Politechnika Warszawska

WYDZIAŁ ELEKTRYCZNY

Instytut Sterowania i Elektroniki Przemysłowej

Praca dyplomowa inżynierska

na kierunku Informatyka Stosowana

w specjalności Inżynieria Danych i Multimedia

Prognozowanie bilansu energetycznego prosumentów na przykładzie Estonii

Daniel Ślusarczyk

numer albumu -

promotor

dr inż. Grzegorz Sarwas

WARSZAWA 2024

Prognozowanie bilansu energetycznego prosumentów na przykładzie Estonii

Streszczenie

Rynek energii elektrycznej ulega nieustannym transformacjom. Jedną z przyczyn takiej sytuacji jest aktualna tendencja do zwiększenia wykorzystania odnawialnych źródeł energii. W rezultacie uczestnikiem sieci elektroenergetycznej przybierającym na znaczeniu jest prosument, który za sprawą zainstalowanej mikroinstalacji jest w stanie nie tylko konsumować energię elektryczną, ale również aktywnie ją produkować. Konsekwencją rozpowszechnienia produkcji energii przez instalacje małej mocy jest powiększenie problemu nierównowagi energetycznej, czyli sytuacji, w której wyprodukowana energia nie pokrywa się z rzeczywistym zapotrzebowaniem. Charakteryzujące się dużą złożonością zużycie energii podmiotów posiadających własne instalacje powoduje problemy logistyczne i finansowe dla przedsiębiorstw energetycznych, a w konsekwencji zwiększenie kosztów operacyjnych i nieefektywne wykorzystanie produkowanej energii. Rozwiązanie tego problemu skutkowałoby zwiększeniem niezawodności sieci, a także poprawieniem integracji mikroinstalacji z centralnymi źródłami energii zarządzanymi przez operatora. Efektywnie działający system stanowiłby dodatkowy atut przemawiający za dalszym rozwojem energii odnawialnej produkowanej na poziomie prywatnej działalności.

Niniejsza praca podejmuje próbę przyczynienia się do rozwiązania problemu niezbilansowania, wykorzystując uczenie maszynowe do skomponowania modelu zdolnego do przewidywania zachowania prosumentów. Celem stworzonej implementacji jest krótkoterminowa predykcja konsumpcji i produkcji energii elektrycznej danej grupy prosumentów na terenie Estonii. Pracę rozpoczyna opisanie, ważnych z punktu widzenia poruszanej tematyki, podstaw teoretycznych i analiza eksploracyjna dostępnych informacji. Eksperymentalna część pracy obejmuje natomiast przedstawienie ośmiu przygotowanych modeli predykcyjnych zaimplementowanych z użyciem lasu losowego, systemu XGBoost i regresji liniowej. Zakończenie badań stanowi selekcja najefektywniejszego rozwiązania pod względem przyjętych kryteriów i zawarcie ostatecznej konkluzji w formie podsumowania. Opracowane rozwiązanie może stanowić efektywne narzędzie z perspektywy operatora sieci i stanowić fundament dalszej eksploracji poruszanej problematyki.

Słowa kluczowe: prosument, uczenie maszynowe, szeregi czasowe, rynek energii elektrycznej

Forecasting the energy balance of prosumers on the example of Estonia

Abstract

The electricity market is constantly transforming. One of the reasons for this situation is the tendency to increase the use of renewable energy sources. As a result, the prosumer as a participant of the power grid is becoming continuously even more important. Prosumers are capable of not only consuming electrical energy but also actively producing it using their own micro-installations. The consequence of the widespread production of energy by low-power devices is an increase in the problem of the energy imbalance, i.e. a situation where the produced energy does not align with the actual demand. The complex and unpredictable energy consumption of individuals with their own installations poses logistical and financial challenges for energy companies, resulting in increased operational costs and inefficient use of produced energy. Solving this problem would allow to increase network reliability as well as improve integration of micro-installations with central energy sources managed by the operator. An effectively functioning system would be an additional advantage supporting further development of renewable energy produced at the level of private activity.

This study attempts to contribute to solving the imbalance problem by using machine learning to create a model capable of predicting prosumer behavior. The purpose of the created implementation is short-term prediction of energy consumption and production of a given group of prosumers in Estonia. The paper begins with a description of the theoretical basis that are important from the point of view of the discussed subject and an exploratory analysis of the available information. The experimental part of the study includes the introduction of eight prepared prediction models implemented using random forest, the XGBoost system and linear regression. The research ends with the selection of the most effective solution in terms of the adopted criteria and the final conclusion. The developed solution can be an effective tool from the perspective of the electrical grid operator and the basis for further research of the subject.

Keywords: prosumer, machine learning, time series, electricity market

Spis treści

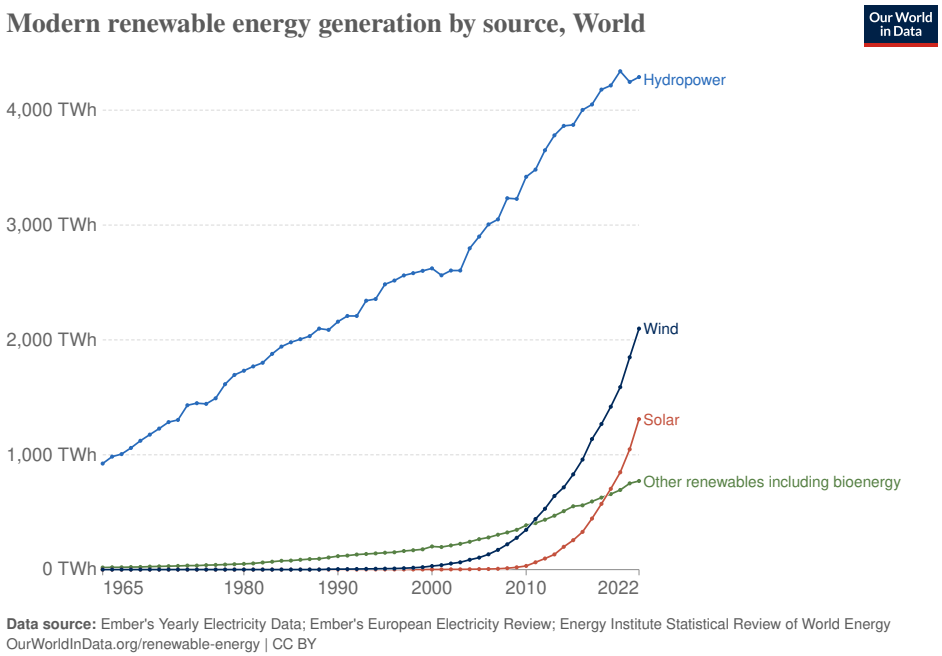
Rozdział 1

Wstęp

Energia elektryczna stała się nieodzownym elementem społeczeństwa i trudno wyobrazić sobie funkcjonowanie w aktualnej rzeczywistości bez dostępu do tak podstawowego zasobu. Dynamiczny rozwój technologii, przyrost liczby ludności i elektryfikacja państw słabiej rozwiniętych to tylko przykłady powodów nieustannie rosnącego zapotrzebowania na energię elektryczną [14]. Konsekwencją takiego działania jest to, że rynek energetyczny na całym świecie ulega ciągłym transformacjom. Znaczenie energii elektrycznej w perspektywie kolejnych lat będzie prawdopodobnie wyłącznie rosnąć, a zapewnienie ludziom dostępu do stabilnego źródła energii będzie stwarzać coraz większe wyzwanie. Jak informuje Międzynarodowa Agencja Energetyczna (ang. *International Energy Agency*, IEA) w raporcie „Global Energy & CO2 Status Report – The latest trends in energy and emissions in 2018" [19], globalne zużycie energii w 2018 r. wzrosło niemal dwukrotnie w porównaniu ze średnią stopą wzrostu od 2010 r. Głównymi przyczynami takiego stanu rzeczy są rozwój gospodarki światowej i liczniejsze potrzeby w zakresie regulacji temperatury pomieszczeń w coraz większej części świata. Wobec przedstawionych informacji, niezaprzeczalna staje się konieczność ciągłego poszukiwania rozwiązań produkcyjnych zdolnych do spełniania wyzwań stawianych przez zapotrzebowanie.

Nieodzownym aspektem poruszanym w kontekście produkcji energii jest również jej wpływ na kryzys związany z globalnym ociepleniem. Energetyka konwencjonalna, które swoje działanie opiera o spalanie nieodnawialnych paliw, wyraźnie przyczynia się do światowej emisji CO2 i dalszych zmian klimatu. Ponadto, ten typ energetyki może być w przyszłości niewystarczający do spełniania potrzeb ludzi ze względu na wyczerpanie zasobów paliw odnawialnych [15]. Warto również zwrócić uwagę na niebagatelny wpływ ingerencji tego typu źródeł w środowisko naturalne, ponieważ wydobycie i spalanie paliw wiąże się z dużymi szkodami w przyrodzie w postaci zanieczyszczonego powietrza, zniszczeń środowiska naturalnego czy zatruwaniem gleby i wody [3]. W świetle przedstawionych argumentów nieuchronne staje się poszukiwanie alternatywnych, odnawialnych źródeł energii, a także rozwój energetyki niekonwencjonalnej.

Kluczowy potencjał jest aktualnie dostrzegany w sektorze energetyki opartym o odnawialne źródła energii (OZE), których wpływ na środowisko i zmiany klimatu jest wyraźnie mniej zauważalny. Zgodnie z przytoczonym na wstępie raportem [19] odnawialne źródła energii i sama energia jądrowa pokryły większość popytu, ale nadal produkcja elektrowni węglowych i gazowych znacznie wzrosła, czego skutkiem było powiększenie emisji CO2 sektora energetycznego w 2018 r. o 2,5%. We wszystkich regionach świata zauważa się tendencję do rozwoju zdolności wytwarzania energii opartej o źródła odnawialne. Potwierdza to raport „Renewable Energy" przygotowany w 2020 r. [28] poruszający temat wzrostu znaczenia energii odnawialnej na przestrzeni lat. Zgodnie z przedstawionymi informacjami ostatnie lata przyniosły dynamiczny rozwój generacji energii w oparciu o wiatr i słońce, co uwidacznia poniższy wykres 1. W oparciu o zaprezentowane na rysunku dane, jasny staje się fakt wzrostu znaczenia OZE w nowoczesnej energetyce, ze szczególnym naciskiem na źródła oparte na słońcu i wietrze, które wykładniczo przybierają na znaczeniu.



Rysunek 1. Podział energii odnawialnej, źródło: [28]

Instalacje fotowoltaiczne i wiatrowe, które odpowiadają za znaczną część produkcji z OZE, są w stanie w dużym stopniu zaspokoić zapotrzebowanie indywidualnego, przeciętnego gospodarstwa domowego, a nawet przedsiębiorstwa. Ponadto charakteryzują się praktycznie dowolną skalowalnością instalacji, co pozwala na dostosowanie inwestycji do własnych potrzeb. Doprowadziło to do wzrostu zainteresowania indywidualnych użytkowników systemu elektroenergetycznego (SEE) własną produkcją energii i powstaniem tzw. energetyki rozproszonej (ER), która polega na wytwarzaniu energii elektrycznej bądź ciepła przez małe jednostki lub obiekty produkcyjne dla użytku lokalnego. W wyniku opisanych transformacji nowym podmiotem, który zyskuje na znaczeniu nieprzerwanie od

wielu lat jest tzw. prosument, który łączy w sobie cechy odbiorcy i wytwórcy energii elektrycznej [32]. Takie rozwiązanie pozwala na wzrost niezależności indywidualnych użytkowników od scentralizowanych źródeł energii, ale z perspektywy SEE taki stan powoduje konieczność przewidywania zachowania prosumentów w celu oszacowania zapotrzebowania względem wszystkich użytkowników. Model zdolny do przewidywania decyzji podejmowanych przez prosumentów odnośnie dysponowania wytwarzaną energią elektryczną byłby narzędziem dla operatorów SEE, które pozwalałoby na efektywniejsze zarządzanie produkcją i dystrybucją energii potrzebnej na pokrycie całego zapotrzebowania, uwzględniając udział prosumentów. Jest to fundamentalne zagadnienie z perspektywy funkcjonowania całego systemu i wszystkich użytkowników, ponieważ wysokie zaburzenia równowagi energetycznej mogą zwiększyć całkowity koszt systemu, zarówno na rynku partnerskim, jak i poza nim, ponieważ operator systemu musi w krótkim czasie podejmować kosztowne działania naprawcze spowodowane zmiennym udziałem prosumentów [7]. Wykorzystanie danych operatorów sieci na temat bilansu energetycznego odbiorców klasyfikowanych jako prosumenci i uczenia maszynowego daje teoretyczne możliwości stworzenia modelu zdolnego zwiększyć wiedzę operatora w zakresie przewidywania zachowania tego rodzaju uczestników sieci.

Celem niniejszej pracy jest opracowanie modelu uczenia maszynowego zdolnego do krótkoterminowego przewidywania zachowania prosumentów (rozumianego jako ilości produkowanej i konsumowanej energii) w perspektywie następnego dnia. Przygotowane rozwiązanie zakłada dostępność następujących informacji dla użytkownika w momencie dokonywania predykcji:

- prognoza pogody,
- historyczne dane pogodowe,
- koszt energii elektrycznej,
- koszt gazu ziemnego,
- cechy danej grupy prosumentów.

Na podstawie dostarczonych informacji (w ograniczonym zakresie czasu), zadaniem modelu będzie dokonanie predykcji przyszłego zachowania prosumentów dla dowolnej kombinacji trzech cech: prowincji Estonii, wartości logicznej określającej, czy predykcja dotyczy działalności gospodarczej (biznesu) i rodzaju umowy obejmującej prosumentów.

Praca została podzielona na pięć fragmentów. Całość rozważań poprzedzono zwięzłym opisem tematu w formie wstępu zawartego w części 1. W rozdziale 2 zostały wprowadzone najważniejsze pojęcia konieczne do prawidłowego zrozumienia problematyki i rozszerzania wiedzy przedstawionej we wstępie. Kolejna część 3 została poświęcona dokładnemu przedstawieniu analizowanych danych i najważniejszych zależności w dostępnych informacjach. Rozdział 4 skupia się na omówieniu metodyki przeprowadzonych eksperymentów i selekcji optymalnego rozwiązania. Ostateczne podsumowanie otrzymanych rezultatów zostało opisane w ostatnim fragmencie pracy 5.

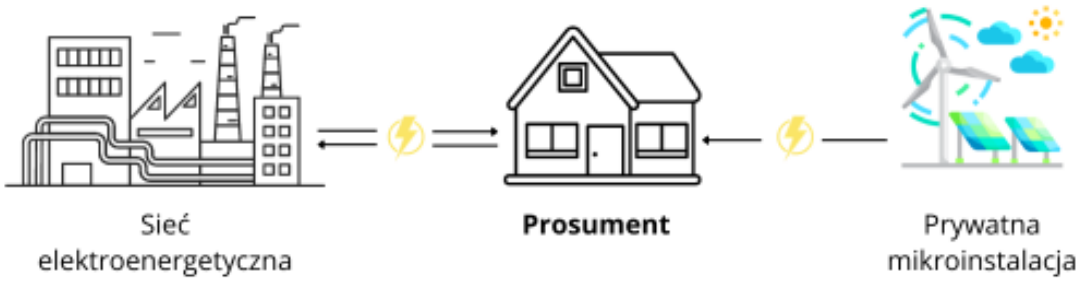
Rozdział 2

Teoria

W niniejszym rozdziale zostały przedstawione istotne koncepcje stojące za prawidłowym postrzeganiem rynku energii i uczenia maszynowego. Pierwsze poruszone zagadnienia dotyczą definiowania i funkcjonowania działalności prosumenckiej, podczas gdy dalsza część rozdziału skupia się na opisaniu teorii szeregów czasowych i dokonywania predykcji. Wprowadzone pojęcia mają kluczowe znaczenie dla dalszych rozważań i przybliżenia istoty problemu.

2.1 Prosument

Etymologia słowa „prosument" odnosi się do połączenia słów „producent" oraz „konsument" [35], co znajduje swoje odzwierciedlenie w kontekście wytwarzania energii odnawialnej. Prosument jest definiowany jako odbiorca końcowy, którego udział w SEE nie ogranicza się wyłącznie do poboru energii, ale również do jej wytwarzania. Dokonuje on zakupu energii elektrycznej w podobny sposób do reszty uczestników, ale w odróżnieniu od nich, jednocześnie wytwarza ją za pomocą własnej mikroinstalacji opartej o odnawialne źródła energii. Prosumenci mogą dokonywać decyzji o sposobie zarządzania wytworzoną samodzielnie energią w sposób całkowicie niezależny zarówno od innych uczestników, jak i głównych dostawców. Decyzja o sprzedaży energii może być podyktowana nadmierną produkcją względem własnych potrzeb, ale również np. względami finansowymi bazującymi na aktualnej sytuacji rynkowej. Dodatkowo wiele instalacji jest wyposażona w możliwość magazynowania energii, co zwiększa elastyczność podejmowanych decyzji. Poniższy schemat 2 przedstawia w sposób poglądowy udziału tego typu podmiotu w działalność całego systemu.

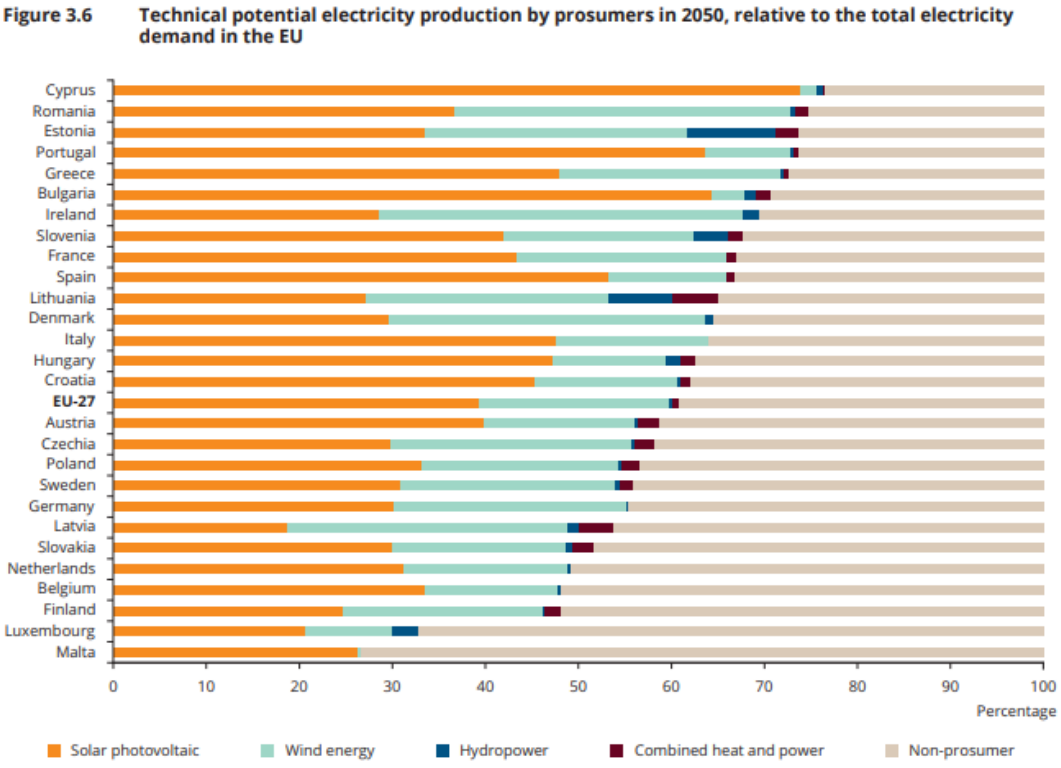


Rysunek 2. Poglądowy schemat działania prosumentów, źródło: *opracowanie własne*

Precyzyjna definicja prosumentów pozostaje po stronie konkretnych państw [11]. Większość państw członkowskich Unii Europejskiej (UE) opiera się w tej kwestii na wielkości lub wydajności instalacji, kierując się jej relatywnie małymi osiągami. W Polsce mikroinstalacja została ograniczona do łącznej zainstalowanej mocy nie większej niż 40 kW i powinna być przyłączona do sieci o napięciu znamionowym mniejszym niż 110 kV lub mocy wytwarzania ciepła w skojarzeniu większego niż 120 kW, ale nie większego niż 600 kW. W Estonii definicja mikroproducenta odnosi się do instalacji jednofazowej o maksymalnej mocy nominalnej 3,68 kW lub instalacji trójfazowej o maksymalnej mocy nominalnej 11 kW. Przyjęte regulacje w poszczególnych krajach zależą często od stopnia rozwinięcia tego typu działalności w danym kraju, technicznych aspektów sieci i wpływu na rynek energetyczny.

Ważnym aspektem jest również potencjał i perspektywy rozwoju działalności prosumenckiej. Aktualnie zainteresowanie aktywnym udziałem w produkcję energii gwałtownie rośnie z całej UE. Jak podaje raport „Energy prosumers in Europe – Citizen participation in the energy transitions" przygotowany przez European Environment Agency (EEA) [10] liczba prosumentów korzystających z paneli fotowoltaicznych w Niderlandach wzrosła z niecałych 500.000 w 2015 r. do ponad 1 milion w 2020 r., analogiczna sytuacja miała miejsce w Portugalii, gdzie liczba instalacji wzrosła z 3.000 do 30.000 w 2019 roku. W Polsce liczba prosumentów zwiększyła się z 510.000 w 2018 r. do 847.000 w 2021 r., a zainstalowana moc wyniosła prawie 6 GW. Biorąc pod uwagę energię słoneczną na mieszkańca, Estonia wyłoniła się jako jeden z nowych liderów, zwiększając 405 W w 2021 na mieszkańca do 596 W na mieszkańca w 2022 r, pozycjonując się na 6 miejscu wśród 27 członków UE [29]. Dodatkowo raport EEA podkreśla ogromny potencjał rozwoju prosumpcji w całej UE. Podaje, że z technicznego punktu widzenia prawie jedna czwarta całego zapotrzebowania mogłaby zostać pokryta z samych dachowych systemów fotowoltaicznych w oparciu o istniejące już zasoby budowlane. W wielu miejscach wiązałoby się to z koniecznością lokalnej bądź regionalnej rozbudowy systemu elektroenergetycznego, ale nie podważa to potencjału dalszego zwiększania się liczby prosumentów. Model opracowany w raporcie [24] szacuje, że w roku 2050 aż 89% zapotrzebowania gospodarstw na energię elektryczną może być pokryte przez nie same. Zauważa, że produkcją energii elektrycznej z paneli fotowoltaicznych ma najwyższy potencjał wzrostu w Europie Południowej, gdzie ponad 70% tego potencjału wynika z możliwości założenia instalacji na powierzchni dachowej, a reszta to instalacje naziemne. Jednocześnie turbiny wiatrowe będące własnością dużych kolektywów

prosumenckich mogą mieć duże znaczenie w krajach, gdzie występuje wystarczająca ilość dostępnej przestrzeni wokół miast. Podobne wnioski zostały przedstawione w raporcie EEA, z którego wynika, że prosumenci mogą dostarczyć 30-70% całkowitej energii elektrycznej, w zależności od państwa członkowskiego UE zgodnie z wykresem 3.



Rysunek 3. Wykres potencjału prosumentów, źródło: [10]

Reasumując przedstawione fakty, można założyć, że działalność prosumencka niesie za sobą szereg korzyści. Najważniejszą zaletą jest zwiększenie niezależności energetycznej od dostawców działających na szczeblu krajowym. W sytuacji awarii systemu dystrybucyjnego lub ograniczenia działania dostawców, własna instalacja daje możliwość zaspokojenia chociażby w części zapotrzebowanie i prawidłowe działanie większości urządzeń w ograniczonym zakresie. Niebagatelne znaczenie ma również zmniejszenie wrażliwości na wahania cen. Jest to ważna zaleta zarówno w przypadku prowadzenia działalności gospodarczej, jak i w kontekście osób prywatnych, ponieważ umożliwia bardziej przewidywalne zarządzanie budżetem związanym z zakupem energii. Kolejnym argumentem przytaczanym w kontekście działalności prosumenckiej jest pozytywny wpływ na środowisko naturalne. Mikroinstalacje oparte na OZE stanowią bezinwazyjną metodę pozyskiwania energii, co jest postrzegane jako potencjalny sposób przeciwdziałania zmianom klimatycznym.

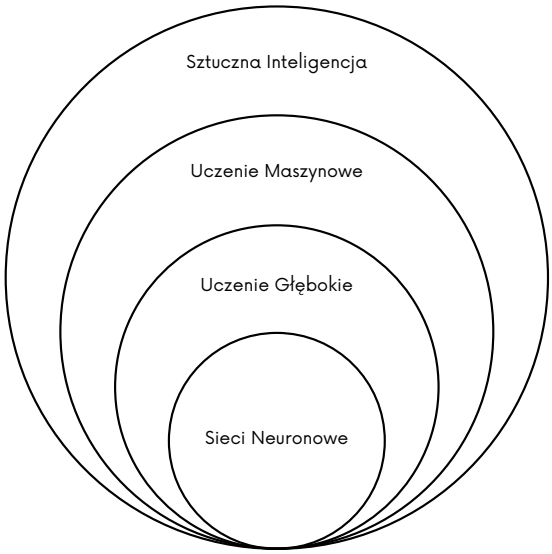
2.2 Sieć Elektroenergetyczna

Terminem systemu elektroenergetycznego (SEE) nazywany jest zespół urządzeń przeznaczonych do wytwarzania, przesyłu i rozdziału energii elektrycznej. SEE odpowiada za proces ciągłej dostawy energii elektrycznej odbiorcom z zachowaniem minimalnych nakładów przeznaczonych na realizację tego celu [34]. Konwencjonalny model energetyczny zakładał istnienie jednokierunkowych systemów, których działanie polegało na wytwarzaniu energii w centralnych elektrowniach, a następnie przesyłanie jej sieciami przesyłowymi i dystrybuowanie do odbiorców końcowych. Umożliwienie prowadzenia działalności prosumenckiej wymagało wprowadzenie szeregu zmian w sposobie działania całego systemu dystrybucji i zarządzania energią elektryczną. Produkcja energii przez prosumentów odbywa się najczęściej bez udziału pośrednika, umożliwiając użytkownikowi sieci swobodne zarządzanie własną instalacją. Taki model działania sieci nazywany jest w literaturze „peer-to-peer" (P2P) i oznacza sytuację, w której producenci energii mogą bezpośrednio sprzedawać lub dzielić się wytworzoną energią z innymi uczestnikami systemu, pomijając tradycyjne struktury dostawców energii. Obrót energią elektryczną w takim trybie zaspokaja dużą część potrzeb zarówno prosumentów, jak i konsumentów oraz zmniejsza wielkość strat na łączach spowodowanych dystrybucją na duże dystanse [16]. Niemniej jednak problem niestabilności zapotrzebowania i podaży energii może wpływać na niezawodność rynku P2P. Prawidłowe prognozowanie bilansu energetycznego prosumentów pozwala na wydajniejsze planowanie i zarządzanie siecią oraz lepszą alokację zasobów. Duże nasycenie danego obszaru mikroinstalacjami niesie za sobą również szereg innych problemów wynikających ze sposobu ich funkcjonowania takich jak: zwrotny przepływ mocy, niestabilność napięcia spowodowana lokalnymi oscylacjami lub mała bezwładność systemu [27]. Możliwość precyzyjnego przewidywania zachowania prosumentów może mieć pozytywny wpływ na rozwiązanie wymienionych trudności działania sieci.

2.3 Uczenie maszynowe

Uczenie maszynowe (ang. *Machine Learning*, UM) jest terminem szczególnie rozpowszechnionym w ostatnim czasie za sprawą prężnego rozwoju sztucznej inteligencji (ang. *Artificial Intelligence*, SI) i osiągnięciu przełomowych rozwiązań w tej dziedzinie nauki, które na zawsze zmieniły postrzeganie tego obszaru wiedzy. Samo pojęcie „sztuczna inteligencja" jest często mylnie utożsamiane jako synonim uczenia maszynowego, a fakt istnienia wielu definicji obu terminów dodatkowo komplikuje jednoznaczną klasyfikację. Jeden z prekursorów tej dziedziny John McCarthy w swoim artykule [22] zdefiniował SI jako naukę polegającą na tworzeniu inteligentnych programów komputerowych. Podsumowując te słowa można stwierdzić, że SI odpowiada za automatyczne wykonywanie zadań, które w jakimś stopniu wymagałyby ludzkiej inteligencji. Jedynymi z metod służącymi do osiągnięcia tego celu jest uczenie maszynowe. W przedstawionej tematyce często pojawiają się również terminy uczenia głębokiego (ang. *deep learning*) i sieci neuronowych (ang. *neural networks*). Oba pojęcia odnoszą się do szczególnego podejścia stosowanego w UM, przy czym sieci neuronowe są jedynie

przykładem uczenia głębokiego. Poniższy schemat 4 obrazuje wzajemne zależności pomiędzy omawianymi zagadnieniami.



Rysunek 4. Schemat wzajemnych zależności między terminami, źródło: *opracowanie własne*

Pojęcie uczenia maszynowego wywodzi się zatem z dziedziny sztucznej inteligencji i dzieli wiele wspólnych cech z dziedziną matematyki – statystyką. Odnosi się do algorytmów i systemów zdolnych do rozszerzania swojej wiedzy i umiejętności poprzez doświadczenie, rozumiane jako samodzielne znajdowanie zależności w analizowanych danych [13]. Precyzyjniejsza definicja została zaproponowana przez Toma Mitchella w 1997 roku [23], który opisał UM jako program komputerowy uczący się na podstawie doświadczenia E w odniesieniu do jakiegoś zadania T i pewnej miary wydajności P, jeśli jego wydajność (mierzona przez P) wobec zadania T wzrasta wraz z nabywaniem doświadczenia E. Sztandarowym zastosowaniem UM jest filtr spamu, czyli program komputerowy wykorzystywany do wykrywania niechcianych wiadomości w skrzynce elektronicznej. Dane, które byłyby potrzebne do wytrenowania odpowiedniego modelu, noszą nazwę zbioru (zestawu) uczącego (ang. *training set*), a każdy pojedynczy przykład w takim zbiorze jest próbką uczącą (przykładem uczącym). W analizowanym przykładzie zbiorem uczącym byłby zestaw historycznych wiadomości (wraz ze wszystkim informacjami o tych wiadomościach), a poszczególne wiadomości byłyby próbkami uczącymi. Zadaniem T w takim przypadku byłoby prawidłowe oznaczenie niechcianych wiadomości, a doświadczeniem E zbiór uczący. Ważnym aspektem pozostaje również sposób wyznaczania miary wydajności P, która wymaga specyficznego podejścia dla każdego problemu. Wydajność filtra spamu może być mierzona przykładowo za pomocą metryki zwaną dokładnością (ang. accuracy), czyli stosunkiem prawidłowo zaklasyfikowanych, niechcianych wiadomości do wiadomości nieprawidłowo zaklasyfikowanych. Jakość wyników otrzymywanych w takim przykładzie będzie zależeć od jakości i stopnia rozbudowania danych stosowanych do nauki, prawidłowego zdefiniowania pomiaru wydajności P, ale również zastosowania odpowiedniego modelu.

Systemy uczenia maszynowego są klasyfikowane ze względu na stopień i rodzaj nadzorowania procesu uczenia. Pod tym względem występują cztery najważniejsze rodzaje uczenia:

- nadzorowane,
- nienadzorowane,
- pół-nadzorowane,
- przez wzmacnianie.

W kontekście analizowanej pracy kluczowe jest uczenie nadzorowane, które polega na dostarczeniu algorytmowi danych uczących zawierających dodatkową etykietę z rozwiązaniem problemu. Model po przeanalizowaniu takich danych i przeprowadzeniu procesu uczenia powinien być w stanie samodzielnie znajdować odpowiedź na nowe dane, które nie pojawiły się wcześniej w zbiorze uczącym.

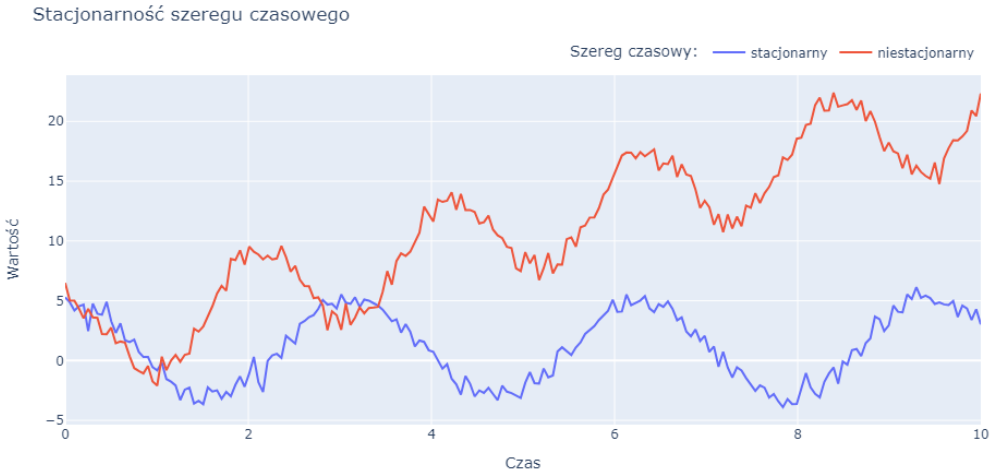
2.3.1 Przewidywanie szeregów czasowych

Jednym z praktycznych zastosowań uczenia maszynowego jest przewidywanie szeregów czasowych. Szereg czasowy jest realizacją pewnego procesu stochastycznego, w której dziedziną jest czas, czyli jest to sekwencja obserwacji zarejestrowanych w pewnych interwałach czasowych [20]. W zależności od przyjętego kroku czasowego wyróżnia się różne kategorie szeregów, niemniej najczęściej spotykanymi szeregami czasowymi są: godzinowe, dzienne, miesięczne, kwartalne i roczne. W problemie przewidywania szeregów czasowych zadaniem jest predykcja wartości w określonych momentach. Przykładowo może być to zadanie przewidywania produkcji paneli fotowoltaicznych z dziennym wyprzedzeniem, dysponując prognozą pogody na przyszły dzień. Dyscyplina uczenia maszynowego i statystyki wypracowały różne modele zdolne do rozwiązywania opisanego problemu, ale skuteczność tych algorytmów zależy od prawidłowego rozumienia ich przeznaczenia i znajomości specyfiki szeregów czasowych.

Analiza danych, w których zmienna objaśniana zależy od czasu, w znacznym stopniu polega na prawidłowym wyznaczeniu czterech najważniejszych komponentów charakteryzujących szeregi czasowe:

- **Wahania sezonowe** (sezonowość) – wzorzec danych, który można zaobserwować w regularnych odstępach czasu. Przykładem może być sezonowość temperatury, która zawsze w miesiącach zimowych przyjmuje mniejsze wartości niż latem.
- **Tendencja rozwojowa** (trend) – długoterminowe, utrzymujące się ukierunkowanie danych. Może być rosnące, malejące, ale również liniowe i nieliniowe. Przykładem może być średnia, roczna temperatura na świecie, która ma tendencję wzrostową z powodu globalnego ocieplenia.
- **Wahania cykliczne** (cykliczność) – komponent występujący w momencie, gdy w danych można zaobserwować wzrosty i spadki o nieregularnych częstotliwościach. Brak stałych interwałów czasowych tych zmian jest elementem odróżniającym cykliczność od sezonowości.
- **Wahania losowe** – szereg czasowy może być naznaczony pewną losowością, której nie da się objąć żadnym modelem.

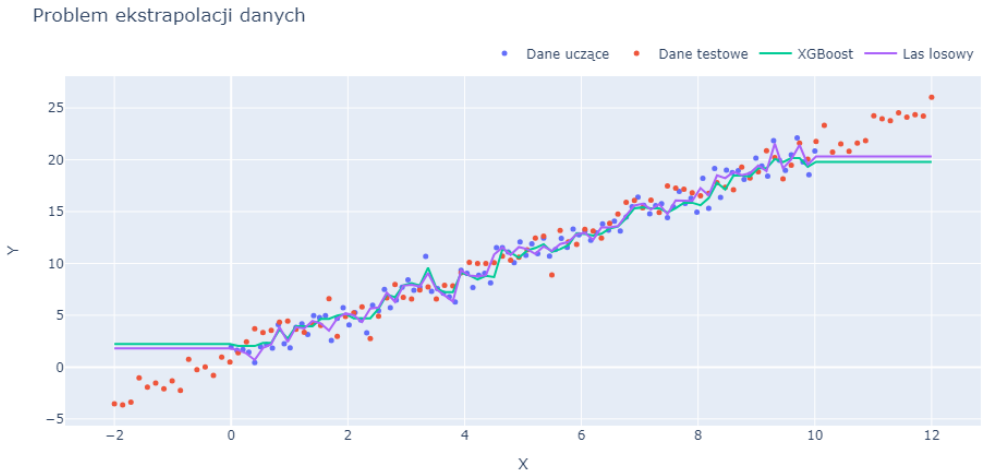
Istotnym pojęciem jest również stacjonarność badanego procesu. W ścisłym sensie stacjonarność procesu stochastycznego jest definiowana jako istnienie takie samego rozkładu prawdopodobieństwa dla dowolnych obserwacji i ich przesunięcia w czasie. W podejściu praktycznym zdefiniowana tzw. silna stacjonarność jest zjawiskiem praktycznie niewystępującym dla procesów rzeczywistych. Z tego powodu empirycznie testowana jest tzw. słaba stacjonarność, która występuje, gdy wartość oczekiwana oraz wariancja procesu są stałe w czasie. Najczęstszą przyczyną zaburzenia stacjonarności jest występowanie trendu. W takiej sytuacji istnieje często możliwość zastosowania przekształceń szeregu czasowego doprowadzając go do modelu stacjonarnego poprzez usunięcie składowej trendu. Poniższy rysunek 5 prezentuje dwa rodzaje procesów:



Rysunek 5. Porównanie stacjonarności procesów, źródło: *opracowanie własne*.

Określenie stacjonarności modelu jest kluczowe dla prawidłowego opracowania modelu predykcji. Wiele modeli powszechnie stosowanych w UM jest dostosowanych do danych, których rozkład prawdopodobieństwa dla dowolnych obserwacji i wybranego zakresu czasowego jest możliwie stały. Szczególnie podatne na problem niestacjonarności danych są modele oparte na strukturze drzew decyzyjnych ze względu na brak możliwości ekstrapolacji danych [5, 37]. Oznacza to, że takie modele nie potrafią dokonywać prawidłowej predykcji w punktach spoza dziedziny zbioru danych uczących. Problem ten wynika z budowy drzew decyzyjnych, których struktura jest tworzona wyłącznie na bazie zakresu danych uczących, więc wartość minimalna, bądź maksymalna, którą może zwrócić dane drzewo jako predykcje jest równa zakresowi zbioru. Opisany problem jest szczególnie istotny w przypadku predykcji szeregów czasowych, w których zauważalne jest występowanie pewnego trendu. Taka sytuacja może skutkować tym, że skuteczność modelu drastycznie spadnie, gdy przyjmowane wartości rzeczywiste zaczną wykraczać poza zakres danych, na których został wyszkolony model. Poniższy wykres 6 przedstawia poglądowy przypadek problemu ekstrapolacji danych na przykładzie modelu lasu losowego i systemu XGBoost. Oba modele zostały nauczone na zbiorze danych,

którego zakres mieścił się w przedziale $[0, 10]$, więc próba dokonania predykcji dla danych z zakresu $[-2, 12]$ prowadzi do sytuacji, w której oba modele nie są w stanie dokonywać poprawnych predykcji w przedziałach $[-2, 0]$ oraz $[10, 12]$.



Rysunek 6. Problem ekstrapolacji danych, źródło: *opracowanie własne*

2.3.2 Modele uczenia maszynowego

Uczenie maszynowe przybiera na znaczeniu nieprzerwanie od lat. Popularność dziedziny SI spowodowała opracowanie wielu nowych modeli przeznaczonych do różnorodnych zastosowań, a potencjał rozwoju tej dyscypliny nieustannie rośnie [6, 21]. Współczesna literatura wyróżnia wiele rozwiązań do przewidywania szeregów czasowych, zarówno z dziedziny uczenia maszynowego, jak i statystyki, których wyniki znacznie się różnią w zależności od analizowanego problemu i sposobu podejścia [8, 26]. W niniejszej pracy szczególna uwaga zostanie poświęcona algorytmom opartym o drzewa decyzyjne takim jak: XGBoost i las losowy, a w ramach porównania wydajność tych podejść zostanie zestawiona z wynikami regresji liniowej.

Regresja liniowa

Model regresji liniowej (ang. *linear regression*) jest jednym z podstawowych podejść stosowanym w UM [17]. Modele liniowe zostały opracowane jeszcze przed rozpowszechnieniem stosowania komputerów w statystyce, ale nawet dzisiaj istnieje szereg powodów do ich stosowania. Charakteryzują się prostą budową i relatywnie łatwym do zinterpretowania pochodzeniem wyniku. Prosta budowa tego modelu nie stanowi jednak przeszkody do osiągania często zbliżonych wyników do bardziej skomplikowanych odpowiedników. W przypadku modelowania problemu nieliniowego zbadanie efektywności tego modelu może stanowić punkt wyjścia do zdobycia orientacji, w jakim stopniu zmienna opisywana jest wyrażona przez atrybuty i stanowić powód do zastosowania innych podejść.

Regresja liniowa polega na próbie wyrażenia zmiennej opisywanej w formie kombinacji liniowej atrybutów zbioru uczącego. Cel ten jest osiągany przed odpowiednie dobranie parametrów cech, a następnie dokonanie predykcji zgodnie ze wzorem 1.

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} X_j \beta_j, \tag{1}$$

gdzie:

$\hat{y}_i$ – predykcja modelu,
β_j – współczynnik przypisany do atrybutu X_j,
β_0 – wyraz wolny,
p – liczba atrybutów.

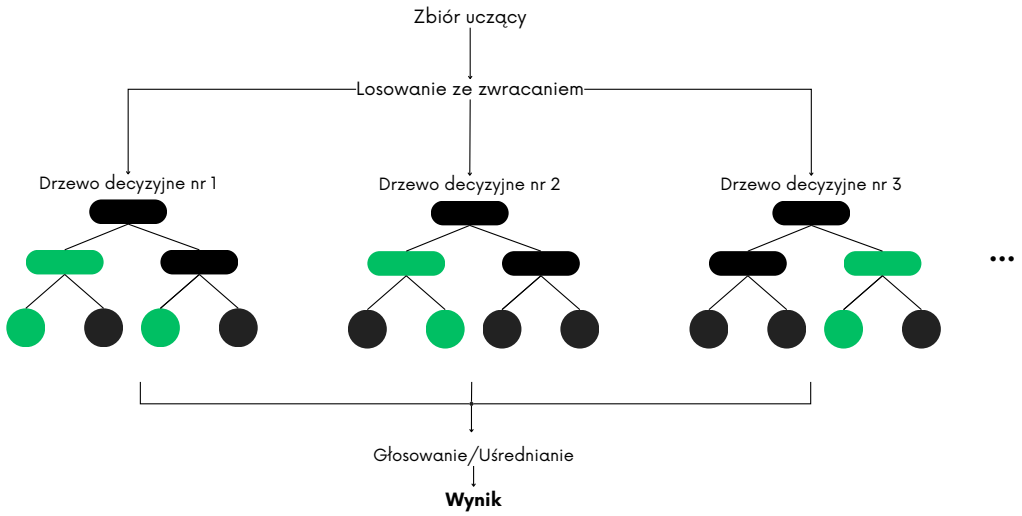
Istotnym aspektem modelu regresji liniowej jest fakt, że zmienna X_j może być bezpośrednią wartością ze zbioru uczącego, ale również dowolną transformacją tej wartości lub kombinacją kilku różnych cech. Zastosowanie nieliniowych przekształceń często znacząco rozszerza możliwości tego modelu.

Dobór wartości parametrów β wzoru 1 polega na minimalizowaniu przyjętej funkcji dopasowania. Najbardziej popularną metodą estymacji jest metoda „najmniejszych kwadratów", która polega na minimalizowaniu kwadratu różnicy pomiędzy wynikiem modelu i wartością rzeczywistą. W przypadku tak zdefiniowanego problemu dokładne wartości parametrów mogą zostać obliczone np. z użyciem rachunku macierzowego i pseudoodwrtoności Moore'a-Penrose'a, co stanowi znacząca przewagę tego modelu, ponieważ sposób obliczeń gwarantuje pewne znalezienie optymalnego rozwiązania.

Model lasu losowego

Model lasu losowego (ang. *random forest*) został przedstawiony pierwszy raz w 1995 roku przez Tin Kam Ho [18, 25] i od tego czasu znalazł zastosowanie w wielu praktycznych dziedzinach, w których używane jest uczenie maszynowe. Literatura podaje wiele przykładów praktycznej implementacji wykorzystującej omawiany model m.in. w bioinformatyce [9], wykrywaniu niebezpiecznego oprogramowania w systemie Android [1] bądź teledetekcji [4].

Algorytm lasu losowego jest istotną modyfikacją mechanizmu „baggingu", której ideą jest tworzenie dużej ilości nieskorelowanych, prostych modeli, w tym wypadku drzew decyzyjnych, a następnie uśrednienie wyniku [17]. Poniżej na schemacie 7 został przedstawiony przykładowy schemat działania modelu.



Rysunek 7. Poglądowy schemat lasu losowego, źródło: *opracowanie własne*

Las losowy opiera swoje działania na uśrednianiu wielu niedokładnych i nieobciążonych modeli, co w konsekwencji prowadzi do zmniejszania ostatecznej wariancji. Drzewa decyzyjne, będące podstawą działania modelu, mogą z powodzeniem być używane w tym procesie, ponieważ są zdolne do uchwycenia złożonych zależności pomiędzy danymi i rosnąc przy tym do znacznych głębokości, potrafią zachować relatywnie niskie obciążenie. Przeprowadzenie uśredniania na takich drzewach pozwala na uzyskanie wielu korzyści. Dodatkowo drzewa są tworzone na podstawie podzbioru danych tworzonych za pomocą tzw. metody bootstrapowej, która polega na losowaniu ze zwracaniem określonej liczby próbek ze zbioru uczącego. W konsekwencji zbudowane drzewa decyzyjne mają zbliżony rozkład danych wejściowych, więc oczekiwany wynik średniej jest taki sam, jak oczekiwany wynik pojedynczych drzew. Oznacza to, że obciążenie modelu jest takie samo jak obciążenie pojedynczych drzew, a takie podejście umożliwia jedynie redukcję wariancji modelu. Kolejnym ważnym aspektem działania drzewa losowego jest sposób doboru atrybutów. Korzystając z faktu, że średnia wariancja B atrybutów posiadających identycznych rozkład z pozytywną korelacją parami ρ wynosi zgodnie ze wzorem 2:

$$\rho\sigma^2 + \frac{1-\rho}{\mathrm{B}}\sigma^2, \tag{2}$$

można zauważyć, że wraz ze wzrostem liczby B drugi składnik sumy zanika, stąd wielkość korelacji par drzew w workach ogranicza korzyści z uśrednienia. Z tego powodu w procesie budowania poszczególnych drzew decyzyjnych wybierane są losowe zmienne w formie danych wejściowych. Przed każdym podziałem dokonywana jest selekcja m cech ze zbioru p zmiennych wejściowych jako kandydaci do podziału. Najczęściej za m przyjmuje się:

- $\lfloor\sqrt{\mathrm{p}}\rfloor$ dla klasyfikacji,
- $\lfloor\frac{\mathrm{p}}{3}\rfloor$ dla regresji.

W praktyce dobranie tej wartości powinno się odbyć eksperymentalnie, ponieważ najlepsza wartość może zależeć od modelowanego problemu. Działanie budowy lasu losowego i obliczenia ostatecznego wyniku odbywa się zgodnie z poniższym algorytmem 1.

Algorytm 1 Algorytm Lasu Losowego

Wymagane $\mathrm{n_{min}}$
Wymagane Zbiór uczący o p zmiennych

dla $\mathrm{b} = 1 : \mathrm{B}$ **wykonaj**
Wylosuj ze zwracaniem podzbiór próbek $\mathrm{Z}*$ rozmiaru N ze zbioru testowego
Rozpocznij budowę drzewa decyzyjnego $\mathrm{T_b}$ na podstawie $\mathrm{Z}*$

dla każdego węzła końcowego **wykonaj**
jeżeli nie osiągnięto minimalnej liczby węzłów $\mathrm{n_{min}}$ **to**
a) Wybierz losowo m zmiennych ze zbioru wszystkich p zmiennych
b) Wybierz zmienną ze zbioru m, która dokonuje najlepszego podziału
c) Dodaj do drzewa dwa węzły-dzieci względem przetwarzanego węzła
koniec jeżeli
koniec dla
koniec dla

Stwórz las losowy z drzew $\{\mathrm{T_b}\}_1^{\mathrm{B}}$

Dla stworzonego lasu losowego algorytmem 1 ostateczną predykcję dokonuje się dla nowej próbki x zgodnie z poniższymi wzorami 3–4:

dla regresji:

$$\hat{\mathrm{f}}_{\mathrm{rf}}^{\mathrm{B}}(\mathrm{x}) = \frac{1}{\mathrm{B}} \sum_{\mathrm{b}=1}^{\mathrm{B}} \mathrm{T_b}(\mathrm{x}), \tag{3}$$

dla klasyfikacji:

$$\hat{\mathrm{C}}_{\mathrm{rf}}^{\mathrm{B}}(\mathrm{x}) = \text{większość głosów} \{\hat{\mathrm{C}}_{\mathrm{b}}(\mathrm{x})\}_1^{\mathrm{B}}, \tag{4}$$

gdzie:

$\hat{\mathrm{C}}_{\mathrm{b}}(\mathrm{x})$ – wynik klasyfikacji b-tego drzewa.

System Extreme Gradient Boosting

Model Extreme Gradient Boosting (XGBoost) został zaprezentowany przez Chen Tianqi i Carlos Gestrina w 2016 roku i od tamtego czasu jego rozwój i ciągłe udoskonalanie było kontynuowane przez kolejne lata za sprawą wielu badań przeprowadzanych przez naukowców [2]. Podstawową ideą stojącą za działaniem modelu XGBoost jest mechanika „boostingu", której założeniem jest stworzenie kombinacji wielu niedokładnych modeli w jeden większy model charakteryzujący się większą dokładnością [12]. Stosując odpowiednie parametry wpływające na sposób uczenia i wzajemnego wpływu pomniejszych modeli XGBoost jest w stanie osiągnąć zadowalającą dokładność przewidywania. Jednak gdy analizowany zbiór danych jest złożony i składa się z wielu próbek i atrybutów może zaistnieć konieczność zastosowania tysięcy mniejszy modeli, których dopiero wspólne działanie zapewnia satysfakcjonujące wyniki [2]. Zgodnie z opracowaniem modelu [36] system XGBoost opiera swoje działanie na przeprowadzaniu procesu boostingu na drzewach decyzyjnych niewielkiej głębokości, z których następnie ostateczna predykcja jest wyliczana sumując wyniki poszczególnych drzew składowych.Dla zbioru danych o n próbkach i m atrybutach $\mathrm{D} = \{(\mathrm{x_i}, \mathrm{y_i})\}, |\mathrm{D}| = \mathrm{n}, \mathrm{x_i} \in \mathbb{R}^{\mathrm{m}}, \mathrm{y_i} \in \mathbb{R}$, używając K funkcji wynikowych pomniejszych modeli, ostateczna predykcja jest wyliczana zgodnie ze wzorami 5–6:

$$\hat{\mathrm{y}}_\mathrm{i} = \phi(\mathrm{x_i}) = \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{f_k}(\mathrm{x_i}), \mathrm{f_k} \in \digamma, \tag{5}$$

$$\digamma = \{\mathrm{f}(\mathrm{x}) = \omega_{\mathrm{q(x)}}\} \left(\mathrm{q} : \mathbb{R}^{\mathrm{m}} -> \mathrm{T}, \omega \in \mathbb{R}^{\mathrm{T}}\right), \tag{6}$$

gdzie:

$\digamma$ – przestrzeń drzew regresyjnych,

q – reprezentacja struktury każdego drzewa przypisującego próbce indeks liścia,

T – liczba liści w drzewie,

$\mathrm{f_k}$ – funkcja odpowiadająca strukturze drzewa q i wagach liści ω.

Proces uczenia modelu opiera się na minimalizowaniu funkcji celu wyrażonej wzorami 7–8:

$$\Gamma(\phi) = \sum_{\mathrm{i}} \mathrm{l}(\hat{\mathrm{y}}_\mathrm{i}, \mathrm{y_i}) + \sum_{\mathrm{k}} \Omega(\mathrm{f_k}), \tag{7}$$

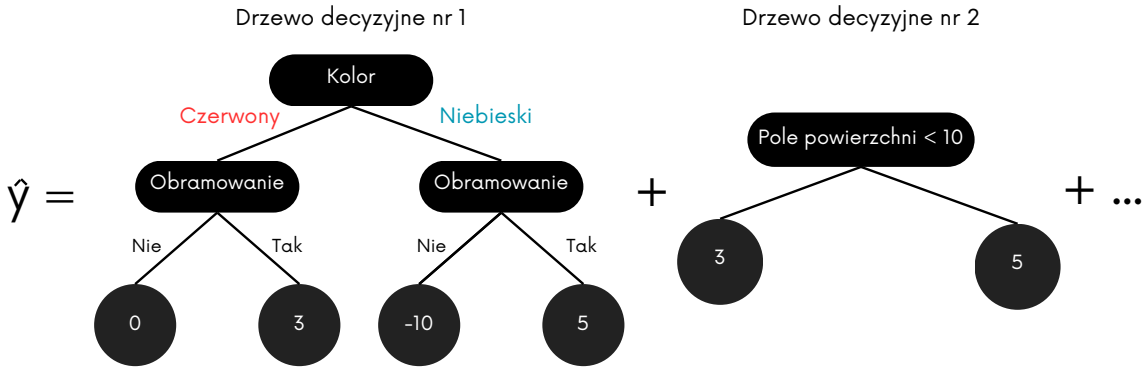
$$\Omega(\mathrm{f}) = \gamma \mathrm{T} + \frac{1}{2}\gamma ||\mathrm{w}||^2, \tag{8}$$

gdzie:

l – funkcja straty obliczana na podstawie różnicy pomiędzy wartości rzeczywistą $\mathrm{y_i}$ i przewidywaną $\hat{\mathrm{y}}_\mathrm{i}$,

Ω – funkcja kary za skomplikowanie modelu – przyjmuje większe wartości dla modeli charakteryzujących się większą rozbudową w celu zachowania maksymalnie prostego rozwiązania.

Poniższy schemat 8 przedstawia w sposób poglądowy metodę obliczania ostatecznego wyniku dla systemu XGBoost. Schemat przedstawia budowę wyłącznie dwóch drzew, ale praktyczne zastosowanie wymaga większej ilości.



Rysunek 8. Poglądowy schemat systemu XGBoost, źródło: *opracowanie własne*

Dokładna budowa modelu jest procesem skomplikowanym, ponieważ system korzysta z wielu technik umożliwiających zarówno zwiększenie efektywności działania, jak i dokładności otrzymywanych wyników. Kompleksowa specyfikacja modelu i omówienie poszczególnych wzorów znajduje się w opracowaniu. Warto zaznaczyć natomiast, że system XGBoost stał się szczególnie popularnym rozwiązaniem w ostatnich latach oraz został powszechnie doceniony w kontekście różnych wyzwań z dziedziny uczenia maszynowego i eksploracji danych. Efektywność tego rozwiązania potwierdzają statystyki wyników platformy z konkursami z dziedziny uczenia maszynowego „Kaggle", gdzie spośród 29 zwycięskich rozwiązań w 2015 r., aż 17 wykorzystywało XGBoost [36].

2.4 Ocena jakości predykcji

W celu obiektywnego porównania badanych modeli niezbędne jest opracowanie spójnej metody porównywania otrzymywanych wyników. Statystyka i dyscyplina uczenia maszynowego wyróżnia wiele metryk umożliwiających kompleksową ocenę jakości otrzymywanych predykcji [30], ale wiele z nich znajduje jedynie zastosowanie w specyficznych przypadkach. Dobierając odpowiednie metody określania jakości modelu należy kierować się specyfiką analizowanych danych, charakterystyką problemu i teoretycznym zastosowaniem każdej miary, ponieważ wiele metryk zostało opracowanych z myślą o konkretnym modelu i może prowadzić do mylnej interpretacji w przypadku użycia niezgodnego z przeznaczeniem. W analizowanym problemie predykcji zachowania prosumentów pod

względem konsumpcji i produkcji energii zostały zastosowane dwie metryki oceny jakości modelu regresji opisane poniżej.

Średni błąd bezwzględny

Średni błąd bezwzględny (ang. *Mean Absolute Error*, MAE) mierzy średnią wielkość błędów w zestawie prognoz bez uwzględniania ich kierunku. Innymi słowy, jest to średnia arytmetyczna z bezwzględnych różnic między przewidywaniami, a faktycznymi obserwacjami, przy czym wszystkie różnice posiadają jednakową wagę. Wynik bliski wartości 0 oznacza, że model jest bardzo dobrze dopasowany do danych. Średni błąd bezwzględny wyrażony jest następującym wzorem 9:

$$\mathrm{MEA} = \frac{1}{\mathrm{N}} \sum_{\mathrm{n}=1}^{\mathrm{N}} |\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i}|, \tag{9}$$

gdzie:

N – liczebność zbioru,

$\mathrm{y_i}$ – rzeczywista wartość próbki,

$\hat{\mathrm{y}}_\mathrm{i}$ – przewidywana wartość próbki.

Pierwiastek błędu średnio-kwadratowego

Pierwiastek błędu średniokwadratowego (ang. *Root Mean Square Error*, RMSE) mierzy pierwiastek średniej wielkości kwadratu błędu w zestawie prognoz. Oznacza to, że obliczana jest średnia arytmetyczna kwadratu różnicy próbek z wartościami przewidywanymi i wartościami rzeczywistymi, a następnie obliczany jest pierwiastek kwadratowy uzyskanej średniej. Im wartość jest bliższa 0, tym model lepiej jest dopasowany do danych. Warto zaznaczyć, że poprzez podnoszenie do kwadratu wartości różnic, błąd średnio-kwadratowy przypisuje stosunkowo wielkie znaczenie błędom o dużych wartościach. Metryka znajduje zatem szczególne zastosowanie, gdy sytuacja występowania nawet pojedynczych dużych błędów jest niekorzystna. Pierwiastek błędu średnio-kwadratowego wyrażony jest wzorem 10:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}} \sum_{\mathrm{n}=1}^{\mathrm{N}} (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2}, \tag{10}$$

gdzie:

N – liczebność zbioru,

$\mathrm{y_i}$ – rzeczywista wartość próbki,

$\hat{\mathrm{y}}_\mathrm{i}$ – przewidywana wartość próbki.

Choć pod względem analitycznym obie metryki wyglądają podobnie, występujące pomiędzy nimi różnice wpływają na ich zastosowanie i interpretacje. Istotną przewagą RMSE nad MAE jest unikanie użycia operatora wartości bezwzględnej, który jest często niepożądany w obliczeniach matematycznych za sprawą prowadzenia do skomplikowanych obliczeń [31].

Stosując powyższe metryki oceny jakości modelu należy mieć również na względzie zachowanie kompromisu między obciążeniem a wariancją (ang. *bias-variance tradeoff*) [33]. Jest to problem leżący u podstaw modelowania w uczeniu maszynowym. Polega na sprzeczności między zwiększaniem obciążenia modelu (dopasowania do danych) i spadkiem wariancji (wrażliwości na dane testowe). Choć z reguły małe obciążenie jest niepożądane to wysoka wariancja również nie może być akceptowana w ramach optymalnego rozwiązania. Sytuacja, w której wszystkie metryki osiągają wartość bliską 0 dla zbioru uczącego może oznaczać małe obciążenie modelu i przeuczenie, a w konsekwencji dużą wariancję i złe wyniki na zbiorze testowym. Stosując metryki jakości modelu należy kierować się nie tylko zmniejszaniem ich wartości, ale przede wszystkim minimalizowaniem błędu w praktycznym zastosowaniu.

Rozdział 3

Dane

Celem niniejszego rozdziału jest przedstawianie dostępnych zbiorów danych z wyszczególnieniem atrybutów i opisów ich znaczenia. Dalsza część tego fragmentu pracy skupia się nad opisaniem najważniejszych zależności w zgromadzonych informacjach, wykorzystując wiedzę teoretyczną na temat działania rynku energii i uczenia maszynowego.

3.1 Zbiory danych

Wszystkie dane analizowane i przetwarzane przez opracowane modele zostały udostępnione przez estoński koncern energetyczny „Eesti Energia AS" z siedzibą w Tallinnie (Estonia) w ramach konkursu „Predict Energy Behavior of Prosumers – Predict Prosumer Energy Patterns and Minimize Imbalance Costs" ogłoszonego na platformie konkursowej gromadzącej społeczność analityków z całego świata „Kaggle". Udostępnione zbiory zostały wyselekcjonowane przez organizatorów konkursu w taki sposób, żeby dotyczyły podobnego zakresu czasowego, ale mogą zawierać dużo informacji, których istotność powinna być obiektem badań.

3.1.1 Produkcja i konsumpcja energii

Dane o produkcji i konsumpcji energii zawierają informacje na temat ilości wyprodukowanej lub zużytej ilości energii przez prosumentów w danej godzinie dla każdej kombinacji prowincji w Estonii i rodzaju umowy prosumenta z wyszczególnieniem klientów będących przedsiębiorstwami. Zbiór danych dotyczy próbek w przedziale od 2021-09-01 00:00 do 2023-05-31 23:00 i zawiera 528 brakujących wartości zmiennej opisywanej, a jego całkowita liczebność wynosi 2.018.352. Opis atrybutów zbioru został umieszczony kolejno w tabelach 1 i 2.

Nazwa atrybutu	Opis
target	Konsumpcja lub produkcja energii dla danej wartości „prediction_unit_id" w danej godzinie (zmienna opisywana). Dokładna jednostka zmiennej nie jest sprecyzowana
is_consumption	Wartość logiczna określająca, czy celem tego wiersza jest zużycie (wartość *True*), czy produkcja (wartość *False*)
county	Identyfikator jednoznacznie przypisany do prowincji Estonii
is_business	Wartość logiczna określająca, czy prosument jest firmą (wartość *True*)
product_type	Identyfikator jednoznacznie przypisany do jednego rodzaju umowy
datetime	Czas – strefa estońska (UTC+02:00)
data_block_id	Zmienna pomocnicza do określania dostępu pomiędzy danymi
row_id	Unikalny identyfikator wiersza
prediction_unit_id	Unikalny identyfikator dla kombinacji kolumn „county", „is_business" i „product_type"

Tabela 1. Opis zbioru danych produkcji i konsumpcji energii

Podstawowe informacje

Nazwa atrybutu	Średnia	Minimum	Maksimum	Odch. Stand.
county	7,297	0,000	15,000	4,781
product_type	1,899	0,000	3,000	1,082
target	274,856	0,000	15.480,274	909,502
datetime	2022-07-20 08:29:25	2021-09-01 00:00:00	2023-05-31 23:00:00	-
data_block_id	321,875	0,000	637,000	182,634
row_id	43,084	-32.768,000	32.767,000	18.970,475
prediction_unit_id	33,045	0,000	68,000	19,591

Tabela 2. Podstawowe metryki danych produkcji i konsumpcji energii

3.1.2 Ceny gazu

Zbiór zawiera informacje o poziomie cen gazu ziemnego na terenie Estonii w danym dniu w zakresie dat od 2021-09-01 do 2023-05-30 i nie zawiera żadnych brakujących danych, a jego liczebność wynosi 637. Opis atrybutów zbioru został umieszczony kolejno w tabelach 3 i 4.

Nazwa atrybutu	Opis
origin_date	Data udostępnienia danych
forecast_date	Data, w których obowiązuje dana cena gazu ziemnego
lowest_price_per_mwh	Najniższa cena gazu ziemnego oferowana na rynku dnia następnego wyrażona w € za odpowiednik kWh
high_price_per_mwh	Najwyższa cena gazu ziemnego oferowana na rynku dnia następnego wyrażona w € za odpowiednik kWh
data_block_id	Zmienna pomocnicza do określania dostępu pomiędzy danymi

Tabela 3. Opis zbioru danych cen gazu

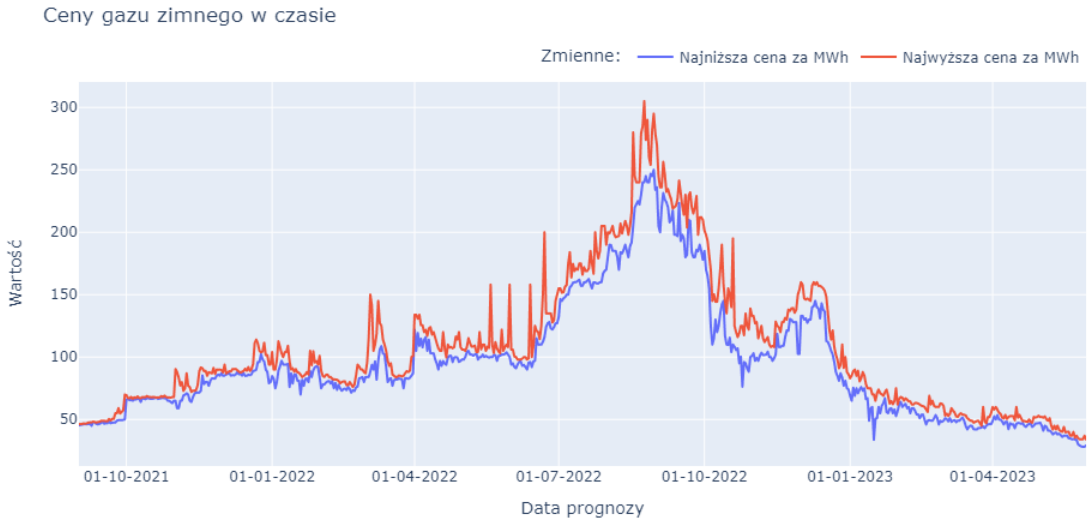
Podstawowe informacje

Nazwa atrybutu	Średnia	Minimum	Maksimum	Odch. Stand.
forecast_date	2022-07-16 00:00:00	2021-09-01 00:00:00	2023-05-30 00:00:00	-
lowest_price_per_mwh	95,037	28,100	250,000	47,552
highest_price_per_mwh	107,755	34,000	305,000	54,744
origin_date	2022-07-15 00:00:00	2021-08-31 00:00:00	2023-05-29 00:00:00	-
data_block_id	319,000	1,000	637,000	184,030

Tabela 4. Podstawowe metryki danych cen gazu

Wykres cen gazu ziemnego

Naniesienie otrzymanych cen gazu ziemnego w czasie zostało przedstawione na wykresie 9.



Rysunek 9. Ceny gazu ziemnego w czasie

3.1.3 Ceny energii elektrycznej

Dane o cenach energii elektrycznej zawierają informacje o kształtowaniu się cen na terenie Estonii z dokładnością do danej godziny w zakresie dat od 2021-08-31 do 2023-05-29. Zbiór nie zawiera żadnych brakujących danych, a jego liczebność wynosi 15.286. Opis atrybutów zbioru został umieszczony kolejno w tabelach 5 i 6.

Nazwa atrybutu	Opis
origin_date	Data udostępnienia danych
forecast_date	Data, w których obowiązuje dana cena gazu ziemnego
euros_per_mwh	Cena energii elektrycznej oferowanej na rynku dnia następnego wyrażona w € za kWh
data_block_id	Zmienna pomocnicza do określania dostępu pomiędzy danymi

Tabela 5. Opis zbioru danych cen energii elektrycznej

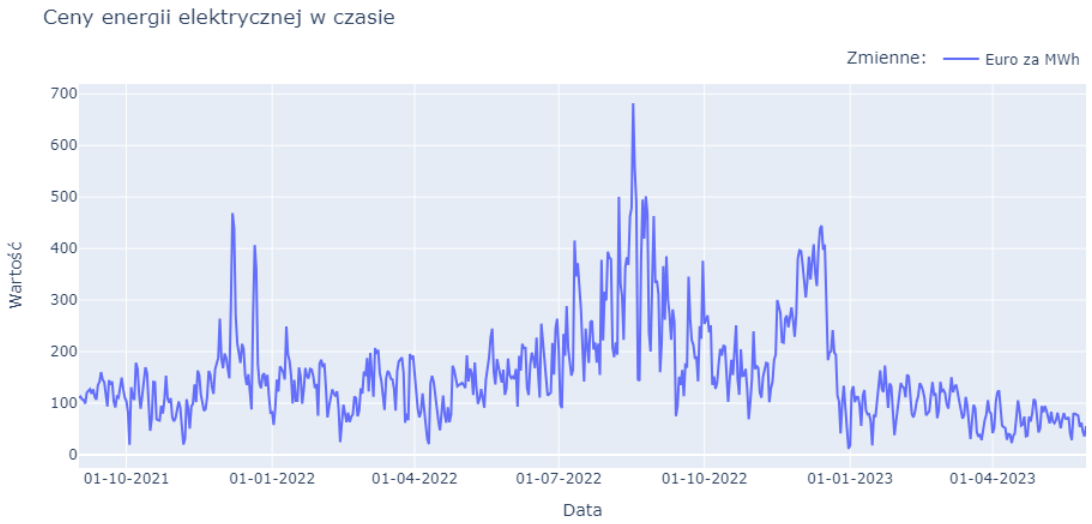
Podstawowe informacje

Nazwa atrybutu	Średnia	Minimum	Maksimum	Odch. Stand.
euros_per_mwh	157,064	-10,060	4.000,000	121,149
origin_date	2022-07-15 11:16:41	2021-08-31 00:00:00	2023-05-29 23:00:00	-
data_block_id	318,991	1,000	637,000	183,890
forecast_date	2022-07-16 11:16:41	2021-09-01 00:00:00	2023-05-30 23:00:00	-

Tabela 6. Podstawowe metryki danych o energii elektrycznej

Wykres cen energii elektrycznej

Naniesienie otrzymanych cen energii w czasie na wykres zostało przedstawione na wykresie 10.



Rysunek 10. Ceny energii elektrycznej w czasie

3.1.4 Historyczne dane o pogodzie

Historyczne dane o pogodzie zawierają pomiary warunków atmosferycznych w kolejnych godzinach dla wyszczególnionych punktów pogodowych na terenie Estonii w zakresie dat od 2021-09-01 00:00 do 2023-05-30 10:00. Zbiór nie zawiera żadnych brakujących danych, a jego liczebność wynosi 1.710.800. Opis atrybutów zbioru został umieszczony kolejno w tabelach 7 i 8.

Nazwa atrybutu	Opis
latitude/longitude	Szerokość i długość geograficzna lokalizacji stacji pogodowej
temperature	Temperatura powietrza 2 metry nad poziomem gruntu wyrażona w stopniach Celsjusza
dewpoint	Temperatura punktu rosy 2 metry nad poziomem gruntu wyrażona w stopniach Celsjusza
snowfall	Opady śniegu w ciągu godziny wyrażone w cm
surface_pressure	Ciśnienie powietrza na powierzchni w hPa
cloudcover_low	Procent pokrycia nieba przez chmury w zakresie wysokości 0–3km
cloudcover_mid	Procent pokrycia nieba przez chmury w zakresie wysokości 3–8km
cloudcover_high	Procent pokrycia nieba przez chmury na wysokości ponad 8km
cloudcover_total	Procent pokrycia nieba przez chmury w całym analizowanym zakresie wysokości
windspeed_10m	Szybkość wiatru na wysokości 10 metrów powyżej powierzchni wyrażona w $\mathrm{m/s}$
winddirection_10m	Kierunek wiatru na wysokości 10 metrów powyżej powierzchni wyrażony w stopniach
shortwave_radiation	Globalne poziome napromieniowanie w $\frac{\mathrm{Wh}}{\mathrm{m}^2}$
direct_solar_radiation	Bezpośrednie natężenia promieniowania słonecznego docierające na powierzchnię w płaszczyźnie prostopadłej do kierunku słońca, skumulowane w ciągu jednej godziny, wyrażone w $\frac{\mathrm{Wh}}{\mathrm{m}^2}$
diffuse_radiation	Rozproszone natężenie promieniowania słonecznego wyrażone w $\frac{\mathrm{Wh}}{\mathrm{m}^2}$
data_block_id	Zmienna pomocnicza do określania dostępu pomiędzy danymi

Tabela 7. Opis zbioru historycznych danych pogodowych

Podstawowe informacje

Nazwa atrybutu	Średnia	Minimum	Maksimum	Odch. Stand.
datetime	2022-07-16 05:00:00	2021-09-01 00:00:00	2023-05-30 10:00:00	-
temperature	5,704	-23,700	32,600	8,013
dewpoint	2,209	-25,900	22,600	7,211
rain	0,049	0,000	16,800	0,206
snowfall	0,016	0,000	2,660	0,075
surface_pressure	1.009,281	942,900	1.049,300	13,099
cloudcover_total	61,021	0,000	100,000	37,755
cloudcover_low	46,822	0,000	100,000	40,760
cloudcover_mid	34,457	0,000	100,000	38,355
cloudcover_high	36,085	0,000	100,000	41,374
windspeed_10m	4,855	0,000	21,750	2,479
winddirection_10m	197,852	0,000	360,000	89,922
shortwave_radiation	105,935	0,000	848,000	179,388
direct_solar_radiation	64,048	0,000	739,000	132,906
diffuse_radiation	41,887	0,000	386,000	61,841
latitude	58,650	57,600	59,700	0,687
longitude	24,950	21,700	28,200	2,016
data_block_id	319,271	1,000	637,000	183,730

Tabela 8. Podstawowe metryki danych historycznej pogody

3.1.5 Prognoza pogody

Dane o prognozie pogody zawierają prognozowane informacje pogodowe dla kolejnych godzin z wyszczególnieniem kilku wybranych punktów znajdujących się na terenie (lub okolicy) Estonii w zakresie dat od 2021-09-01 00:00 do 2023-05-30 10:00. Zbiór nie zawiera żadnych brakujących danych, a jego liczebność wynosi 1.710.800. Opis atrybutów zbioru został umieszczony kolejno w tabelach 9 i 10.

Nazwa atrybutu	Opis
latitude/longitude	Szerokość i długość geograficzna lokalizacji stacji pogodowej
origin_datetime	Stempel czasowy momentu generowania danych
hours_ahead	Liczba godzin pomiędzy generowaniem danych a momentem dla, którego dane są generowane
temperature	Temperatura powietrza 2 metry nad poziomem gruntu wyrażona w stopniach Celsjusza
dewpoint	Temperatura punktu rosy 2 metry nad poziomem gruntu wyrażona w stopniach Celsjusza
cloudcover_[low/mid/high]	Procent pokrycia nieba przez chmury na wysokości 0–2km/2–6km/ponad 6km
cloudcover_total	Procent pokrycia nieba przez chmury w całym zakresie wysokości
10_metre_[u/v]_wind	Wschodnia/Północna składowa prędkości wiatru mierzona 10 metrów nad powierzchnią wyrażona w $\mathrm{m/s}$
data_block_id	Zmienna pomocnicza do określania dostępu pomiędzy danymi
forecast_datetime	Stempel czasowy momentu, dla którego generowana jest pogoda
direct_solar_rad	Bezpośrednie natężenia promieniowania słonecznego docierające na powierzchnię w płaszczyźnie prostopadłej do kierunku słońca, skumulowane w ciągu jednej godziny, wyrażone w $\frac{\mathrm{Wh}}{\mathrm{m}^2}$
surface_solar_rad_d	Bezpośrednie i rozproszone natężenie promieniowania słonczego docierające do płaszczyzny poziomej powierzchni Ziemi, wyrażone w $\frac{\mathrm{Wh}}{\mathrm{m}^2}$
snowfall	Opady śniegu w ciągu godziny wyrażone poprzez równowartość metrów wody
total_precipitation	Nagromadzona ciecz powstała w wyniku opadów deszczu i śniegu, która spadła na powierzchnię Ziemi w ciągu godziny, wyrażona w metrach

Tabela 9. Opis zbioru danych prognozy pogody

Podstawowe informacje

Nazwa atrybutu	Średnia	Minimum	Maksimum	Odch. Stand.
datetime	2022-07-16 05:00:00	2021-09-01 00:00:00	2023-05-30 10:00:00	-
temperature	5,704	-23,700	32,600	8,013
dewpoint	2,209	-25,900	22,600	7,211
rain	0,049	0,000	16,800	0,206
snowfall	0,016	0,000	2,660	0,075
surface_pressure	1.009,281	942,900	1.049,300	13,099
cloudcover_total	61,021	0,000	100,000	37,755
cloudcover_low	46,822	0,000	100,000	40,760
cloudcover_mid	34,457	0,000	100,000	38,355
cloudcover_high	36,085	0,000	100,000	41,374
windspeed_10m	4,855	0,000	21,750	2,479
winddirection_10m	197,852	0,000	360,000	89,922
shortwave_radiation	105,935	0,000	848,000	179,388
direct_solar_radiation	64,048	0,000	739,000	132,906
diffuse_radiation	41,887	0,000	386,000	61,841
latitude	58,650	57,600	59,700	0,687
longitude	24,950	21,700	28,200	2,016
data_block_id	319,271	1,000	637,000	183,730

Tabela 10. Podstawowe metryki danych prognozy pogody

3.1.6 Cechy prosumentów

Dane o prosumentach zawierają dodatkowe informacje charakteryzujące daną grupę prosumentów w określonym czasie z dokładnością do dnia w zakresie od 2021-09-01 00:00 do 2023-05-29 00:00. Zbiór nie zawiera żadnych brakujących danych, a jego liczebność wynosi 41.919. Opis atrybutów zbioru został umieszczony kolejno w tabelach 11 i 12.

Nazwa atrybutu	Opis
product_type	Identyfikator jednoznacznie przypisany do jednego rodzaju umowy
county	Identyfikator jednoznacznie przypisany do prowincji w Estonii
is_business	Wartość logiczna określająca, czy prosument jest firmą (wartość *True*)
data_block_id	Zmienna pomocnicza do określania dostępu pomiędzy danymi
eic_count	Łączna liczba punktów poboru – European Indetifier Code (EIC)
installed_capacity	Zainstalowana moc paneli fotowoltaicznych wyrażona w kW
date	Data gromadzenia danych

Tabela 11. Opis zbioru danych prosumentów

Podstawowe informacje

Nazwa atrybutu	Średnia	Minimum	Maksimum	Odch. Stand.
product_type	1,899	0,000	3,000	1,082
county	7,297	0,000	15,000	4,781
eic_count	73,345	5,000	1.517,000	144,064
installed_capacity	1.450,771	5,500	19.314,310	2.422,233
date	2022-07-18 21:34:22	2021-09-01 00:00:00	2023-05-29 00:00:00	-
data_block_id	322,899	2,000	637,000	182,076

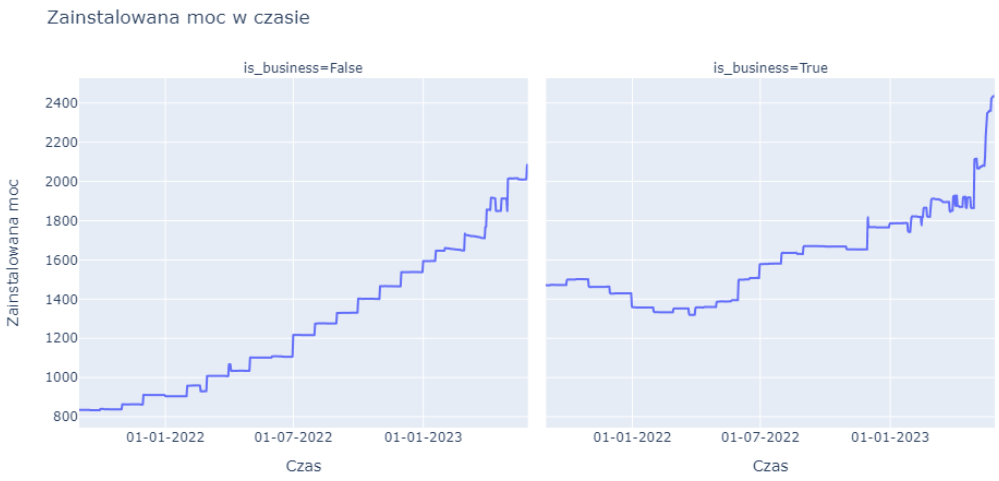
Tabela 12. Podstawowe metryki danych o prosumentach

3.2 Eksploracyjna Analiza Danych

Udostępnione dane zawierają wiele informacji, które mogą znaleźć zastosowanie w predykcji zachowania prosumentów. W celu lepszego zrozumienia problemu i dostępnych atrybutów niezbędne jest przeprowadzenie eksploracyjnej analizy danych. Pozwoli to na znalezienie zależności pomiędzy cechami i opracowanie sposobu na połączenie danych w jeden zbiór uczący.

Zainstalowana moc w czasie

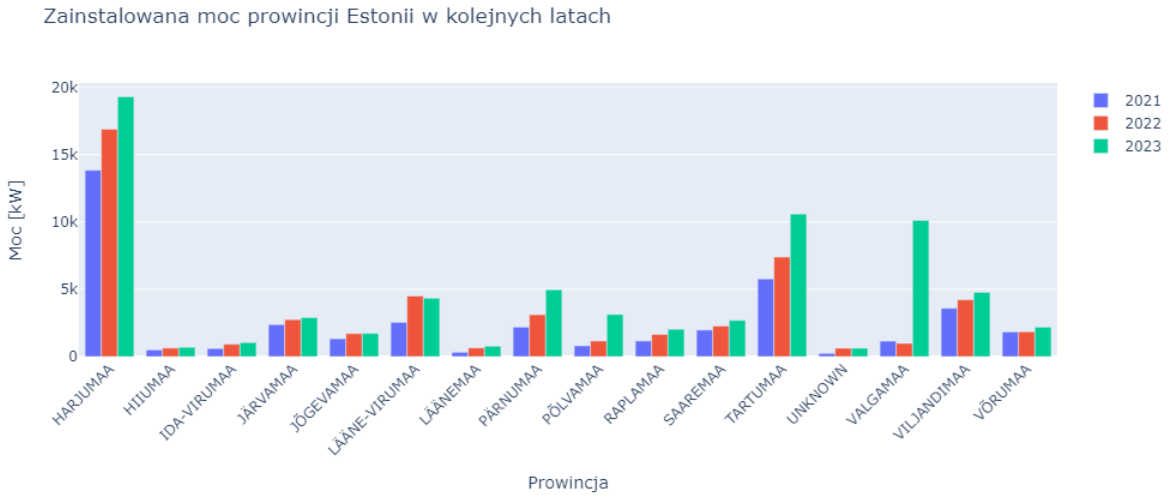
Dysponując informacjami na temat grup prosumentów możliwe jest zbadanie rozkładu zainstalowanej mocy na przestrzeni analizowanego okresu. Wartość uśredniona dla wszystkich prowincji została przedstawiona na wykresie 11. Przebieg danych pozwala zauważyć trend zgodny z informacjami przedstawionymi na wstępie pracy. Zainstalowana moc prosumentów w Estonii ulega zwiększeniu, co jest widoczne rozpatrując nawet wąski przedział czasowy objęty przez zbiór.



Rysunek 11. Zainstalowana moc prosumentów w czasie

Zainstalowana moc w prowincjach Estonii

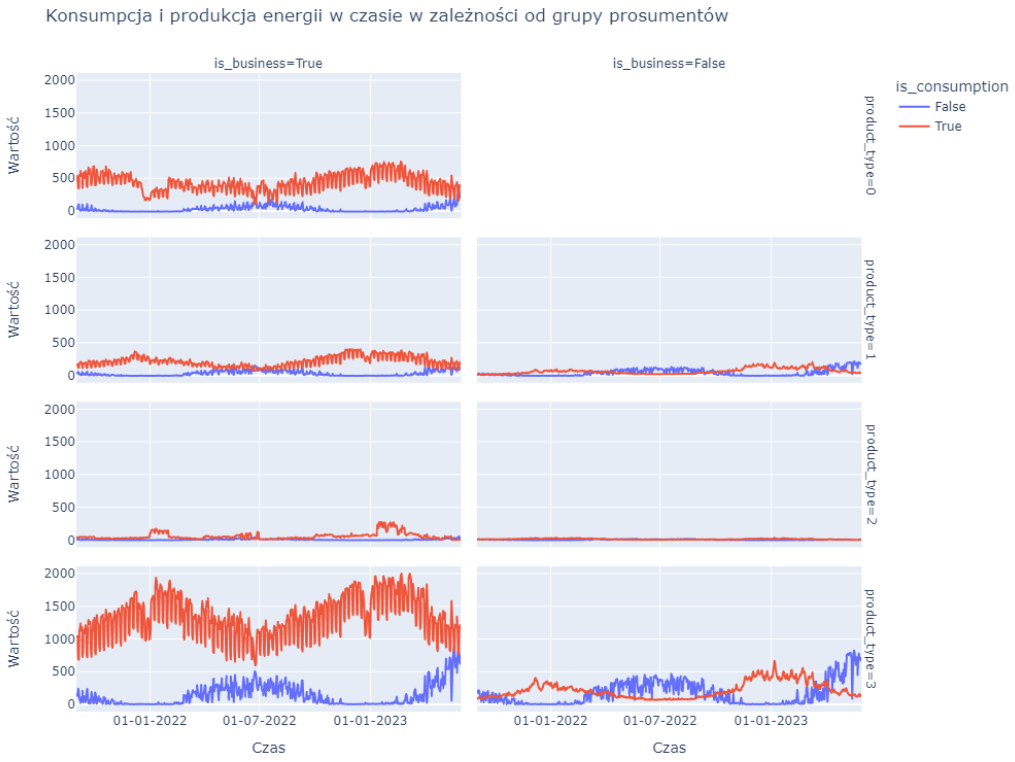
Istotnym czynnikiem jest również zbadanie rozłożenia zainstalowanej mocy w poszczególnych prowincjach. W tym celu został stworzony wykres 12 prezentujący rozkład maksymalnej zainstalowanej mocy paneli fotowoltaicznej w poszczególnych prowincjach na przestrzeni lat. Analiza pozwala zauważyć, że każda prowincja charakteryzuje się przyrostem dysponowanej mocy w kolejnych latach, ale istnieją przypadki, takie jak „Harjumaa", „Tartumaa" i „Valgamaa", które wyraźnie przodują w tym zakresie.



Rysunek 12. Zainstalowana moc w poszczególnych prowincjach

Konsumpcja i produkcja energii w czasie

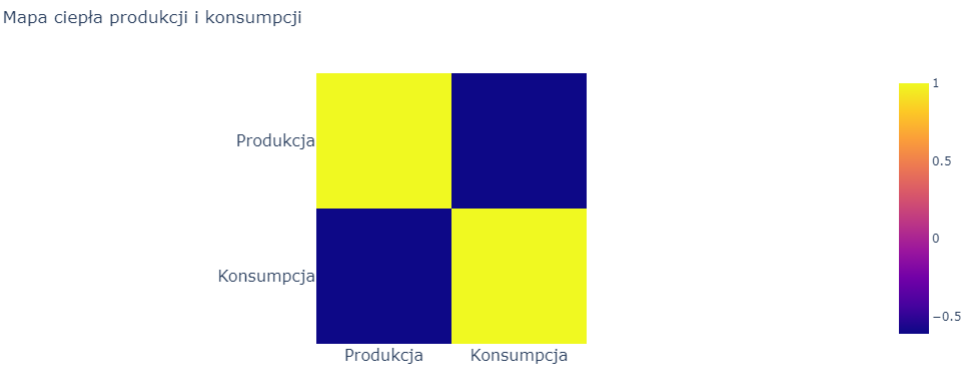
Naniesienie danych na wykres 13 prezentujący przebieg produkcji i konsumpcji w analizowanym zakresie czasu pozwala zauważyć charakterystyczne cechy zachowania poszczególnych grup prosumentów. Średnia ilość produkowanej i konsumowanej energii wyraźnie różni się dla poszczególnych typów rozliczeń prosumentów. W przypadku rozliczeń typu „Spot" (product_type=3) przyjmowane są największe wartości zarówno dla grupy prosumentów oznaczonych, jak i nieoznaczonych jako biznes. Dodatkowo można zauważyć, że typ „Combined" (product_type=0) występuje tylko w przypadku przedsiębiorstw. Niezależnie od grupy prosumentów wyraźny jest również trend wzrostu ilości produkcji i konsumpcji w analogicznym okresie następnego roku.



Rysunek 13. Produkcja i konsumpcja energii dla wyszczególnionych grup prosumentów

Korelacja konsumpcji i produkcji

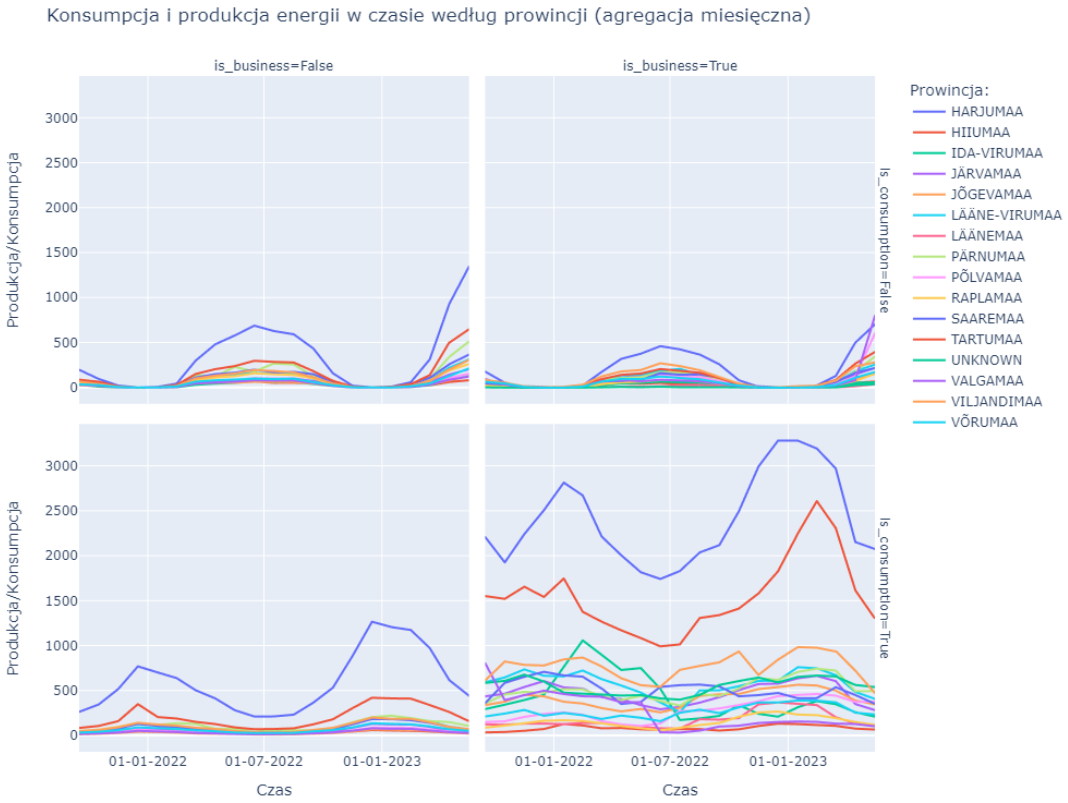
Analiza wykresu konsumpcji i produkcji w czasie na rysunku 13 pozwala zauważyć odwrotnie proporcjonalną relację pomiędzy tymi zmiennymi. Jest to zgodne z oczekiwaniami, ponieważ rosnąca produkcja pokrywa część zapotrzebowania prosumentów. Poniższa mapa ciepła 14 pozwala stwierdzić dokładną wartość korelacji tych cech na poziomie -0,6.



Rysunek 14. Mapa ciepła korelacji pomiędzy produkcją i konsumpcją

Produkcja i konsumpcja według prowincji

Naniesienie na wykres 15 wartości produkcji i konsumpcji dla poszczególnych prowincji pozwala zauważyć podobieństwo pomiędzy nimi w zakresie zachowania prosumentów. Największymi wartościami charakteryzują się prowincje o największej zainstalowanej mocy. Dodatkowo można stwierdzić, że konsumpcja energii przedsiębiorstw jest najbardziej nieregularnym przebiegiem.

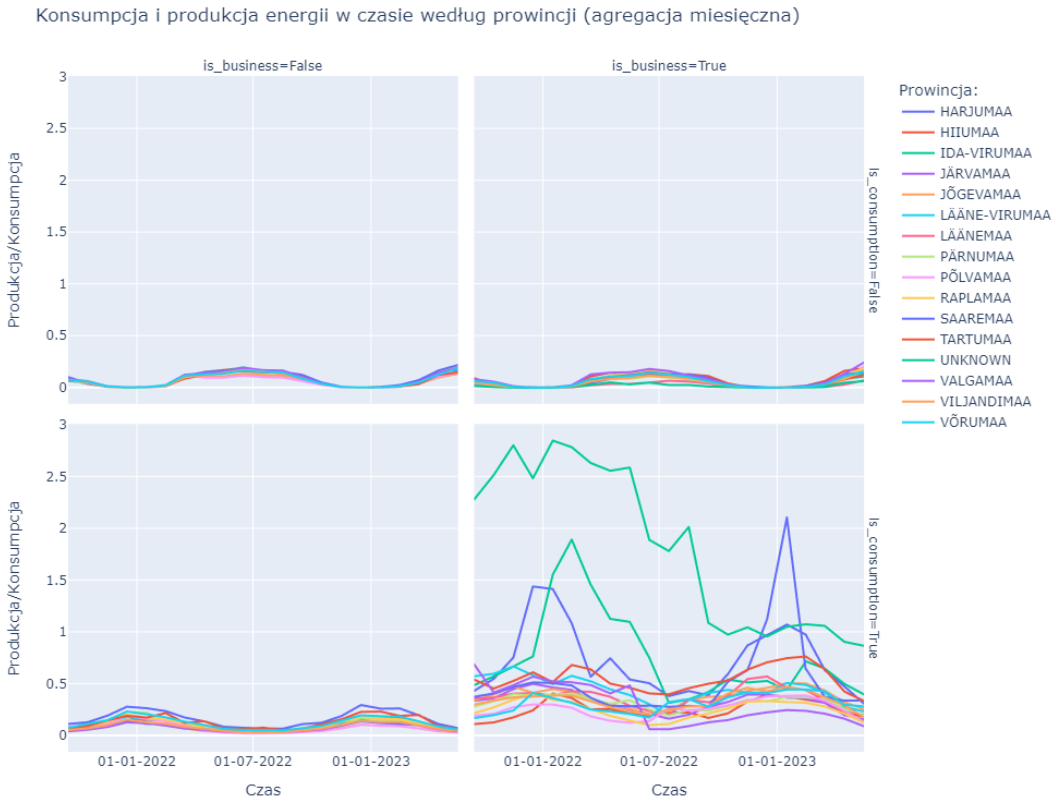


Rysunek 15. Produkcja i konsumpcja według prowincji

Występowanie trendu

Analiza przebiegu produkcji i konsumpcji w zakresie, który zawiera zbiór danych, pozwala zauważyć tendencję do wzrostu tych cech w kolejnych latach. Jest to zachowanie zgodne z perspektywami rozwoju rynku energii, ale wiąże się z wystąpieniem problemu ekstrapolacji danych. Teoria modeli opartych o drzewa decyzyjne podkreśla ich ograniczenia w możliwości dokonywania predykcji w zakresie wykraczającym poza zbiór uczący. W konsekwencji model bazujący na takim zbiorze uczącym mógłby działać efektywnie wyłącznie przez ograniczony czas do momentu, kiedy dane rzeczywiste nie zaczęłyby przyjmować wartości wykraczających poza nauczony zakres. W celu uniknięcia takiej sytuacji została podjęta próba transformacji danych uczących w taki sposób, żeby ograniczyć występowania tego zjawiska. Korzystając z obserwacji przebiegu zainstalowanej mocy

w czasie, zmienna opisywana została podzielona przez jej wartość. W przypadku produkcji takie postępowania jest uargumentowane faktem, że ilość produkcji powinna być wprost proporcjonalna do zainstalowanej mocy. Wysoka ujemna korelacja pomiędzy produkcją i konsumpcją pozwala natomiast przeprowadzić podobne działanie dla samej konsumpcji, ale ze względu na dostęp wyłącznie do danych z zakresu od 2021-09 do 2023-05 korzystne byłoby sprawdzenie utrzymania tej zależności w szerszym zakresie. Wykres 16 prezentuje przebieg obu szeregów w czasie po transformacji danych. Zaproponowane podejście pozwala otrzymać rozkłady na podobnym poziomie w kolejnych latach, a w konsekwencji dużo większe prawdopodobieństwo zachowania efektywności modelu opartego o drzewa decyzyjne nauczonego na takich danych w dłuższej perspektywie. Problemem jest jedynie zapotrzebowanie grupy prosumentów oznaczonych flagą biznesu, ponieważ taka transformacja doprowadziła do większej nieregularności przebiegów w przypadku niektórych prowincji.



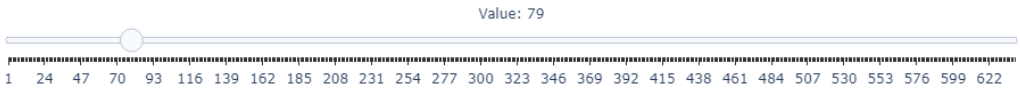
Rysunek 16. Przebieg produkcji i konsumpcji po transformacji

Dostępność danych

Ważnym aspektem predykcji szeregów czasowych jest stwierdzenie dostępności informacji w celu uniknięcia przypadku, w którym model do stworzenia przewidywań używa wartości, które nie będą dostępne dla późniejszego użytkownika. W analizowanych zbiorach istnieje specjalny atrybut pomocniczy „data_block_id". Taka sama wartość tej zmiennej w obrębie zestawu oznacza, że dana grupa danych jest dostępna w tym samym momencie. Dodatkowo wartość tego atrybutu rośnie wraz z przyrostem kolejnych rekordów, więc wszystkie wartości mniejsze od aktualnie analizowanego „data_block_id" są również dostępne w formie próbek historycznych. Poniższa tabela 17 przedstawia przykładową wartość tej cechy i zakres poszczególnych zbiorów, które są oznaczone taką samą wartością. Analiza tabeli pozwala zauważyć jakim zakresem dat dla poszczególnych zestawów można się posługiwać tworząc predykcję dla zmiennej opisywanej („target") ze zbioru oznaczonego w tabeli jako „Produkcja/Konsumpcja". Przykładowo konkretna wartość 79 oznacza tworzenie predykcji od 2021-11-19 00:00:00 do 2021-11-19 23:00:00. W tabeli pokazane są przedziały dat z innych zbiorów, które można w tym celu stosować, ponieważ również są oznaczone wartością 79.

Data block id: 79

data_block_id	Początek	Koniec	Liczebność	Zbiór danych
79	2021-11-17T11:00:00	2021-11-18T10:00:00	2688	Historyczna pogoda
79	2021-11-18T02:00:00	2021-11-20T01:00:00	720	Prognoza pogody
79	2021-11-18	2021-11-18	24	Ceny energii elektrycznej
79	2021-11-18T00:00:00	2021-11-18T00:00:00	1	Ceny gazu ziemnego
79	2021-11-19T00:00:00	2021-11-19T23:00:00	3024	Produkcja/Konsumpcja
79	2021-11-17T00:00:00	2021-11-17T00:00:00	63	Dane o prosumentach



Rysunek 17. Tabla prezentująca przeznaczenie atrybutu data_block_id

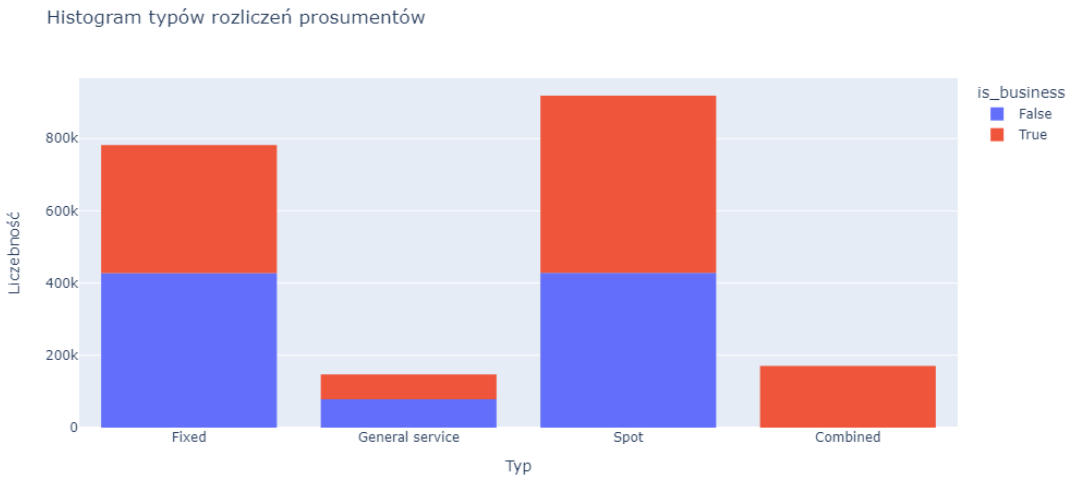
Rodzaje rozliczeń prosumentów

Zbiór danych zakłada możliwość przyjęcia jednego z czterech możliwych typów rozliczeń:

- Combined (product_type = 0),
- Fixed (product_type = 1),
- General service (product_type = 2),
- Spot (product_type = 3).

W zależności od wybranego wariantu prosument dokonuje handlu energią w inny sposób. W przypadku „Fixed" umowa określa stałą, niezmienną cenę za dostarczoną energię przez ustalony czas. Takie rozwiązanie zapewnia konsumentowi stabilność i przewidywalność, która jest niezależna od wahań

rynku. Opcja „Spot" polega na handlu energią po aktualnej cenie rynkowej. Cena jest ustalana na podstawie bieżącej sytuacji rynku, więc jest podatna na wahania. Trzeci wariant „Combined" polega na zobowiązaniu prosumenta poprzez przyjęcie wybranych warunków z innych umów i może obejmować połączenie stałych i zmiennych cech w zależności od przyjętej umowy. Ostatni rodzaj „General service" gwarantuje konsumentowi jedynie podstawowy poziom usług energetycznych. Może mieć z góry określone warunki, a struktura cen może być stała lub zmienna. Wykres 18 przedstawia liczebność poszczególnych wariantów w zbiorze uczącym z podziałem na rekordy dotyczące przedsiębiorstw. Wykres uwidacznia fakt, że rodzaje „Fixed", „General service" i „Spot" są równomiernie rozłożone pomiędzy przedsiębiorstwa i zwykłych prosumentów, ale rodzaj „Combined" dotyczy wyłącznie grupy przedsiębiorców. Dodatkowo rodzaj „Spot" i „Fixed" jest znacząco bardziej popularny od innych.

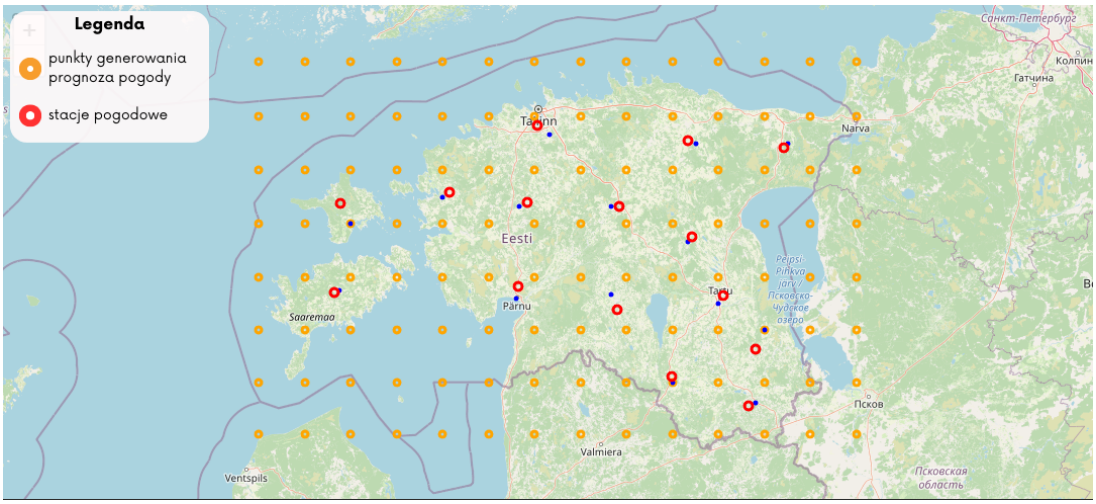


Rysunek 18. Histogram rodzajów rozliczeń prosumentów

Rozmieszczenie punktów pogodowych

Przygotowane zbiory danych pogodowych zawierają wyłącznie informacje na temat współrzędnych geograficznych miejsc zbierania informacji. Zestaw prognozy pogody został przygotowany przez portal, który umożliwia generowanie przewidywanych wartości parametrów atmosferycznych dla dowolnych punktów geograficznych. Historyczne dane pogodowe zostały natomiast stworzone przez fizyczny pomiar określonych warunków atmosferycznych przez rzeczywiste stacje pogodowe. Naniesienie współrzędnych geograficznych na mapę 19 pozwala lepiej zrozumieć rozmieszczenie informacji pogodowych. Na jej podstawie można zauważyć, że dane dotyczące prognozy pogody zostały wygenerowane za pomocą siatki 8×14 pokrywającej cały teren Estonii, a punkty historycznej pogodowy pochodzą z miejsc, w których znajdują się prawdziwe stacje pogodowe. Taka obserwacja pozwala stwierdzić konieczność wykluczenia punktów pogodowych wykraczających poza obszar Estonii i opracowania sposobu na przypisanie występującym kombinacjom szerokości i długości

geograficznej odpowiednich prowincji. Dodatkowo część prowincji posiada więcej niż jeden punkt pogodowych, więc wymagane jest uśrednienie wartości w obrębie takich punktów.



Rysunek 19. Punkty pogodowe naniesione na mapę Estonii

3.3 Podsumowanie

Udostępnione dane zawierają kompleksowy zbiór informacji, znajdujący zastosowanie w przewidywaniu produkcji i konsumpcji prosumentów. Niewielka ilość brakujących danych i relatywnie duża ilość atrybutów stwarza perspektywy na opracowanie modelu o dużej efektywności. Eksploracyjna analiza danych pozwoliła na potwierdzenie zgodności rzeczywistych wartości z teoretycznymi oczekiwaniami. Ponadto, naniesienie danych na wykresy uwidoczniło wiele zależności, które wymagają uwzględnienia podczas opracowywania optymalnego modelu.

Rozdział 4

Modele predykcyjne

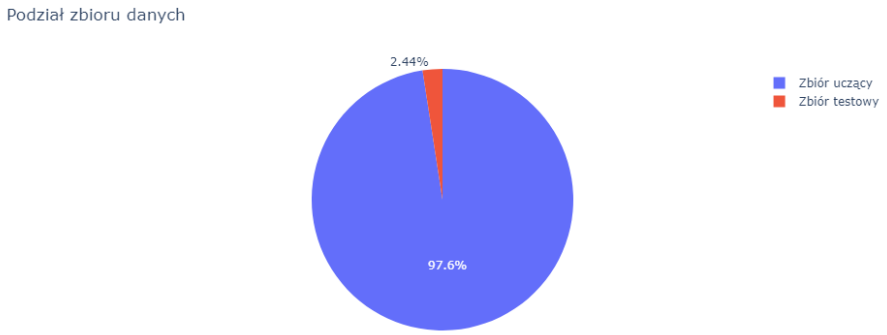
Rozdział 4 opisuje metodykę skomponowania najlepszego modelu predykcyjnego. Proces ten został podzielony na kilka części. Pierwsze przeprowadzone eksperymenty służyły opracowaniu jak największej liczby atrybutów, które pozwoliłyby na efektywniejsze opisanie zmiennej przewidywanej. W kolejnych krokach została podjęta próba wybrania najlepszych parametrów oraz atrybutów, rozwiązania problemu potencjalnej niestacjonarności i przeprowadzenie dodatkowych eksperymentów. Najważniejsze z opracowanych rezultatów zostały opisane w formie kolejnych ośmiu modeli.

4.1 Zbiór danych

Na podstawie przeanalizowanych informacji zawartych w poszczególnych zbiorach danych została przygotowana implementacja, której zadaniem było scalenie wszystkich podzbiorów w jeden zbiór uczący. Każdy zbiór uczący wymagał indywidualnego dobrania kolumn, który jednoznacznie dodawał atrybuty dla konkretnej grupy prosumentów. Ze względu na podobieństwo nazewnicze atrybutów ze zbiorów pogodowych, cechy historycznych danych pogodowych zostały poprzedzone prefiksem „h_". Wiersze zawierające brakujące dane zostały usunięte, a ostateczny zbiór danych (bez opracowania dodatkowych atrybutów) zawierał 56 kolumn i 2.018.352 wierszy.

4.2 Podział danych

Podział danych na zbiór danych uczących i testowych został przeprowadzony zgodnie z podejściem stosowanym dla przewidywania szeregów czasowych. Dane uczące zostały posortowane, a następnie podzielone na dwa podzbiory za pomocą wybranej daty. Ze względu na przeznaczenie modelu do predykcji krótkoterminowej i niewielki zakres dat zbioru uczącego, za datę podziału został wybrany dzień 2023-05-17. Stosunek zbioru uczącego do testowego został przedstawiony na wykresie 20.



Rysunek 20. Stosunek zbioru testowego do uczącego

4.3 Narzędzia i architektura

Proces tworzenia modeli uczenia maszynowego wymaga znaczących nakładów obliczeniowych. W szczególności problem ten zyskuje na istotności podczas pracy ze skomplikowaną strukturą danych. Z tego powodu do eksperymentów zostały wybrany modele lasu losowego i XGBoost, które charakteryzują się dużą wydajnością obliczeń, umożliwiającą prowadzenie działań na standardowym komputerze stacjonarnym. Dokładna specyfikacja użytego urządzenia przedstawia się następująco:

Procesor:	Intel Core i7-13700F,
Karta Graficzna:	MSI GeForce GTX 1060 6GB,
Pamięć RAM:	Patriot 32GB (4x8GB) 3600MHz cl14,
System:	Microsoft Windows 11 Education.

Wszystkie implementacje zostały przygotowane z użyciem języka programowania „Python" w wersji *3.10.12*. Modele uczenia maszynowego zostały zaczerpnięte z gotowych pakietów zgodnie z informacjami w tabeli 13.

Model	Pakiet	Wersja
Las losowy	sklearn.ensemble.RandomForestClassifier	1.3.2
XGBoost	xgboost	2.0.3
Regresja liniowa	sklearn.linear_model.LinearRegression	1.3.2

Tabela 13. Tabela ze specyfikacją użytych pakietów

4.4 Ocena modeli

Obiektywne porównywanie kolejnych modeli zostało zagwarantowane przez tożsame podejście do oceny wszystkich rozwiązań. W tym celu została przygotowana implementacja, której ocena dzieli się na cztery etapy:

- obliczenie średniego błędu bezwzględnego (MAE) i pierwiastka błędu średniokwadratowego (RMSE) dla całego zbioru testowego,
- obliczenie błędów dla poszczególnych prowincji,
- prezentacja predykcji dla przykładowego tygodnia ze zbioru testowego,
- prezentacja 25 najważniejszych atrybutów.

4.5 Prace eksperymentalne

Model nr 1

Podstawowy model

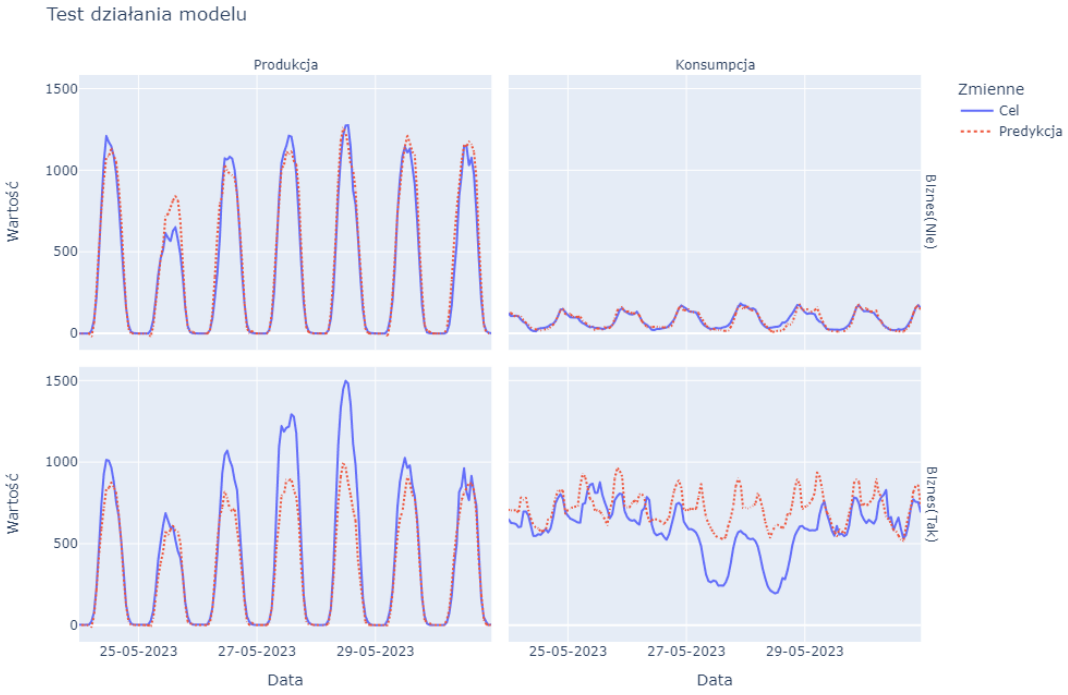
Pierwszym przeprowadzonym eksperymentem było zastosowanie modelu XGBoost (ze względu na szybszy proces uczenia) i użycie wszystkich dostępnych atrybutów. Takie podejście gwarantuje relatywnie dobre wyniki uwzględniając niewielki nakład pracy. Wyniki błędów modelu nr 1 zostały przedstawione w tabelach 14–15, a przykładowa predykcja tygodnia i wagi atrybutów na wykresach 21–22. Na podstawie wykresu dla przykładowego tygodnia można zauważyć, że model najsłabiej radzi sobie z przewidywaniem konsumpcji dla grupy prosumentów biznesowych.

Zbiór	MAE	RMSE
Uczący	55,79853	206,19108
Testowy	127,30929	484,04636

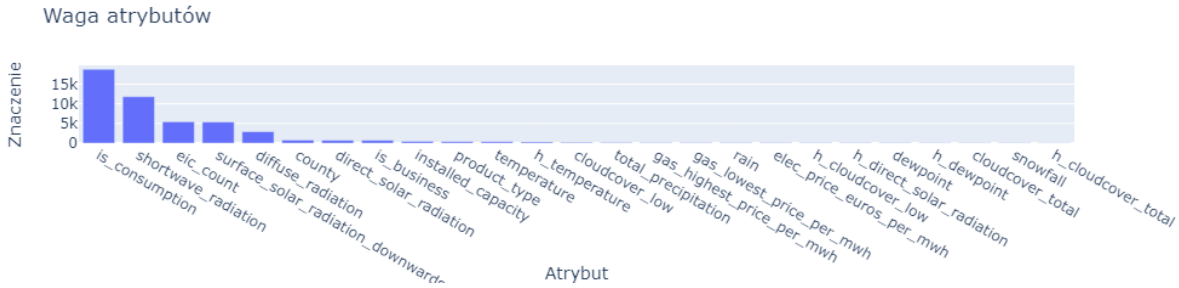
Tabela 14. Wyniki błędów dla modelu nr 1

Prowincja	MAE	RMSE
HARJUMAA	14,604288	213,285216
HIIUMAA	0,411312	0,169177
IDA-VIRUMAA	2,633005	6,932717
JÄRVAMAA	3,599736	12,958100
JÕGEVAMAA	7,157673	51,232282
LÄÄNE-VIRUMAA	4,603243	21,189846
LÄÄNEMAA	11,634515	135,361930
PÄRNUMAA	44,061362	1941,403653
PÕLVAMAA	8,635167	74,566101
RAPLAMAA	9,594132	92,047365
SAAREMAA	8,888839	79,011461
TARTUMAA	9,046372	81,836848
VALGAMAA	196,922743	38778,566870
VILJANDIMAA	26,615459	708,382684
VÕRUMAA	1,791367	3,208996

Tabela 15. Wyniki błędów poszczególnych prowincji dla modelu nr 1



Rysunek 21. Predykcja dla przykładowego tygodnia – model nr 1



Rysunek 22. Znaczenie atrybutów – model nr 1

Model nr 2

Dodanie do modelu rodzaju dnia i wcześniejszych wyników

W kolejnym przeprowadzonym eksperymencie została podjęta próba dodania do modelu atrybutów pozwalających lepiej modelować trend danych i uwzględniać dane wynikające z rodzaju dnia. Analiza danych wykazuje, że produkcja i konsumpcja przyjmuje odmienny przebieg dla dni od poniedziałku do piątku (dni robocze), oraz od soboty do niedzieli (weekend). Dodatkowo różnica jest również widoczna dla dni, które są ustawowymi dniami wolnymi w Estonii. Z tego powodu do modelu zostały dodane dwa nowe atrybuty:

- is_holiday - flaga oznaczająca, czy dany dzień jest ustawowo dniem wolnym w Estonii,
- is_weekend - flaga oznaczająca, czy dany dzień wypada w sobotę bądź niedzielę.

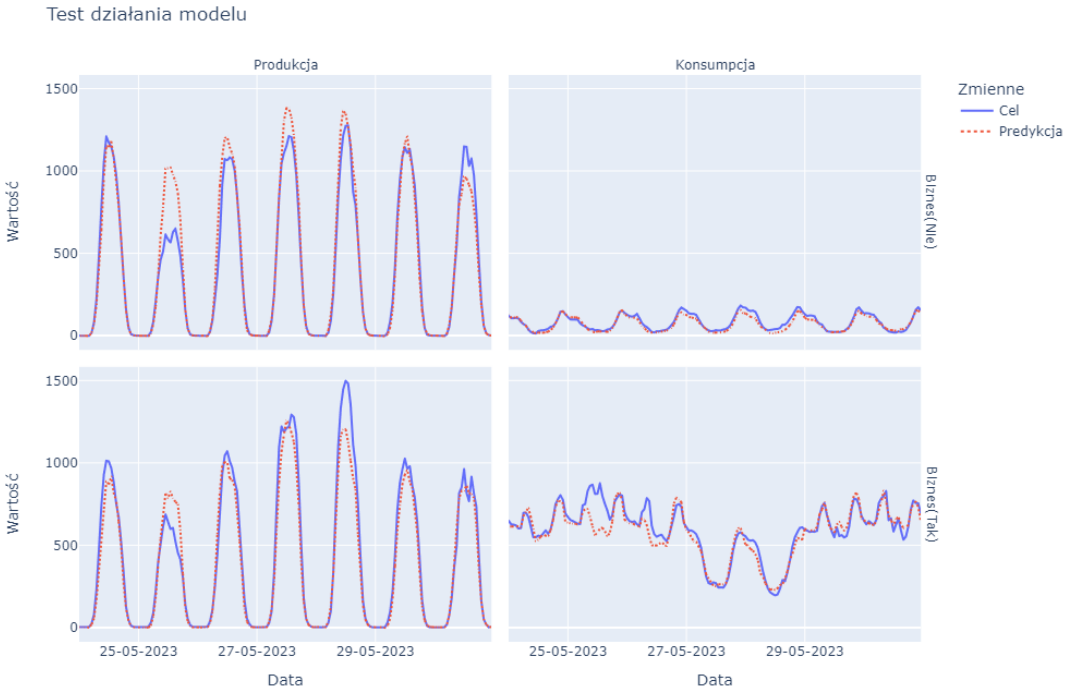
Pomocnymi atrybutami z perspektywy modelowania zmienności są również wartości zmiennej opisywanej, które są już dostępne dla użytkownika. Z tego powodu została dodana grupa atrybutów „target_..._days_ago", które zawiera informacje jaka była wartość cechy w analogicznym czasie od 2 do 14 dni temu. Wyniki błędów modelu nr 2 zostały przedstawione w tabelach 16–17, a przykładowa predykcja tygodnia i wagi atrybutów na wykresach 23–24. Na podstawie wyników można stwierdzić, że opracowanie tego modelu poskutkowało znaczącym obniżeniem błędów dla wszystkich grup prosumentów na poziomie spadku błędu MAE z 127,3 do 74,9 jednostek. Wprowadzenie nowych cech znacząco zmodyfikowało również wykres przedstawiający wagi atrybutów, co dodatkowo podkreśla ich znaczenie z perspektywy efektywności rozwiązania.

Zbiór	MAE	RMSE
Uczący	37,26217	136,62682
Testowy	74,90508	304,51557

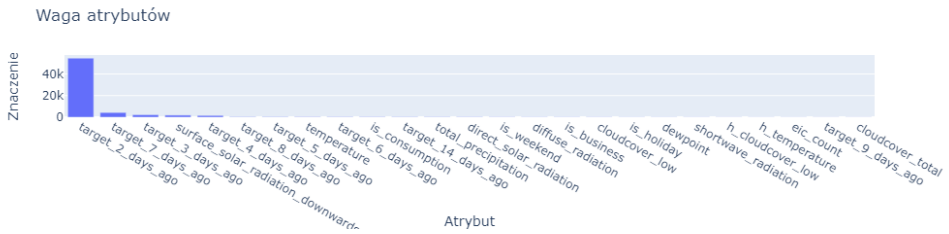
Tabela 16. Wyniki błędów dla modelu nr 2

Prowincja	MAE	RMSE
HARJUMAA	7,191905	51,723496
HIIUMAA	0,479880	0,230285
IDA-VIRUMAA	0,833809	0,695237
JÄRVAMAA	10,551148	111,326716
JÕGEVAMAA	4,186760	17,528963
LÄÄNE-VIRUMAA	4,700697	22,096549
LÄÄNEMAA	3,266753	10,671674
PÄRNUMAA	17,924485	321,287176
PÕLVAMAA	3,589125	12,881821
RAPLAMAA	4,840765	23,433008
SAAREMAA	5,878237	34,553671
TARTUMAA	0,095384	0,009098
VALGAMAA	166,604765	27757,147817
VILJANDIMAA	5,665343	32,096108
VÕRUMAA	3,886269	15,103090

Tabela 17. Wyniki błędów poszczególnych prowincji dla modelu nr 2



Rysunek 23. Predykcja dla przykładowego tygodnia – model nr 2



Rysunek 24. Znaczenie atrybutów – model nr 2

Model nr 3

Uwzględnienie sezonowości

Produkcja i konsumpcja naniesiona na wykres 13 wyraźnie pokazuje, że jej przebieg charakteryzuje się sezonowością zarówno w obrębie godzin, jak i miesięcy. Aby kształtować taki szereg pomocne jest dodanie atrybutów zawierających informacje o aktualnej porze dnia, tygodnia, miesiąca i roku. Niemniej jednak zalecane jest również zwrócenie uwagi na prawidłowe kodowanie takich atrybutów. Transformacja cech na zmienne cykliczne pozwala modelowi precyzyjniej interpretować informacje ze względu na zachowanie ciągłości danych zawierających sąsiednie wartości. Z tego powodu dodane

zmienne zostały poddane transformacji za pomocą funkcji sinus i cosinus zgodnie z wzorami 11–12:

$$\mathrm{x}_{\sin} = \sin\left(\frac{2\pi \mathrm{x}}{\max(\mathrm{X})}\right), \tag{11}$$

$$\mathrm{x}_{\cos} = \cos\left(\frac{2\pi \mathrm{x}}{\max(\mathrm{X})}\right). \tag{12}$$

Ostatecznie do zbioru atrybutów zostały dodane cechy z informacjami o godzinie, miesiącu, dniu tygodnia, dniu roku i roku po wcześniejszym dokonaniu transformacji na zmienne cykliczne (z pominięciem roku):

datetime_hour_sin	godzina rekordu – składowa sinus	$\max(\mathrm{X}) = 24$
datetime_hour_cos	godzina rekordu – składowa cosinus	
datetime_month_sin	miesiąc rekordu – składowa sinus	$\max(\mathrm{X}) = 31$
datetime_month_cos	miesiąc rekordu – składowa cosinus	
datetime_weekday_sin	dzień tygodnia rekordu – składowa sinus	$\max(\mathrm{X}) = 7$
datetime_weekday_cos	dzień tygodnia rekordu – składowa cosinus	
datetime_yearday_sin	dzień roku rekordu – składowa sinus	$\max(\mathrm{X}) = 365$
datetime_yearday_cos	dzień roku rekordu – składowa cosinus	
datetime_year	rok rekordu	-

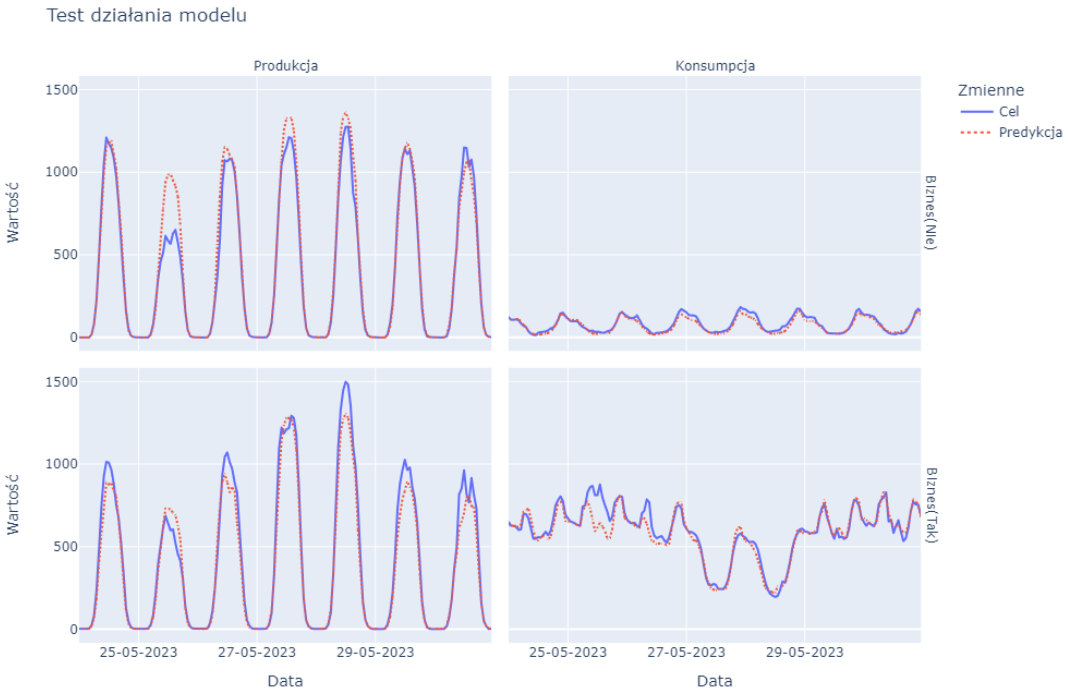
Wyniki błędów modelu nr 3 zostały przedstawione w tabelach 18–19, a przykładowa predykcja tygodnia i wagi atrybutów na wykresach 25–26. Podsumowując zebrane w nich informacje, można stwierdzić, że takie podejście zapewniło obniżenie średniego błędu bezwzględnego o 4 jednostki i pierwiastka błędu średnio-kwadratowego o ponad 1 jednostkę. Prawie wszystkie nowe atrybuty znalazły się na wykresie 25 najważniejszych atrybutów.

Zbiór	MAE	RMSE
Uczący	18,31389	70,45060
Testowy	70,45061	303,32660

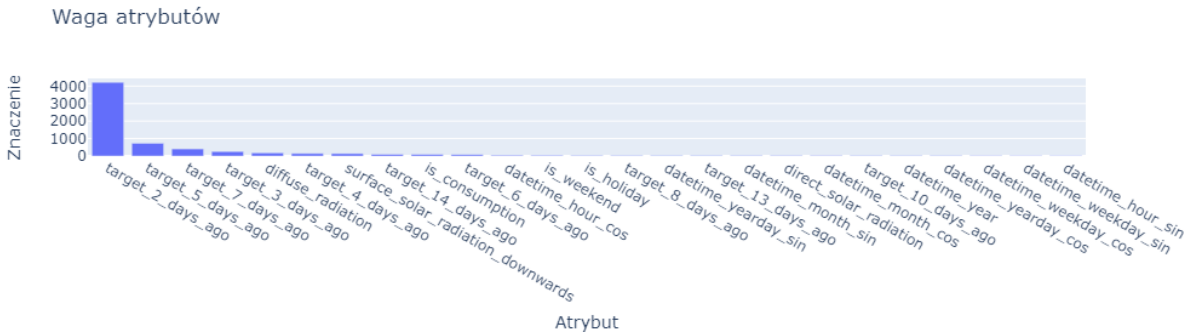
Tabela 18. Wyniki błędów dla modelu nr 3

Prowincja	MAE	RMSE
HARJUMAA	8,452923	71,451906
HIIUMAA	2,210350	4,885647
IDA-VIRUMAA	0,693257	0,480605
JÄRVAMAA	9,364120	87,686735
JÕGEVAMAA	2,976905	8,861963
LÄÄNE-VIRUMAA	7,299325	53,280146
LÄÄNEMAA	0,620745	0,385325
PÄRNUMAA	12,159288	147,848295
PÕLVAMAA	2,345044	5,499232
RAPLAMAA	7,385059	54,539102
SAAREMAA	11,882235	141,187499
TARTUMAA	5,485313	30,088661
VALGAMAA	197,633169	39058,869517
VILJANDIMAA	8,105995	65,707156
VÕRUMAA	3,118584	9,725566

Tabela 19. Wyniki błędów poszczególnych prowincji dla modelu nr 3



Rysunek 25. Predykcja dla przykładowego tygodnia – model nr 3



Rysunek 26. Znaczenie atrybutów – model nr 3

Model nr 4

Uwzględnienie problemu trendu

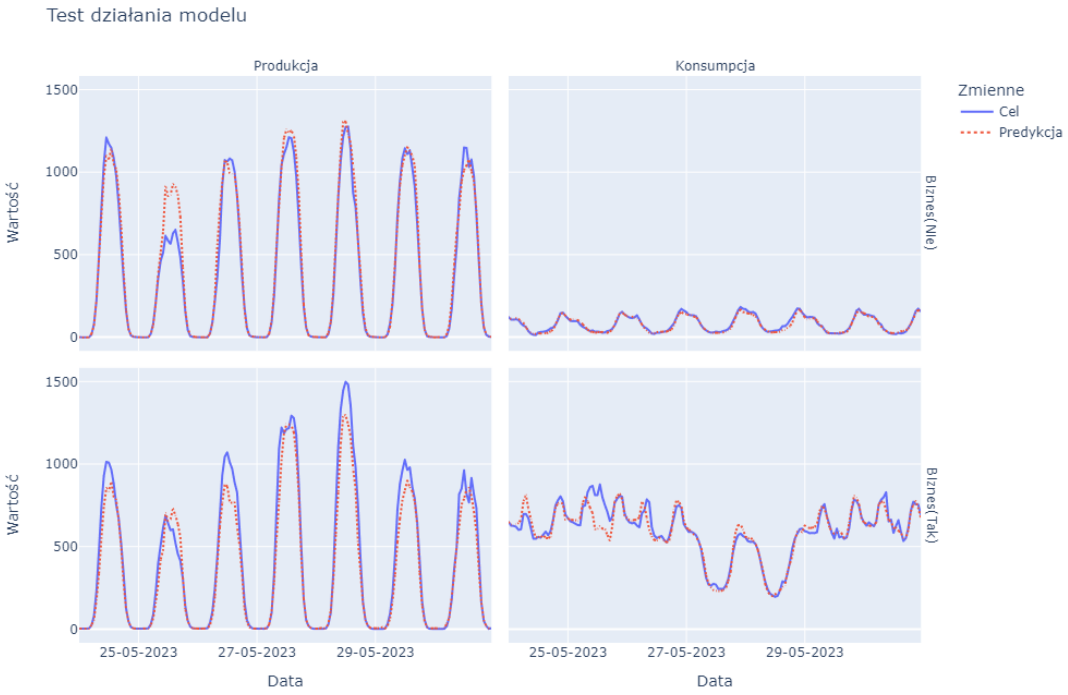
Uwzględnienie sezonowości w poprzednim modelu pozwoliło na zwiększenie dokładności otrzymywanych wyników. Analiza danych produkcji i konsumpcja na przestrzeni okresu, który zawiera dostępny zbiór danych wykazuje występowanie również innego komponentu szeregów czasowych – trendu rosnącego zmiennej przewidywanej. Bazując na aktualnej sytuacji na rynku i informacjach zawartych w eksploracyjnej analizie danych można stwierdzić wysokie prawdopodobieństwo wystąpienia sytuacji, w której zarówno produkcja i konsumpcja prosumentów będzie przyjmowała coraz większe wartości w każdym kolejnym roku. Skorzystanie z przeliczania wartości opisanego w analizie danych (3.2) pozwoliło na opracowanie rozwiązania, którego wyniki są jedynie minimalnie gorsze od wcześniejszych wersji, ale ma dużo większe prawdopodobieństwo do prawidłowego działania w dłuższej perspektywie czasu. Wyniki błędów modelu nr 4 zostały przedstawione w tabelach 20–21, a przykładowa predykcja tygodnia i wagi atrybutów na wykresach 27–28. Uzyskany błąd MAE wzrósł nieznacznie z 70,5 do 71.9 jednostek, a błędy poszczególnych prowincji pozostają w podobnych relacjach. Znaczenie poszczególnych atrybutów również zostało w większości zachowane w niezmienionej formie. Niemniej jednak obiektywne sprawdzenie poprawności wprowadzonych transformacji wymagałoby zbadania dokładności modelu na danych, które staną się dostępne w kolejnych latach.

Zbiór	MAE	RMSE
Uczący	29,47217	102,93714
Testowy	71,90585	306,09208

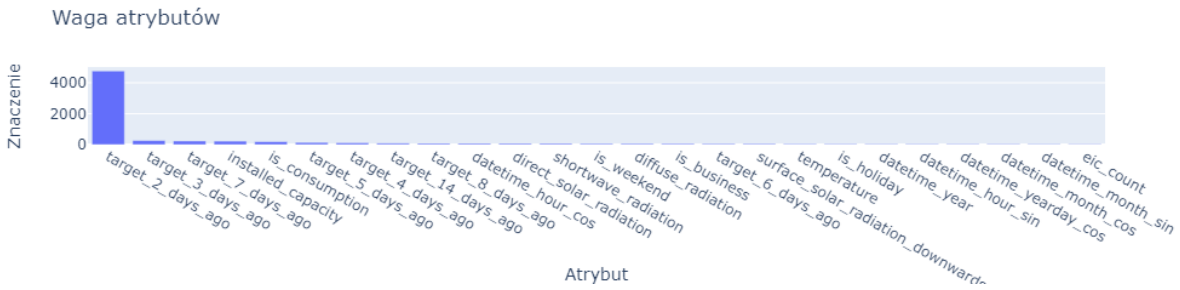
Tabela 20. Wyniki błędów dla modelu nr 4

Prowincja	MAE	RMSE
HARJUMAA	28,215380	796,107664
HIIUMAA	1,790438	3,205669
IDA-VIRUMAA	2,817064	7,935847
JÄRVAMAA	4,896966	23,980279
JÕGEVAMAA	3,534678	12,493950
LÄÄNE-VIRUMAA	1,596660	2,549323
LÄÄNEMAA	3,904000	15,241217
PÄRNUMAA	20,335660	413,539073
PÕLVAMAA	9,545223	91,111279
RAPLAMAA	1,949055	3,798815
SAAREMAA	7,286097	53,087205
TARTUMAA	7,324062	53,641890
VALGAMAA	192,847096	37190,002624
VILJANDIMAA	0,175824	0,030914
VÕRUMAA	3,133141	9,816571

Tabela 21. Wyniki błędów poszczególnych prowincji dla modelu nr 4



Rysunek 27. Predykcja dla przykładowego tygodnia – model nr 4

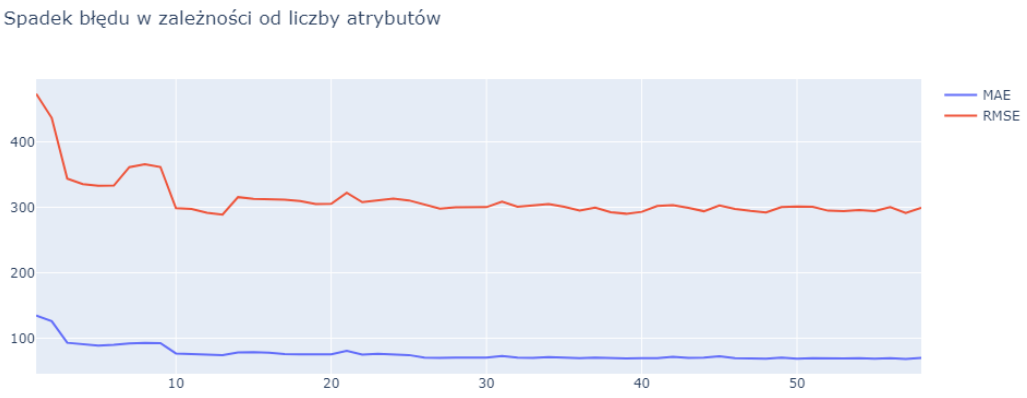


Rysunek 28. Znaczenie atrybutów – model nr 4

Model nr 5

Ograniczenie liczby atrybutów

Ważnym krokiem w tworzeniu praktycznego modelu jest ograniczenie liczby wymaganych atrybutów, tym bardziej, że wykresy wag atrybutów poprzednich metod uwidaczniają fakt, że tylko niewielka część z dostarczonych zmiennych ma znaczenie dla modelu. Założeniem kolejnego podejścia było ograniczenie liczby 59 wymaganych atrybutów modelu nr 4 do minimalnej liczby gwarantującej poprawne wyniki. Znalezienie optymalnej liczby cech zostało dokonane w oparciu o iteracyjne dodawanie jednego atrybutu do modelu ze zbioru najważniejszych cech wybranych przez poprzedni model. Wyniki wpływu liczby atrybutów na działanie rozwiązania zostało przedstawione na wykresie 29. Ostatecznie liczba zmiennych została ustalona na 36, co skutkowało skróceniem czasu uczenia i minimalnym polepszeniem rezultatów. Wyniki błędów modelu nr 5 zostały przedstawione w tabelach 22–23, a przykładowa predykcja tygodnia i wagi atrybutów na wykresach 30–31. Uzyskane wartości potwierdzają pozytywny wpływ ograniczenia złożoności modelu, ponieważ mniejsza liczba cech pozwoliła na zaobserwowanie spadku błędu MAE do najmniejszego uzyskanego poziomu.



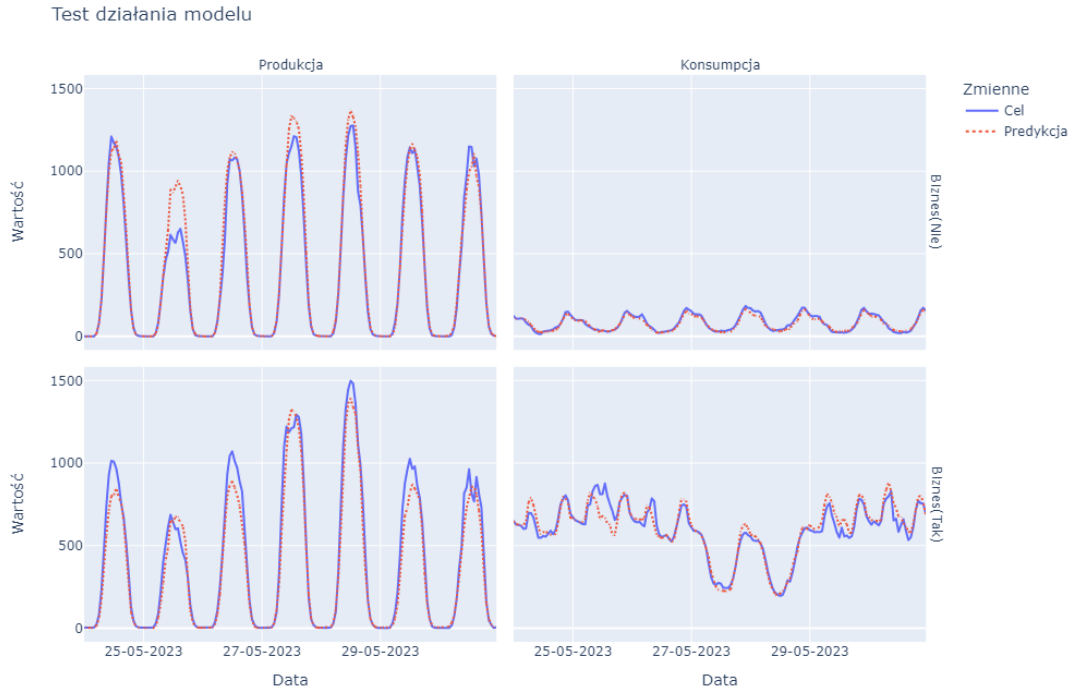
Rysunek 29. Zależność efektywności od liczby atrybutów

Zbiór	MAE	RMSE
Uczący	24,10513	83,67761
Testowy	70,435803	300,7492

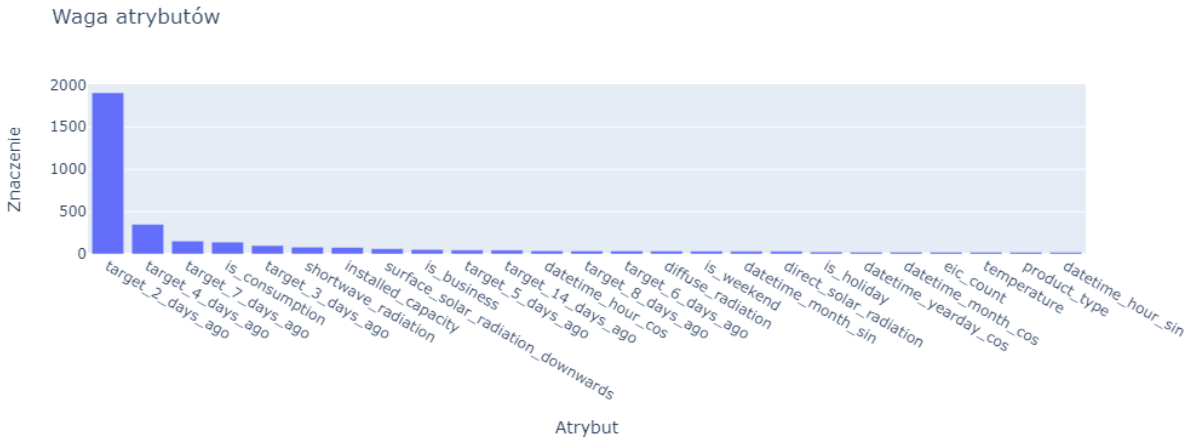
Tabela 22. Wyniki błędów dla modelu nr 5

Prowincja	MAE	RMSE
HARJUMAA	33,633818	1131,233697
HIIUMAA	1,001835	1,003674
IDA-VIRUMAA	2,634605	6,941142
JÄRVAMAA	12,351874	152,568799
JÕGEVAMAA	0,636802	0,405517
LÄÄNE-VIRUMAA	3,977651	15,821709
LÄÄNEMAA	4,243842	18,010198
PÄRNUMAA	14,985336	224,560295
PÕLVAMAA	7,373189	54,363923
RAPLAMAA	1,497105	2,241322
SAAREMAA	6,798903	46,225079
TARTUMAA	10,089873	101,805537
VALGAMAA	185,187117	34294,268381
VILJANDIMAA	3,602353	12,976948
VÕRUMAA	0,327891	0,107512

Tabela 23. Wyniki błędów poszczególnych prowincji dla modelu nr 5



Rysunek 30. Predykcja dla przykładowego tygodnia – model nr 5



Rysunek 31. Znaczenie atrybutów – model nr 5

Model nr 6

Model lasu losowego

Po opracowaniu odpowiednich atrybutów i rozwiązaniu problemu trendu dane zostały użyte do nauczenia modelu lasu losowego. Podobne podstawy teoretyczne działania obu modeli pozwalają użyć

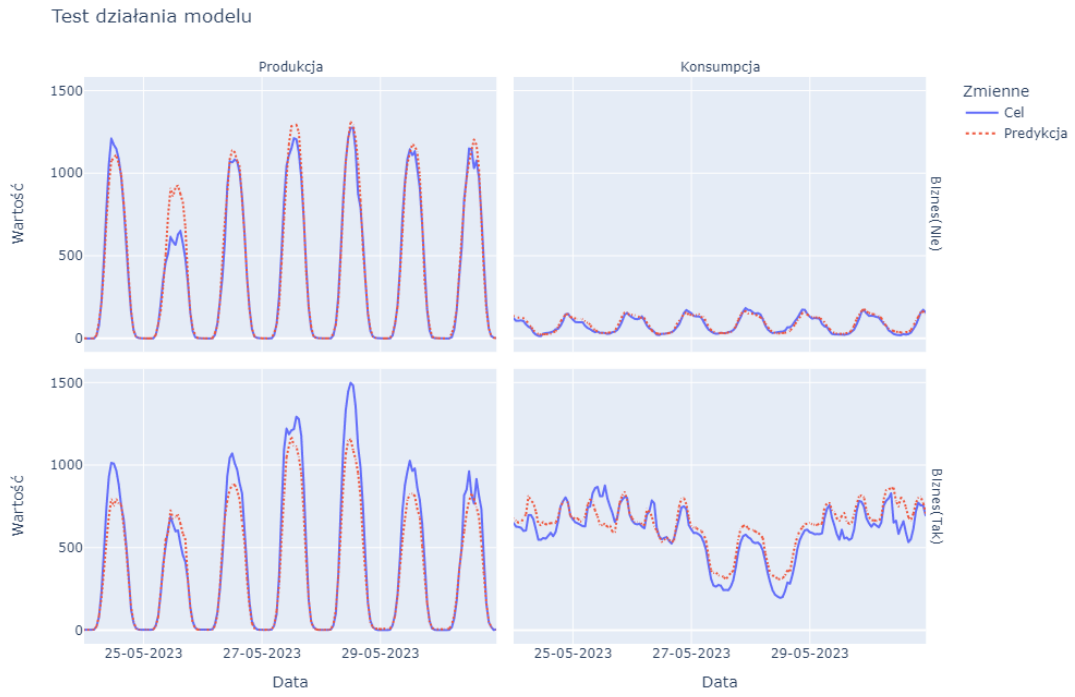
danych bez konieczności stosowania przekształceń. Wyniki błędów modelu nr 6 zostały przedstawione w tabelach 24–25, a przykładowa predykcja tygodnia i wagi atrybutów na wykresach 32–33. Opracowane rozwiązanie pogorszyło dokładność przewidywań i znacząco wydłużyło proces uczenia (stworzenie lasu losowego składającego się z 40 drzew decyzyjnych zajęło aż 32 minuty). Ten eksperyment podkreśla wydajność modelu xGBoost w przypadku złożonych danych. Interesujące są również niewielkie błędy dla zbioru uczącego, które mogą sugerować przeuczenie modelu, ale zastosowanie innych parametrów lasu losowego prowadziło do pogorszenia wyników w obrębie obu zbiorów.

Zbiór	MAE	RMSE
Uczący	9,41400	31,435570
Testowy	80,073272	323,132437

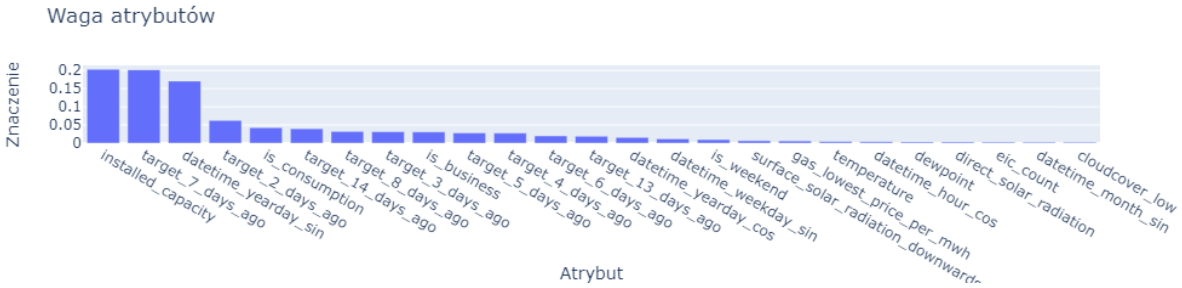
Tabela 24. Wyniki błędów dla modelu nr 6

Prowincja	MAE	RMSE
HARJUMAA	59,291682	3515,503562
HIIUMAA	1,618805	2,620529
IDA-VIRUMAA	1,895594	3,593277
JÄRVAMAA	16,642975	276,988628
JÕGEVAMAA	14,056564	197,586994
LÄÄNE-VIRUMAA	19,929387	397,180470
LÄÄNEMAA	7,122415	50,728798
PÄRNUMAA	1,568335	2,459675
PÕLVAMAA	6,699872	44,888283
RAPLAMAA	0,602460	0,362958
SAAREMAA	7,140509	50,986869
TARTUMAA	14,035563	196,997039
VALGAMAA	188,417704	35501,231257
VILJANDIMAA	4,074041	16,597813
VÕRUMAA	1,974001	3,896681

Tabela 25. Wyniki błędów poszczególnych prowincji dla modelu nr 6



Rysunek 32. Predykcja dla przykładowego tygodnia – model nr 6



Rysunek 33. Znaczenie atrybutów – model nr 6

Model nr 7

Osobny model produkcji i konsumpcji

Większość badanych powyżej implementacji podejmowała próbę ujęcia w jeden model przewidywania konsumpcji i produkcji. Niemniej odmienny charakter obu przebiegów i czynników, które wpływają na ich wartości skłania do podjęcia próby opracowania osobnego modelu obu szeregów. W tym celu zbiór uczący został podzielony na dwa podzbiory bazując na wartości atrybutu „is_consumption", a następnie został przeprowadzony analogiczny proces implementacji. Wyniki błędów modelu nr 7 zostały przedstawione w tabelach 26–27, a przykładowa predykcja tygodnia i wagi atrybutów na

wykresach 34–35. Analiza otrzymanych wyników pozwala stwierdzić, że przewidywana konsumpcja charakteryzuje się prawie dwukrotnie mniejszymi błędami MAE i RMSE od produkcji. Niemniej jednak średnia arytmetyczna z obu wyników prowadzi do otrzymania wartości tożsamej z wartością najlepszego rozwiązania ujmującego jednocześnie oba przebiegi pomimo, że przebieg predykcji dla przykładowego tygodnia jest dokładniej dopasowany. Prawdopodobną przyczyną takiego stanu jest fakt, że pojedynczy model jest zdolny do wystarczającego ujęcia złożoności danych produkcji i konsumpcji jednocześnie.

Zbiór	Cel	MAE	RMSE
Uczący	Konsumpcja	21,82356	67,02650
Testowy		43,951607	142,696597
Uczący	Produkcja	8,69249	39,33973
Testowy		93,13925	328,04627

Tabela 26. Wyniki błędów dla modelu nr 7

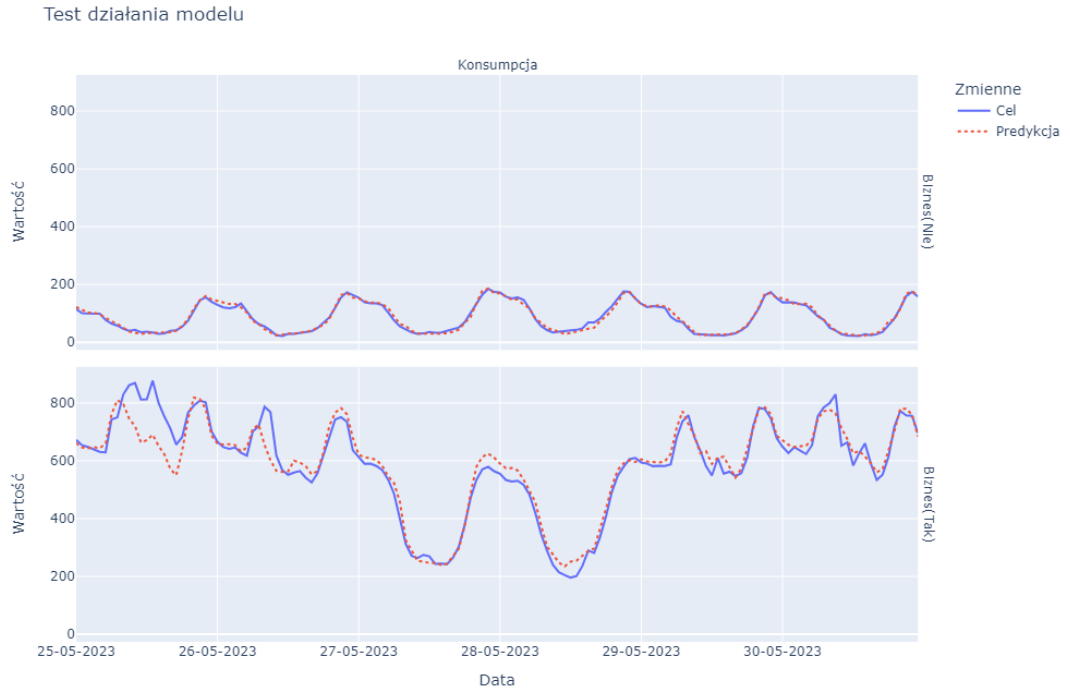
Prowincja	MAE	RMSE
HARJUMAA	12,206785	149,005594
HIIUMAA	2,445358	5,979774
IDA-VIRUMAA	12,856324	165,285058
JÄRVAMAA	4,661116	21,726003
JÕGEVAMAA	5,749766	33,059806
LÄÄNE-VIRUMAA	2,531535	6,408670
LÄÄNEMAA	10,975243	120,455961
PÄRNUMAA	23,312911	543,491803
PÕLVAMAA	8,778660	77,064873
RAPLAMAA	5,559451	30,907494
SAAREMAA	2,270648	5,155841
TARTUMAA	11,898124	141,565351
VALGAMAA	42,624202	1816,822589
VILJANDIMAA	0,107575	0,011572
VÕRUMAA	0,971442	0,943699

(a) Model konsumpcji

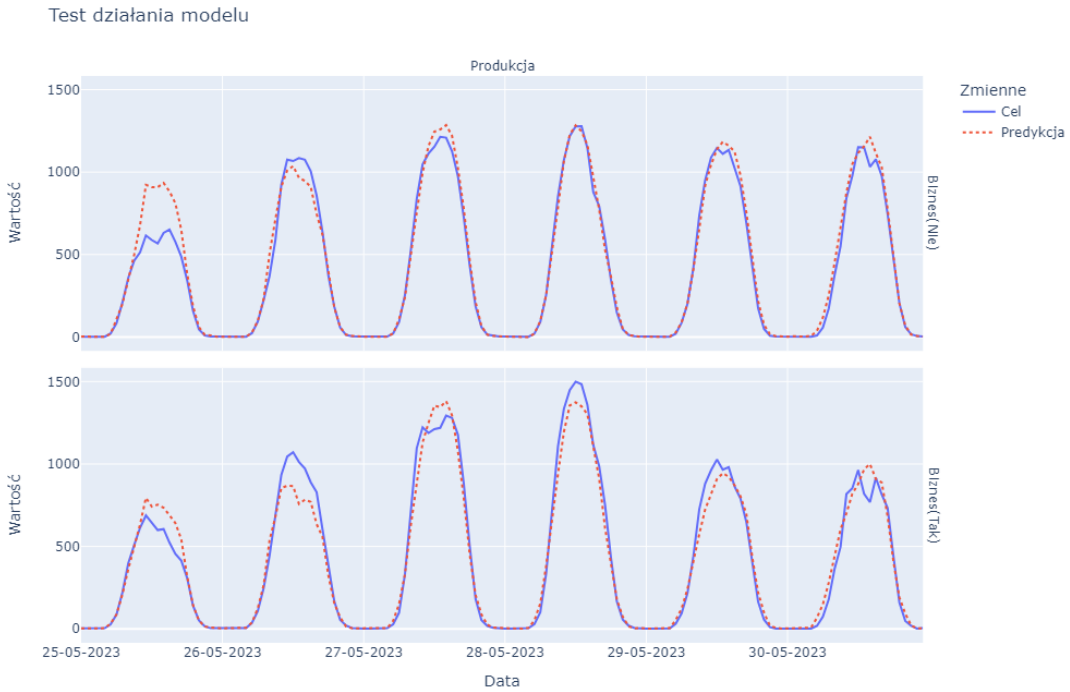
MAE	RMSE
37,931052	1438,764729
5,263081	27,700021
1,078122	1,162346
21,386381	457,377282
12,087119	146,098456
21,785619	474,613196
15,062066	226,865825
128,251749	16448,511185
6,459547	41,725749
6,148192	37,800263
22,884626	523,706097
47,752467	2280,298076
89,155803	7948,757139
15,989773	255,672842
4,986876	24,868929

(b) Model produkcji

Tabela 27. Wyniki błędów poszczególnych prowincji dla modelu nr 7

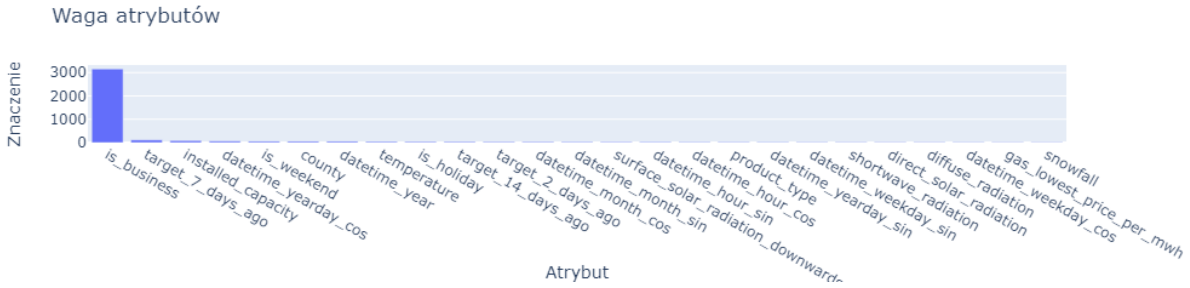


(a) Model konsumpcji energii

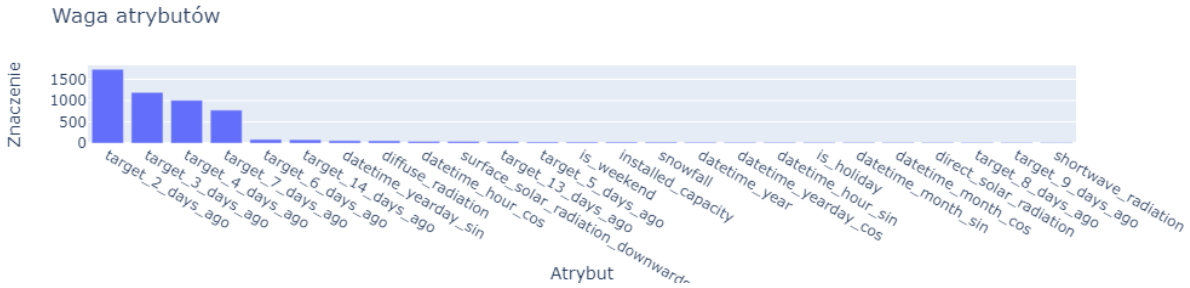


(b) Model produkcji energii

Rysunek 34. Predykcja dla przykładowego tygodnia – model nr 7



(a) Model konsumpcji energii



(b) Model produkcji energii

Rysunek 35. Znaczenie atrybutów – model nr 7

Model nr 8

Regresja liniowa

Regresja liniowa jest jednym z fundamentalnych podejść uczenia maszynowego i znajduje zastosowanie również w problemie poruszanym w niniejszej pracy. W celu użycia tego modelu zbiór uczący został prawidłowo spreparowany. Wszystkie atrybuty zostały przeskalowane do przedziału $[0, 1]$ zgodnie ze wzorem 13, aby umożliwić prawidłową interpretację istotności poszczególnych cech. Ze względu na niewrażliwość modelu na występowanie trendu, dane uczące nie podlegały innym transformacjom.

$$y_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{13}$$

Wyniki błędów modelu nr 8 zostały przedstawione w tabelach 28–29, a przykładowa predykcja tygodnia i wagi atrybutów na wykresach 36–37. Otrzymane wyniki błędów potwierdzają relatywnie dużą efektywność modelu regresji liniowej. Wyróżniającą się zaletą tego podejścia na tle reszty modeli jest prosty proces uczenia modelu, który sprowadza się do wyliczenia odpowiednich współczynników bez konieczności iteracyjnego poszukiwania rozwiązania optymalnego. Otrzymany ostatecznie model nie przewyższa najlepszego modelu opracowanego z użyciem systemu XGboost, ale wymaga podkreślenia, że tego typu podejście nie jest wrażliwe na problem ekstrapolacji danych, więc w dłuższej perspektywie czasu może prowadzić do zadowalających wyników bez konieczności dodatkowej weryfikacji, które wymagają modele oparte na drzewach decyzyjnych. Dodatkowo nie jest wykluczona możliwość, w której opracowanie dodatkowych atrybutów, które byłyby

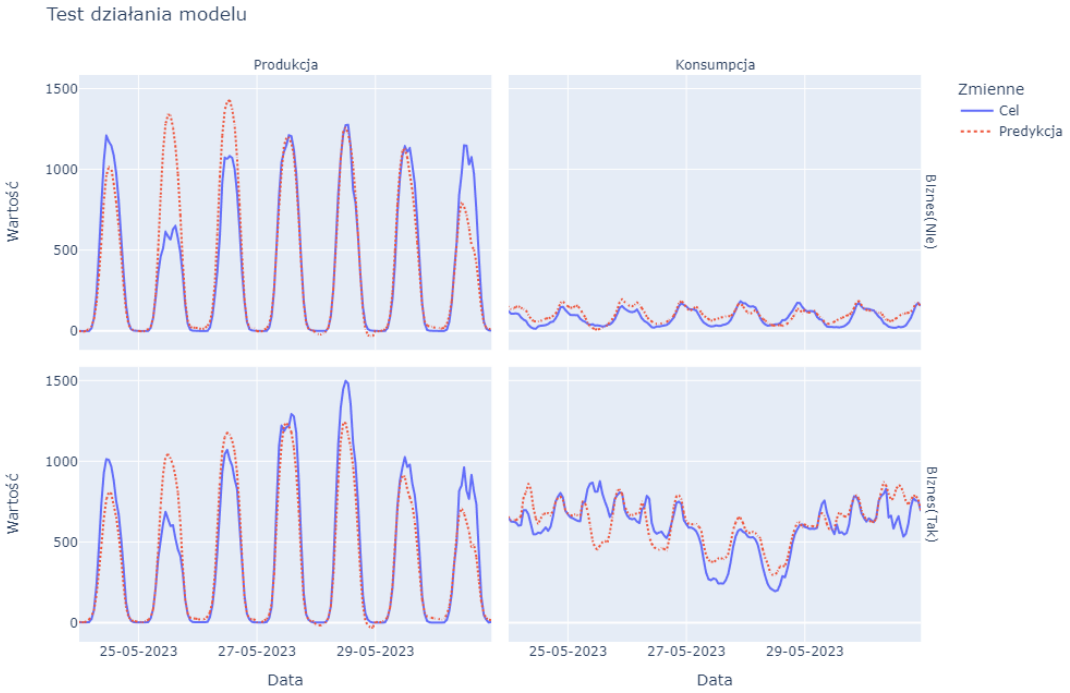
przekształceniem lub kombinacją już istniejących cech, pozwoliłoby na poprawienie otrzymywanych błędów. Podsumowując, model regresji liniowej zapewnia efektywne rozwiązanie w podstawowej formie i posiada potencjał dalszego rozwoju.

Zbiór	MAE	RMSE
Uczący	59,429758	188,30417
Testowy	100,87253	351,03840

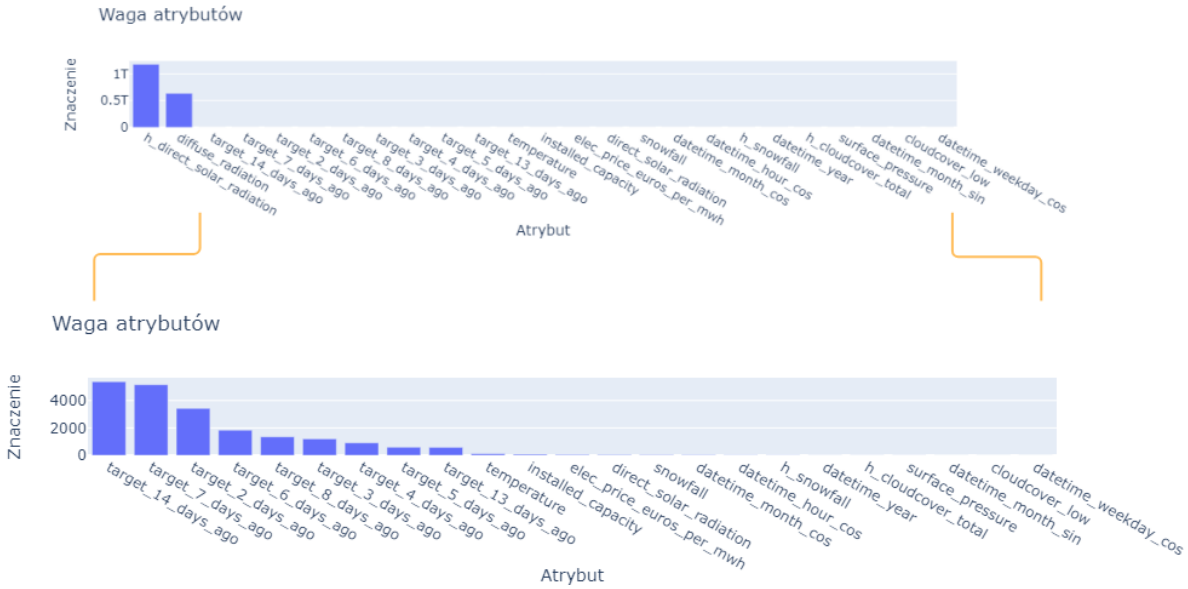
Tabela 28. Wyniki błędów dla modelu nr 8

Prowincja	MAE	RMSE
HARJUMAA	20,860494	435,160228
HIIUMAA	11,019624	121,432104
IDA-VIRUMAA	8,321978	69,255325
JÄRVAMAA	23,585352	556,268829
JÕGEVAMAA	17,479699	305,539884
LÄÄNE-VIRUMAA	18.758155	351,868366
LÄÄNEMAA	13,484913	181,842885
PÄRNUMAA	13,590023	184,688714
PÕLVAMAA	9,033674	81,607269
RAPLAMAA	8,943121	79,979408
SAAREMAA	8,073514	65,181633
TARTUMAA	12,301641	151,330372
VALGAMAA	151,480644	22946,385621
VILJANDIMAA	9,984729	99,694817
VÕRUMAA	7,668031	58,798703

Tabela 29. Wyniki błędów poszczególnych prowincji dla modelu nr 8



Rysunek 36. Predykcja dla przykładowego tygodnia – model nr 8

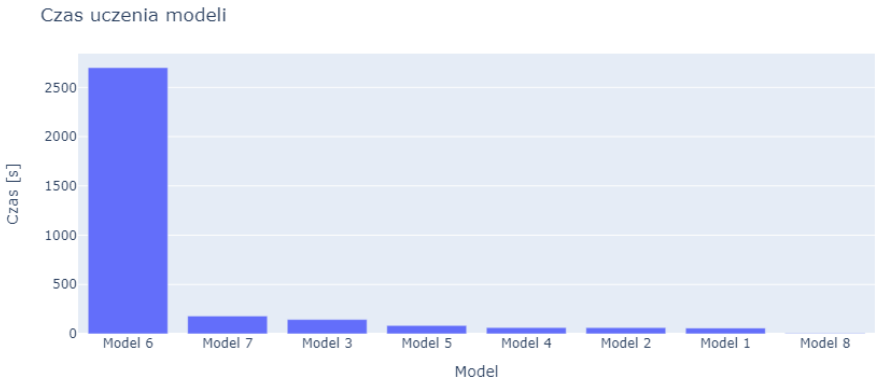


Rysunek 37. Znaczenie atrybutów – model nr 8

4.6 Wnioski prac eksperymentalnych

Założeniem opisanych prac eksperymentalnych było wyłonienie najefektywniejszej metody do przewidywania produkcji i konsumpcji podanej grupy estońskich prosumentów. Przeprowadzone badania objęły wiele różnych podejść stosowanych podczas użycia modeli opartych o drzewa decyzyjne (system XGBoost i las losowy) i podstawową regresję liniową. Najważniejsze opracowane rezultaty zostały opisane w formie ośmiu różnych rozwiązań, z których każde stanowiło istotny element ostatecznej konkluzji.

Wyniki błędów MAE i RMSE obliczone na zbiorach testowych pozwalają stwierdzić, że model nr 5 oparty o system XGBoost prezentuje najlepsze wyniki z błędem MAE na poziomie 70,4 i RMSE wynoszącym 300,7. Wybranie tej metody jest poparte również faktem ograniczenia liczby niezbędnych atrybutów do przeprowadzenia fazy uczenia i zastosowaniem transformacji danych uczących, która zwiększa stacjonarność badanego szeregu czasowego. Przeprowadzone przekształcenie podbudowuje perspektywy użycia modelu przez dłuższy czas bez konieczności przeprowadzania kolejnej fazy uczenia. Błędy w przypadku poszczególnych prowincji utrzymują się na podobnym poziomie z wyjątkiem przypadków Harjumaa i Valgamaa, których wyniki odbiegają od reszty obszarów w przypadku każdego podejścia. Prawdopodobną przyczyną tego stanu rzeczy jest odstający od standardowego przebieg konsumpcji i produkcji dla prosumentów z tych jednostek administracyjnych. Istotnym aspektem końcowego wyboru jest również długość procesu uczenia. Wykres 38 przedstawiający długość fazy uczenia poszczególnych algorytmów pozwala zauważyć podobny czas wszystkich systemów XGBoost. Negatywnie wyróżnia się czas wymagany przez las losowy (nr 6), który był wielokrotnie większy od reszty rozwiązań, a pozytywnie regresja liniowa (nr 8), której czas obliczenia optymalnych współczynników wyniósł zaledwie 5,4 s. Opisane wyniki potwierdzają wydajność systemu XGBoost dla dużych zbiorów uczących i stanowią dodatkowy atut na rzecz wybranego modelu. Niemniej jednak warta szczególnej uwagi jest również regresja liniowa, która pomimo gorszych wyników błędów posiada potencjał ze względu na możliwość opracowania dodatkowych atrybutów i niewrażliwość na występowanie tendencji rozwojowej.



Rysunek 38. Długość uczenia poszczególnych modeli

Rozdział 5

Podsumowanie

Celem niniejszej pracy dyplomowej było podjęcie problematyki opracowania modelu przewidywania szeregów czasowych produkcji i konsumpcji energii przez użytkowników sieci elektroenergetycznej Estonii, którzy wytwarzają energię korzystając z własnych mikroinstalacji (prosumentów). Istotność badanego problemu jest uzasadniona transformacjami rynku energii i perspektywami dalszego rozwoju. Wiele opublikowanych raportów międzynarodowych organizacji wskazuję, że odnawialne źródła energii, zarządzane przez prosumentów w sposób niezależny od innych uczestników, będą stanowić coraz większy udział w całości produkowanej energii. W konsekwencji problemy spowodowane dużym nasyceniem mikroinstalacjami danego obszaru będą przybierać na znaczeniu, a prawidłowo przewidywane zachowanie prosumentów może stanowić istotne narzędzie w przeciwdziałaniu problemom i pozwalać na dalszy rozwój rynku energii.

Zrealizowane eksperymenty odmiennych koncepcji opracowania modelu predykcyjnego pozwoliły na porównanie efektywności implementacji opartych o drzewa decyzyjne (XGBoost, las losowy) z fundamentalnym podejściem regresji liniowej. Wszystkie przeprowadzone doświadczenia zawierają pośredni lub bezpośredni udział w uzyskaniu rozwiązania charakteryzującego się najlepszymi wynikami. Mnogość przeprowadzonych badań świadczy o złożoności przewidywanego problemu i konieczności zrozumienia wielu idei stojących za predykcją szeregów czasowych i uczenia maszynowego. Niemniej jednak uzyskany ostatecznie regresor stanowi narzędzie pozwalające z relatywnie dużą dokładnością przewidywać zachowanie danej grupy prosumentów Estonii. Istotnym czynnikiem, który musiałby zostać zbadany podczas używania przygotowanej metody jest przebieg tendencji rozwojowej przewidywanego szeregu czasowego. Przeprowadzona analiza wskazała bowiem wysokie prawdopodobieństwo wzrostu konsumpcji i produkcji prosumentów w każdym kolejnym roku. Spowodowana tym zmiana zakresu zmiennej opisywanej może prowadzić do problemu ekstrapolacji danych dotyczącego algorytmów opartych o drzewa decyzyjne. Zaproponowana transformacja danych uczących stanowi sposób na poprawienie stacjonarności przewidywanego procesu, ale poprawność przyjętych założeń wymaga potwierdzenia w dłuższej perspektywie czasu.

Problem ekstrapolacji danych i przeprowadzona eksploracyjna analiza danych może stanowić dalszą perspektywę rozwoju dla badanej problematyki. Korzystne z punktu widzenia efektywności

modelu w dłuższym okresie może być również zastosowanie metod opartych o sztuczne sieci neuronowe lub przeprowadzenie dalszych badań w zakresie transformacji produkcji i konsumpcji energii elektrycznej do procesu stacjonarnego. Pożądane może być również podjęcie próby opracowania dodatkowych atrybutów, które mogą wpływać na działanie mikroinstalacji i decyzji podejmowanych przez prosumentów w celu kompleksowego opisania zmiennej endogenicznej. Zwiększenie zakresu zbioru danych w przyszłości może także stworzyć możliwość zauważenia większej liczby zależności pomiędzy atrybutami. Warto zaznaczyć jednak, że proces produkcji i konsumpcji w skali całej grupy prosumentów charakteryzuje się pewną losowością, która nie zostanie objęta przez żaden model.

Podsumowując uzyskane rezultaty pracy, można jednoznacznie stwierdzić, że cel pracy został spełniony. Opracowane modele potwierdzają możliwość wykorzystania metod uczenia maszynowego do opracowania narzędzia zdolnego przewidywać zachowanie prosumentów. Zrealizowane prace mogą być istotnym krokiem w poprawieniu sytuacji rynku energii i stanowić fundament w opracowaniu analogicznych rozwiązań dla innych państw, które dysponują podobnymi danymi.

Bibliografia

[1] Alam, M. S. i Vuong, S. T., „Random Forest Classification for Detecting Android Malware", IEEE Computer Society, 2013.

[2] Alim, M., Ye, G.-H., Guan, P., Huang, D.-S., Zhou, B.-S. i Wu, W., „Comparison of ARIMA model and XGBoost model for prediction of human brucellosis in mainland China: a time-series study", *BMJ Open*, grud. 2020.

[3] Ansari, M. A., College, J. J. i Koderma, J., „Enviromental impacts of non-renewable energy sources", *IJCRT*, s. 57–60, mar. 2017.

[4] Belgiu, M. i Drăguţ, L., „Random forest in remote sensing: A review of applications and future directions", *ISPRS Journal of Photogrammetry and Remote Sensing*, s. 24–31, 2016.

[5] Brockwell, P. J. i Davis, R. A., „Introduction to Time Series and Forecasting, Second Edition", w Springer, 2016.

[6] Buradkar, V. T. i More, M., „Introduction to Machine Learning and Its Applications: A Survey", *Journal of Artificial Intelligence, Machine Learning and Soft Computing*, s. 9–10, 2020.

[7] Capper, T., Kukriakose, J. i Sharmina, M., „Impact of Energy Imbalance on Financial Rewards in Peer-to-Peer Electricity Markets", *IEEE*, maj 2022.

[8] Cerqueira, V., Torgo, L. i Mozetic, I., „Evaluating time series forecasting models: an empirical study on performance estimation methods", *Spriner*, lip. 2020.

[9] Che, D., Liu, Q., Rasheed, K. i Tao, X., „Decision tree and ensemble learning algorithms with their applications in bioinformatics", *NIH*, 2011.

[10] EEA, „Energy prosumers in Europe - Citizen participation in the energy transition", EEA, spraw. tech., sty. 2022.

[11] European Commission, „Study on "Residential Prosumers in the European Energy Union"", European Commission, spraw. tech., maj 2017.

[12] Ferreira, A. i Figueiredo, M., *Boosting Algorithms: A Review of Methods, Theory, and Applications*.

[13] Flach, P., „Machine Learning - The Art and Science of Algorithms that Make Sense of Data", w Cambridge, 2012.

[14] Gates, B., „How to Avoid a Climate Disaster: The Solutions We Have and the Breakthroughs We Need", w Agora, 2021, rozd. 2, s. 28–50.

[15] ——, „How to Avoid a Climate Disaster: The Solutions We Have and the Breakthroughs We Need", w Agora, 2021, rozd. 1, s. 9–28.

[16] Hassan, H. G., Shahin, A. A. i Ziedan, I. E., „Energy consumption forecast in peer to peer energy trading", *SN Applied Sciences*, lip. 2023.

[17] Hastie, T., Tibshirani, R. i Friedman, J., „The Elements of Statistical Learning", w Springer, 2001, rozd. 15, s. 587–602.

[18] Ho, T. K., „Random Decision Forests", *AT&T Bell Laboratories*, 1995.

[19] Internation Energry Agency, „Global Energy & CO2 Status Report - The latest trends in energy and emissions in 2018", *IEA*, 2019.

[20] Jose, J., „Introduction to time series analysis and its applications", *Resarch Gate*, sierp. 2022.

[21] Koteluk, O., Wartecki, A., Mazurel, S., Kołodziejczak, I. i Mackiewicz, A., „How Do Machines Learn? Artificial Intelligence as a New Era in Medicine", *Journal of Personalized Medicine*, s. 1–14, sty. 2021, <https://www.mdpi.com/2075-4426/11/1/32>, dostęp uzyskano 2023-12-17.

[22] McCarthy, J., „What is artificial intelligence?", *Stanford University*, list. 2007.

[23] Mitchell, T., „Machine Learning", w McGraw-Hill Science, 1997.

[24] Naber, N., Kampman, B., Scholten, T., Vendrik, J. i Water, S. van de, „Potential of prosumer technologies in the EU", CE Delft, spraw. tech., mar. 2021.

[25] Nordfjell, O. i Ring, G., „Investigating the Performance of Random Forest Classification for Stock Trading", *KTH Royal Institute of Technology*, 2023.

[26] Paliari, I., Karanikola, A. i Kotsiantis, S., „A comparison of the optimized LSTM, XGBOOST and ARIMA in Time Series forecasting", *Journal of Artificial Intelligence, Machine Learning and Soft Computing*, lip. 2021.

[27] Popławski, T., Dudzik, S. i Szeląg, P., „Forecasting of Energy Balance in Prosumer Micro-Installations Using Machine Learning Models", *MDPI*, wrz. 2023.

[28] Ritchie, H., Roser, M. i Rosado, P., „Renewable Energy", *Our World in Data*, 2020.

[29] Schmela, M. i Solar Power Europe, „EU Market Outlook for Solar Power 2022-2026", Solar Power Europe, spraw. tech., maj 2022.

[30] Steurer, M., Robert J. Hill i Pfeifer, N., „Metrics for Evaluating the Performance of Machine Learning", IARIW, 2021.

[31] T. Chai i R. R Draxler, „Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature", *Copernicus Publications*, s. 2–3, czer. 2014.

[32] *Udział prosumentów w rozwoju sieci smart grid*, Politechnika Rzeszowska, grud. 2015.

[33] Washington, U. of, *Understanding the Bias-Variance Tradeoff*.

[34] Wasiak, I., „ELEKTROENERGETYKA W ZARYSIE - Przesył i rozdział energii elektrycznej", w Politechnika Łódzka, 2010, rozd. 1, s. 11–19.

[35] Wolny, R., „Prosumpcja i prosument na rynku e-usług", w Uniwersytet Ekonomiczny w Katowicach, 2013, s. 152–153.

[36] *XGBoost: A Scalable Tree Boosting System*, University of Washington, 2016.

[37] Zhang, H., Nettleton, D. i Zhu, Z., „Regression-Enhanced Random Forests", *JSM*, 2017.

Wykaz skrótów i symboli

EEA European Environment Agency 14, 15

EIC European Indetifier Code 39

ER Energetyka Rozproszona 10

ISEP Instytut Sterowania i Elektroniki Przemysłowej 1

MAE Mean Absolute Error 26, 51, 54, 58, 60, 65, 71

OZE Odnawialne Źródła Energii 10, 15

RMSE Root Mean Square Error 26, 51, 65, 71

SEE System Elektroenergetyczny 10, 13, 16

SI Sztuczna Inteligencja 16, 20

UE Unia Europejska 14, 15

UM Uczenie Maszynowe 16

Spis rysunków

Spis tabel