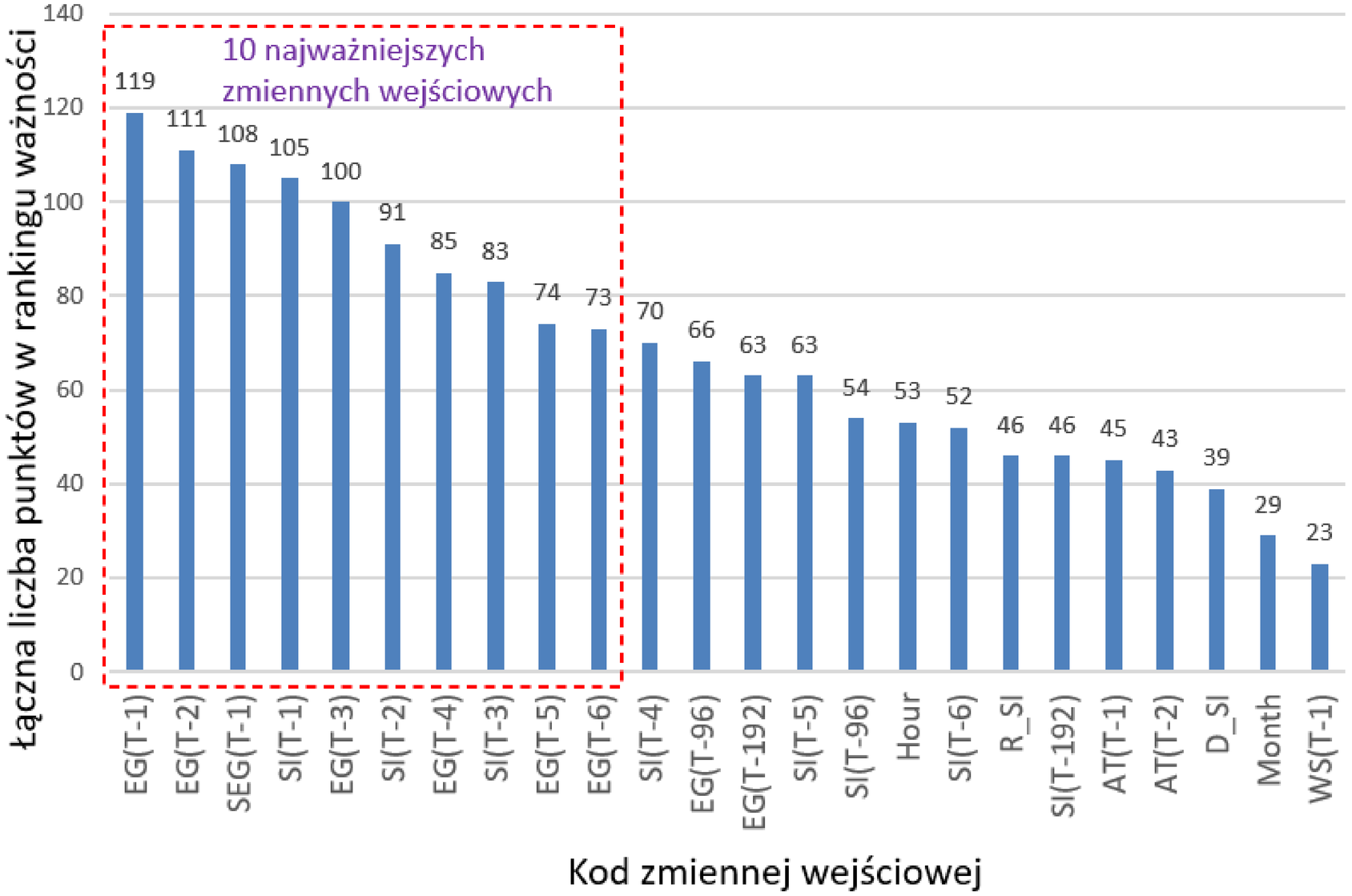
# Ranking ważności zmiennych wejściowych



| Kod zmiennej | R | MLP | RF | GBDT | MLR | Suma punktów |
|---|---|---|---|---|---|---|
| EG(T-1) | 23 | 24 | 24 | 24 | 24 | 119 |
| EG(T-2) | 22 | 22 | 21 | 22 | 24 | 111 |
| SEG(T-1) | 24 | 23 | 23 | 23 | 15 | 108 |
| SI(T-1) | 21 | 20 | 22 | 21 | 21 | 105 |
| EG(T-3) | 20 | 17 | 20 | 20 | 23 | 100 |
| SI(T-2) | 19 | 21 | 17 | 19 | 15 | 91 |
| EG(T-4) | 18 | 12 | 19 | 18 | 18 | 85 |
| SI(T-3) | 17 | 18 | 16 | 17 | 15 | 83 |
| EG(T-5) | 16 | 14 | 13 | 16 | 15 | 74 |
| EG(T-6) | 14 | 15 | 15 | 14 | 15 | 73 |
| SI(T-4) | 15 | 9 | 18 | 13 | 15 | 70 |
| EG(T-96) | 11 | 5 | 11 | 15 | 24 | 66 |
| EG(T-192) | 9 | 10 | 9 | 11 | 24 | 63 |
| SI(T-5) | 13 | 11 | 14 | 10 | 15 | 63 |
| SI(T-96) | 10 | 7 | 10 | 12 | 15 | 54 |
| Hour | 5 | 19 | 7 | 7 | 15 | 53 |
| SI(T-6) | 12 | 3 | 12 | 8 | 17 | 52 |
| R_SI | 4 | 13 | 2 | 3 | 24 | 46 |
| SI(T-192) | 8 | 6 | 8 | 9 | 15 | 46 |
| AT(T-1) | 7 | 4 | 6 | 6 | 22 | 45 |
| AT(T-2) | 6 | 8 | 5 | 4 | 20 | 43 |
| D_SI | 3 | 16 | 3 | 2 | 15 | 39 |
| Month | 2 | 2 | 4 | 5 | 16 | 29 |
| WS(T-1) | 1 | 1 | 1 | 1 | 19 | 23 |

# Ranking ważności zmiennych wejściowych





# Ważność zmiennych wejściowych - wnioski



- Dla wszystkich modeli uczenia maszynowego zmienna EG(T-1) czyli ostatnia wartość prognozowanego szeregu czasowego sprzed okresu prognozy była najważniejszą zmienną wejściową.
- W przypadku prędkość wiatru w ostatnim okresie WST(T-1) wszystkie metody jednoznacznie wskazały, że jest to zmienna najmniej przydatna. Wykluczono więc potencjalną przesłankę co do zdolności odwzorowania przez modele uczenia maszynowego chłodzenia paneli wiatrem.
- Na uwagę zwraca również fakt, że dla modelu MLP zmienne o kodach R_SI oraz D_SI (sygnatury kierunku zmian prognozowanego procesu: wzrost/spadek) były znacznie ważniejsze niż dla pozostałych modeli prognostycznych. Podobne zjawisko wystąpiło dla zmiennej o kodzie Hour.
- W przypadku RF oraz GBDT wystąpiło silne podobieństwo w ważności poszczególnych zmiennych wejściowych – oba modele mają pewne cechy wspólne oraz są metodami zespołowymi.

# Prognozy generacji energii z PV



- Jako metodę naiwną zastosowano  ostatni znany pomiar generacji (EG(T-1))
- Jako metody prognostyczne zastosowano drzewa decyzyjne wzmacniane gradientowo (GBDT), lasy losowe (RF) sieć neuronową typu MLP, regresję liniową wieloraką (MLR)
- Całościowo pozyskano dane za 1 rok, z kwantyzacją 15-min
- Jako miary błędu prognoz zastosowano nRMSE [j.w.], nMAE[j.w.], nAPEmax[%]

# Prognozy generacji energii z PV



- Zebrane dane zostały podzielone  na zbiory treningowe, walidacyjne i testowe tak, by odzwierciedlić w prognozach charakter czterech sezonów klimatycznych. Wiosna trwała od pierwszego marca do pierwszego czerwca, lato kolejno do pierwszego września, jesień do pierwszego grudnia,  a pozostała część roku była traktowana jako zima.
- Ostatni tydzień każdego sezonu był przypisany do danych testowych, zaś pozostałe dane zostały podzielone na treningowo-walidacyjne w proporcji 80 % do 20%.
- Dane testowe stanowiły więc około 17 % całości dostępnego rocznego zbioru danych.
- Wykorzystano wszystkie zmienne wejściowe

# Modele prognostyczne: hiperparametry



GBDT:

- Liczba drzew 50-100
- Maksymalna głębokość drzew 2-6
- Współczynnik uczenia 0,1 – 0,25

RF:

- Liczba drzew 50-300
- Maksymalna głębokość drzew 3-30
- Liczbę próbek w węźle umożliwiającą podział 50-400,
- maksymalna głębokość 5-15,
- maksymalna liczbę węzłów 100/200.

MLP :

- Liczba  neuronów w pierwszej warstwie ukrytej: 10 do 30
- Funkcja aktywacji warstwy ukrytej:  tanh/sigmoid/exp
- Liczba iteracji: 38
- Optymalizator LBFGS

MLR: z członem stałym i bez członu stałego

# Wyniki prognoz generacji energii w systemie PV



| Model | nRMSE [j.w.] | nMAE [j.w.] | nAPEmax [%] |
|---|---|---|---|
| **RF** | 0,02123 | 0,01026 | 18,7999 |
| **MLP** | 0,02160 | 0,01007 | 18,5677 |
| **GBDT** | 0,02164 | 0,01090 | 18,21088 |
| **MLR** | 0,02233 | 0,01114 | 20,2689 |
| **NAIVE** | 0,02374 | 0,00988 | 20,2918 |

# Wnioski



- Zastosowanie kilku miar dokładności umożliwiło równoczesne zmniejszenie błędów znacząco dużych modeli predykcyjnych (nRMSE) oraz oczekiwanej maksymalnej niedokładności prognoz (nAPEmax). Największą poprawę uzyskał model MLP, dla którego błąd nRMSE zmalał o 9% , a nAPEmax o 8,5% w stosunku do metody naiwnej.
- Metody RF, GBDT oraz MLR bardziej koncentrowały się na redukcji błędów dużych, niż na redukcji błędu średniego. Model liniowy MLR miał większy błąd nMAE o prawie 13%, a GBDT błąd większy o 10,2%. Dla MLP i RF błąd nMAE był większy jedynie o odpowiednio 1,9% i 3,8% w stosunku do błędu metody naiwnej.
- Model RF wykazał największa poprawę błędu nRMSE (o 10,6% w stosunku do metody naiwnej). Modele MLP oraz RF mogą więc stanowić skuteczne narzędzie optymalizacji wielokryterialnej dla analizowanego problemu. Dla odmiany model GBDT oraz model MLR kompensują poprawę błędów dużych, wzrostem wartości błędu średniego.

# Wnioski c.d.



- Najważniejszymi zmiennymi wejściowymi dla modeli prognostycznych są wartości generacji energii elektrycznej oraz wartości natężenia promieniowania słonecznego z kilku ostatnich 15-minutowych okresów przed okresem prognozy.
- Najmniej wartościowymi zmiennymi wejściowymi są prędkość wiatru oraz wskaźnik sezonowości (Month).
- Metoda MLP jest bardzo wrażliwa na podanie jawnych informacji o fragmencie cyklu dobowego (godziny i markerów spadku oraz wzrostu nasłonecznienia. Dla modeli RF oraz GBDT nie są to informacje ważne. Różnice w ważności informacji mogą świadczyć o potencjale łączenia zbadanych modeli prognostycznych w metody zespołowe.

# Wnioski c.d.



- Dla modelu MLR najlepsze wyniki uzyskano dla modeli bez członu stałego.
- Dla modelu MLP najlepsze wyniki osiągnięto dla 16 neuronów w warstwie ukrytej z funkcją aktywacji tanh, liniową funkcją aktywacji warstwy wyjściowej oraz dla liczby epok równej 38.
- Dla modelu GBDT najlepsze wyniki osiągnięto dla liczby drzew, maksymalnej głębokości drzew oraz współczynnika uczenia równych kolejno 85, 3 oraz 0,1.
- Dla modelu RF najlepsze wyniki osiągnięto dla 300 drzew, minimalnej liczbie próbek w węźle umożliwiającej podział = 100, maksymalnej głębokości drzew =10 oraz maksymalnej liczbie węzłów równej 100.

# Podziękowania



Badania były finansowane ze środków NCBiR przeznaczonych na implementację międzynarodowego projektu badawczego “Digital Energy Path for Planning and Operation of sustainable grid, products and society” (akronim: DIEGO). Projekt Diego był fundowany przez ERA-Net Smart Energy Systems on Digital Transformation for Green Energy Transition (EnerDigit).

This paper is financed from the funds of the National Center for Research and Development for the implementation of the international research project entitled “Digital Energy Path for Planning and Operation of sustainable grid, products and society” (acronym: DIEGO). DIEGO project is funded through the ERA-Net Smart Energy Systems on Digital Transformation for Green Energy Transition (EnerDigit).

Wydział Elektryczny
POLITECHNIKA WARSZAWSKA

*Dziękuję za uwagę*

Politechnika Warszawska

DIEGO
DIGITAL SOLUTIONS