

Politechnika Warszawska

W Y D Z I A Ł M E C H A T R O N I K I



Instytut Automatyki i Robotyki

Praca dyplomowa magisterska

na kierunku Automatyka, Robotyka i Informatyka Przemysłowa

Przygotowanie zbioru danych wejściowych do opracowania modelu do predykcji cen energii elektrycznej na rynku dnia następnego w Polsce

inż. Ivan Kaliankovich

numer albumu 335662

promotor

prof. dr inż. Paweł Wnuk

WARSZAWA 2025

PRACA DYPLOMOWA magisterska	
Kierunek studiów/specjalność: Automatyka, Robotyka i Informatyka Przemysłowa	
Instytut dyplomujący: Instytut Automatyki i Robotyki	
Temat pracy: Przygotowanie zbioru danych wejściowych do opracowania modelu do predykcji cen energii elektrycznej na rynku dnia następnego w Polsce	
Zakres pracy: 1. Analiza i przygotowanie zbioru danych do prognozowania cen energii elektrycznej na Polskim rynku dnia następnego 2. Wykorzystanie modeli statystycznych i uczenia maszynowego 3. Ocena skuteczności modeli oraz przydatności zbioru danych	
Literatura: 1. Rafał Weron "Electricity price forecasting: A review of the state-of-the-art with a look into the future" 2. K. Maciejowska, B. Uniejewski, R. Weron: „Forecasting Electricity Prices” 3. Sean J. Taylor, Benjamin Letham: “Forecasting at Scale”	
Słowa kluczowe: Rynek dnia następnego, predykcja cen energii, metody modelowania	
Praca dyplomowa jest realizowana we współpracy z przemysłem	
Tak/Nie *,	
Nazwa firmy:	
<i>Ivan Kalianovich</i> Ivan Kalianovich	<i>Paweł Wnuk</i> Paweł Wnuk
Imię i nazwisko konsultanta:	
Temat wydano dnia: 01 05 2024	Termin ukończenia pracy: 11 05 2025
Zatwierdzenie tematu	
<i>Ivan Kalianovich</i> Opiekun kierunku/specjalności	<i>Paweł Wnuk</i> Z-ca Dyrektora Instytutu
Przyjęcie zakresu pracy	
<i>Ivan Kalianovich</i> Student	<i>Paweł Wnuk</i> Promotor

*Niepotrzebne skreślić

Przygotowanie zbioru danych wejściowych do opracowania modelu do predykcji cen energii elektrycznej na rynku dnia następnego w Polsce

Streszczenie

Niniejsza praca poświęcona jest analizie i prognozowaniu cen energii elektrycznej na Rynku Dnia Następnego w Polsce w latach 2016-2023. Badanie skupia się na zrozumieniu dynamiki cen na tym rynku, który odgrywa kluczową rolę w zarządzaniu systemem elektroenergetycznym, umożliwiając elastyczne dostosowanie podaży do popytu w krótkim horyzoncie czasowym. Centralnym elementem pracy jest przygotowanie kompleksowej bazy danych, która obejmuje zarówno okresy stabilności cenowej, jak i momenty wysokiej zmienności, wynikające z czynników ekonomicznych, geopolitycznych oraz pogodowych. Uwzględniono szeroki zestaw danych, takich jak historyczne ceny energii, zapotrzebowanie, generacja energii z odnawialnych źródeł, ceny paliw, emisje CO₂, a także warunki atmosferyczne, co pozwoliło na holistyczne podejście do analizy.

Dobór zmiennych do analizy został oparty na gruntownym przeglądzie literatury, który wskazał na znaczenie takich czynników, jak opóźnione ceny energii, zapotrzebowanie systemowe, generacja energii odnawialnej oraz zmienne sezonowe i ekonomiczne, w tym efekty kalendarzowe i wahania cen surowców. Uwzględnienie tych aspektów umożliwiło dostosowanie badania do specyficznych uwarunkowań polskiego rynku energii, który charakteryzuje się unikalnymi wyzwaniami, takimi jak zależność od tradycyjnych źródeł energii i rosnąca rola odnawialnych źródeł energii. Analiza została przeprowadzona w dwóch różnych okresach, co pozwoliło na ocenę skuteczności stosowanych metod w zróżnicowanych warunkach rynkowych, od stabilnych po te o dużej niestabilności.

W badaniu wykorzystano zarówno modele statystyczne, jak i techniki uczenia maszynowego, dostosowując ich parametry, aby zwiększyć zdolność do przewidywania cen. Wyniki poddano ocenie za pomocą standardowych miar, analizując wpływ różnych zmiennych na jakość prognoz oraz porównując efektywność modeli w zależności od użytych danych. Praca rzuca światło na trudności związane z prognozowaniem w niestabilnych warunkach rynkowych, szczególnie w kontekście nagłych zmian cen wywołanych czynnikami zewnętrznymi, i wskazuje na konieczność dalszego rozwoju zaawansowanych metod analitycznych, takich jak podejście hybrydowe, które lepiej radzą sobie z nielinowością danych. Opracowanie może stanowić wsparcie dla uczestników rynku energii w podejmowaniu decyzji handlowych, zarządzaniu ryzykiem finansowym oraz planowaniu strategii w dynamicznie zmieniającym się środowisku energetycznym.

Słowa kluczowe: Rynek Dnia Następnego, prognozowanie cen energii, zbiór danych

Preparation of an input dataset for the development of a model for the prediction of electricity prices in the day-ahead market in Poland

Abstract

The following study focuses on the analysis and forecasting of electricity prices on the Day-Ahead Market in Poland over the period 2016-2023. The study aims to understand the price dynamics on this market, which plays a crucial role in managing the power system by enabling flexible adjustments of supply to demand in the short term. A central element of the research is the development of a comprehensive database covering both periods of price stability and high volatility, driven by economic, geopolitical, and weather-related factors. The dataset includes a wide range of variables such as historical electricity prices, demand, generation from renewable energy sources, fuel prices, CO₂ emissions, and weather conditions, facilitating a holistic approach to the analysis.

The selection of variables was guided by an in-depth literature review, highlighting the importance of factors such as lagged energy prices, system demand, renewable energy generation, and seasonal and economic variables, including calendar effects and raw material price fluctuations. Incorporating these aspects allowed the study to address the specific conditions of the Polish energy market, which faces unique challenges, such as reliance on traditional energy sources and the growing role of renewables. The analysis was conducted across two distinct periods, enabling an evaluation of model performance under varying market conditions, from stable to highly volatile.

The study employed both statistical models and machine learning techniques, adjusting their parameters to enhance predictive capabilities. The results were assessed using standard metrics, analyzing the impact of different variables on forecast accuracy and comparing model effectiveness based on the datasets used. The work sheds light on the challenges of forecasting in unstable market conditions, particularly in the context of sudden price changes triggered by external factors, and underscores the need for further development of advanced analytical methods, such as hybrid approaches, which better handle data nonlinearity. This research may support market participants in making trading decisions, managing financial risk, and planning strategies in a dynamically changing energy environment.

Keywords: Day-Ahead Market, electricity price forecasting, dataset

Spis treści

1 Wstęp	11
2 Przegląd literatury	14
2.1 Zmienne wejściowe	14
2.2 Modele prognozowania	16
3 Dane i zmienne	19
3.1 Zmienna zależna	19
3.2 Zbiór zmiennych niezależnych	21
3.2.1 Dane pogodowe	21
3.2.2 Produkcja energii z wybranych źródeł	27
3.2.3 Handel energią z państwami sąsiednimi	28
3.2.4 Ceny paliw kopalnych i emisji CO ₂	30
3.2.5 Straty mocy w systemie elektroenergetycznym	32
3.2.6 Zapotrzebowanie na energię i wolumen handlu	33
3.2.7 Rynek bilansujący	34
3.2.8 Inne zmienne	35
4 Eksploracja danych	39
4.1 Wstępna obróbka danych	39
4.2 Analiza korelacji	40
4.3 Podział danych	44
4.4 Przygotowanie danych	45
5 Metodologia	47
5.1 Ocena jakości prognoz	47
5.1.1 Mean Absolute Error (MAE)	47
5.1.2 Root Mean Squared Error (RMSE)	48
5.1.3 Mean Absolute Percentage Error (MAPE)	49
5.1.4 Symmetric Mean Absolute Percentage Error (sMAPE)	49
5.1.5 Współczynnik determinacji	50
5.2 Wybrane metody weryfikacji zbioru danych	51

5.2.1	Regresja liniowa	51
5.2.2	Regresja grzbietowa (Ridge)	52
5.2.3	Prophet	53
5.2.4	Wielowarstwowy perceptron (MLP)	56
6	Estymacja przykładowych modeli	58
6.1	Okres stabilny	58
6.2	Okres niestabilny	67
7	Analiza wyników i podsumowanie	76
7.1	Analiza wyników	76
7.2	Podsumowanie	79
Bibliografia		81
Wykaz skrótów i symboli		83
Spis rysunków		84
Spis tabel		86

Rozdział 1

Wstęp

Wprowadzenie

Rynek energii elektrycznej to jeden z filarów współczesnej gospodarki, a jego sprawne funkcjonowanie ma kluczowe znaczenie dla stabilności systemów elektroenergetycznych, przedsiębiorstw i codziennego życia konsumentów. W centrum rynku znajduje się Rynek Dnia Następnego (RDN), który działa jako platforma handlu energią na dzień przed jej dostarczeniem. RDN jest miejscem, gdzie producenci energii od elektrowni węglowych po farmy wiatrowe spotykają się z odbiorcami, takimi jak dostawcy energii dla domów mieszkalnych czy duże zakłady przemysłowe, w celu ustalenia ceny i volumenów energii na każdą godzinę kolejnego dnia. Mechanizm działania RDN opiera się na systemie aukcyjnym: uczestnicy składają oferty kupna i sprzedaży, określając, ile energii są w stanie dostarczyć lub kupić oraz po jakiej cenie. System następnie dopasowuje te oferty, ustalając cenę równowagi, która obowiązuje dla danej godziny [9].

Taki model handlu pozwala na elastyczne reagowanie na zmieniające się warunki - zarówno po stronie podaży, jak i popytu. Na przykład, jeśli prognozy wskazują na silny wiatr, producenci energii wiatrowej mogą zwiększyć podaż, co może obniżyć ceny; z kolei fala upałów może zwiększyć zapotrzebowanie na energię do klimatyzacji, podnosząc ceny. RDN działa w wielu krajach, choć jego specyfika różni się w zależności od regionu. W Europie, w tym w Polsce, rynek ten jest częścią szerszego systemu integracji rynków energii, który ma na celu zapewnienie płynności i efektywności handlu. W Stanach Zjednoczonych RDN funkcjonuje w ramach regionalnych rynków, takich jak PJM Interconnection czy California ISO (CAISO), gdzie handel energią jest dodatkowo skomplikowany przez różnice regulacyjne między stanami. Niezależnie od regionu, RDN jest kluczowym narzędziem w zarządzaniu systemem elektroenergetycznym, umożliwiając szybkie dostosowanie podaży do popytu w krótkim horyzoncie czasowym [9].

Jednak handel na RDN to nie tylko szansa na zysk dla wszystkich biorących udział, ale i ogromne ryzyko finansowe, które wynika z nieprzewidywalności cen energii. Na rynku amerykańskim, gdzie mechanizm licytacji między kupującymi a sprzedającymi jest szczególnie rozwinięty, ryzyko to jest wyjątkowo widoczne. Uczestnicy rynku muszą składać oferty w czasie rzeczywistym, próbując przewidzieć, jak zachowa się cena w danej godzinie. Jeśli producent energii zaoferuje

zbyt wysoką cenę, jego energia może nie zostać kupiona, co oznacza utratę przychodów; z kolei kupujący, który zaoferuje zbyt niską cenę, może zostać zmuszony do zakupu energii po znacznie wyższej cenie rynkowej. Przykładem skali tego ryzyka jest kryzys w Teksasie w lutym 2021 roku [2]. W wyniku ekstremalnych mrozów i awarii systemu elektroenergetycznego ceny na rynku ERCOT (Electric Reliability Council of Texas) wzrosły do 9000 USD/MWh - poziomu, który dla wielu uczestników rynku oznaczał straty liczone w dziesiątkach milionów dolarów. W takich warunkach dokładna predykcja cen staje się nie tylko narzędziem do optymalizacji handlu, ale wręcz koniecznością, by uniknąć katastrofalnych strat. Na rynkach krajów rozwiniętych, gdzie Odnawialne Źródła Energii (OZE) odgrywają coraz większą rolę, ceny mogą spadać do wartości ujemnych w godzinach nadprodukcji np. z energii słonecznej, by kilka godzin później gwałtownie wzrosnąć, gdy zapotrzebowanie przewyższa podaż. Ta zmienność sprawia, że handel na RDN przypomina grę o wysoką stawkę, w której każdy ruch musi być dokładnie przemyślany.

Rynek bilansujący stanowi kolejny istotny element systemu elektroenergetycznego, uzupełniając funkcjonowanie RDN. Działa on w czasie rzeczywistym, umożliwiając operatorom systemu elektroenergetycznego (w Polsce jest to Polskie Sieci Elektroenergetyczne, PSE [15]) równoważenie podaży i popytu w sytuacjach, gdy rzeczywiste zużycie energii odbiega od prognoz ustalanych na RDN. Na rynku bilansującym uczestnicy mogą zgłaszać oferty na dostawy energii w bardzo krótkim horyzoncie czasowym, nawet w ciągu kilkunastu minut, co pozwala na szybkie reagowanie na nagłe zmiany, takie jak awarie elektrowni, nieoczekiwane skoki zapotrzebowania czy zmienność produkcji z OZE. Ceny na rynku bilansującym są często bardziej zmienne niż na RDN, co dodatkowo zwiększa ryzyko finansowe dla uczestników, ale jednocześnie podkreśla znaczenie precyzyjnych prognoz cen, które mogą pomóc w lepszym zarządzaniu tymi krótkoterminowymi wahaniami.

W niniejszej pracy magisterskiej skupiam się na analizie RDN w Polsce, który choć różni się od rynku amerykańskiego pod względem skali i regulacji, również zmaga się z podobnymi wyzwaniem. W Europie, w tym w Polsce, RDN jest częścią zintegrowanego systemu handlu energią, który opiera się na współpracy między krajami i dąży do harmonizacji rynków energii w ramach Unii Europejskiej. Mechanizm ustalania cen na RDN w Polsce opiera się na zasadzie jednolitej ceny, gdzie cena rozliczeniowa jest wyznaczana na podstawie równowagi popytu i podaży dla każdej godziny. Oznacza to, że wszyscy uczestnicy, których oferty zostały zaakceptowane, rozliczają się po tej samej cenie, co zapewnia przejrzystość i efektywność handlu. W Polsce RDN jest prowadzony przez Towarową Giełdę Energii (TGE), która od 2000 roku pełni rolę kluczowej platformy handlu energią. TGE organizuje aukcje na RDN, umożliwiając uczestnikom składanie ofert na każdą godzinę kolejnego dnia. W latach 2016-2023, które obejmują analizowany w pracy okres, ceny na RDN w Polsce wały się od 200 PLN/MWh aż do ponad 3000 PLN/MWh, co odzwierciedla zarówno lokalne uwarunkowania (np. zależność od węgla, rosnąca rola OZE), jak i globalne trendy. Ważną rolą w takiej zmienności odegrały także niespodziewane czynniki zewnętrzne, zwane czarnymi łabędziami. Z przykładów takich czynników można wymienić wybuch pandemii i wprowadzenie w związku z tym restrykcji na pracę oraz kryzys energetyczny w 2022 roku wywołany konfliktem zbrojnym na Ukrainie. Wszystkie te czynniki sprawiają, że prognozowanie cen energii elektrycznej na RDN w Polsce staje się niezwykle

złożonym zadaniem, które wymaga uwzględnienia wielu zmiennych i zastosowania zaawansowanych metod analizy danych.

Cel pracy

Większość dotychczasowych badań w literaturze naukowej dotycząca prognozowania cen energii elektrycznej, skupia się głównie na rynkach amerykańskich, europejskich oraz azjatyckich, pomijając specyfikę polskiego rynku energii. W związku z tym, celem niniejszej pracy jest opracowanie obszernej bazy danych z polskiego rynku energii, obejmującej zarówno okres umiarkowanej zmienności cen, jak i okres dużej zmienności cen i brak stabilności rynkowej. Baza danych zostanie przygotowana w celu uwzględnienia kluczowych czynników wpływających na ceny energii elektrycznej, charakterystycznych dla polskiego rynku. Następnie, w celu przetestowania użyteczności opracowanej bazy danych, przeprowadzono analizę z wykorzystaniem wybranych modeli uczenia maszynowego: regresji liniowej, regresji grzbietowej, modelu Prophet oraz wielowarstwowej sieci neuronowej.

Rozdział 2

Przegląd literatury

Rozdział przedstawia przegląd literatury dotyczącej prognozowania cen energii elektrycznej, ze szczególnym uwzględnieniem zmiennych wejściowych oraz zwykle używanych modeli prognozowania. W analizie odwołano się zarówno do prac profesora Rafała Werona z Politechniki Wrocławskiej, który jest uznany ekspertem w dziedzinie EPF, jak i do badań innych autorów, aby zapewnić kompleksowy kontekst dla przeprowadzonego badania.

2.1 Zmienne wejściowe

W przeglądzie literatury dotyczącym prognozowania cen energii elektrycznej na rynkach dnia następnego, artykuł napisany przez Jesus Lago o współautorstwie prof. Werona [6] może stanowić istotny punkt odniesienia pod względem wyboru zmiennych wejściowych. Autorzy przyjęli zestaw dostępnych danych wejściowych dla prognozowania godzinowych cen na rynkach energii elektrycznej Nord Pool, EPEX-BE, EPEX-FR, EPEX-DE oraz PJM. Niezależnie od zastosowanych przez autorów modeli (LEAR czy DNN), dla wszystkich rynków stworzono wektory zawierające po 24 wartości odpowiadające każdej godzinie dla cen z poprzednich trzech dni oraz tygodnia wstępco. Do tych wektorów dodano wektor określający dzień tygodnia w celu uwzględnienia sezonowości dziennej. Dodatkowo pod uwagę wzięto zmienne fundamentalne, różniące się w zależności od analizowanego rynku.

Dla rynku Nord Pool (NP) były to prognozy zapotrzebowania na moc oraz generacji energii wiatrowej.

Dla rynku PJM w Stanach Zjednoczonych uwzględniono dwie serie prognoz zapotrzebowania na moc: prognozę zapotrzebowania dla całego systemu oraz prognozę dla strefy Commonwealth Edison.

Dla rynku EPEX-BE w Belgii wykorzystano prognozę zapotrzebowania na moc we Francji oraz prognozę generacji we Francji, ponieważ wcześniejsze badania wykazują, że właśnie te dwie zmienne są najlepszymi predyktormi cen belgijskich.

Dla rynku EPEX-FR we Francji brano pod uwagę prognozę zapotrzebowania na moc oraz prognozę generacji.

Dla rynku EPEX-DE w Niemczech uwzględniono prognozę zapotrzebowania na moc oraz zagrego-

wane prognozy generacji energii wiatrowej i słonecznej.

Do wszystkich wymienionych zmiennych fundamentalnych dodano wektory godzinowe tych zmiennych na dzień oraz tydzień wstecz.

Kolejnym artykułem wartym uwagi jest artykuł skupiony na postprocesingu [8] w prognozowaniu cen energii elektrycznej metodami probabilistycznymi. Do osiągnięcia celu zostały wykorzystane następujące zmienne:

- Historyczne ceny energii elektrycznej z poprzednich jednego, dwóch, trzech i siedmiu dni
- Prognoza systemowego zapotrzebowania na moc
- Prognoza generacji energii z odnawialnych źródeł będąca sumą prognoz generacji energii wiatrowej i słonecznej
- Ceny uprawnień do emisji dwutlenku węgla
- Ceny gazu ziemnego
- Ceny ropy naftowej Brent
- Cena emisji węgla
- Zmienne czasowe reprezentujące dni tygodnia
- Zmienne czasowe, czyli zmienne binarne reprezentujące dni tygodnia

Warto podkreślić, że autorzy nie używali surowych danych do modelowania, lecz zastosowali transformację hiperbolicznego sinusa do zmiennych wyjściowych w celu stabilizacji wariancji przed modelowaniem. Transformacja ta jest opisana wzorem:

$$\sinh(x) = \frac{e^x - e^{-x}}{2}$$

gdzie e to liczba Eulera.

W innym artykule o współautorstwie prof. Werona [19] podkreśla się, że opóźnione ceny oraz zmienne określające obciążenie systemowe oraz strefowe, czyli zapotrzebowanie na moc w poszczególnych strefach, bądź w całym systemie energetycznym w stosunku do całej dostępnej mocy, mają bardzo istotny wpływ na prognozowanie cen. Korzystając ze zbiorów danych z rynków Nord Pool oraz rynku Brytyjskiego, autorzy użyli używali tych zmiennych do prognozowania cen energii elektrycznej.

Artykuł prof. Ziel [22] analizujący niemiecki RDN, potwierdza poprzednie tezy, wskazując na znaczenie cen energii opóźnionych, obciążenia sieci i energii z OZE. Opisujący przez niego model autoregresyjny opisuje każdą z tych zmiennych jako zależną od jej własnych opóźnionych wartości. Wprowadza skomplikowany zestaw cech opóźnionych jak i również opóźnione zależności niektórych z cech:

- Opóźnienia cen energii elektrycznej o 1, 2, ..., 361, 504, 505, 672, 673, 840, 841, 1008, 1009 godzin
- Opóźnienia obciążenia sieci o 1, 2, ..., 361, 504, 505, 672, 673, 840, 841, 1008, 1009 godzin
- Opóźnienia generacji energii z OZE o 1, 2, ..., 361 godzine
- Opóźnienia obciążenia zależnego od ceny

- Opóźnienia ceny zależnej od obciążenia o 1, 2, ..., 361, 504, 505, 672, 673, 840, 841, 1008, 1009 godzin
- Opóźnienia ceny zależnej od generacji energii z OZE o 1, 2, ..., 49 godzin
- Opóźnienia obciążenia zależnego od generacji energii z OZE o 1, 2, ..., 49 godzin

Oprócz opóźnionych wartości zmiennych, model uwzględnia również trend i efekty sezonowe, takie jak efekty czasowe i kalendarzowe, godziny w ciągu dnia, dni tygodnia, święta publiczne (w tym krajowe i regionalne), zmiany czasu letniego.

Innym przykładem jest praca inżynierska napisana przez Mikołaj Kalisz i Adam Mantiuk na Politechnice Warszawskiej [5]. Analizując rynek energii elektrycznej w Polsce w latach 2021-2023, autorzy analizują zmienne średnią temperaturę godzinową w Polsce, handel międzynarodowy, kurs polskiego złotego względem euro oraz dolara oraz poziom inflacji miesiąc do miesiąca i rok do roku oraz rezerwy mocy ponad i poniżej zapotrzebowania. Po analizie korelacji z wielkiego zestawu cech wybierają dziesięć zmiennych, które mają największy wpływ na cenę energii w wybranym okresie. Są to: produkcja energii z fotowoltaiki, rezerwa mocy poniżej zapotrzebowania, cena uprawnień emisyjnych (CO₂), inflacja w porównaniu do poprzedniego roku, cena gazu oraz historyczne ceny prądu opóźnione od dwóch do sześciu dni.

Podsumowując, w literaturze dotyczącej prognozowania cen energii elektrycznej na RDN zmienne wejściowe są różnorodne i zależą od specyfiki danego rynku. Wybrane badania podkreślają znaczenie historycznych cen energii, zapotrzebowanie oraz generacji energii z OZE jako kluczowych zmiennych wpływających na prognozy. Dodatkowo, zmienne fundamentalne, takie jak ceny surowców energetycznych czy zmienne czasowe, również odgrywają istotną rolę. Wszystkie z kluczowych cech są wykorzystywane w niniejszej pracy.

2.2 Modele prognozowania

Wśród stosowanych metod do predykcji cen energii elektrycznej wyróżnia się różne podejścia. Rafał Weron w swojej ponadczasowej pracy z 2014 roku "Electricity price forecasting: A review of the state-of-the-art with a look into the future" [20] dokonuje przeglądu literatury z ubiegłych wtedy 15 lat, systematyzując szybko rosnącą liczbę publikacji w tej dziedzinie. Weron wyjaśnia mechanizmy kształtowania się cen na rynkach energii elektrycznej, koncentrując się na cenach dnia następnego. Klasyfikuje techniki predykcyjne pod względem horyzontu czasowego i zastosowanej metodologii. Wymienia następujące kategorie modeli:

- multi-agent
- fundamentalne,
- reduced-form,
- statystyczne,
- computational intelligence,
- hybrydowe.

Modele multi-agent symulują zachowanie uczestników rynku energii, takich jak producenci i konsumenci, w celu przewidywania cen. Weron [20] wskazuje, że modele oparte na równowadze Nasha-Cournota czy równowadze funkcji podaży, są przydatne w analizie długoterminowej, ale nie uwzględniają strategiczne obstawianie cen przez uczestników rynku. W związku z tym tego rodzaju metody pasują najbardziej do stabilnych rynków bez dużych wahań cenowych.

Modele fundamentalne opierają się na analizie czynników ekonomicznych i fizycznych, takich jak ceny paliw, emisje CO₂ czy zapotrzebowanie. Weron [20] dzieli je na parameter-rich (uwzględniające wiele zmiennych) i parsimonious structural (uproszczone). Głównymi wyzwaniami takich modeli są ich złożoność oraz duże ilości danych, które często mogą być niedostępne w czasie rzeczywistym.

Modele reduced-form opisują dynamikę cen za pomocą procesów stochastycznych, takich jak jump-diffusions czy Markov regime-switching. Weron [20] wskazuje, że są one użyteczne w modelowaniu dziennej zmienności cen, ale mogą nie być dokładne w próbie dokładnego liczenia cen godzinowych, gdyż nie uwzględniają wpływu zmiennych fundamentalnych, takich jak sezonowość czy zmiany w podaży i popicie.

Modele statystyczne, takie jak ARIMA i GARCH, są szeroko stosowane w krótkoterminowym EPF. Modele autoregresywne (AR) wykorzystują liniową kombinację przeszłych wartości zmiennej do prognozowania przyszłych wartości. Modele średniej ruchomej (MA) prognozują zmienną na podstawie liniowej kombinacji przeszłych błędów prognoz. Modele ARMA łączy te dwa podejścia. Modele ARIMA dodają różnicowanie, co pozwala uwzględnić niestacjonarność danych, czyli ich zmienność w czasie. W celu uwzględnienia sezonowości wykorzystuje się również model SARIMA. Zgodnie z [12] modele SARIMA wykazują dobre wyniki w prognozowaniu cen energii elektrycznej, ale ich skuteczność może być ograniczona w przypadku danych o dużej zmienności. Weron [20] wspomina również o modelach ARX i ARMAX, gdzie symbol X oznacza zmienne zewnętrzne, takie jak pogoda czy zapotrzebowanie, które pomagają wyjaśnić zmiany cen. Popularny również w literaturze Model GARCH jest szczególnie przydatny w modelowaniu zmienności cen energii elektrycznej, ponieważ potrafi uwzględnić nierównomierne wahania cen (heteroskedastyczność) oraz fakt, że te wahania zmieniają się w czasie, zależąc od wcześniejszych zmian cen i ich niestabilności.

Modele computational intelligence, oparte na technikach uczenia maszynowego, zyskały na popularności w nowszych badaniach. Wśród popularnych metod wyróżniają się metody rozmyte, metody wektorów nośnych, LSTM oraz CNN. Jednym z przykładów zastosowania takiego modelu jest wspomniana przeze mnie praca inżynierska na Politechnice Warszawskiej [5]. W celu stworzenia modelu do skomplikowanego okresu rynkowego od 2021 do 2023 roku, autorzy używają Perceptronu wielowarstwowego (MLP). W innej pracy Grzegorza Marcjasza [10], proponowany jest dobór hiperparametrów do głębszej sieci neuronowej (DNN). Badanie porównuje DNN do statystycznego modelu LASSO i podkreśla, iż wyniki DNN są lepsze, co może wskazywać na przewagę nowoczesnego uczenia maszynowego nad tradycyjnymi metodami statystycznymi.

Modele hybrydowe łączą elementy różnych kategorii, aby wykorzystać ich zalety. Jinliang Zhang jest przykładem takiego podejścia. W pracy [17] łączy transformację falkową z ARIMA i GARCH. Argumentuje to w sposób, że połączenie WT z modelami ARIMA i GARCH pozwala na skuteczne

Rozdział 2. Przegląd literatury

modelowanie złożonych cech cen energii elektrycznej, takich jak niestacjonarność, nieliniowość i wysoka zmienność. W późniejszej pracy Zhang [21] łączy wspomniane transformację falkową oraz ARIMA z metodą najmniejszych kwadratów maszyn wektorów nośnych (LSSVM). Potwierdzając skuteczność takich metod, Weron [20] podaje, że popularnym podejściem hybrydowym jest łączenie modeli statystycznych z sieciami neuronowymi, co pozwala na modelowanie zarówno liniowych, jak i nieliniowych zależności.

W niniejszej pracy stworzony zbiór danych jest analizowany za pomocą regresji liniowej oraz grzbietowej, jak i modelami Prophet i Wielowarstwowy Perceptron (MLP). Według klasyfikacji Werona, regresja liniowa oraz regresja grzbietowa należą do zbiorów modeli statystycznych. Prophet również odpowiada kategorii modeli statystycznych, ale jego algorytm jest bardziej złożony od prostej regresji bądź szeroko używanych modeli ARIMA. Z kolei perceptron wielowarstwowy to przykład modelu computational intelligence, który wykorzystuje sieci neuronowe o różnej architekturze do prognozowania zmiennej docelowej.

Rozdział 3

Dane i zmienne

Dane odgrywają kluczową rolę w analizie i prognozowaniu cen energii na RDN. Zbiór danych przygotowany w pracy obejmuje okres od 1 stycznia 2016 roku do 31 grudnia 2023 roku i zawiera granulację godzinową. Zbiór danych składa się z różnorodnych zmiennych, które można podzielić na kilka kategorii, które są opisane w tym rozdziale. Duża ilość danych pochodzi z raportów Polskie Sieci Elektroenergetyczne (PSE) [14]. Warto zauważyć, że od 14 czerwca 2024 rok sposób raportowania danych przez PSE został zmieniony i przeniesiony na nową stronę [16]. Pierwotnie, zbiór danych zawierał dane z 2024 roku, ale w związku ze zmienionym sposobem raportowania danych i zmiany rozdzielczości danych na 15-minutową, zdecydowano się na usunięcie tych danych z analizy. Wiele z danych zostały pobrane również ze strony instrat [4]. Jest to strona, która pobiera dane z platformy PSE i udostępnia je w sposób wygodniejszy, dzięki czemu arkusze csv z dowolnego okresu czasowego o dowolnej częstotliwości można pobrać jednym przyciskiem myszy. Dane dotyczące cen paliw kopalnych oraz kursów walut zostały pozyskane z innych źródeł, takich jak publiczne bazy danych rynkowych lub platformy finansowe. Niniejszy rozdział szczegółowo opisuje zmienną zależną i zmienne niezależne podzielone na kategorie, a także prezentuje kluczowe cechy danych za pomocą tabel i wykresów, co wyjaśnia dobór zmiennych i pozwala na lepsze zrozumienie ich specyfiki i wyzwań związanych z modelowaniem.

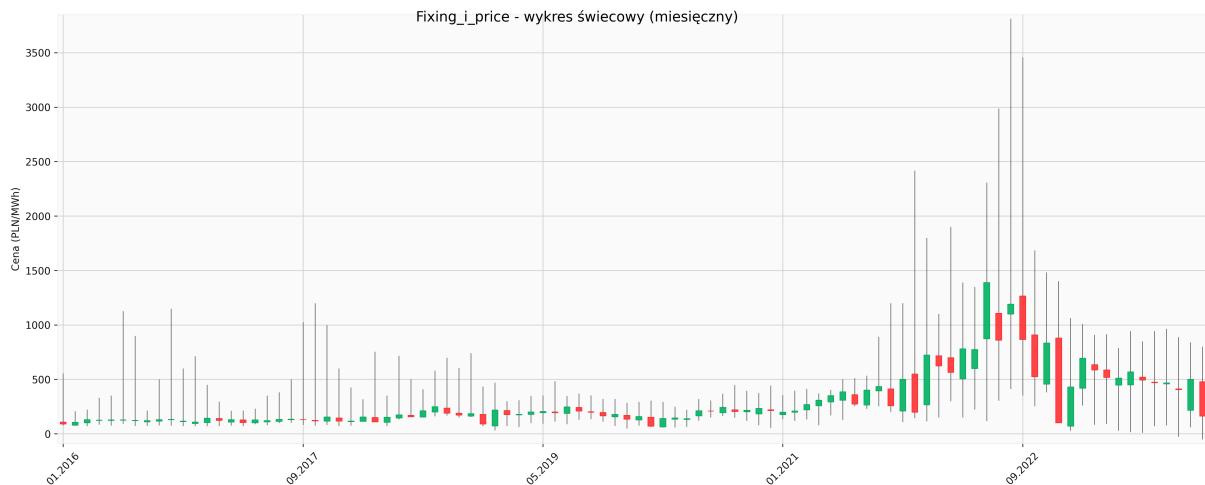
3.1 Zmienna zależna

Zmienna zależna w niniejszej pracy to `fixing_i_price`, czyli cena energii elektrycznej na Rynku Dnia Następnego w Polsce, wyrażona w PLN/MWh. Dane dotyczące tej zmiennej zostały pobrane z wymienionej platformy instrat w granulacji godzinowej. Zbiór danych obejmuje okres od 1 stycznia 2016 roku do 31 grudnia 2023 roku. Statystyki opisowe zmiennej `fixing_i_price` przedstawiono w Tabeli poniżej.

Statystyka	Wartość
Średnia	334.82 PLN/MWh
Odchylenie std.	272.48 PLN/MWh
Minimum	-50.00 PLN/MWh
25% (Q1)	170.79 PLN/MWh
Mediania	235.00 PLN/MWh
75% (Q3)	417.70 PLN/MWh
Maksimum	3812.45 PLN/MWh

Tabela 1. Podstawowe statystyki zmiennej fixing_i_price

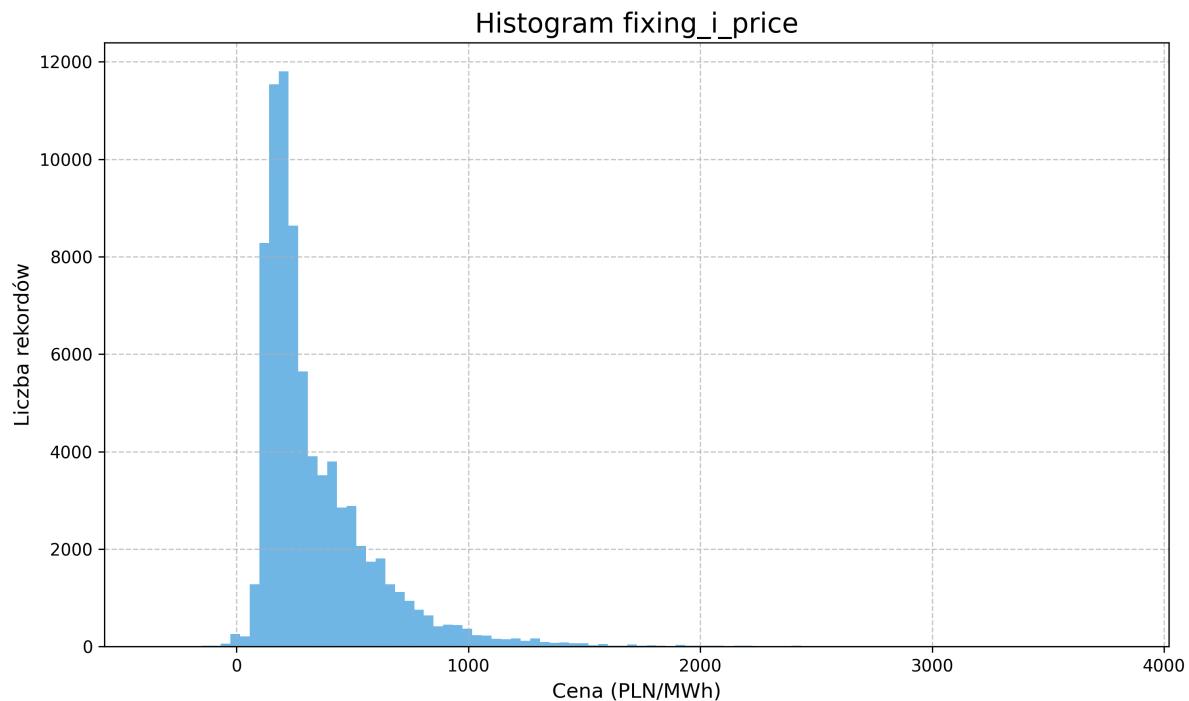
Aby lepiej zrozumieć dynamikę cen energii na RDN, przeanalizowano ich zmienność w całym okresie badania. Rysunek poniżej przedstawia zmienność cen energii w czasie zebranych danych za pomocą wykresu świecowego.



Rysunek 1. Ceny energii elektrycznej na RDN w latach 2016-2023. Opracowanie własne na podstawie danych instrat.

Wyraźnie widoczna jest różnica w poziomie cen w różnych okresach: od 2016 do Q4 roku 2020 ceny były stosunkowo stabilne z rzadkimi szczytami. Większa część cen oscyluje w przedziale 100-300 PLN/MWh. Sytuacja zmieniła się w 2020 roku, gdy zaczęły pojawiać się pierwsze skoki cenowe z powodu poważnych obostrzeń z powodu pandemii, a w 2022 roku, w wyniku kryzysu energetycznego wywołanego wojną na Ukrainie i ograniczeniami dostaw paliw kopalnych, ceny osiągnęły rekordowe poziomy. Pierwszy okres zostanie określony jako okres stabilności cenowej, a drugi jako okres skoków cenowych. Drugi z okresów pokazuje, jak niekorzystne sytuacje gospodarcze mogą wpływać na dynamikę cen energii, co ma istotne implikacje dla modelowania i prognozowania.

Rysunek 2 przedstawia histogram rozkładu zmiennej fixing_i_price. Rozkład jest wyraźnie asymetryczny, z długim prawym ogonem, co odzwierciedla występowanie skoków cenowych, takich



Rysunek 2. Histogram rozkładu zmiennej fixing_i_price. Opracowanie własne na podstawie danych instrat.

jak te w 2022 roku. Ujemne ceny, choć rzadkie (ok. 0,4% rekordów), są widoczne w lewej części histogramu, co potwierdza specyficzne cechy danych i rynku.

3.2 Zbiór zmiennych niezależnych

W niniejszej pracy wykorzystano różnorodne zmienne niezależne, które można podzielić na kilka kategorii. Obejmują one dane pogodowe, zapotrzebowanie, straty sieciowe, bilans wymiany transgranicznej, dane o produkcji energii przez poszczególne typy generatorów, ceny paliw kopalnych, emisji CO₂ i inne. Wybór tych zmiennych oparty jest na ich potencjalnym wpływie na ceny energii elektrycznej. Poniżej przedstawiono szczegółowy opis każdej z kategorii zmiennych niezależnych, które zostały uwzględnione w analizie.

3.2.1 Dane pogodowe

Pierwotnie zbiór danych miał być zestawiony z danych dostępnych za pomocą oficjalnej strony Instytutu Meteorologii i Gospodarki Wodnej, natomiast dostępne dane historyczne z lat 2016-2023 mają ograniczoną rozdzeliczość. Dane IMGW są zbierane przez wiele stacji meteorologicznych rozproszonych po całym kraju, ale jedynie w godzinach 6:00, 12:00 oraz 18:00. Aproxymować dane pogodowe w godzinach nocnych jest zadaniem nie do wykonania, szczególnie w przypadku sezonów zimowych, gdzie temperatura w nocy może drastycznie spadać w ciągu godziny. Z tego powodu jako

Rozdział 3. Dane i zmienne

źródło danych pogodowych wykorzystano stronę open-meteo [11]. Jest to strona, która zbiera dane godzinowe z dokładnością do jednego kilometra. Dostępne są dane pogodowe dla dowolnego okresu czasu i lokalizacji.

Dane pogodowe zostały pobrane dla czterech lokalizacji w Polsce: Warszawy (WAW), Koszalina (KSZ), Krakowa (KRK) i Babimostu (BAB). Wybór miast został podkutowany ich zróżnicowaniem geograficznym i klimatycznym, co pozwala uwzględnić regionalne różnice w warunkach pogodowych wpływających na produkcję i zapotrzebowanie na energię.

Warszawa, jako stolica i największe miasto Polski, reprezentuje centralny region kraju o wysokim zapotrzebowaniu na energię, szczególnie w okresach zimowych i letnich. Koszalin, położony na Pomorzu, jest kluczowy ze względu na bliskość farm wiatrowych na Morzu Bałtyckim, co czyni go istotnym punktem dla analizy produkcji energii wiatrowej. Kraków, znajdujący się w południowej Polsce, charakteryzuje się większym udziałem energii słonecznej w miksie energetycznym, a także wysokim zapotrzebowaniem na energię w sezonie grzewczym. Babimost, zlokalizowany w zachodniej Polsce, jest istotny ze względu na swoje położenie na zachodzie Polski w pobliżu granicy z Niemcami.

Parametry pogodowe zostały wybrane z uwzględnieniem ich bezpośredniego wpływu na rynek energii.

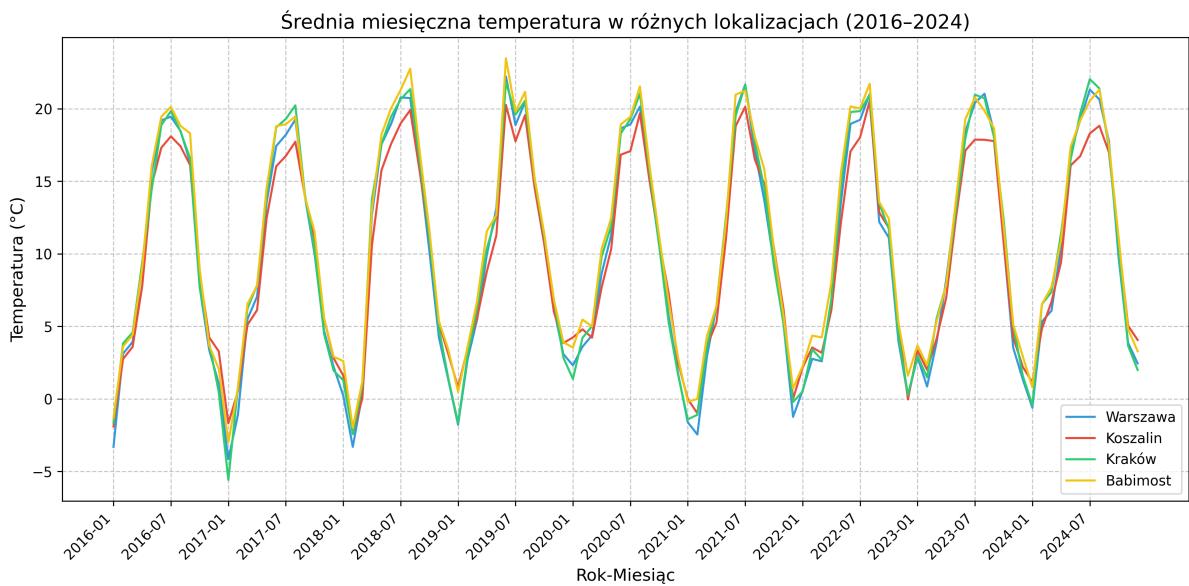
Temperatura jest kluczowym czynnikiem, ponieważ bezpośrednio wpływa na zapotrzebowanie na energię. Niskie temperatury zwiększą zużycie energii na ogrzewanie, natomiast bardzo wysokie temperatury latem mogą podnosić zapotrzebowanie na klimatyzację.

Pobrane dane prędkości wiatru dotyczą prędkości wiatru na wysokości 100 metrów nad powierzchnią ziemi. Taka wysokość została wybrana, ponieważ jest przeciętna dla turbin wiatrowych w Polsce, co pozwala dokładniej oszacować potencjalną produkcję energii z farm wiatrowych.

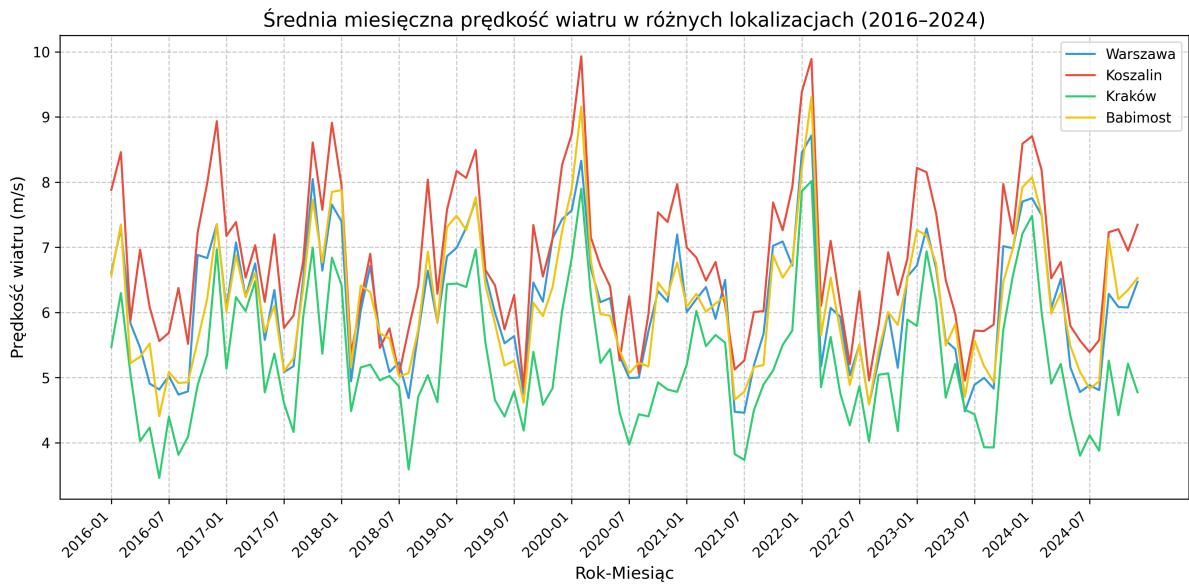
Promieniowanie słoneczne jest istotnym czynnikiem dla produkcji energii z paneli fotowoltaicznych. Zachmurzenie zostało uwzględnione, ponieważ wysoki poziom zachmurzenia zmniejsza efektywność paneli słonecznych. Wybór tych parametrów pozwala na kompleksową analizę wpływu pogody na ceny energii na RDN.

Poniżej przedstawię wykresy dla każdego z parametrów pogodowych, które zostały uwzględnione w analizie. Wykresy przedstawiają zmienność danych pogodowych w czasie. Zachmurzenie jest wyrażone w oktantach (0-8), gdzie 0 oznacza brak zachmurzenia, a 8 oznacza całkowite zachmurzenie.

3.2. Zbiór zmiennych niezależnych

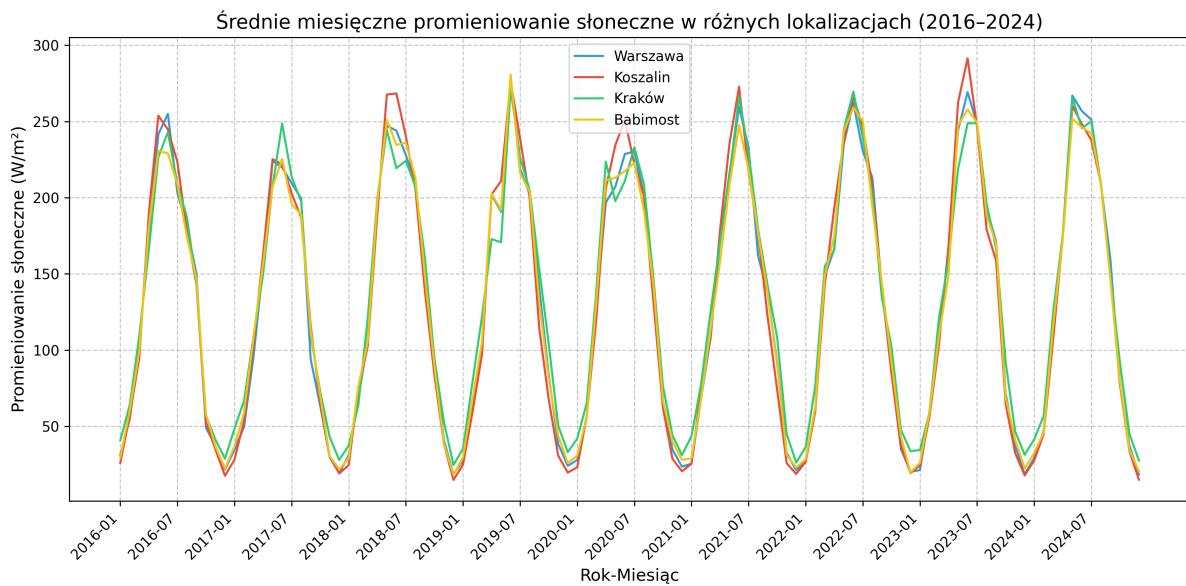


Rysunek 3. Zmiennaść temperatury w czasie (2016-2023). Opracowanie własne na podstawie danych open-meteo.

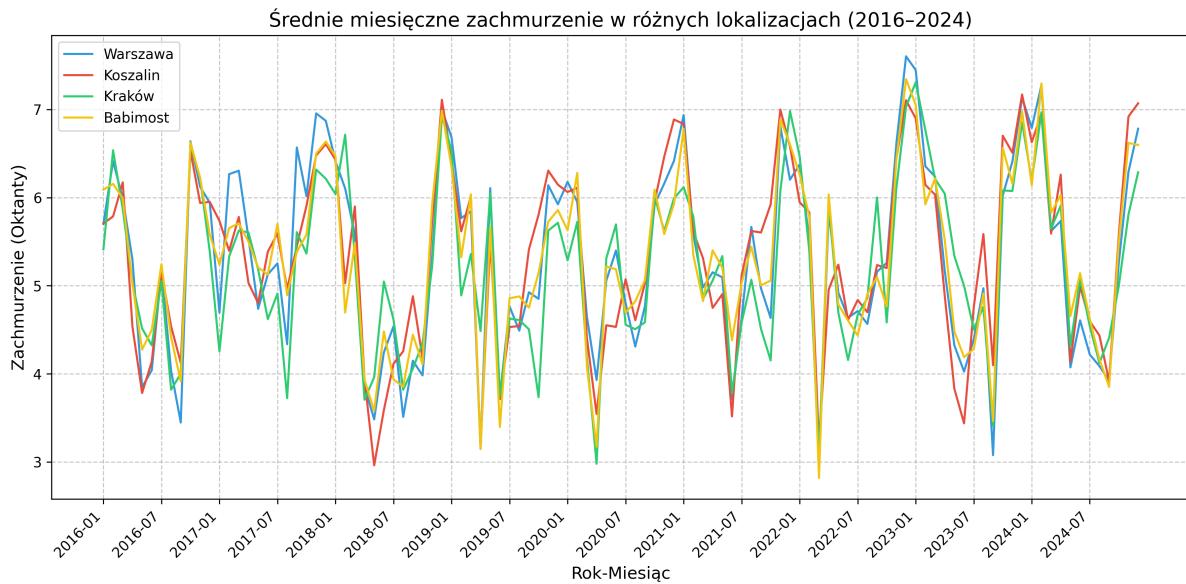


Rysunek 4. Zmiennaść prędkości wiatru w czasie (2016-2023) Opracowanie własne na podstawie danych open-meteo.

Rozdział 3. Dane i zmienne



Rysunek 5. Zmienność promieniowania słonecznego w czasie (2016-2023). Opracowanie własne na podstawie danych open-meteo.

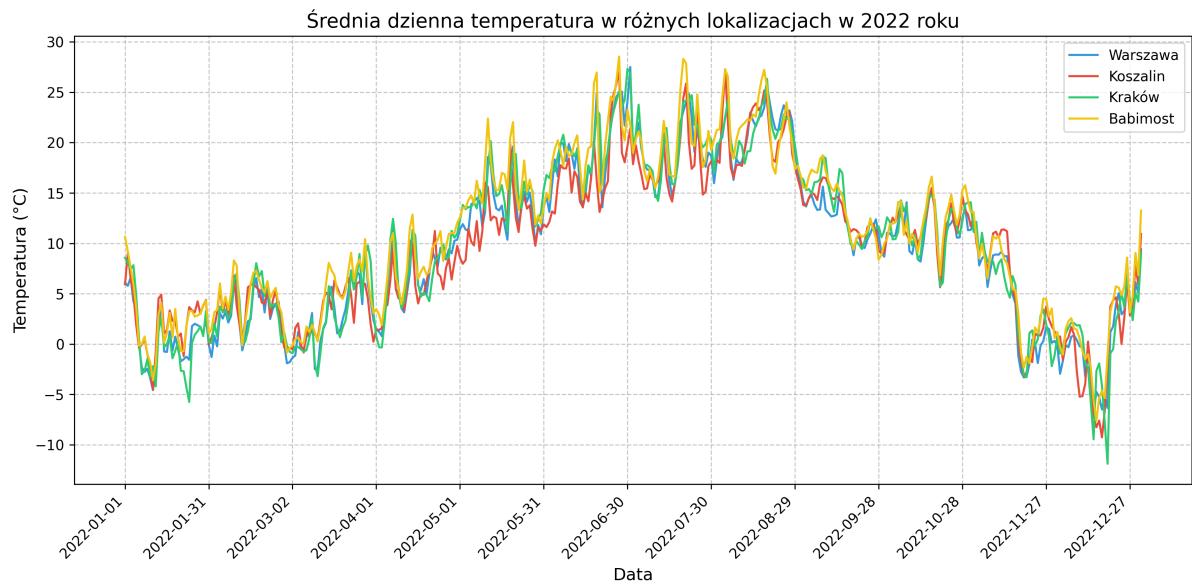


Rysunek 6. Zmienność zachmurzenia w czasie (2016-2023). Opracowanie własne na podstawie danych open-meteo.

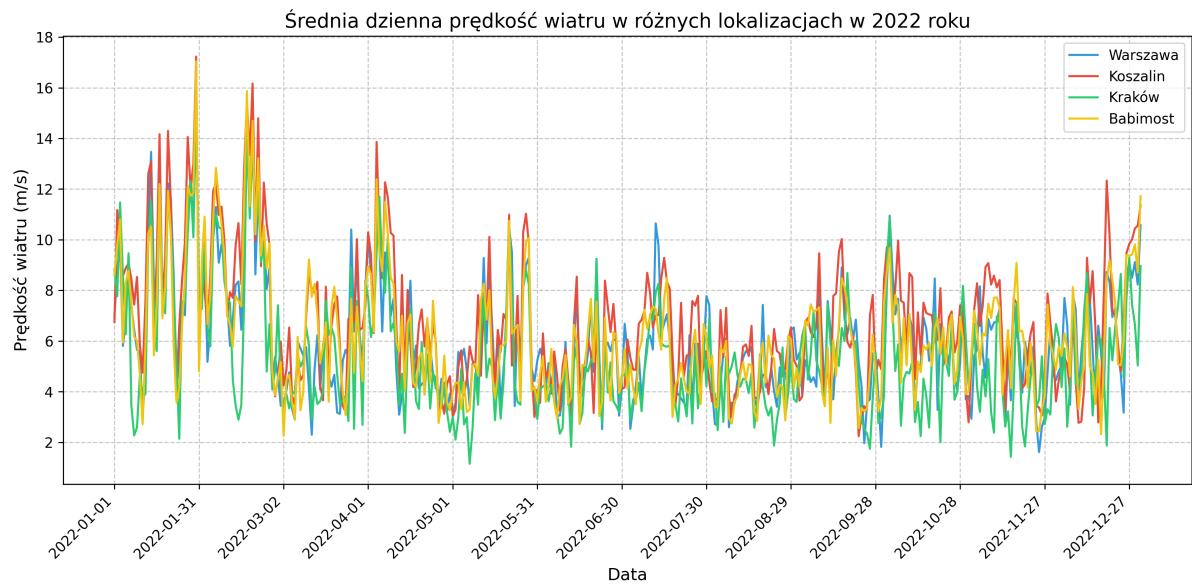
Każdy z wykresów przedstawia zmienność danego parametru pogodowego w wybranych lokalizacjach w przeciągu okresu badawczego. Wyraźnie widać sezonowe wahania parametrów pogodowych, co jest typowe dla klimatu Polski. Temperatura i promieniowanie słoneczne mają wyraźnie większe wartości w sezonach letnich, prędkość wiatru w sezonach zimowych, z kolei zachmurzenie ma bardziej zróżnicowany charakter.

3.2. Zbiór zmiennych niezależnych

Dodatkowo, w celu przedstawienia wahań zmiennych pogodowych w przeciągu roku, poniżej zostały załączone wykresy zmiennych pogodowych za rok 2022.

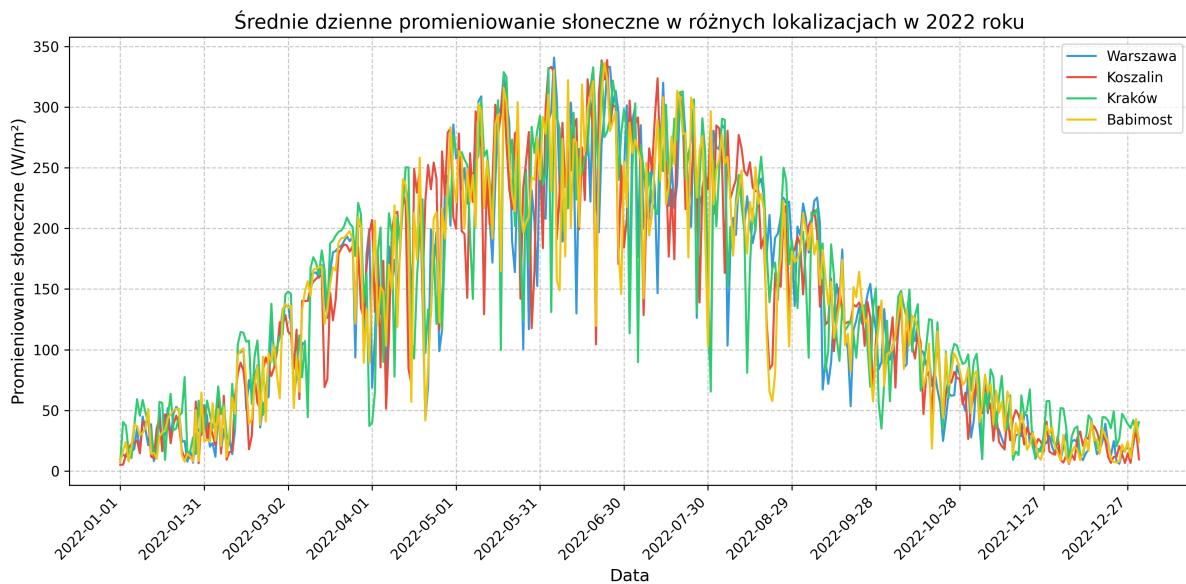


Rysunek 7. Zmienność temperatury w czasie (2022). Opracowanie własne na podstawie danych open-meteo.

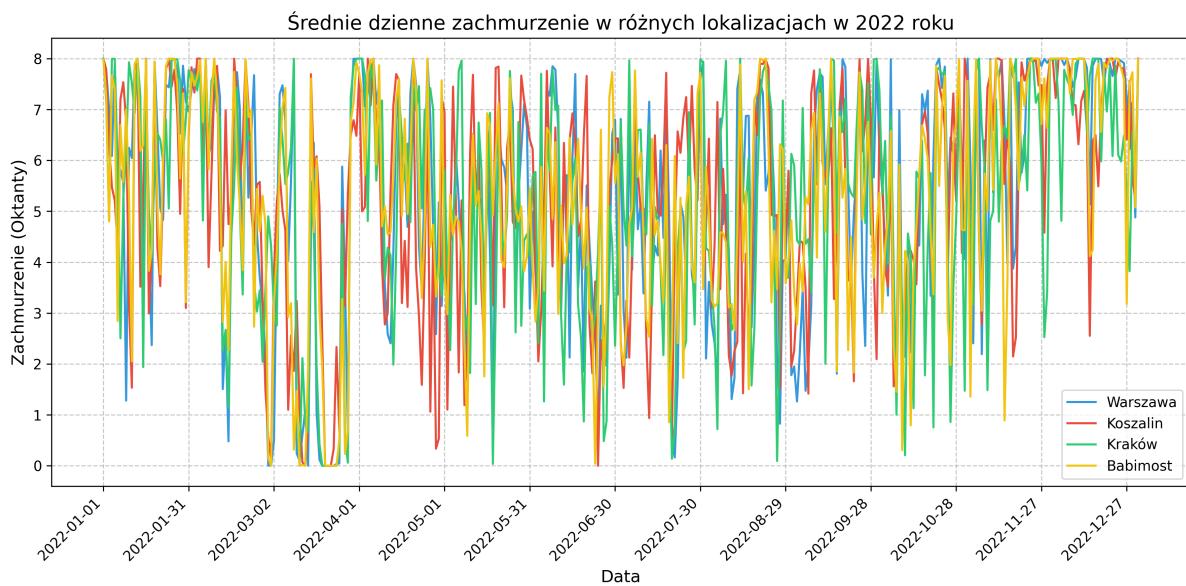


Rysunek 8. Zmienność prędkości wiatru w czasie (2022). Opracowanie własne na podstawie danych open-meteo.

Rozdział 3. Dane i zmienne



Rysunek 9. Zmienna promieniowania słonecznego w czasie (2022). Opracowanie własne na podstawie danych open-meteo.



Rysunek 10. Zmienna zachmurzenia w czasie (2022). Opracowanie własne na podstawie danych open-meteo.

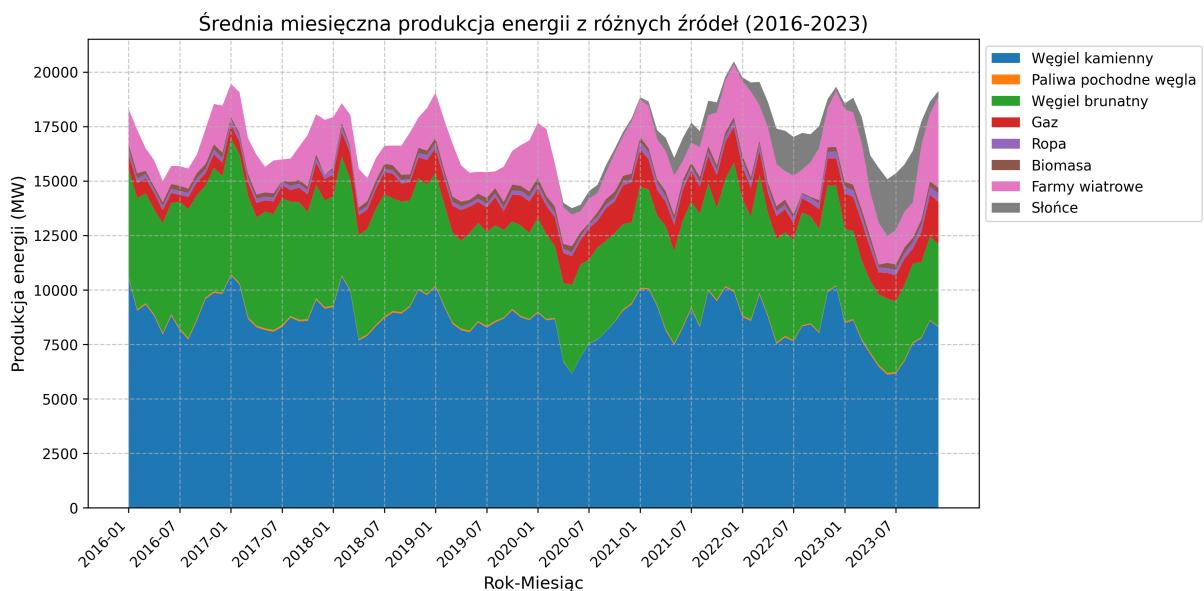
Temperatura jest rzadko uwzględniana w modelach progностycznych cen energii, ponieważ jej związek z produkcją energii nie jest bezpośredni. W kontekście Polski, gdzie znaczna część energii elektrycznej pochodzi z elektrowni węglowych, temperatura odgrywa jednak istotną rolę pośrednią, wpływając na zapotrzebowanie na energię, zwłaszcza w okresie sezonu grzewczego. Włączenie zmiennej do analizy może więc dostarczyć cennych informacji o sezonowych wzorcach konsumpcji i ich wpływie na dynamikę cen energii

3.2.2 Produkcja energii z wybranych źródeł

Zmienne dotyczące produkcji energii z różnych źródeł odgrywają kluczową rolę w analizie cen energii na RDN, ponieważ odzwierciedlają strukturę podaży energii w Polsce, która ma bezpośredni wpływ na dynamikę cen. W niniejszej pracy uwzględniono osiem zmiennych opisujących produkcję energii:

- **hard_coal** - produkcja z węgla kamiennego (MWh),
- **coal_derived** - produkcja z paliw pochodnych węgla (MWh),
- **lignite** - produkcja z węgla brunatnego (MWh),
- **gas** - produkcja z gazu ziemnego (MWh),
- **oil** - produkcja z ropy naftowej lub jej pochodnych (MWh),
- **biomass** - produkcja z biomasy (MWh),
- **wind** - produkcja z elektrowni wiatrowych lądowych (MWh),
- **solar** - produkcja z paneli fotowoltaicznych (MWh).

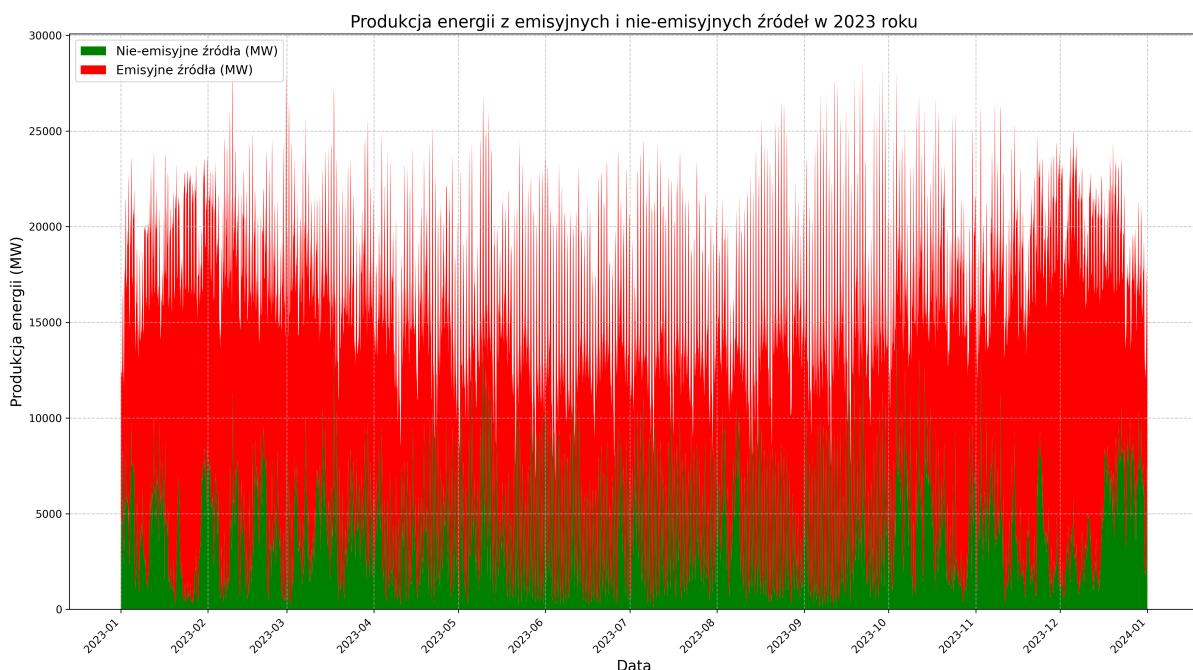
Dane te pochodzą z Polskich Sieci Elektroenergetycznych (PSE) [14] i zostały dopasowane do godzinowego formatu danych RDN, co pozwoliło na ich integrację z pozostałymi zmiennymi.



Rysunek 11. Zmienna produkcji energii z różnych źródeł w czasie. Opracowanie własne na podstawie danych PSE.

Wykres powyżej przedstawia średnią miesięczną produkcję energii z różnych źródeł w Polsce w latach 2016-2023, co jest istotne w kontekście prognozowania cen energii na RDN. Wykres obszarowy ukazuje dominację węgla kamiennego i węgla brunatnego, które do tej pory odpowiadają za większość produkcji energii elektrycznej. Biomasa oraz paliwa pochodne węgla stanowią znikomą część podaży i mogą być pominięte dla redukcji ilości zmiennych. Produkcja z gazu i ropy stale posiada niewielką, ale istotną część produkcji energii. Warto zauważyć, że produkcja z OZE, szczególnie ze słońca znaczco rośnie w ostatnich latach, co może mieć istotny wpływ na ceny energii

i umiejętności jej prognozowania. Do poprzednio omówionych zmiennych dodano również zmienną `non_emissive_sources_percentage`, która reprezentuje procentowy udział energii produkowanej z OZE w całkowitej produkcji energii elektrycznej w Polsce. Warto zaznaczyć, że średnia wartość produkcji energii z OZE w całym zakresie danych wynosi 14%, gdyż średnia wartość w 2023 wynosi już 23%. Wskazuje to wyraźną dynamikę wzrostu udziału OZE w systemie energetycznym kraju.



Rysunek 12. Porównanie produkcji energii z emisyjnych i bezemisyjnych źródeł w 2023 roku. Opracowanie własne na podstawie danych PSE.

W niektórych godzinach udział OZE w produkcji energii może przekraczać udział źródeł emisyjnych, co zwykle prowadzi do spadku cen. Warto również zauważać, że produkcja z węgla kamiennego i brunatnego jest bardziej stabilna i przewidywalna niż produkcja z OZE, co wpływa na dokładność prognoz. W związku z tym, zmienne dotyczące produkcji energii z różnych źródeł są istotnym elementem modelowania cen energii.

3.2.3 Handel energią z państwami sąsiednimi

Zmienne dotyczące wymiany energii z innymi krajami są kolejnym istotnym elementem analizy cen energii na Rynku Dnia Następnego, ponieważ pozwalają na uwzględnienie wpływu handlu międzynarodowego na ceny energii w Polsce. W niniejszej pracy uwzględnione zostały bilanse handlu energią z Państwami sąsiednimi, z którymi Polska ma połączenia transgraniczne. Są to Państwa takie jak:

- Niemcy (DE),
- Czechy (CZ),

- Słowacja (SK),
- Litwa (LT),
- Szwecja (SE),
- Ukraina (UA).

Dane dotyczące wymiany energii z tymi krajami pochodzą z Polskich Sieci Elektroenergetycznych i są dostępne w granulacji godzinowej. Warto zauważyć, że Polska jest jednym z kluczowych graczy na rynku energii w Europie Środkowo-Wschodniej, co sprawia, że wymiana energii zwykle występuje z każdym krajem w każdej godzinie czasu rzeczywistego. W okresach wysokiego zapotrzebowania na energię lub w sytuacjach poważnych awarii, Polska sięga po energię z innych krajów, co prowadzi do wzrostu cen. Z drugiej strony, w okresach niskiego zapotrzebowania, Polska może eksportować, czyli sprzedawać nadwyżki energii, w momentach spadków cen. Prawdopodobnie, taki aktywny handel może również wynikać z różnic cenowych pomiędzy krajami. Poniżej przedstawiam tabele wymiany energii z sąsiadami w latach 2016-2023.

Rok	Niemcy (MW)	Czechy (MW)	Litwa (MW)	Słowacja (MW)	Szwecja (MW)	Ukraina (MW)
2016	994.26	-761.81	68.21	-475.47	294.36	108.99
2017	836.24	-612.73	119.92	-499.21	340.96	102.45
2018	803.00	-373.78	102.53	-366.14	310.83	160.96
2019	1149.22	-312.31	216.38	-367.32	329.80	157.17
2020	1275.75	-264.90	201.93	-348.95	428.81	167.98
2021	957.79	-901.92	110.04	-519.59	364.69	93.55
2022	918.62	-1000.89	86.53	-684.25	429.17	126.95
2023	851.27	-526.46	58.40	-382.18	431.60	5.76

Tabela 2. Średni bilans wymiany energii z sąsiadami w latach 2016-2023. Opracowanie własne na podstawie danych PSE.

Wartości te są średniogodzinnymi bilansami wymiany energii z sąsiadami w latach 2016-2023. Wartości dodatnie oznaczają eksport energii, a wartości ujemne import energii. Polska ma dodatni bilans wymiany energii z Niemcami, co oznacza, że regularnie eksportuje energię do Niemiec. Największe ujemne bilanse Polska posiada z Czechami, prawdopodobnie spowodowane jest to funkcjonującymi w Czechami elektrowniami atomowymi. Przepływy transgraniczne są warte uwagi i uwzględnienia w zbiorze.

Dodatkowo, w analizie uwzględniono dane cenowe z rynków sąsiednich, które odzwierciedlają koszty energii w tych krajach i mogą wpływać na decyzje o importie lub eksportie. Zebrane dane cenowe obejmują cztery obszary: Czechy (CZ), Słowację (SK), Litwę (LT) oraz Szwecję (SE4). Szwedzki rynek energii elektrycznej jest podzielony na cztery obszary i właśnie ten ostatni - południowy jest połączony z systemem Polskim. Dane cenowe zostały pobrane ze strony transparency entsoe [3], będącej oficjalnym źródłem danych ENTSO-E, czyli Europejskiej Sieci Operatorów Systemów Przesyłowych Energii Elektrycznej. Dane dostępne są w rozdzielcości godzinowej podobnie do Polskiego rynku.

Dane cenowe z rynku ukraińskiego (UA) są dostępne dopiero od grudnia 2020 roku, co uniemożliwiło

ich uwzględnienie w analizie dla całego okresu. Podobnie, dane z rynku niemieckiego (DE) są dostępne dopiero od października 2018 roku, dlatego również zostały pominięte. Brak tych danych wynika z ograniczeń dostępności historycznych informacji na wspomnianej platformie.

Wszystkie dane cenowe z rynków sąsiednich, pierwotnie wyrażone w euro na megawatogodzinę, zostały przekonwertowane na PLN/MWh zgodnie z kursem średnim Narodowego Banku Polskiego dla odpowiednich dat. Przeliczenie to zapewniło ujednolicenie jednostek walutowych z innymi zmiennymi w zbiorze, takimi jak zmienna zależna. Włączenie tych cen do modelu pozwala na analizę wpływów różnic cenowych między krajami na handel energią oraz na dynamikę cen w Polsce, co jest szczególnie istotne w okresach zmiennej dostępności energii odnawialnej lub w sytuacjach kryzysowych. Poniżej załączam tabele statystyk cenowych z rynków sąsiednich w latach 2016-2023 podobnie do tych przedstawionych dla zmiennej zależnej 1.

Statystyka	Szwecja (SE4)	Słowacja (SK)	Czechy (CZ)	Litwa (LT)
Średnia	265.92 PLN/MWh	380.30 PLN/MWh	361.74 PLN/MWh	349.51 PLN/MWh
Odchylenie std.	332.26 PLN/MWh	453.45 PLN/MWh	428.63 PLN/MWh	418.34 PLN/MWh
Minimum	-267.23 PLN/MWh	-293.49 PLN/MWh	-310.33 PLN/MWh	-251.70 PLN/MWh
25% (Q1)	115.20 PLN/MWh	141.34 PLN/MWh	138.38 PLN/MWh	141.41 PLN/MWh
Median	165.57 PLN/MWh	210.72 PLN/MWh	201.64 PLN/MWh	202.91 PLN/MWh
75% (Q3)	260.83 PLN/MWh	420.55 PLN/MWh	405.60 PLN/MWh	373.50 PLN/MWh
Maksimum	3786.10 PLN/MWh	4260.61 PLN/MWh	4139.78 PLN/MWh	9991.76 PLN/MWh

Tabela 3. Statystyki cen energii (PLN/MWh) na rynkach sąsiednich w okresie 2016-2023. Opracowanie własne na podstawie danych ENTSO-E.

Statystyki pokazują, że ceny energii na rynkach sąsiednich są bardziej zmienne niż ceny na RDN, co może być wynikiem różnic w strukturze rynku, dostępności źródeł energii oraz polityki energetycznej w poszczególnych krajach oraz waluty euro, w której są wyrażone. Wartości minimalne i maksymalne wskazują na występowanie ekstremalnych zdarzeń cenowych, które mogą mieć istotny wpływ na handel energią oraz na ceny w Polsce. Na przykład, w sierpniu 2022 roku ceny energii na Litwie osiągały rekordowy poziom 9991.76 PLN/MWh.

3.2.4 Ceny paliw kopalnych i emisji CO₂

Zmienne dotyczące cen paliw kopalnych oraz emisji CO₂ odgrywają kluczową rolę w analizie cen energii na RDN, ponieważ koszty paliw i emisji mają bezpośredni wpływ na ceny energii elektrycznej w Polsce, gdzie dominującym źródłem energii jest węgiel. W niniejszej pracy uwzględniono cztery zmienne, które zostały opisane w Tabeli 4.

3.2. Zbiór zmiennych niezależnych

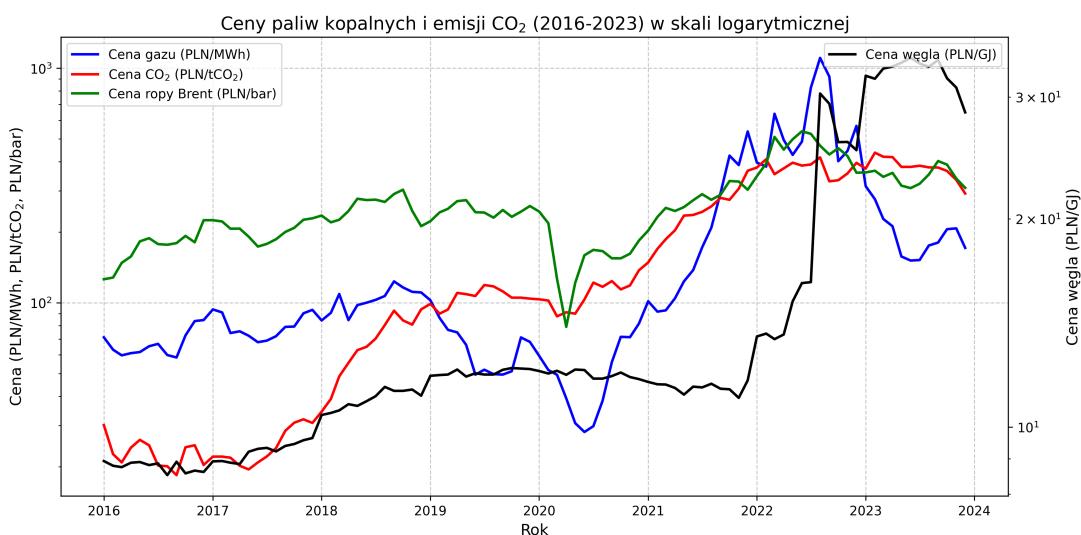
Nazwa zmiennej	Opis	Źródło danych	Rozdzielcość
gas_price	Cena gazu ziemnego (PLN/MWh)	instrat [4]	Dzienna
coal_price	Cena węgla kamiennego (PLN/GJ)	instrat [4]	miesięczna
co2_price	Cena uprawnień do emisji CO ₂ (PLN/tCO ₂)	instrat [4]	tygodniowa
brent_price	Cena ropy Brent Europe (PLN/bar)	EIA [1]	Dzienna

Tabela 4. Opis zmiennych dotyczących cen paliw kopalnych i emisji CO₂.

Dane te zostały dopasowane do godzinowego formatu danych RDN, w taki sposób, że w dniach pracy giełdy i znanych cen paliw, cena stale jest przypisana do godzin, a w dniach wolnych od pracy lub bez znanych cen, wartości są interpolowane.

Ceny paliw kopalnych i emisji CO₂ są istotne w kontekście prognozowania cen energii, ponieważ wpływają na koszty produkcji energii w elektrowniach konwencjonalnych, które dominują w polskim miescie energetycznym. Na przykład, wzrost ceny węgla (coal_price) lub uprawnień do emisji CO₂ (co2_price) zwiększa koszty produkcji energii w elektrowniach węglowych. Cena gazu (gas_price) jest kluczowa dla elektrowni gazowych, które pełnią rolę bilansującą w systemie. Cena ropy Brent (brent_price) ma mniejszy bezpośredni wpływ na produkcję energii w Polsce, ale jest istotna w kontekście globalnych trendów cen paliw i skorelowana z innymi cenami surowców energetycznych.

Aby zilustrować dynamikę tych zmiennych, na rysunku poniżej przedstawiono zmiany cen paliw kopalnych i emisji CO₂ w latach 2016-2023 w ujęciu miesięcznym. Dla lepszego przedstawienia danych użyłem skali logarytmicznej.



Rysunek 13. Ceny paliw kopalnych i emisji CO₂ w latach w ujęciu miesięcznym. Źródło: Opracowanie własne na podstawie danych rynkowych.

Ceny paliw kopalnych w okresie wysokiej zmienności szybko rosną, co jest jedną z przyczyn rosnących cen energii. Największy wzrost ma cena gazu, która od rozpoczęcia konfliktu zbrojnego wzrosła ponad 10-krotnie w ciągu roku.

3.2.5 Straty mocy w systemie elektroenergetycznym

Zmienne dotyczące strat mocy w systemie elektroenergetycznym odgrywają istotną rolę w analizie cen energii na Rynku Dnia Następnego (RDN), ponieważ wpływają na dostępność energii w systemie oraz koszty jej przesyłu i dystrybucji. W niniejszej pracy uwzględniono następujące dane zbierane przez systemy PSE: `power_loss` (utrata mocy w wyniku awarii w MW) oraz `network_loss` (utrata mocy w sieci w MW). Dane te są dostępne w rozdzielcości godzinowej i pochodzą z Polskich Sieci Elektroenergetycznych [14].

Aby zilustrować dynamikę tych zmiennych, na rysunku 14 przedstawiono zmiany strat mocy w wyniku awarii (`power_loss`) oraz strat mocy w sieci (`network_loss`) w latach 2016-2023 w ujęciu miesięcznym.



Rysunek 14. Straty mocy w wyniku awarii i w sieci w ujęciu miesięcznym. Źródło: Opracowanie własne na podstawie danych z PSE.

Analiza rysunku 14 ujawnia wyraźne różnice w dynamice obu zmiennych. Straty mocy w wyniku awarii (`power_loss`, czerwona linia) wykazują znaczną zmienność w czasie, osiągając wartości od około 2 MW do szczytowych wartości przekraczających 10 000 MW. Największe szczyty można obserwować w roku 2022, co może być powiązane z kryzysem energetycznym oraz zwiększym zapotrzebowaniem na energię w Europie, prowadzącym do przeciążeń systemu.

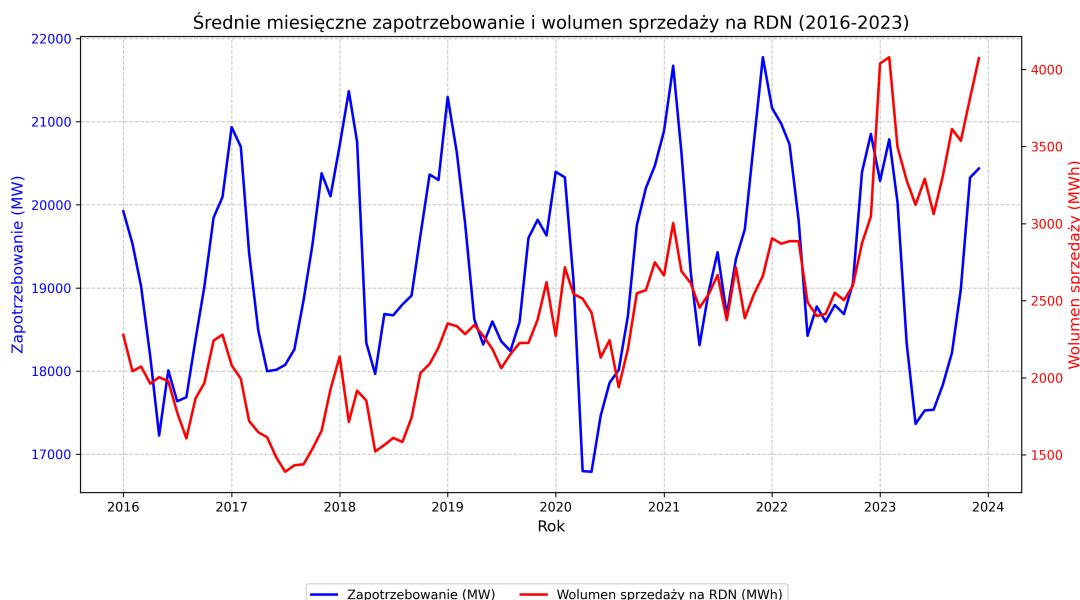
Straty mocy w sieci (`network_loss`, niebieska linia) są znacznie bardziej stabilne. Warto zauważyć, że w 54.09% przypadków zmienna `network_loss` przyjmuje wartość 0, co wskazuje na brak strat mocy w sieci w ponad połowie analizowanych godzin. Stabilność strat sieciowych wynika z ich zależności od fizycznych właściwości systemu przesyłowego, takich jak długość linii, opór elektryczny czy poziom obciążenia, które zmieniają się w sposób bardziej przewidywalny niż losowe awarie.

Porównanie obu zmiennych wskazuje, że straty w wyniku awarii mają większy wpływ na krótkoterminowe wahania dostępności energii w systemie, co może bezpośrednio przekładać się na wzrost cen na RDN w okresach wysokich strat. Uwzględnienie tych zmiennych w modelowaniu cen energii pozwala na lepsze zrozumienie wpływu czynników technicznych na dynamikę rynku.

3.2.6 Zapotrzebowanie na energię i wolumen handlu

Kolejnymi z ważnych zmiennych są zapotrzebowanie na energię i wolumen handlu. Przedstawiają one popyt na rynku, przez co można je określić podstawowymi czynnikami kształtującym ceny. W niniejszej pracy uwzględniono następujące zmienne: `load` (zapotrzebowanie na energię w MWh) oraz `trade_volume` (wolumen handlu w MWh). Obie zmienne pochodzą z Polskich Sieci Elektroenergetycznych (PSE) i są dostępne w granulacji godzinowej.

Poniżej na wspólnym wykresie przedstawię różnice pomiędzy zapotrzebowaniem a wolumenem handlu.



Rysunek 15. Zapotrzebowanie na energię i wolumen handlu w latach 2016-2023 (w ujęciu miesięcznym). Źródło: Opracowanie własne na podstawie danych z PSE.

Rozbieżność między zapotrzebowaniem a wolumenem sprzedaży na RDN ma swój powód. Wolumen sprzedaży na RDN zwykle nie przekracza 5000 MWh. Porównanie do zapotrzebowania na poziomie ponad 15 000 MWh wskazuje, że RDN pokrywa jedynie część całkowitego zapotrzebowania. Pozostała część jest zaspakajana przez: (1) kontrakty bilateralne (OTC), (2) rynek bilansujący, na którym PSE obsługuje zakup oraz sprzedaż energii w czasie rzeczywistym, (3) import energii z sąsiednich krajów, jak pokazano w podrozdziale 3.2.3 oraz innego rodzaju transakcje.

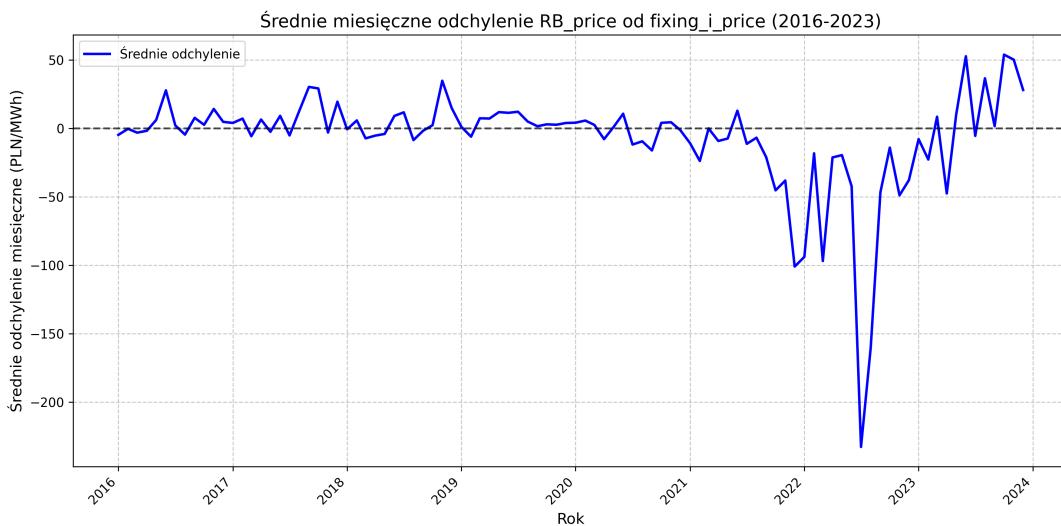
Wyraźnie widoczne są duże szczyty zapotrzebowania zapotrzebowania w okresie zimowym. Wynika to z zwiększonego zapotrzebowania na energię elektryczną w okresie grzewczym, co jest typowe dla klimatu Polski.

3.2.7 Rynek bilansujący

Rynek bilansujący odgrywa kluczową rolę w systemie elektroenergetycznym, zapewniając równowagę między podażą a popytem na energię elektryczną w czasie rzeczywistym. W niniejszej analizie uwzględniono zmienną `RB_price` (cena na rynku bilansującym w PLN/MWh), pochodzącą z danych Polskich Sieci Elektroenergetycznych (PSE) [14], dostępną w rozdzielcości godzinowej. Rynek Bilansujący (RB) służy jako mechanizm korygujący, umożliwiając operatorowi systemu (PSE) zarządzanie nagłymi niedoborami lub nadwyżkami energii, wynikającymi np. z awarii, zmian w zapotrzebowaniu lub fluktuacji produkcji ze źródeł odnawialnych. W przeciwieństwie do Rynku Dnia Następnego (RDN), gdzie ceny ustalane są z wyprzedzeniem (`fixing_i_price`), ceny na rynku bilansującym odzwierciedlają bieżące warunki rynkowe.

Ceny z rynku bilansującego mogą wpływać na ceny RDN, ponieważ różnica między nimi może powodować zmiany w zachowaniach uczestników rynku. Na przykład, jeśli cena na rynku bilansującym jest znacznie niższa niż na RDN, może to skłonić uczestników rynku do zakupu energii na rynku bilansującym, co z kolei wpłynie na ceny na RDN.

W celu oceny różnic cenowych między rynkami przeprowadzono analizę odchyлеń ceny na rynku bilansującym (`rb_price`) od ceny na RDN (`fixing_i_price`). Obliczone średnie odchylenie na całym zbiorze danych wynosi -7.29 PLN/MWh, co wskazuje, że ceny na rynku bilansującym są średnio niższe niż na RDN w analizowanym okresie (2016-2023). Taka wartość może sugerować efektywność mechanizmów równoważenia w systemie elektroenergetycznym. Dynamikę tych różnic zilustrowano na wykresie poniżej.



Rysunek 16. Średnie miesięczne odchylenie ceny na rynku bilansującym (rb_price) od ceny na Rynku Dnia Następnego (fixing_i_price) w latach 2016-2023 (w PLN/MWh). Źródło: Opracowanie własne na podstawie danych z PSE.

W okresach największych wahań cenowych na rynku energii w 2022 roku, ceny na rynku bilansującym niższe niż na RDN. Wartości te mogą być wynikiem większej elastyczności rynku bilansującego, który może szybko reagować na zmiany w zapotrzebowaniu i podaży energii. Wartości ujemne sugerują, że w tych okresach operator systemu mógł być w stanie zaspokoić zapotrzebowanie na energię po niższych kosztach niż te ustalone na RDN.

3.2.8 Inne zmienne

W swoim zbiorze danych uwzględnione zostały inne zmienne, które mogą mieć wpływ na ceny energii na RDN.

Zmienne sezonowe

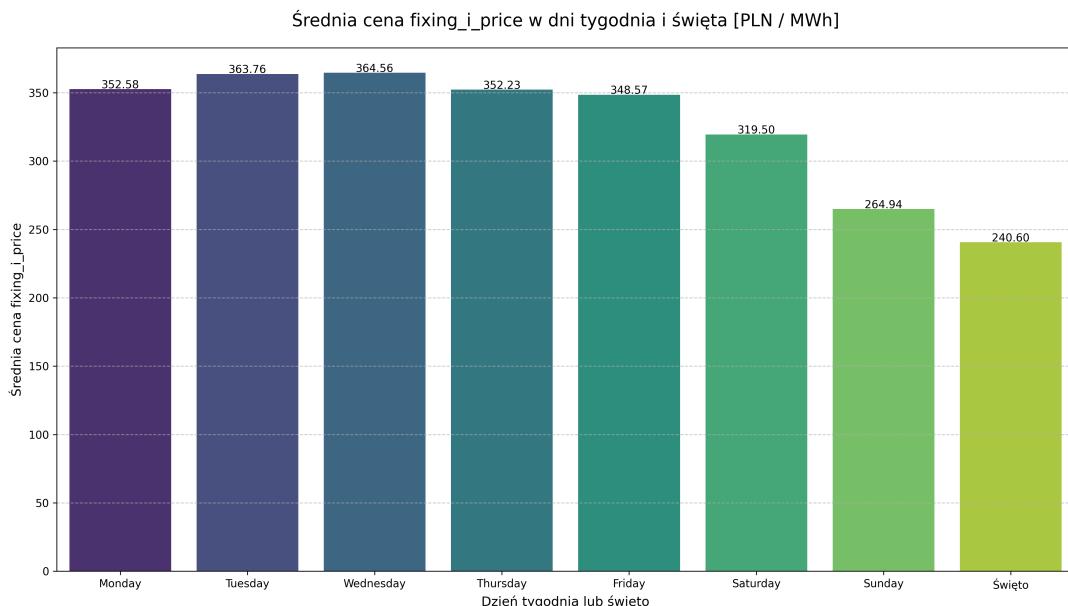
Zmienne sezonowe, takie jak day_of_week, month i is_holiday, zostały wprowadzone do zbioru danych w celu uchwycenia cykliczności i wzorców sezonowych w cenach energii na Rynku Dnia Następnego.

Zmienna month reprezentuje miesiąc roku (1-12, gdzie 1 to styczeń, a 12 to grudzień). Na temat cykliczności miesięcznej wspomniałem wcześniej w podrozdziałach 3.2.4 i 3.2.6. Niemniej jednak, wprowadzenie zmiennych day_of_week i month do modelu pozwala lepiej uchwycić cykliczność i sezonowość w cenach energii, na poszczególne godziny dnia o czym świadczy poniższy wykres korelacji.

Zmienna day_of_week reprezentuje dzień tygodnia w zakresie od 1 do 7. Zmienna pozwala uwzględnić modelowi cykliczność tygodniową w cenach energii. Z wykresu poniżej wynika, że zapotrzebowanie na energię jest zazwyczaj wyższe w dni robocze, gdy działa przemysł i biura,

Rozdział 3. Dane i zmienne

a niższe w weekendy i święta, gdy aktywność gospodarcza jest mniejsza. Ta cykliczność przekłada się na ceny energii na RDN: w dni robocze ceny są zazwyczaj wyższe, szczególnie w godzinach szczytu. Wprowadzone zmienne sezonowe pozwalają modelowi lepiej uchwycić te wzorce, co jest istotne w dokładnym modelowaniu.



Rysunek 17. Korelacja dni tygodnia z cenami energii na RDN. Opracowanie własne.

Dodatkowo wprowadzono zmienną `peak_hour`, która jest zmienną binarną określającą godziny szczytu, w których występuje zwiększone zapotrzebowanie na energię elektryczną. Przyjmuje wartość 1, jeśli spełniony jest jeden z dwóch warunków, w przeciwnym razie 0:

- W dni wolne lub dni weekendowe, zapotrzebowanie przekracza 18 000 MW, a godzina należy do przedziałów porannych (7:00-9:00) lub wieczornych (16:00-18:00).
- W dni robocze, zapotrzebowanie przekracza 23 000 MW i posiada takie same przedziały godzinowe jak w dni wolne.

Definicja ta odzwierciedla wyższe zapotrzebowanie na energię w godzinach szczytu, z różnymi programami dla dni roboczych i wolnych, co jest zgodne z charakterystyką polskiego rynku energii.

Ceny historyczne

Zmienne historyczne `fixing_i_price_lag24`, `fixing_i_price_lag48`, `fixing_i_price_lag72`, `fixing_i_price_lag96`, `fixing_i_price_lag120`, `fixing_i_price_lag144` oraz `fixing_i_price_lag168`, zostały wprowadzone do zbioru danych w celu uchwycenia autokorelacji w cenach energii na RDN. Zmienne te zostały wygenerowane na podstawie kolumny `fixing_i_price`, która reprezentuje cenę energii na RDN w danej godzinie (w PLN/MWh), poprzez przesunięcie wartości o odpowiednią liczbę godzin.

Dodatkowo, zostały wprowadzone średnie kroczące z ostatnich 24 i 48 godzin. W wyniku tego do datasetu zostały dodane zmienne `fixing_i_price_mean24` i `fixing_i_price_mean48`, które reprezentują średnie ceny energii na RDN w ostatnich 24 i 48 godzinach. Wartości te zostały obliczone na podstawie danych z kolumny `fixing_i_price` z przesunięciem o jedną godzinę, żeby nie w średniej kroczącej nie brać pod uwagę obecnej wartości.

Wprowadzenie zmiennych historycznych do modelu pozwala lepiej uchwycić autokorelację i cykliczność w cenach energii, co jest kluczowe dla poprawy dokładności prognoz. Zgodnie z literaturą, ceny historyczne są często jednymi z najważniejszych predyktorów w modelach prognozowania cen energii.

Kurs wymiany walut

Kursy walut

Zmienne `pln_usd` oraz `pln_eur` reprezentują odpowiednio kursy wymiany złotego polskiego względem dolara amerykańskiego (USD/PLN) oraz euro (EUR/PLN) w danej godzinie. Dane te zostały pozyskane z oficjalnej strony Narodowego Banku Polskiego (NBP) [13] w granulacji dziennej. W celu dopasowania danych do godzinowego formatu Rynku Dnia Następnego (RDN), wartości kursów zostały przypisane do wszystkich godzin w danym dniu. W przypadku dni wolnych od pracy lub braku dostępnych danych w weekendy lub święta, wartości kursów zostały interpolowane metodą liniową, co pozwoliło na uzyskanie ciągłości danych.

Rok	Średni kurs PLN/USD	Średni kurs PLN/EUR
2016	3.94	4.36
2017	3.78	4.26
2018	3.61	4.26
2019	3.84	4.30
2020	3.90	4.44
2021	3.86	4.57
2022	4.46	4.69
2023	4.20	4.54

Tabela 5. Średni kurs wymiany PLN/USD i PLN/EUR w latach 2016-2023. Źródło: Opracowanie własne na podstawie danych NBP.

Kursy wymiany walut są istotnymi czynnikami w kontekście prognozowania cen energii na Rynku Dnia Następnego, ponieważ wpływają na koszty importu surowców energetycznych oraz handel międzynarodowy. Zmienna `pln_usd` jest kluczowa, ponieważ Polska importuje znaczną część paliw kopalnych, takich jak gaz ziemny i ropa naftowa (np. Brent, której cena jest wyrażona w dolarach amerykańskich i przeliczona na PLN w zmiennej `brent_price`). Wzrost kursu PLN/USD, czyli osłabienie złotego względem dolara, zwiększa koszty importu tych surowców, co może prowadzić do wzrostu kosztów produkcji energii, a w konsekwencji do wyższych cen na RDN.

Rozdział 3. Dane i zmienne

Z kolei zmienna `pln_eur` odgrywa istotną rolę w kontekście handlu energią z krajami sąsiednimi, ponieważ ceny energii na rynkach europejskich, takich jak Czechy (`cz_price`), Słowacja (`sk_price`), Litwa (`lt_price`) czy Szwecja (`se_price`), są wyrażone w euro. W niniejszej analizie ceny te zostały przeliczone na złotówki przy użyciu kursu `pln_eur`, co pozwoliło na ujednolicenie jednostek walutowych i bezpośrednie porównanie z cenami na polskim RDN (`fixing_i_price`). Wahania kursu PLN/EUR mogą wpływać na decyzje dotyczące importu lub eksportu energii - na przykład, osłabienie złotego względem euro zwiększa koszt importu energii z rynków sąsiednich, co może podnieść ceny w Polsce.

Kursy polskiego złotego odzwierciedlają również zmiany w gospodarce krajowej i globalnej. Próbując stabilizować kurs, Narodowy Bank Polski zmienia politykę monetarną kraju, co może wpływać na koszt produkcji energii (np. poprzez zmiany stóp procentowych, które oddziałują na koszty finansowania inwestycji energetycznych). W efekcie zarówno kurs PLN/USD, jak i PLN/EUR mają pośredni wpływ na dynamikę cen energii na Rynku Dnia Następnego.

Rozdział 4

Eksploracja danych

W niniejszym rozdziale przedstawiono proces przygotowania omówionego zestawu danych. Proces eksploracji danych obejmuje obsługę braków wartości, dostosowanie rozdzielczości zmiennych, analizę korelacji, podział danych na okresy o różnej zmienności, podział na zbiory treningowy i testowy oraz preprocessing danych przed modelowaniem.

4.1 Wstępna obróbka danych

Obsługa braków wartości

Pierwszym etapem eksploracji danych jest analiza i obróbka brakujących wartości w zbiorze danych, co stanowi kluczowy krok w zapewnieniu jego jakości dla dalszej analizy. Brakujące wartości (NaN) w zmiennych objaśniających mogą wynikać z braku pomiarów w określonych godzinach, błędów w procesie gromadzenia danych lub ich celowego pominięcia. Początkowa ocena wykazała obecność 830 brakujących wartości oraz 120 zduplikowanych w zbiorze danych, który docelowo powinien obejmować 70128 rekordów godzinowych. Łączna liczba zduplikowanych i brakujących rekordów (950) stanowiła 1.35% w stosunku do oczekiwanej liczby rekordów.

Brakujące wartości rzadko stanowiły okres dłuższy, niż dwie godziny. Z tego powodu, w procesie obróbki zastosowano interpolację liniową do wypełnienia brakujących wartości, zapewniając ciągłość danych w ujęciu godzinowym. 120 zduplikowanych timestampów, z których usunięto drugie wystąpienie każdej pary, co pozwoliło na uzyskanie unikalnego zbioru 70128 rekordów. Takie podejście zminimalizowało utratę informacji i zapewniło spójność danych, umożliwiając ich dalsze wykorzystanie w modelowaniu.

Obsługa różnic w rozdzielczości danych

Dane użyte w pracy charakteryzowały się różną rozdzielczością czasową. PSE udostępnia dane dotyczące sieci energii elektrycznej, w tym handlu w rozdzielczości godzinowej. Zmienne pogodowe również mają rozdzielcość dzienną. Natomiast zmienne makroekonomiczne, takie jak ceny paliw, są dostępne w częstotliwości dniowej lub tygodniowej.

Aby ujednolicić rozdzielczość do poziomu godzinowego, zastosowano dwie techniki. W przypadku zmiennych o rozdzielczości dziennej, takich jak ceny gazu ziemnego, założono, że wartości w ciągu doby nie ulegają zmianie. Wartości dzienne przypisano więc każdej godzinie danego dnia, co pozwoliło na zachowanie prostoty przy jednocośnym dostosowaniu danych do godzinowej rozdzielczości ceny energii.

Dla zmiennych o rzadszej rozdzielczości, na przykład dla ceny węgla (miesięczna) oraz emisji CO₂ (tygodniowa), przeprowadzono uzupełnianie danych metodą interpolacji liniowej między sąsiednimi wartościami. Poniższy przykład dla rozdzielczości tygodniowej.

$$\text{cena w dniu } d = \text{cena w tygodniu } t + \left(\frac{\text{cena w tygodniu } t+1 - \text{cena w tygodniu } t}{6} \right) \times d$$

, gdzie $d \in [t, t+1]$

Następnie, podobnie jak w przypadku danych dziennych, wartości te przypisano każdej godzinie w danej dobie. Podejście to umożliwiło ujednolicenie wszystkich zmiennych do godzinowej rozdzielczości danych docelowych.

4.2 Analiza korelacji

Aby zbadać zależności między zmiennymi objaśniającymi a zmienną docelową, przeprowadzono analizę korelacji, wykorzystując dwa współczynniki: Pearsona i Spearmana. Wybór odpowiedniego współczynnika dla każdej zmiennej oparto na charakterze jej relacji z `fixing_i_price`, co pozwoliło na bardziej precyzyjne oszacowanie siły i rodzaju zależności.

Współczynnik korelacji Pearsona (r) mierzy liniową zależność między dwiema zmiennymi. Jest on zdefiniowany wzorem:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

gdzie x_i i y_i to wartości zmiennych i danej zmiennej objaśniającej), \bar{x} i \bar{y} to ich średnie, a n to liczba obserwacji. Współczynnik Pearsona przyjmuje wartości w przedziale $[-1, 1]$, gdzie $r = 1$ oznacza doskonałą dodatnią zależność liniową, $r = -1$ doskonałą ujemną zależność liniową, a $r = 0$ brak liniowej zależności.

Pearson jest odpowiedni dla zmiennych, których relacja jest liniowa. Na przykład wzrost zapotrzebowania zwykle większa ceny energii liniowo.

Współczynnik korelacji Spearmana (ρ) mierzy monotoniczną zależność między zmiennymi, co czyni go bardziej odpowiednim dla relacji nieliniowych. Spearman opiera się na rangach wartości zmiennych, a jego wzór to:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

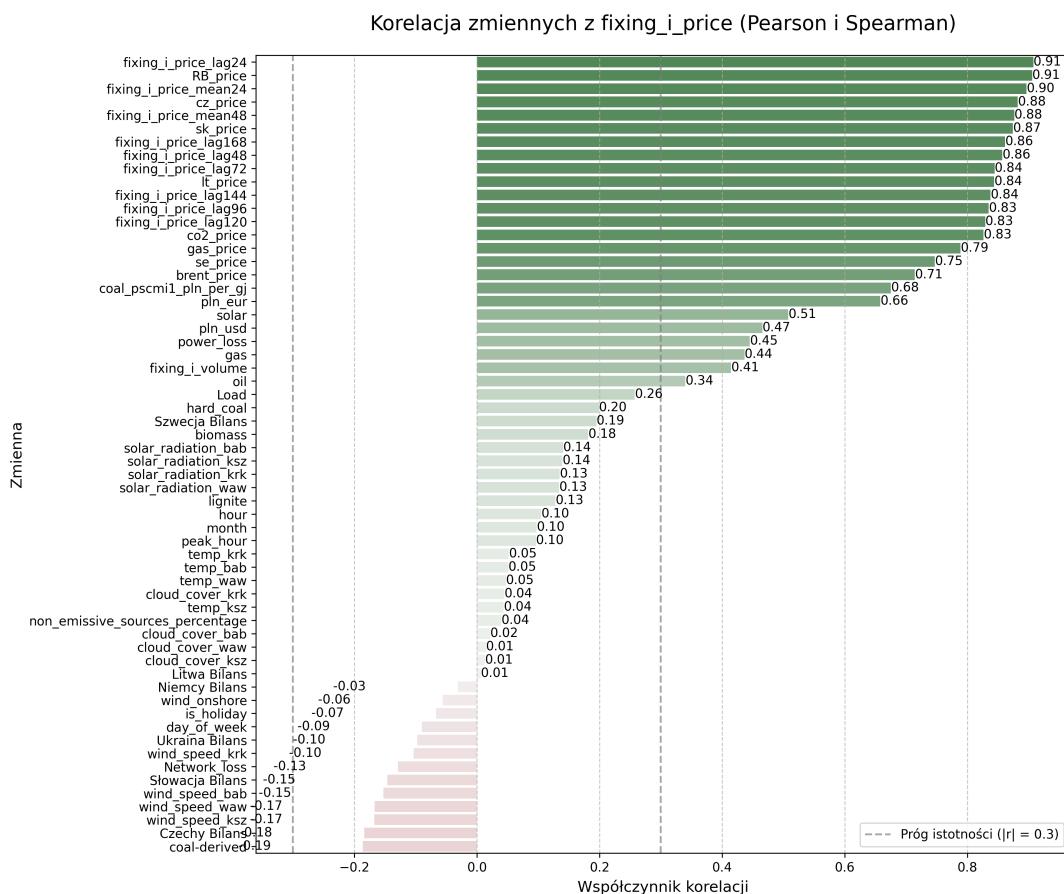
gdzie d_i to różnica między rangami wartości x_i i y_i . Współczynnik Spearmana również przyjmuje wartości w przedziale $[-1, 1]$, ale nie zakłada liniowości relacji - wystarczy, że wzrost jednej zmiennej odpowiada wzrostowi (lub spadkowi) drugiej w sposób monotoniczny.

Spearman jest szczególnie użyteczny dla zmiennych o nieliniowej relacji z `fixing_i_price`. Na przykład wzrost cen gazu w 2022 roku prowadził do nieproporcjonalnego wzrostu cen energii, co lepiej oddaje Spearman niż Pearson.

W celu wyboru odpowiedniego współczynnika korelacji obliczono zarówno korelację Pearsona, jak i Spearmana dla wszystkich zmiennych objaśniających względem `fixing_i_price`. Następnie obliczono bezwzględną różnicę między tymi współczynnikami ($|\rho - r|$). Zmienne, dla których różnica była większa niż 0,1, uznano za posiadające nieliniową relację z `fixing_i_price`, stosując dla nich korelację Spearmana. W pozostałych przypadkach wybrano korelację Pearsona, zakładając liniową zależność.

Analiza wykazała, że zmienne takie jak `co2_price` (różnica 0,114), `coal_pscmi1_pln_per_gj` (0,161), `pln_usd` (0,112), `solar` (0,172), `gas` (0,153), `biomass` (0,174), `coal-derived` (0,142) oraz zmienne związane z promieniowaniem słonecznym (różnice 0,106-0,115) mają nieliniową relację z `fixing_i_price`. Nieliniowość wynika z charakteru tych zmiennych - na przykład wysoka produkcja energii z fotowoltaiki w miesiącach letnich obniża ceny energii w sposób nieproporcjonalny, co prowadzi do bardzo niskich, a czasem ujemnych cen. Podobnie gwałtowny wzrost cen CO2 w okresie niespokojnym (2020-2023) zwiększa koszty produkcji energii w elektrowniach węglowych, ale efekt ten był wzmacniany przez inne czynniki, takie jak spekulacje rynkowe.

W wyniku tego powstał następujący wykres korelacji poniżej



Rysunek 18. Wykres korelacji zmiennych objaśniających wzgldem zmiennej docelowej fixing_i_price. Opracowanie własne.

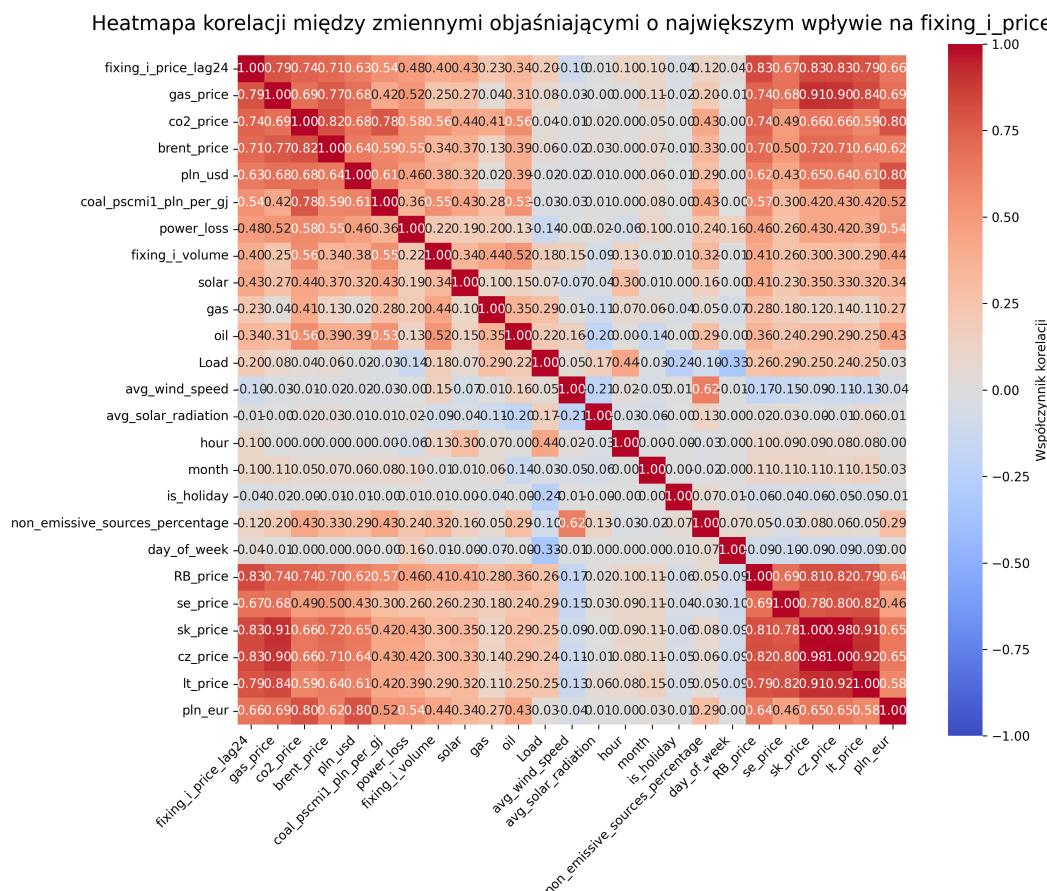
Największą korelację ze zmienną objaśnianą mają zmienne reprezentujące ceny, takie jak ceny na RB, ceny na rynkach sąsiednich, ceny opóźnione oraz średnie kroczące. Kolejnymi zmiennymi wykazującymi silną korelację są ceny na surowce, kurs polskiego złotego oraz zapotrzebowanie i wolumen. Sprzeczna z logiką może być pozytywna korelacja ceny z zmienną solar, która wskazuje na produkcję energii z paneli fotowoltaicznych. Wynika to prawdopodobnie z faktu, w momentach, gdy świeci słońce i produkcja energii z OZE jest wysoka, zapotrzebowanie również jest zwiększone i to powoduje wzrost cen. Z kolei wartości reprezentujące parametry pogodowe nie wykazują istotnej korelacji.

Produkcja energii z odnawialnych źródeł i gazu również odgrywa rolę. Zmienna solar (Spearman: 0,490) wskazuje, że wysoka produkcja energii z fotowoltaiki obniża ceny energii, szczególnie w miesiącach letnich, gdzie nadpodaż energii z OZE może prowadzić do bardzo niskich cen. gas (Spearman: 0,449) pokazuje, że produkcja energii z gazu ma nieliniowy wpływ, zależny od cen gazu i dostępności innych źródeł energii. Ponadto zmienne takie jak fixing_i_volume (Spearman: 0,442), power_loss (Spearman: 0,441), oil (Spearman: 0,349) oraz Load (Pearson: 0,256) mają

umiarkowany wpływ, odzwierciedlając znaczenie wolumenu obrotu, strat w sieci, produkcji z oleju oraz zapotrzebowania na energię.

Zbiór danych skrócony

Na podstawie analizy korelacji stworzony został skrócony zbiór danych, który z pierwotnych 54 regresorów zostawia najbardziej istotne. Tymi zmiennymi zostały wszystkie przekraczające próg istotności na poziomie 0.2, zmienne sezonowe określające dzień tygodnia, miesiąc, godzinę oraz świętą, średnia arytmetyczna ze zmiennych objaśniających prędkość wiatru w Polsce, gdyż wykazują one najmocniejszą odwrotną korelację oraz zmienna objaśniająca generację OZE w procentach, ponieważ jest to ważna zmienna z punktu widzenia literatury. W wyniku tego powstał zbiór danych z 33 zmiennymi objaśniającymi. Taki zbiór danych został określony jako zbiór danych skrócony i będzie użyty w dalszej części pracy do analizy w celu sprawdzenia istotności zbioru danych o największej ilości parametrów. Poniżej przedstawiono mapę cieplną dla zmiennych objaśniających w zbiorze danych skróconym bez uwzględnienia zmiennych opóźnionych i średnich.



Rysunek 19. Mapa cieplna korelacji zmiennych objaśniających w skróconym zbiorze danych. Opracowanie własne.

Z powyższej mapy cieplnej wynika, że zmienne o największej korelacji z ceną energii w dużym stopniu są również skorelowane pomiędzy sobą.

4.3 Podział danych

Podział na okresy spokojny i niespokojny

Dane zostały podzielone na dwa okresy w celu uwzględnienia różnych warunków rynkowych i ich wpływu na ceny energii. Okres spokojny (2016-2019) charakteryzuje się stabilnymi cenami energii, wynikającymi z braku znaczących szoków podażowych, łatwiej przewidywalnych cen paliw oraz łagodnego wzrostu cen CO₂ w ramach polityki klimatycznej UE. W tym okresie nie występowały większe kryzysy geopolityczne ani pandemie, co pozwoliło na utrzymanie cen w stosunkowo wąskim zakresie.

Okres niespokojny (2020-2023) został zdominowany przez szereg wydarzeń, które drastycznie wpłynęły na rynek energii. Pandemia COVID-19 w latach 2020-2021 początkowo obniżyła zapotrzebowanie na energię, ale ożywienie gospodarcze w 2021 roku spowodowało gwałtowny wzrost cen. Kryzys energetyczny w latach 2021-2022, związany z ograniczoną podażą gazu, rekordowymi cenami CO₂ oraz wysokimi cenami węgla, doprowadził do ekstremalnych skoków cen energii. Wybuch wojny na Ukrainie w 2022 roku dodatkowo zaostrzył sytuację, powodując przerwanie dostaw gazu z Rosji, sankcje i spekulacje rynkowe, co przełożyło się na rekordowe ceny energii. W tym okresie pojawiły się również ujemne ceny, wynikające z nadpodaży energii z OZE i ograniczonej elastyczności systemu elektroenergetycznego.

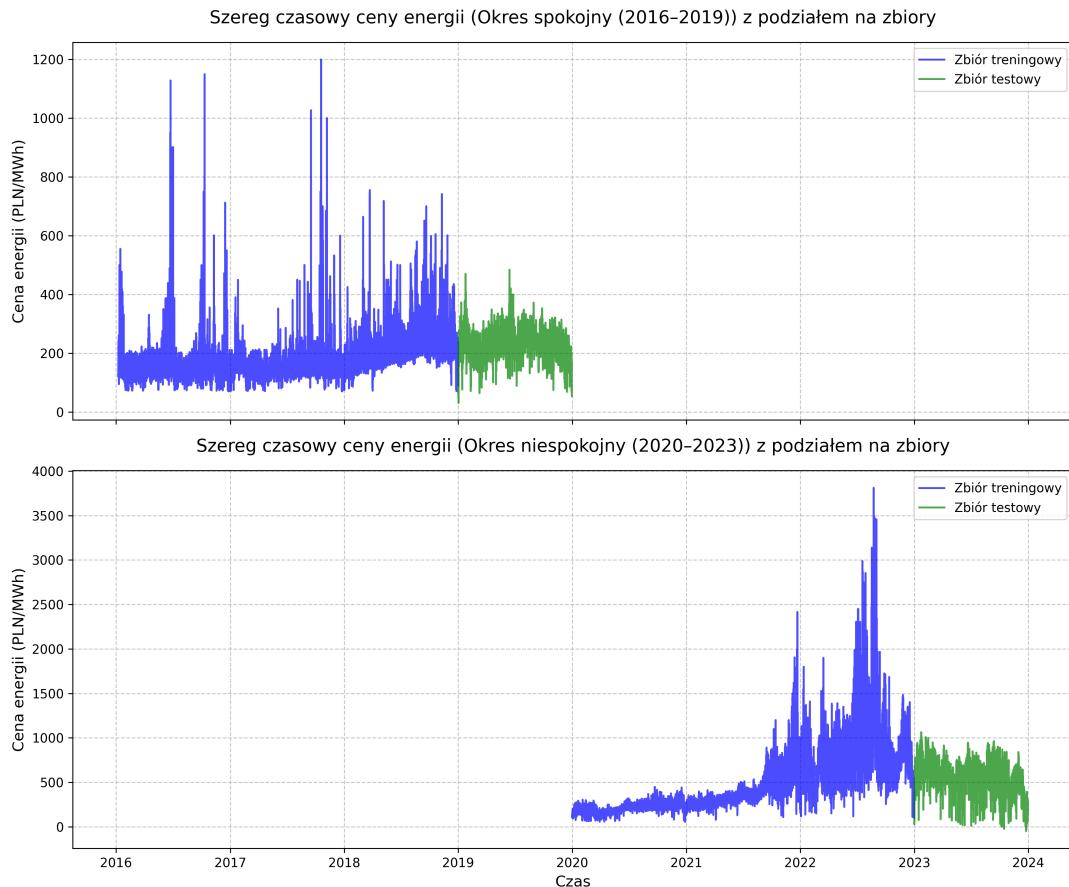
Statystyki opisowe dla obu okresów przedstawiono w tabeli 6. Okres spokojny charakteryzuje się niższą średnią ceną, mniejszą zmiennością i dużo mniejszym odchyleniem standardowym, co odzwierciedla stosunkowo stabilne warunki rynkowe. W okresie niespokojnym średnia cena wzrosła, a współczynnik zmienności i odchylenie znaczco rosną. W latach 2020-2023 pojawiły się również ujemne ceny oraz rekordowe maksima. Podział na te dwa okresy i poddanie ich osobnej analizie pozwoli lepiej ocenić skuteczność wybranych cech i modeli w różnych warunkach rynkowych.

Tabela 6. Porównanie statystyk opisowych cen energii w okresach spokojnym (2016-2019) i niespokojnym (2020-2023).

Miara	Okres spokojny	Okres niespokojny
Średnia (PLN/MWh)	193,51	478,06
Mediana (PLN/MWh)	182,00	412,00
Odchylenie standardowe (PLN/MWh)	70,67	321,39
Współczynnik zmienności (%)	36,52	67,23
Kwartyl Q1 (25%) (PLN/MWh)	143,66	246,41
Kwartyl Q3 (75%) (PLN/MWh)	229,35	609,00
Minimum (PLN/MWh)	31,00	-50,00
Maksimum (PLN/MWh)	1199,53	3812,45
Procent dni z ceną powyżej 500 PLN/MWh (%)	0,41	36,94

Podział danych na zbiory treningowe i testowe

Dane zostały podzielone na zbiory treningowe i testowe w obrębie każdego z dwóch okresów, aby uwzględnić strukturę szeregu czasowego. Podział został przeprowadzony sekwencyjnie w proporcji 75/25, co zapewnia dużą ilość danych do treningu (3 lata) oraz odpowiednią ilość danych do testowania (1 rok). W przypadku takiego podziału w zbiorach testowych można przetestować wszystkie okresy sezonowe.



Rysunek 20. Podział szeregu czasowego cen energii na zbiory treningowe i testowe. Opracowanie własne.

4.4 Przygotowanie danych

Przed przystąpieniem do modelowania dane zostały poddane szeregu kroków preprocessingu, aby zapewnić ich odpowiednią jakość i format dla wybranych modeli.

Pierwszym krokiem jest kodowanie zmiennych cyklicznych. Zmienne sezonowe mają charakter cykliczny (np. po godzinie 23 następuje 0, po grudniu następuje styczeń). Aby uwzględnić tę cykliczność, zastosowano kodowanie za pomocą funkcji sinusoidalnych:

- Dla `day_of_week`: $\sin\left(\frac{2\pi \cdot \text{day_of_week}}{7}\right)$,

- Dla month: $\sin\left(\frac{2\pi \cdot \text{month}}{12}\right)$,
- Dla hour: $\sin\left(\frac{2\pi \cdot \text{hour}}{24}\right)$.

Oryginalne zmienne zostały usunięte, a ich zakodowane wersje dodano do zbioru danych. Kodowanie sinusoidalne pozwala modelom lepiej uchwycić cykliczność danych, w przeciwieństwie do kodowania typu one-hot, które zwiększyłyby wymiarowość danych i nie uwzględniało cykliczności.

Następnie, wszystkie zmienne numeryczne niebinarne zostały poddane standaryzacji StandardScaler z biblioteki `sklearn`. Standaryzacja polega na przekształceniu zmiennych, aby miały średnią 0 i odchylenie standardowe 1. Wartości zmiennych zostały przekształcone według wzoru:

$$z = \frac{x - \mu}{\sigma},$$

gdzie z to wartość po standaryzacji, x to wartość przed standaryzacją, μ to średnia zmiennej, a σ to odchylenie standardowe.

Standaryzacja jest kluczowym krokiem w preprocessingu danych, ponieważ większość algorytmów uczenia maszynowego zakłada, że dane mają podobną skalę. W przeciwnym razie algorytmy mogą być wrażliwe na różnice w skali zmiennych, co prowadzi do nieoptimalnych wyników. Proces standaryzacji został przeprowadzony osobno dla każdego z okresów spokojnego i niespokojnego osobno, aby uwzględnić różnice w rozkładach danych między tymi okresami. W obrębie każdego okresu parametry standaryzacji (średnia i odchylenie standardowe) obliczono na zbiorze treningowym i zastosowano zarówno do danych treningowych, jak i testowych, aby uniknąć wycieku informacji.

Wartości odstające (np. ekstremalnie wysokie ceny w okresie niespokojnym, takie jak 3812,45 PLN/MWh) nie zostały zmodyfikowane, ponieważ odzwierciedlają rzeczywiste zjawiska rynkowe (np. kryzys energetyczny w 2022 roku). Ich wpływ na modele będzie monitorowany podczas analizy wyników.

Rozdział 5

Metodologia

W tym rozdziale przedstawiono metodologię przeprowadzonych badań. Rozdział składa się z dwóch sekcji. W pierwszej przedstawiono metodykę oceny jakości prognoz, a w drugiej omówiono metodykę prognozowania cen energii elektrycznej.

5.1 Ocena jakości prognoz

Ocena jakości modeli prognozowania cen energii elektrycznej jest kluczowym etapem analizy, ponieważ pozwala na porównanie skuteczności różnych podejść. W niniejszej pracy zastosowano następujące popularne metryki oceny: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (sMAPE) oraz R^2 . Wszystkie z tych metryk są omawiane w literaturze porównania skuteczności modeli [7]. W pracy prof. Werona [20] podano, że nie ma standardu obliczenia metryk EPF i wspomina o innych metrykach stosowanych przez innych autorów, między innymi wymieniono - Ważony Średni Błąd Bezwzględny (WMAE), średni błąd dniowy (MDE) i tygodniowy (MWE). Niemniej jednak, w tej pracy skupiono się na tych najszerzej stosowanych metrykach. Każda z tych metryk ma swoje zalety i ograniczenia, których omówienie jest przedstawione poniżej, wraz z ich matematycznymi definicjami i przykładami zastosowania w EPF.

5.1.1 Mean Absolute Error (MAE)

Mean Absolute Error jest jedną z najprostszych i najczęściej stosowanych metryk w prognozowaniu szeregów czasowych, w tym w EPF. MAE mierzy średnią wartość bezwzględnych błędów prognoz, co pozwala na ocenę dokładności modelu bez uwzględniania kierunku błędu (nad-lub niedoszacowania).

Matematyczna definicja MAE jest następująca:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|$$

gdzie:

- y_t to rzeczywista cena energii w godzinie t ,
- \hat{y}_t to przewidywana cena energii w godzinie t ,
- n to liczba obserwacji w zbiorze testowym.

MAE jest wyrażane w tej samej jednostce co prognozowane wartości (w omawianym przypadku jest to PLN/MWh), co łatwo interpretować. Na przykład, jeśli MAE wynosi 10 PLN/MWh, oznacza to, że średni błąd prognozy wynosi 10 PLN na każdą megawatogodzinę.

Zalety MAE:

- Prosta interpretacja i obliczenia.
- Równomierne traktowanie wszystkich błędów, niezależnie od ich kierunku.

Ograniczenia MAE:

- Nie uwzględnia kwadratu błędów, przez co nie penalizuje większych odchyleń w sposób szczególny, co może być problematyczne w EPF, gdzie duże skoki cen (np. w godzinach szczytu) są istotne.

5.1.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) jest kolejną popularną metryką w EPF, która uwzględnia kwadrat błędów, co powoduje większy wpływ dużych odchyleń między wartościami rzeczywistymi a przewidywanymi. RMSE jest szczególnie użyteczne w sytuacjach, gdzie duże błędy prognoz mogą mieć poważne konsekwencje ekonomiczne.

Definicja RMSE jest następująca:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}$$

gdzie:

- y_t , \hat{y}_t i n mają takie same znaczenie jak w MAE.

RMSE jest również wyrażane w jednostkach oryginalnych danych, co ułatwia interpretację. Na przykład, RMSE równe 15 PLN/MWh oznacza, że typowy błąd prognozy w sensie średniego kwadratu wynosi 15 PLN na megawatogodzinę.

Zalety RMSE:

- Większa wrażliwość na duże błędy, co jest istotne w EPF, gdzie skoki cen mogą być kosztowne.

Ograniczenia RMSE:

- Wrażliwość na wartości odstające - pojedyncze duże błędy mogą znaczco zawyżyć wartość RMSE.
- Mniej intuicyjne w interpretacji niż MAE, ponieważ kwadrat błędów zmienia skalę.

5.1.3 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) jest metryką wyrażającą błąd prognozy jako procent rzeczywistej wartości, co czyni ją szczególnie użyteczną w porównaniach między różnymi zbiorami danych lub rynkami o różnych poziomach cen.

Definicja MAPE jest następująca:

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \times 100$$

gdzie:

- y_t , \hat{y}_t i n mają takie same znaczenie jak wcześniej.

MAPE jest wyrażane w procentach, co ułatwia interpretację. Na przykład, MAPE równe 5% oznacza, że średni błąd prognozy wynosi 5% rzeczywistej ceny. W kontekście RDN, jeśli cena energii wynosi 200 PLN/MWh, a MAPE wynosi 5%, średni błąd wynosi 10 PLN/MWh.

Zalety MAPE:

- Intuicyjna interpretacja w procentach, nie trzeba zastanawiać się nad jednostkami bądź kursami walutowymi.

Ograniczenia MAPE:

- Problemy z wartościami bliskimi zera - jeśli y_t jest bardzo małe, co jest możliwe w godzinach nocnych, dzielenie przez y_t prowadzi do bardzo dużych wartości procentowych, a nawet do matematycznego błędu dzielenia przez zero.
- Asymetria - MAPE bardziej penalizuje niedoszacowania niż przeszacowania, co może prowadzić do nieobiektywnej oceny.

5.1.4 Symmetric Mean Absolute Percentage Error (sMAPE)

Symmetric Mean Absolute Percentage Error (sMAPE) jest zmodyfikowaną wersją MAPE, która rozwiązuje problem asymetrii i dzielenia przez zero. sMAPE uwzględnia zarówno rzeczywiste, jak i przewidywane wartości w mianowniku, co czyni ją bardziej stabilną w sytuacjach, gdy ceny energii są niskie.

Definicja sMAPE jest następująca:

$$\text{sMAPE} = \frac{1}{n} \sum_{t=1}^n \frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2} \times 100$$

gdzie:

- y_t , \hat{y}_t i n mają takie same znaczenie jak wcześniej.

Podobnie jak MAPE, sMAPE jest wyrażane w procentach. Na przykład, sMAPE równe 4% oznacza, że średni błąd symetryczny wynosi 4% średniej wartości rzeczywistej i przewidywanej ceny.

Zalety sMAPE:

- Rozwiązuje problem dzielenia przez zero, co jest istotne w EPF, gdzie ceny mogą być bliskie zera.

- Symetria - traktuje oszacowanie w dół i w górę w bardziej zrównoważony sposób niż MAPE.

Ograniczenia sMAPE:

- Nadal może być wrażliwe na skrajne wartości, choć w mniejszym stopniu niż MAPE.
- Interpretacja jest mniej intuicyjna niż w przypadku MAE czy RMSE, ponieważ uwzględnia zarówno y_t , jak i \hat{y}_t w mianowniku.

5.1.5 Współczynnik determinacji

Współczynnik determinacji, oznaczany jako R^2 , jest metryką powszechnie stosowaną w analizie regresji i prognozowaniu. R^2 mierzy, jak dobrze model wyjaśnia zmienność danych rzeczywistych, czyli jaki procent wariancji zmiennej zależnej jest wyjaśniony przez model prognostyczny. Jest to metryka szczególnie użyteczna w ocenie modeli liniowych, ale znajduje zastosowanie również w bardziej złożonych modelach, w celu ogólnej oceny ich dopasowania do danych.

Definicja R^2 jest następująca:

$$R^2 = 1 - \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

gdzie:

- y_t , \hat{y}_t i n mają takie same znaczenie jak wcześniej,
- n to liczba obserwacji w zbiorze testowym.

Licznik w wyrażeniu $\sum_{t=1}^n (y_t - \hat{y}_t)^2$ to suma kwadratów reszt, czyli całkowity błąd modelu, natomiast mianownik $\sum_{t=1}^n (y_t - \bar{y})^2$ to całkowita suma kwadratów, czyli całkowita wariancja danych względem ich średniej. R^2 przyjmuje wartości w przedziale od 0 do 1, gdzie:

- $R^2 = 1$ oznacza, że model idealnie przewiduje wszystkie wartości (błąd wynosi 0),
- $R^2 = 0$ oznacza, że model nie wyjaśnia żadnej zmienności danych i jest równoważny prostemu modelowi średniej ($\hat{y}_t = \bar{y}$).

W kontekście EPF, na przykład na RDN, R^2 równe 0,85 oznaczałoby, że model wyjaśnia 85% zmienności cen energii.

Zalety R^2 :

- Intuicyjna interpretacja - R^2 jasno wskazuje, jaki procent zmienności danych jest wyjaśniony przez model.
- Bez jednostek - umożliwia porównanie modeli na różnych zbiorach danych, niezależnie od skali cen (np. PLN/MWh na RDN vs. EUR/MWh na EEX).

Ograniczenia R^2 :

- Wrażliwość na przeuczenie - R^2 może być zawyżone w modelach o dużej liczbie parametrów, szczególnie w przypadku małych zbiorów danych, co może prowadzić do mylnego wniosku o dobrym dopasowaniu modelu.

- Brak informacji o kierunku błędów - R^2 nie rozróżnia, czy model przecenia czy niedocenia wartości, co w EPF może być istotne z ekonomicznego punktu widzenia.

W niniejszej pracy R^2 zostanie wykorzystane jako dodatkowa metryka oceny, aby uzupełnić analizę opartą na MAE, RMSE, MAPE, sMAPE.

5.2 Wybrane metody weryfikacji zbioru danych

Stworzony zbiór danych z cechami objaśniającymi ceny energii elektrycznej należy zweryfikować pod kątem jego skuteczności. W związku z tym zostały wybrane cztery metody prognozowania.

5.2.1 Regresja liniowa

Opis metody

Regresja liniowa jest jednym z najprostszych i najczęściej stosowanych modeli statystycznych w analizie zbiorów danych. Zakłada liniową zależność między zmienną zależną, a zestawem zmiennych niezależnych (predyktorów). W kontekście EPF regresja liniowa jest często stosowana jako model bazowy, który pozwala na szybkie uzyskanie prognoz i ocenę wpływu poszczególnych zmiennych na ceny energii. Jej zaletą jest prostota interpretacji oraz niski koszt obliczeniowy, co czyni ją odpowiednią do analiz dużych zbiorów danych, co czyni ją odpowiednią do tej pracy.

Wzór modelu

Model regresji liniowej można zapisać jako:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad (1)$$

gdzie:

- y - zmienna zależna, cena energii elektrycznej
- β_0 - wyraz wolny,
- $\beta_1, \beta_2, \dots, \beta_p$ - współczynniki regresji dla zmiennych niezależnych,
- x_1, x_2, \dots, x_p - zmienne niezależne (predyktoři, np. zmienne związane z zapotrzebowaniem, cenami paliw czy danymi kalendarzowymi),
- ϵ - składnik losowy (błąd), zakładany jako $\epsilon \sim N(0, \sigma^2)$.

W macierzowej formie model przyjmuje postać:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (2)$$

gdzie:

- \mathbf{y} - wektor obserwacji zmiennej zależnej,
- \mathbf{X} - macierz projektowa zawierająca wartości zmiennych niezależnych,
- $\boldsymbol{\beta}$ - wektor współczynników regresji,
- $\boldsymbol{\epsilon}$ - wektor błędów.

Estymacja parametrów

Parametry modelu β są estymowane za pomocą metody najmniejszych kwadratów (OLS), która minimalizuje sumę kwadratów błędów:

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3)$$

gdzie $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ to przewidywana wartość dla i -tej obserwacji. Rozwiążanie analityczne to:

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (4)$$

Istotne parametry modelu

W niniejszej pracy regresja liniowa została zaimplementowana za pomocą biblioteki scikit-learn w Pythonie. Kluczowe parametry modelu obejmują:

- `fit_intercept=True`: Włączenie wyrazu wolnego (β_0).
- `normalize=False`: Brak normalizacji zmiennych przed estymacją.
- `solver='auto'`: Automatyczny wybór algorytmu estymacji, domyślnie OLS.

Zalety i ograniczenia

Regresja liniowa jest łatwa do interpretacji, ponieważ współczynniki β_j wskazują, o ile zmieni się cena energii przy wzroście zmiennej x_j o jednostkę (przy założeniu stałości pozostałych zmiennych). Jednak model zakłada liniowe zależności między zmiennymi, co może być ograniczeniem w przypadku bardziej złożonych, nieliniowych wzorców w danych cen energii, szczególnie w okresie niespokojnym.

5.2.2 Regresja grzbietowa (Ridge)

Opis metody

Regresja grzbietowa (ang. Ridge Regression) jest rozszerzeniem regresji liniowej, które wprowadza regularyzację L2, aby zapobiec przeuczeniu i poprawić stabilność modelu w przypadku współliniowości między zmiennymi objaśniającymi. W prognozowaniu cen energii regresja grzbietowa jest szczególnie użyteczna, gdy zestaw danych zawiera wiele zmiennych, które mogą być skorelowane. Regularizacja pozwala na zmniejszenie wpływu mniej istotnych zmiennych, co poprawia generalizację modelu.

Wzór modelu

Model regresji grzbietowej opiera się na tej samej zależności liniowej co regresja liniowa:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad (5)$$

Jednak estymacja parametrów uwzględnia dodatkową karę regularizacyjną L2. Funkcja kosztu w regresji grzbietowej to:

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \quad (6)$$

gdzie:

- Pierwsza część ($\sum_{i=1}^n (y_i - \hat{y}_i)^2$) to suma kwadratów błędów, jak w OLS.
- Druga część ($\lambda \sum_{j=1}^p \beta_j^2$) to kara L2 na wielkość współczynników β_j .
- $\lambda \geq 0$ - parametr regularizacji, który kontroluje siłę kary (większe λ oznacza silniejszą regularizację).

W macierzowej formie funkcja kosztu to:

$$\min \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \quad (7)$$

Rozwiążanie analityczne dla parametrów to:

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (8)$$

gdzie \mathbf{I} to macierz jednostkowa.

Istotne parametry modelu

Regresja grzbietowa została zaimplementowana w Pythonie za pomocą biblioteki scikit-learn.

Kluczowe parametry modelu to:

- `alpha=1.0`: Domyślna wartość parametru regularizacji λ . W pracy metodą empiryczną spróbuje się różnych wartości `alpha` (np. 0.1, 1.0, 10.0, 100.0) za pomocą walidacji krzyżowej, aby wybrać optymalną.
- `fit_intercept=True`: Włączenie wyrazu wolnego (β_0).
- `normalize=False`: Brak normalizacji zmiennych przed estymacją (zmienne przeskalowane wcześniej za pomocą StandardScaler).
- `solver='auto'`: Automatyczny wybór algorytmu (domyślnie Cholesky dla małych zbiorów danych lub SAG dla dużych).

Zalety i ograniczenia

Regresja grzbietowa jest bardziej odporna na współliniowość i przeuczenie od regresji liniowej, co czyni ją odpowiednią do zestawów danych z dużą liczbą zmiennych objaśniających. Jest to szczególnie przydatne, gdyż w pracy uwzględnione zostają parametry temperatury z całej Polski, które zdecydowanie mają korelację. Jednak, podobnie jak regresja liniowa, zakłada liniowe zależności, co może ograniczać jej skuteczność w modelowaniu bardziej złożonych wzorców, szczególnie w niestabilnych okresach rynkowych.

5.2.3 Prophet

Opis metody

Prophet [18] to model prognozowania szeregów czasowych opracowany przez Facebooka, zaprojektowany do analizy danych z wyraźną sezonowością i trendami, które mogą ulegać zmianom w czasie. W kontekście prognozowania cen energii elektrycznej (EPF) Prophet jest szczególnie ciekawy ze względu na zdolność do modelowania cyklicznych wzorców jakie występują na tym rynku oraz uwzględniania efektów specjalnych, takich jak święta. Model jest oparty na addytywnym

podejściu, które rozkłada szereg czasowy na składowe trendu, sezonowości i efektów dodatkowych. Jego intuicyjna parametryzacja i możliwość automatycznego dopasowania do danych czynią go atrakcyjnym narzędziem w analizie dużych zbiorów danych, takich jak te wykorzystane w niniejszej pracy.

Wzór modelu

Prophet modeluje zmienną zależną jako sumę trzech głównych składowych plus składnik losowy:

$$y(t) = g(t) + s(t) + h(t) + r(t) + \epsilon_t \quad (9)$$

gdzie:

- $y(t)$ - wartość prognozowana,
- $g(t)$ - składowa trendu, modelująca długoterminowe zmiany w danych,
- $s(t)$ - składowa sezonowości, modelująca cykliczne wzorce (np. dobowe, tygodniowe),
- $h(t)$ - składowa efektów specjalnych takich jak święta,
- $r(t)$ - składowa zmiennych objaśniających, uwzględniająca wpływ dodatkowych regresorów, takich jak zapotrzebowanie czy dane pogodowe,
- ϵ_t - składnik losowy (błąd), zakładany jako $\epsilon_t \sim N(0, \sigma^2)$.

Składowa trendu ($g(t)$)

Trend w modelu Prophet jest modelowany za pomocą nieliniowej funkcji z punktami zmiany (ang. changepoints), które pozwalają na elastyczne dopasowanie do nagłych zmian w danych. Standardowo używa się funkcji liniowej z punktami zmiany:

$$g(t) = (k + \mathbf{a}(t)^T \boldsymbol{\delta})t + (m + \mathbf{a}(t)^T \boldsymbol{\gamma}) \quad (10)$$

gdzie:

- k - współczynnik nachylenia trendu,
- m - wyraz wolny,
- $\mathbf{a}(t)$ - wektor binarny wskazujący punkty zmiany,
- $\boldsymbol{\delta}$ - wektor zmian nachylenia w punktach zmiany,
- $\boldsymbol{\gamma}$ - wektor przesunięć dla ciągłości trendu w punktach zmiany.

Składowa sezonowości ($s(t)$)

Sezonowość jest modelowana za pomocą szeregu Fouriera, który aproksymuje cykliczne wzorce:

$$s(t) = \sum_{n=1}^N \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \sin\left(\frac{2\pi n t}{P}\right) \right) \quad (11)$$

gdzie:

- P - okres sezonowości (np. 24 godziny dla sezonowości dobowej, 168 godzin dla tygodniowej),
- a_n, b_n - współczynniki szeregu Fouriera,
- N - liczba składników szeregu (kontrolowana przez parametr `fourier_order`).

Składowa efektów specjalnych ($h(t)$)

Efekty specjalne, takie jak święta, są modelowane jako:

$$h(t) = \mathbf{Z}(t)\boldsymbol{\kappa} \quad (12)$$

gdzie:

- $\mathbf{Z}(t)$ - macierz binarna wskazująca wystąpienie efektów specjalnych (np. 1 dla dni świątecznych, 0 w pozostałych),
- $\boldsymbol{\kappa}$ - wektor efektów dla każdego zdarzenia.

Składowa zmiennych objaśniających ($r(t)$)

Zmienne objaśniające, takie jak zapotrzebowanie, dane pogodowe czy bilanse handlowe, są uwzględniane jako dodatkowa już znana składowa liniowa:

$$r(t) = \beta_1 x_1(t) + \beta_2 x_2(t) + \cdots + \beta_p x_p(t) \quad (13)$$

Estymacja parametrów

Parametry modelu ($k, m, \delta, \gamma, a_n, b_n, \boldsymbol{\kappa}, \beta_1, \dots, \beta_p$) są estymowane za pomocą maksymalizacji funkcji wiarogodności lub metod bayesowskich. Prophet wykorzystuje algorytm L-BFGS do optymalizacji w trybie domyślnym, co zapewnia szybkie dopasowanie modelu. Punkty zmiany są automatycznie wykrywane, a ich liczba i rozmieszczenie są kontrolowane przez parametry modelu, takie jak `n_changepoints` i `changepoint_prior_scale`.

Istotne parametry modelu

W niniejszej pracy model Prophet został zaimplementowany w Pythonie za pomocą biblioteki `prophet`. Dane wejściowe zostały przygotowane w formacie wymaganym przez Prophet, gdzie kolumna `ds` zawiera znaczniki czasowe (`timestamp`), a kolumna `y` zawiera ceny energii (`fixing_i_price`). Kluczowe parametry modelu to:

- `n_changepoints`: Liczba punktów zmiany trendu, umożliwiająca dopasowanie do potencjalnych zmian w danych (domyślna wartość 25).
- `changepoint_prior_scale`: Siła regularizacji punktów zmiany (domyślna wartość 0.05).
- `yearly_seasonality=False`: Wyłączenie sezonowości rocznej, ponieważ dane cen energii wykazują głównie sezonowość dobową i tygodniową.
- `weekly_seasonality=True`: Włączenie sezonowości tygodniowej.
- `daily_seasonality=True`: Włączenie sezonowości dobowej.
- `fourier_order`: Liczba składników szeregu Fouriera dla każdej sezonowości (domyślna wartość 10).
- `holidays`: Włączono efekty dni świątecznych na podstawie zmiennej `is_holiday` z danych.
- Dodatkowe regresory: W przypadku pełnego zestawu danych wszystkie zmienne objaśniające, takie jak zapotrzebowanie, dane pogodowe, bilanse handlowe czy ceny paliw, zostały dodane za pomocą funkcji `add_regressor`.

Zalety i ograniczenia

Prophet jest intuicyjny i dobrze radzi sobie z danymi o wyraźnej sezonowości, co czyni go odpowiednim do modelowania cen energii w stabilnych okresach. Automatyczne wykrywanie punktów zmiany, obsługa efektów specjalnych oraz możliwośćłączenia wszystkich zmiennych objaśniających ułatwiają jego stosowanie w praktyce. Jednak model może mieć trudności z modelowaniem bardzo dużych wahań cen, takich jak te obserwowane w okresie niespokojnym (2020-2023), szczególnie jeśli zmienne objaśniające nie w pełni tłumaczą zmienność. Ponadto Prophet zakłada addytywną strukturę szeregu czasowego, co może ograniczać jego zdolność do wychwytywania bardziej złożonych, nieliniowych zależności.

5.2.4 Wielowarstwowy perceptron (MLP)

Opis metody

Wielowarstwowy perceptron (MLP) to rodzaj sztucznej sieci neuronowej wykorzystywany w zadaniach uczenia maszynowego. Składa się z warstw neuronów: wejściowej, ukrytych i wyjściowej, które są w pełni połączone.

Wzór modelu

MLP przekształca wektor zmiennych wejściowych $\mathbf{x} = [x_1, x_2, \dots, x_p]^T$ w wartość prognozowaną \hat{y} (cenę energii, oznaczaną w pracy jako `fixing_i_price`) poprzez sekwencję warstw neuronów. Dla sieci z jedną warstwą ukrytą model można zapisać jako:

$$\hat{y} = f_o(\mathbf{w}_o^T \mathbf{h} + b_o) \quad (14)$$

gdzie:

- $\mathbf{h} = f_h(\mathbf{W}_h \mathbf{x} + \mathbf{b}_h)$ - wektor aktywacji warstwy ukrytej,
- $\mathbf{x} = [x_1, x_2, \dots, x_p]^T$ - wektor zmiennych objaśniających,
- \mathbf{W}_h - macierz wag między warstwą wejściową a ukrytą,
- \mathbf{b}_h - wektor biasów warstwy ukrytej,
- $f_h(\cdot)$ - funkcja aktywacji warstwy ukrytej (np. `tanh`),
- \mathbf{w}_o - wektor wag między warstwą ukrytą a wyjściową,
- b_o - bias warstwy wyjściowej,
- $f_o(\cdot)$ - funkcja aktywacji warstwy wyjściowej (dla regresji zazwyczaj liniowa, tj. $f_o(z) = z$).

Dla sieci z wieloma warstwami ukrytymi proces jest analogiczny, z kolejnymi przekształceniami dla każdej warstwy:

$$\mathbf{h}_k = f_k(\mathbf{W}_k \mathbf{h}_{k-1} + \mathbf{b}_k), \quad k = 1, 2, \dots, K \quad (15)$$

gdzie $\mathbf{h}_0 = \mathbf{x}$, K to liczba warstw ukrytych, a \mathbf{h}_K to wejście do warstwy wyjściowej.

W części analitycznej pracy są przedstawione wyniki z różną ilością warstw ukrytych.

Estymacja parametrów

Parametry modelu ($\mathbf{W}_k, \mathbf{b}_k$ dla każdej warstwy oraz \mathbf{w}_o, b_o) są estymowane przez minimalizację

funkcji kosztu, czyli średniego błędu kwadratowego (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (16)$$

gdzie y_i to rzeczywista cena energii, a \hat{y}_i to przewidywana wartość dla i -tej obserwacji.

Estymacja parametrów odbywa się za pomocą algorytmu wstecznej propagacji błędu (backpropagation) w połączeniu z optymalizatorem, takim jak Adam. Proces treningu polega na iteracyjnym dostosowywaniu wag i biasów w celu zmniejszenia błędu na zbiorze treningowym, z uwzględnieniem walidacji na oddzielnym zbiorze danych w celu uniknięcia przeuczenia.

Istotne parametry modelu

W niniejszej pracy model MLP został zaimplementowany w Pythonie za pomocą modułu Keras z biblioteki TensorFlow. Kluczowe parametry modelu, wraz z ich domyślnymi wartościami w bibliotece Keras, obejmują:

- `units`: Liczba neuronów w każdej warstwie ukrytej.
- `activation`: Funkcja aktywacji dla warstw ukrytych (domyślnie `relu` dla warstw gęstych).
- `optimizer`: Algorytm optymalizacji (domyślnie `rmsprop` dla modelu sekwencyjnego).
- `learning_rate`: Szybkość uczenia dla optymalizatora (domyślnie 0.001 dla optimizera Adam).
- `batch_size`: Rozmiar partii danych w każdej iteracji treningu (domyślnie 32 w metodzie `fit`).
- `epochs`: Maksymalna liczba epok treningu.

Zalety i ograniczenia

MLP jest elastycznym modelem zdolnym do wychwytywania nieliniowych zależności w danych, co czyni go odpowiednim do modelowania cen energii w okresach o dużej zmienności. Możliwość dostosowania architektury sieci i hiperparametrów pozwala na optymalizację modelu pod kątem specyfiki danych. Jednak MLP wymaga starannego doboru hiperparametrów i preprocessingu danych, a jego trening jest bardziej kosztowny obliczeniowo niż w przypadku modeli statystycznych. Ponadto model może być podatny na przeuczenie, jeśli liczba warstw lub neuronów jest zbyt duża w stosunku do dostępnych danych.

Rozdział 6

Estymacja przykładowych modeli

W niniejszym rozdziale przedstawiono wyniki modelowania cen energii za pomocą czterech modeli: regresji liniowej, regresji Ridge, Propheta oraz MLP. Analiza została podzielona na dwa podrozdziały, odpowiadające okresom stabilnemu i niestabilnemu. Zbiór danych został poddany testom w celu oceny skuteczności. W celu uzyskania najlepszych wyników dla modeli przeprowadzono strojenie hiperparametrów.

6.1 Okres stabilny

Regresja liniowa i Ridge

Przeanalizowano modele regresji liniowej oraz regresji Ridge na danych z okresu stabilnego. Modele trenowano na danych z lat 2016–2018, a testowano na danych z 2019 roku.

Wyniki dla pełnego zbioru o 60 parametrach wejściowych przedstawiono w tabeli 7. Regresja Ridge osiągnęła lepsze wyniki niż regresja liniowa, co jest widoczne w tabeli poniżej na podstawie przedstawionych wcześniej metryk 5.1. Wynika to z powodu regularizacji L2 zastosowanej w modelu Ridge.

Tabela 7. Wyniki regresji liniowej i grzebietowej dla pełnego zbioru danych w okresie stabilnym (2019).

Model	MAE	RMSE	MAPE (%)	sMAPE (%)	R^2
Regresja liniowa	15.18	19.79	7.25	7.16	0.8413
Regresja Ridge	15.09	19.64	7.19	7.09	0.8437

Najlepsza wartość hiperparametru α w regresji Ridge wyniosła 500.0. Wysoka wartość $\alpha = 500.0$ sugeruje, że w pełnym zbiorze danych występuje istotna współliniowość między zmiennymi objaśniającymi. Potwierdza to również analiza macierzy korelacji 18.

Następnie przeprowadzono analizę skróconego zbioru danych opisanego w 4.2. Wyniki dla skróconego zbioru danych są przedstawione w tabeli 8, wraz z różnicami w metrykach względem pełnego zbioru danych. Różnice w metrykach wskazują, że dodatkowe zmienne w pełnym zbiorze

danych wnoszą informację, mimo niższego poziomu korelacji ze zmienną objaśnianą. Różnice metryk nie są bardzo duże (np. MAPE różni się o 0.16% dla regresji Ridge), co może sugerować, że skrócony zbiór danych nadal zawiera najważniejsze zmienne objaśniające.

Tabela 8. Wyniki regresji liniowej i Ridge dla skróconego zbioru danych w okresie stabilnym (2019) wraz z różnicami względem pełnego zbioru.

Model	MAE	RMSE	MAPE (%)	sMAPE (%)	R ²	Różnica MAPE (%)
Regresja liniowa	15.61	20.31	7.33	7.32	0.8328	+0.08
Regresja Ridge	15.67	20.40	7.35	7.34	0.8314	+0.16

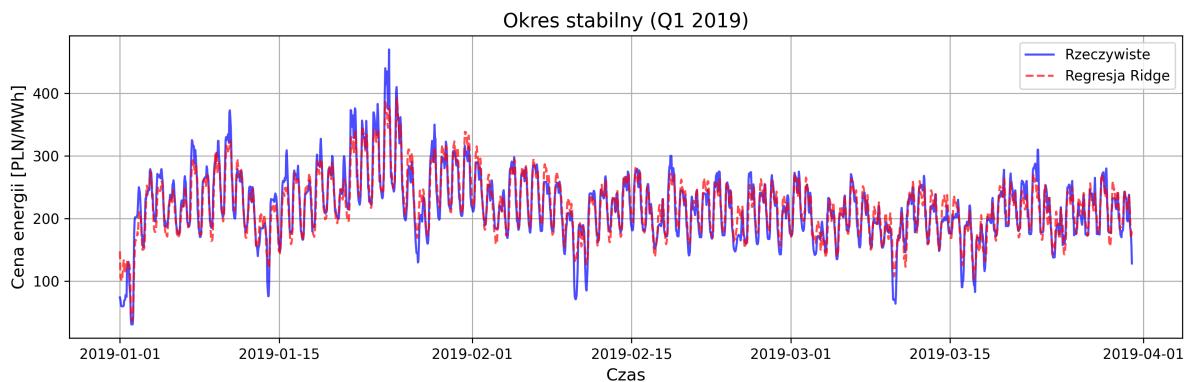
W celu potencjalnego polepszenia wyników zastosowano logarytmizację zmiennej wyjściowej (*fixing_i_price*), co miało na celu zmniejszenie skośności rozkładu cen i poprawę dopasowania modelu. Wyniki z logarytmizacją przedstawiono w tabeli 9.

Tabela 9. Wyniki regresji liniowej i Ridge z logarytmizacją dla okresu stabilnego (2019).

Model i zbiór danych	MAE	RMSE	MAPE (%)	sMAPE (%)	R ²
Regresja liniowa (pełny)	21.63	28.36	9.85	9.16	0.6736
Regresja Ridge (pełny)	19.99	26.06	9.21	8.65	0.7244
Regresja liniowa (skrócony)	26.36	33.07	11.95	10.94	0.5563
Regresja Ridge (skrócony)	23.29	29.38	10.69	9.89	0.6498

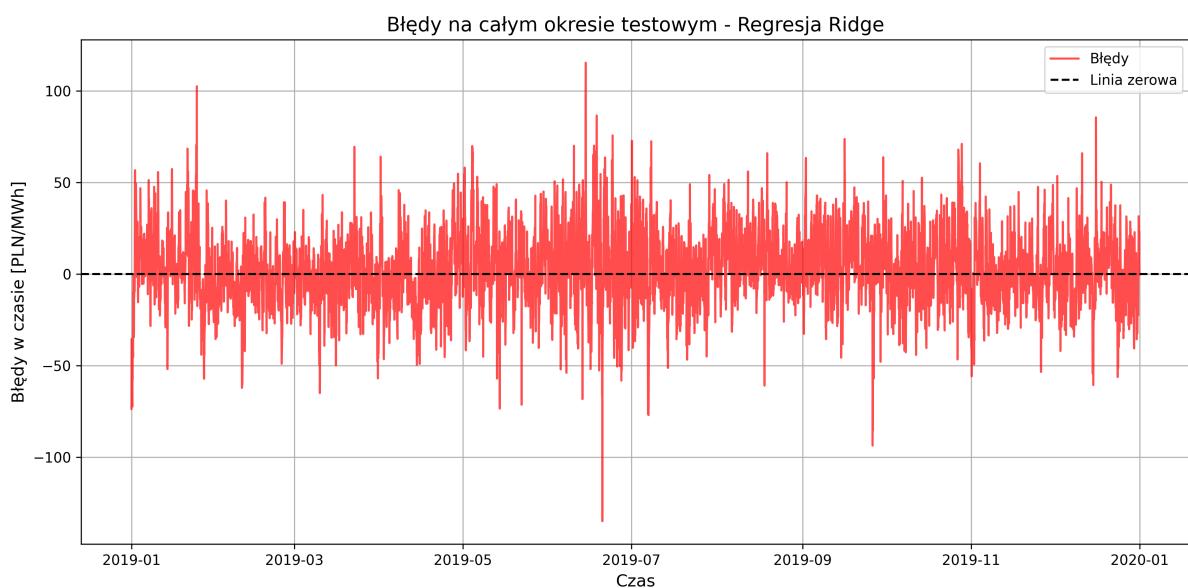
Logarytmizacja nie przyniosła spodziewanych korzyści i pogorszyła wyniki we wszystkich metrykach. Największe pogorszenie zaobserwowano dla skróconego zbioru danych, gdzie MAPE dla regresji liniowej wzrosło do 11.95%, a R^2 spadło do 0.5563. Przyczyną jest najprawdopodobniej fakt, że logarytmizacja wprowadziła niepotrzebne nielinowości, które utrudniły dopasowanie modeli liniowych. Dodatkowo, odwrócenie transformacji logarytmicznej może amplifyfikować błędy predykcji, co wpłynęło na wzrost RMSE i MAE.

Regresja Ridge lepiej przewiduje ceny w porównaniu do regresji liniowej, co potwierdzają wyniki metryk. Poniżej umieszczam wykres rzeczywistych i przewidywanych wartości dla regresji Ridge w okresie stabilnym. Załączony został wykres z pierwszego kwartału 2019 roku, ponieważ wykres z całego 2019 nie jest czytelny z powodu dużej liczby punktów.



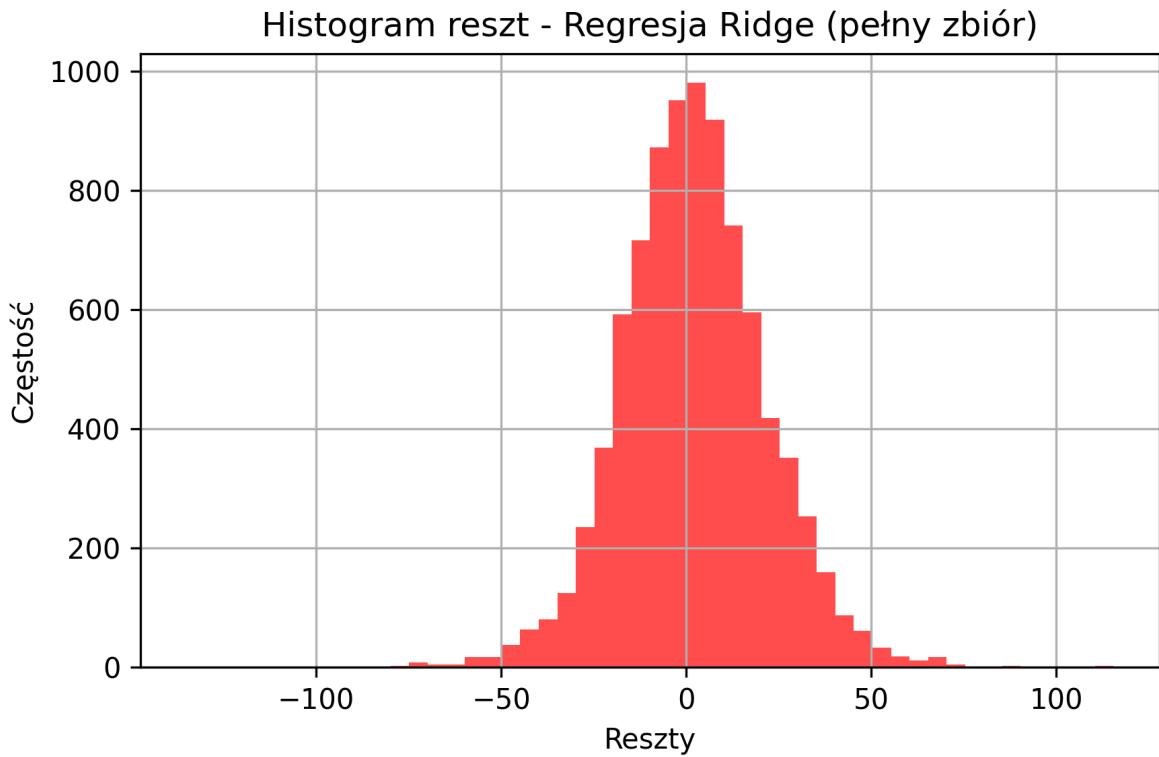
Rysunek 21. Porównanie rzeczywistych i przewidywanych wartości cen energii dla regresji Ridge w okresie stabilnym. Opracowanie własne.

Widoczne na wykresie jest to, że model nie zawsze nadąża za dużymi skokami cen energii, prawdopodobnie z powodu ich nielinowości. Z kolei w przypadku okresów mniejszych zmian i stabilniejszych cen, model oczekuje większej zmienności, co prowadzi do przeszacowania prognoz. Poniżej załączony jest wykres błędów dla regresji Ridge.



Rysunek 22. Błędy prognoz dla regresji Ridge w okresie stabilnym. Opracowanie własne.

Błędy prognoz dla regresji Ridge w okresie stabilnym są rozproszone wokół zera, co sugeruje, że model dobrze radzi sobie z przewidywaniem cen energii elektrycznej. Wartości błędów przeważnie nie przekraczają 50 PLN/MWh. Widoczne są szczyty przekraczające 100 PLN/MWh, które mogą być wynikiem dużych skoków cen energii elektrycznej. Poniżej przedstawiony jest histogram reszt, żeby zobrazować rozkład błędów prognoz.



Rysunek 23. Histogram reszt dla regresji Ridge w okresie stabilnym. Opracowanie własne.

Histogram posiada szczyt w okolicy zera i większość błędów skupia się w okolicy zera. Rozkład reszt jest zbliżony do normalnego bez istotnych odchyleń. Wartości reszt są rozproszone w okolicy zera, co sugeruje, że model dobrze radzi sobie z przewidywaniem cen energii elektrycznej w okresie stabilnym. Widoczne na wykresie 23 szczyty dochodzące do 100 PLN/MWh nie są widoczne na histogramie, co sugeruje, że są to pojedyncze przypadki dużych błędów prognozowania. Wartości tych błędów nie mają wyraźnych wzorców czasowych ani sezonowych. Pojawienie się dużych błędów wynika z fluktuacji cen energii elektrycznej, wynikających z czynników nie wyjaśnialnych przez zbiór zmiennych objaśniających lub czynników zewnętrznych.

Prophet

Model Prophet został skonfigurowany z trybem addytywnym (`seasonality_mode='additive'`), ponieważ wstępne testy wykazały, że tryb mnożnikowy (`seasonality_mode='multiplicative'`) działa dużo gorzej w przypadku zebranych danych. Tryb mnożnikowy zakłada proporcjonalne skalowanie efektów sezonowych i trendów względem wartości zmiennej objaśnianej, co nie jest odpowiednie dla cen energii elektrycznej, gdzie efekty sezonowe (np. różnice między dniami roboczymi a weekendami) mają raczej charakter addytywny, a nie proporcjonalny.

Testowano różne kombinacje pozostałych hiperparametrów opisanych w rozdziale 5.2.3, aby znaleźć optymalne ustawienia. Najlepsze rezultaty uzyskano dla kombinacji

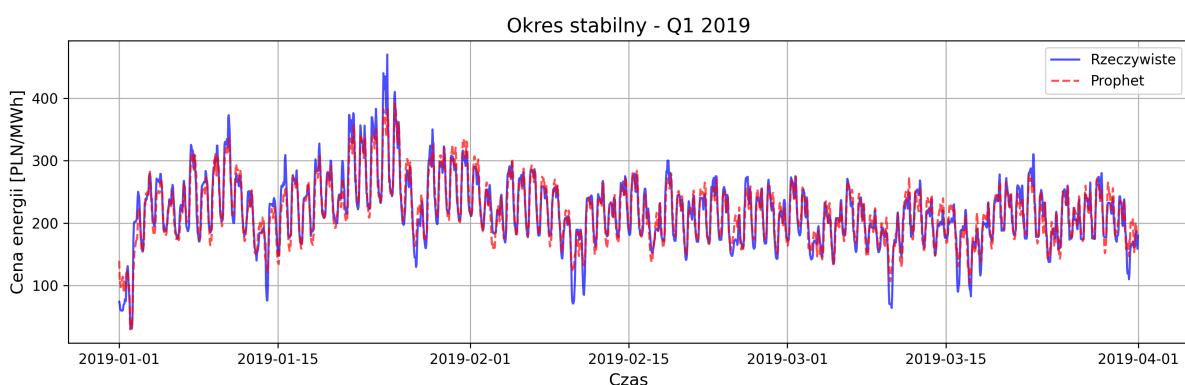
(`changepoint_prior_scale=0.100, seasonality_prior_scale=20.0, holidays_prior_scale=0.1`), gdzie gdzie MAE wyniosło 15.60, RMSE 20.17, MAPE 7.42%, a sMAPE 7.33%. Kombinacja ta charakteryzuje się umiarkowaną elastycznością trendu (`changepoint_prior_scale=0.100`), co pozwala modelowi wychwytywać zmiany w cenach energii (np. stopniowe wzrosty), oraz wysoką elastycznością sezonowości (`seasonality_prior_scale=20.0`), co dobrze oddaje zmienne wzorce dzienne i tygodniowe. Niski wpływ świąt (`holidays_prior_scale=0.1`) sugeruje, że polskie święta mają ograniczony wpływ na ceny energii w tym okresie. Pełne wyniki dla pełnego zbioru danych przedstawiono w tabeli 10.

Tabela 10. Wyniki modelu Prophet dla pełnego i skróconego zbioru danych w okresie stabilnym (2019).

Zbiór danych	MAE	RMSE	MAPE (%)	sMAPE (%)	R^2
Pełny	15.60	20.17	7.42	7.33	0.835011
Skrócony	17.75	22.85	8.17	8.32	0.788273

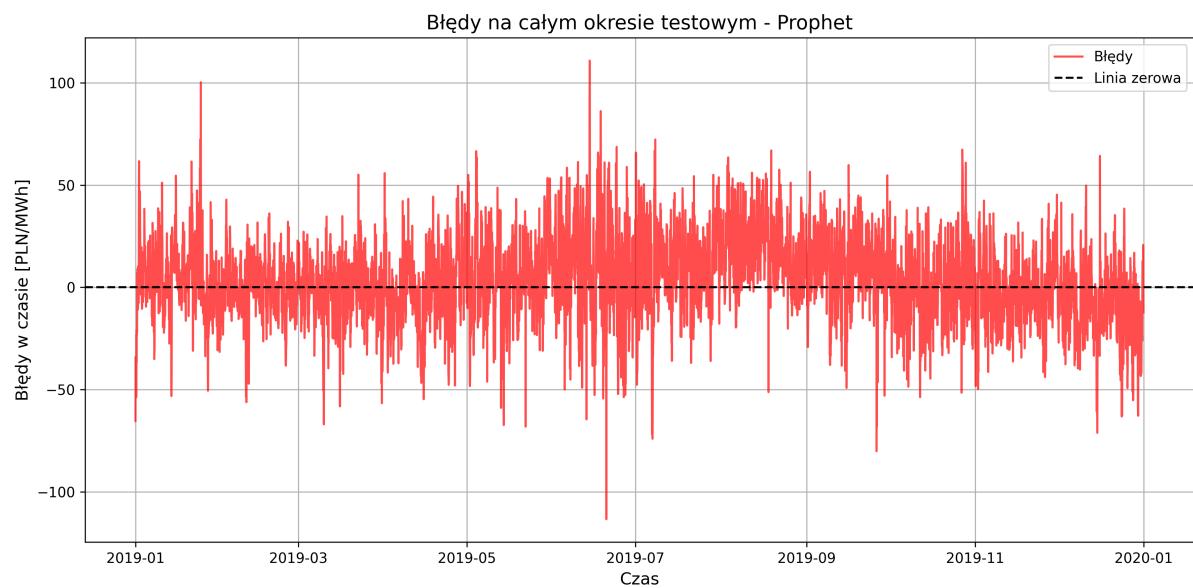
Pełny zbiór danych znowu osiągnął lepsze wyniki od zbioru skróconego. Różnice są większe, niż w przypadku regresji liniowej i Ridge, co sugeruje, że dodatkowe zmienne w pełnym zbiorze danych mają większy wpływ na prognozy modelu Prophet.

Ponizej przedstawiam wykresy rzeczywistych i przewidywanych wartości dla modelu Prophet dla pierwszego kwartału 2019 roku. Wykres jest bardzo podobny do wykresu 21 dla regresji Ridge.



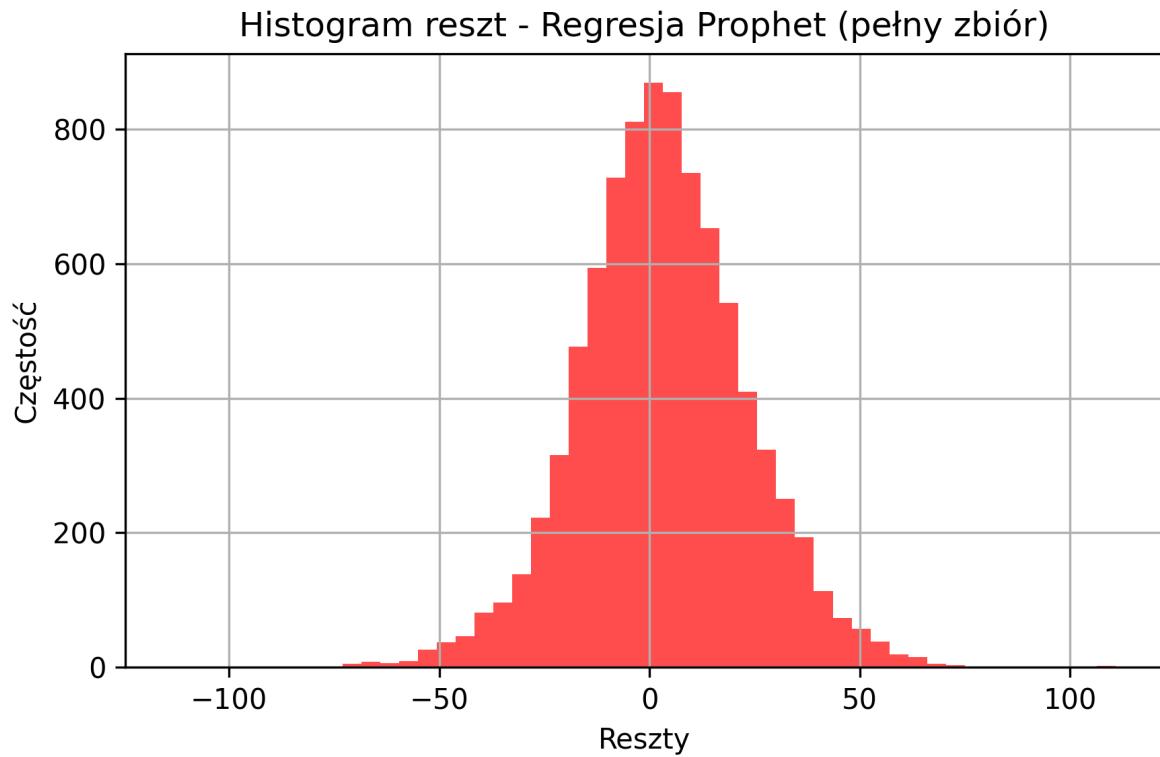
Rysunek 24. Porównanie rzeczywistych i przewidywanych wartości cen energii dla modelu Prophet w okresie stabilnym. Opracowane własne

Wykres błędów prognoz dla modelu Prophet w okresie stabilnym jest również podobny do wykresu regresji Ridge 22.



Rysunek 25. Błędy prognoz dla modelu Prophet w okresie stabilnym.

Histogram reszt dla modelu Prophet wykazuje podobieństwo do histogramu reszt dla regresji Ridge 23, ale lepiej wyjaśnia drobne różnice. Ma niższy szczyt w okolicy zera, ale łagodniejsze zejście w kierunku wartości skrajnych.



Rysunek 26. Histogram reszt dla modelu Prophet w okresie stabilnym. Opracowanie własne.

Największe błędy prognoz występujące w obu modelach są podobne. Największym wspólnym błędem prognoz dla modeli regresji Ridge oraz Prophet jest przeszacowanie ceny energii elektrycznej w dniu 10 Marca o godzinie 5 rano, gdzie rzeczywista cena wynosi 64 PLN/MWh, gdzie prognozy wynoszą 129 PLN/MWh oraz 131 PLN/MWh dla modelu Ridge i Prophet. Prawdopodobnie zmienne objaśniające sygnalizują wzrost cen, która nie miała miejsca, co prowadzi do przeszacowania prognoz.

Analizując wyniki modelu Prophet w odniesieniu do wyników regresji liniowej i grzbietowej, można stwierdzić, że model Prophet osiąga porównywalne wyniki. Wynik MAPE jest o 0.23% wyższy od regresji Ridge, co sugeruje, że model Prophet nie jest w stanie przewidzieć cen energii elektrycznej lepiej. Natomiast czas wykonania programu w języku programowania Python dla modelu Prophet wynosił ponad 40 sekund, gdzie czas wykonania dla regresji Ridge wynosił 5. W kontekście analizy zebranego zbioru danych, należy zauważać, że model Prophet nie przynosi żadnych korzyści w porównaniu do regresji Ridge lub liniowej w przypadku prognozowania w okresie stabilnym.

MLP

W ramach badania skuteczności modelu MLP w okresie stabilnym, obejmującym rok 2019, przetestowano różne konfiguracje sieci neuronowych. Najlepsze wyniki uzyskano dla architektury składającej się z pięciu warstw ukrytych, o strukturze z odpowiednio 64, 64, 32, 16, 8 neuronami w kolejnych warstwach. Model został zaprojektowany w sposób sekwencyjny, gdzie każda warstwa

ukryta korzystała z funkcji aktywacji ReLU, a dodatkowo zastosowano regularyzację L2 z parametrem 0,1, aby zapobiec przeuczeniu. Między warstwami dodano mechanizm Dropout z wartością 0,2, który losowo dezaktywuje część neuronów podczas treningu, co pomaga w uogólnianiu modelu. Ostatnia warstwa wyjściowa zawierała jeden neuron, odpowiadający za przewidywanie wartości ceny energii.

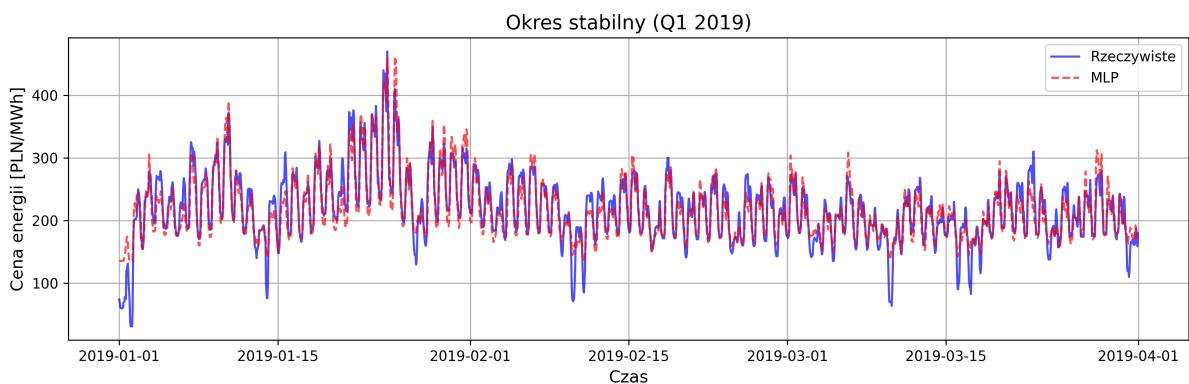
Model został skompilowany z użyciem optymalizatora Adam, skonfigurowanego z bardzo niską szybkością uczenia na poziomie 0,0001, co pozwoliło na stopniowe i stabilne dostosowywanie wag. Jako funkcję straty wybrano średni błąd kwadratowy (MSE), która dobrze mierzy różnice między przewidywaniami a rzeczywistymi wartościami. Trening modelu przeprowadzono na danych z lat 2016-2018, z maksymalną liczbą pięciuset epok i rozmiarem partii wynoszącym sto dwadzieścia osiem, co umożliwiło efektywne przetwarzanie danych w mniejszych porcjach.

W tabeli poniżej przedstawiam wyniki dla pełnego i skróconego zbiorów danych.

Tabela 11. Wyniki modelu MLP dla pełnego i skróconego zbioru danych w okresie stabilnym (2019).

Zbiór danych	MAE	RMSE	MAPE (%)	sMAPE (%)	R ²
Pełny	17.09	22.75	8.43	7.87	0.79
Skrócony	34.55	53.58	14.87	13.56	-0.16

Uzyskane wyniki dla zbioru skróconego są znacznie gorsze niż dla zbioru pełnego, co sugeruje, że model MLP nie jest w stanie dobrze przewidzieć cen energii elektrycznej na podstawie ograniczonej liczby zmiennych wyjaśniających. Wartości MAPE są o ponad procent wyższe niż w przypadku regresji liniowej i Ridge, co sugeruje, że model MLP nie jest w stanie przewidzieć cen energii elektrycznej lepiej niż modele statystyczne. Inne zestawy parametrów lub architektury sieci neuronowej nie przyniosły lepszych wyników.

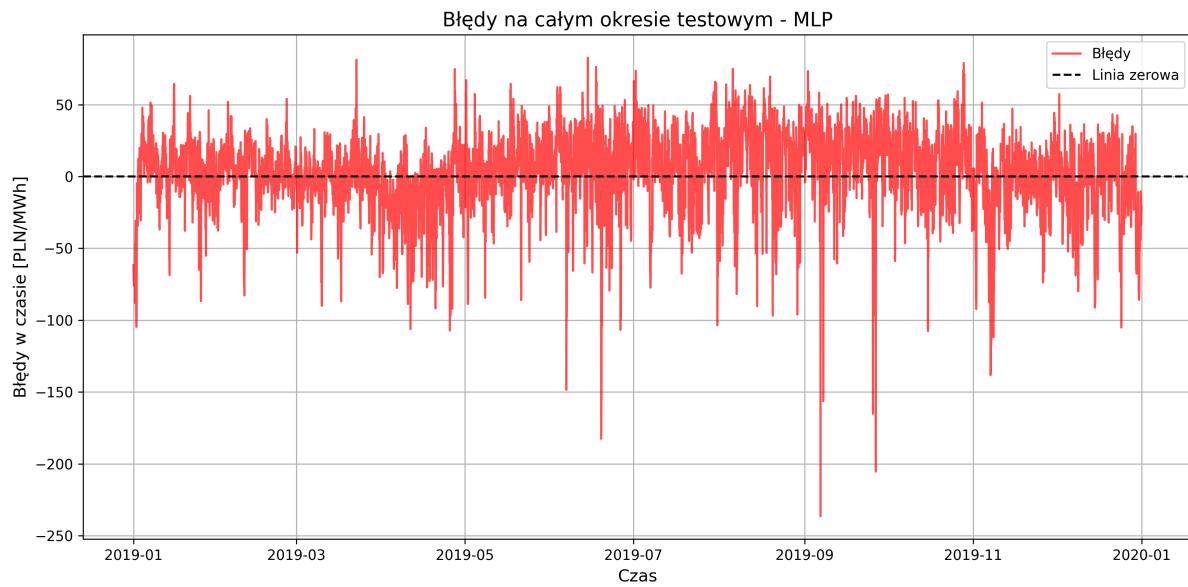


Rysunek 27. Porównanie rzeczywistych i przewidywanych wartości cen energii dla modelu MLP w okresie stabilnym. Opracowanie własne.

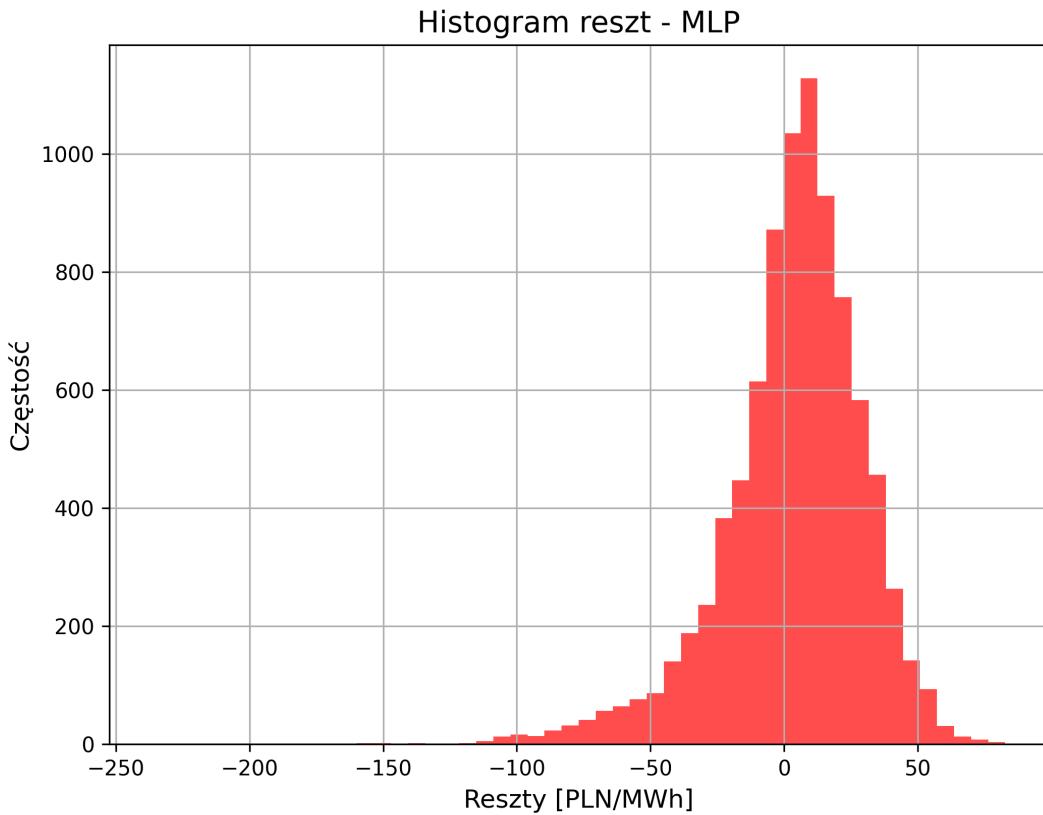
Wykres 27 nieco różni się od analogicznych wykresów dla modeli statystycznych. Prognozy dobrze podążają za rzeczywistymi wartościami, ale widać, że model MLP ma tendencję do przeszacowywania cen energii elektrycznej w momentach wzrostów ceny. Model oczekuje na większe wzrosty, które nie

następują.

Poniżej przedstawione są wykresy błędów prognoz dla modelu MLP oraz histogram reszt dla okresu stabilnego.



Rysunek 28. Błędy prognoz dla modelu MLP w okresie stabilnym. Opracowanie własne.



Rysunek 29. Histogram reszt dla modelu MLP w okresie stabilnym. Opracowanie własne.

Wykres 28 błędów prognozy oscyluje dookoła zera poza okresem w kwietniu, gdzie większość błędów ma wartości ujemne. Widoczne jest, że model MLP ma tendencję do zbyt niskiego oszacowania cen energii elektrycznej w tym okresie. Wartości błędów są znacznie większe niż w przypadku modeli statystycznych, co sugeruje, że model MLP nie radzi sobie dobrze z przewidywaniem cen energii elektrycznej w okresie stabilnym.

Histogram reszt 29 potwierdza te założenia, gdyż mniej przypomina rozkład normalny od poprzednich histogramów. Ma wyższy szczyt, ale jest przesunięty od zera w stronę wartości dodatnich. Większa liczba błędów skupia się w okolicy zera w porównaniu do histogramu reszt dla regresji Ridge lub Prophet.

6.2 Okres niestabilny

Okres niestabilny obejmuje lata 2020-2023, w których ceny energii elektrycznej były znacznie bardziej zmienne niż w okresie stabilnym. W związku z tym, okres ten może być bardziej wymagający dla modeli prognozujących.

Regresja liniowa i Ridge

W analizie wyników regresji liniowej i Ridge dla okresu niestabilnego (2023) dla pełnego i skróconego zbioru danych obserwuje się zbliżone wartości metryk, co wskazuje na ograniczone różnice w skuteczności obu metod w tym okresie. Tabela 12 przedstawia szczegółowe wyniki dla regresji liniowej i Ridge. Zbiór skrócony podobnie do okresu stabilnego nie przyniósł znaczących różnic w metrykach, ale dodatkowe zmienne w pełnym zbiorze danych lekko poprawiły wyniki niewielkim kosztem obliczeniowym.

Wysokie wartości MAPE (powyżej 179%) sugerują problem z wartościami bliskimi zeru 5.1.3, które pojawiają się w zbiorze danych z okresu niestabilnego. Z tego powodu, MAPE może nie być najlepszą metryką do oceny skuteczności modeli na zbiorze okresu niestabilnego. Z tego powodu większej uwadze poświęcono metrykom MAE, RMSE, sMAPE oraz R^2 . Wartości MAE i RMSE są stosunkowo wysokie, co wskazuje na duże błędy prognozowania w jednostkach absolutnych. Błąd na poziomie 59.6 PLN/MWh dla MAE oznacza, że prognozy różnią się średnio o 59.6 PLN od rzeczywistych wartości. Może to prowadzić do znacznych strat finansowych, szczególnie w przypadku dużych transakcji. Wartości sMAPE są dwukrotnie wyższe niż w przypadku okresu stabilnego, co sugeruje, że modele mają trudności z przewidywaniem cen energii w okresach dużej zmienności. Pomimo tego, R^2 nie wzrosło znacząco względem okresu stabilnego, co sugeruje, że modele nadal dobrze wyjaśniają zmienność cen energii, mimo dużych błędów prognozowania.

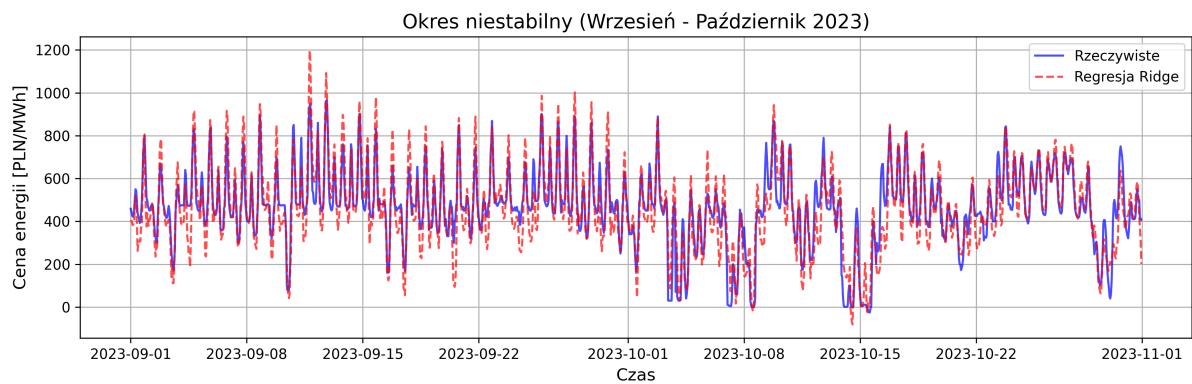
Najlepsza wartość hiperparametru α w regresji Ridge dla okresu niestabilnego wyniosła 0.1, co sugeruje, że w tym przypadku współliniowość między zmiennymi objaśniającymi nie jest tak istotna jak w przypadku pełnego zbioru danych w okresie stabilnym. Wartość ta jest znacznie niższa niż w przypadku pełnego zbioru danych w okresie stabilnym (500.0), co może sugerować, że w okresie niestabilnym modele są bardziej elastyczne.

Tabela 12. Wyniki metryk dla regresji liniowej i Ridge w okresie niestabilnym (2023). Opracowanie własne.

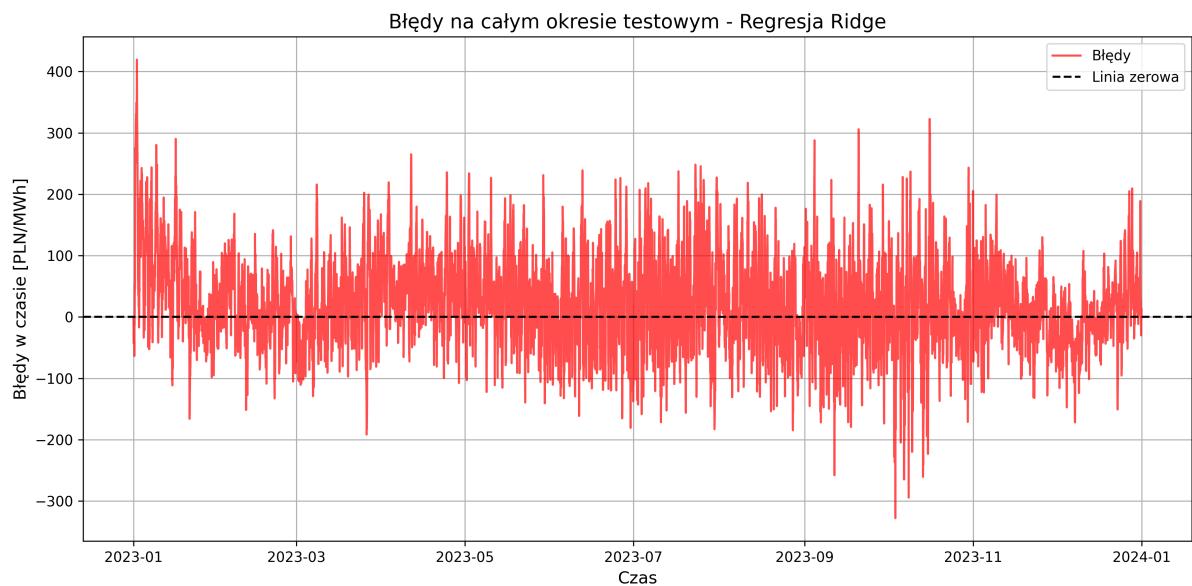
Model	MAE	RMSE	MAPE	sMAPE	R^2
Regresja liniowa (pełny zbiór)	59.61	77.73	179.38	16.51	0.8111
Regresja liniowa (skrócony zbiór)	59.76	78.07	188.24	16.68	0.8095
Regresja Ridge (pełny zbiór)	59.63	77.75	179.79	16.52	0.8110
Regresja Ridge (skrócony zbiór)	59.78	78.08	186.22	16.69	0.8094

Wartości metryk są niemal identyczne dla obu modeli, co sugeruje, że w tym przypadku regularyzacja L2 nie przynosi znaczących korzyści.

Dla dobrego zobrazowania prognoz na wykresie wybrano okres o największej zmienności cen w ramach 2023 roku, czyli od 1 września do 31 października. Wartości prognoz dla regresji Ridge w porównaniu do rzeczywistych cen energii elektrycznej przedstawiono na rysunku 30. Wykres reszt dla modelu Ridge w okresie niestabilnym przedstawiono na rysunku 32.

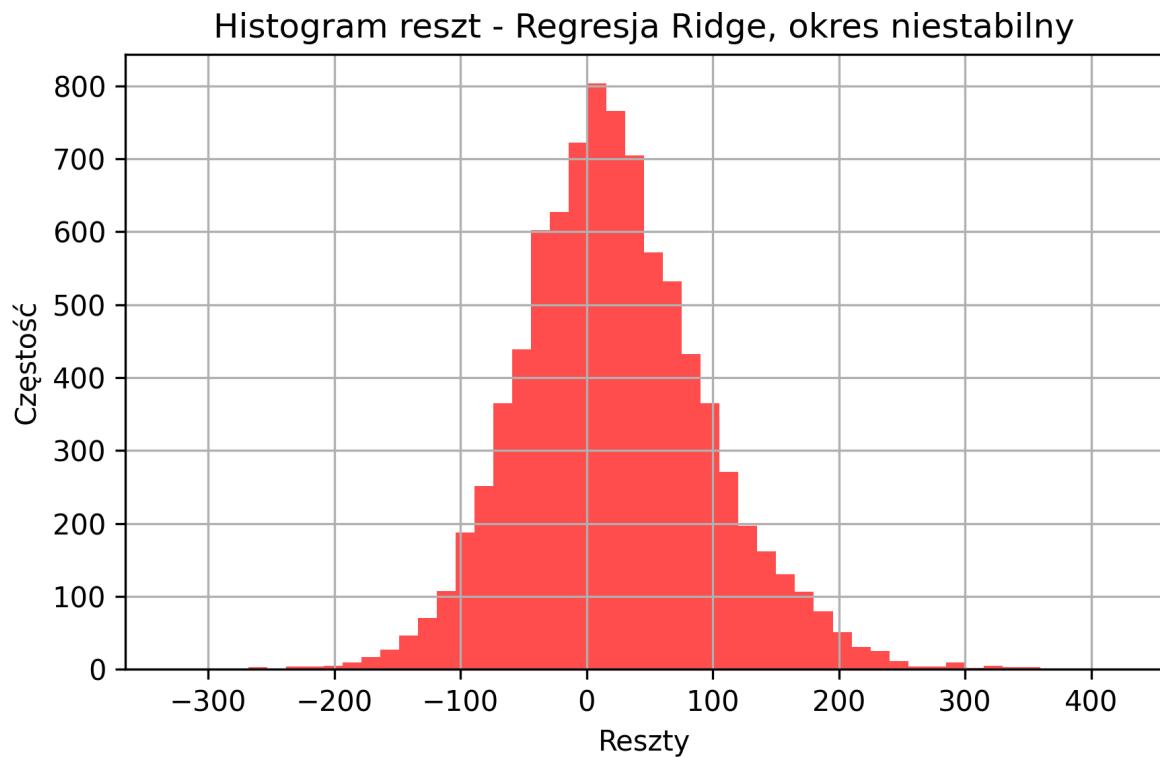


Rysunek 30. Prognozy modelu Ridge w porównaniu do rzeczywistych cen energii w okresie niestabilnym (2023). Opracowanie własne.



Rysunek 31. Błędy prognoz dla modelu Ridge w okresie niestabilnym (2023). Opracowanie własne.

Na wykresie 30 widać, że model jest skłonny przeszacowywać zmienność cen zarówno w górę, jak i w dół. Błędy prognoz dla modelu Ridge w okresie niestabilnym są znacznie większe niż w przypadku okresu stabilnego. Większość błędów oscyluje się od -100 do 100 PLN/MWh. Największy szczyt można zobaczyć na początku okresu testowego, gdzie obserwuje się szczyt powyżej 400 PLN/MWh. Wartości prognoz są znacznie bardziej rozproszone niż w przypadku okresu stabilnego. Histogram reszt dla modelu Ridge w okresie niestabilnym przedstawiono na rysunku 32.



Rysunek 32. Histogram reszt dla modelu Ridge w okresie niestabilnym (2023). Opracowanie własne.

Na podstawie histogramu reszt można zauważyć, że rozkład reszt jest podobny do rozkładu normalnego i ma szczyt w okolicy zera. Porównując histogram reszt z histogramem reszt 23 dla okresu stabilnego, można zauważyć, że histogram reszt w okresie niestabilnym ma szerszy rozkład z większymi ogonami, co potwierdza większe błędy prognozowania w wartościach absolutnych. Histogram reszt jest lekko przesunięty w kierunku wartości dodatnich, co sugeruje, że model ma tendencję do przeszacowywania cen.

Prophet

Model Prophet lekko poprawił wyniki w porównaniu do regresji liniowej i Ridge. Najlepszymi parametrami do okresu niestabilnego okazały się (`changepoint_prior_scale=0.001`, `seasonality_prior_scale=50.0`, `holidays_prior_scale=0.1`). Niska wartość `changepoint_prior_scale` sugeruje, że model preferuje bardziej stabilne trendy, unikając nadmiernego dopasowania do szumów w danych treningowych. Zwiększona wartość `seasonality_prior_scale` pozwala modelowi lepiej uchwycić sezonowe wzorce w danych, co jest istotne w przypadku cen energii elektrycznej, które mogą wykazywać silne sezonowe fluktuacje. Wartość `holidays_prior_scale` jest taka sama jak w przypadku okresu stabilnego, co sugeruje, że wpływ świąt na ceny energii wciąż nie jest istotny. Wyniki dla pełnego zbioru danych przedstawiono w tabeli 13.

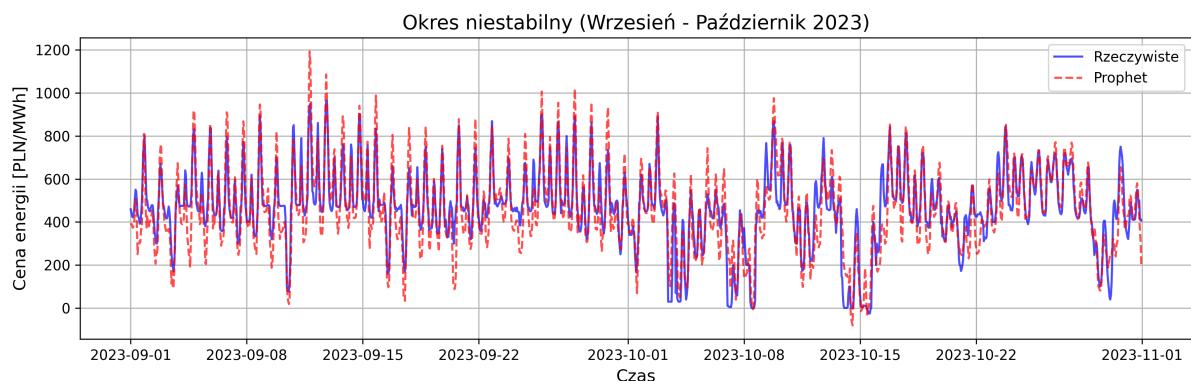
6.2. Okres niestabilny

Tabela 13. Wyniki modelu Prophet dla pełnego i skróconego zbioru danych w okresie niestabilnym (2023).

Zbiór danych	MAE	RMSE	MAPE (%)	sMAPE (%)	R^2
Pełny	57.87	74.39	163.57	16.35	0.827021
Skrócony	61.10	78.56	202.72	18.17	0.80709

Wartość parametru `changepoint_prior_scale` jest najbardziej znacząca w procesie doboru najlepszych hiperparametrów. Zwiększenie wartości tego parametru do 0.1 prowadzi do znacznego pogorszenia wyników na poziomie **MAE = 138** oraz **sMAPE = 20.0%**. Oznacza to, że model zaczyna być bardziej podatny na wykrywanie zmiany w trendzie i próbuje dopasować się do lokalnych fluktuacji. Zmiana innych hiperparametrów zwiększa sMAPE o 1%.

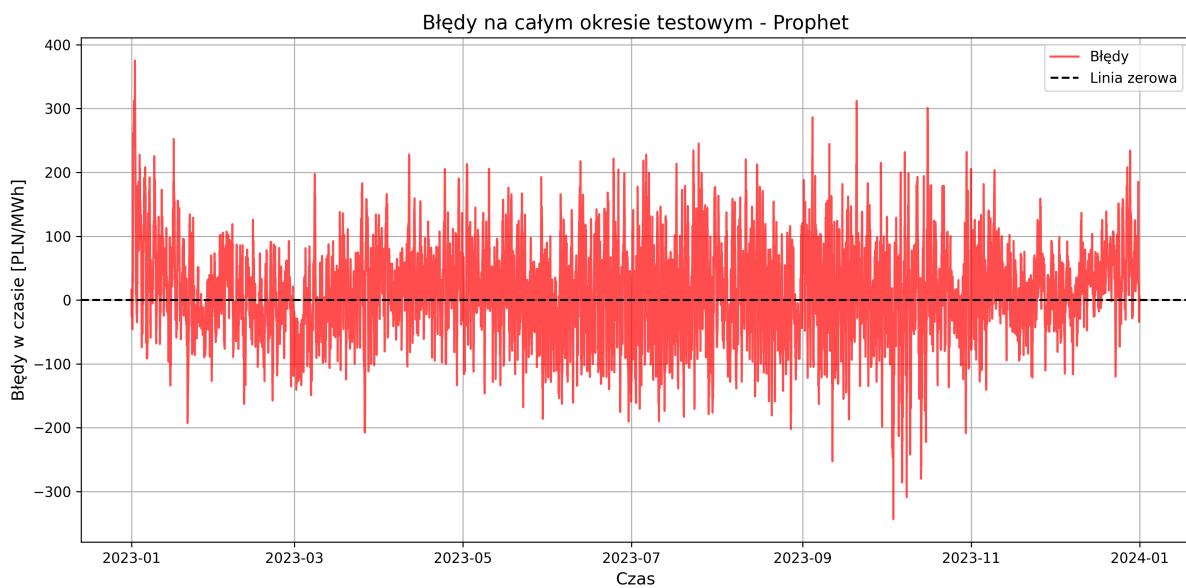
Wynik modelu prophet w porównaniu z regresją Ridge jest nieznaczowo lepszy z punktu widzenia sMAPE, ale średni błąd MAE jest o 1.76 PLN/MWh niższy, co może być bardzo istotne z punktu finansowego w przypadku dużych transakcji. Wartości RMSE są również niższe, co sugeruje, że model Prophet lepiej radzi sobie z przewidywaniem cen energii elektrycznej w okresie niestabilnym.



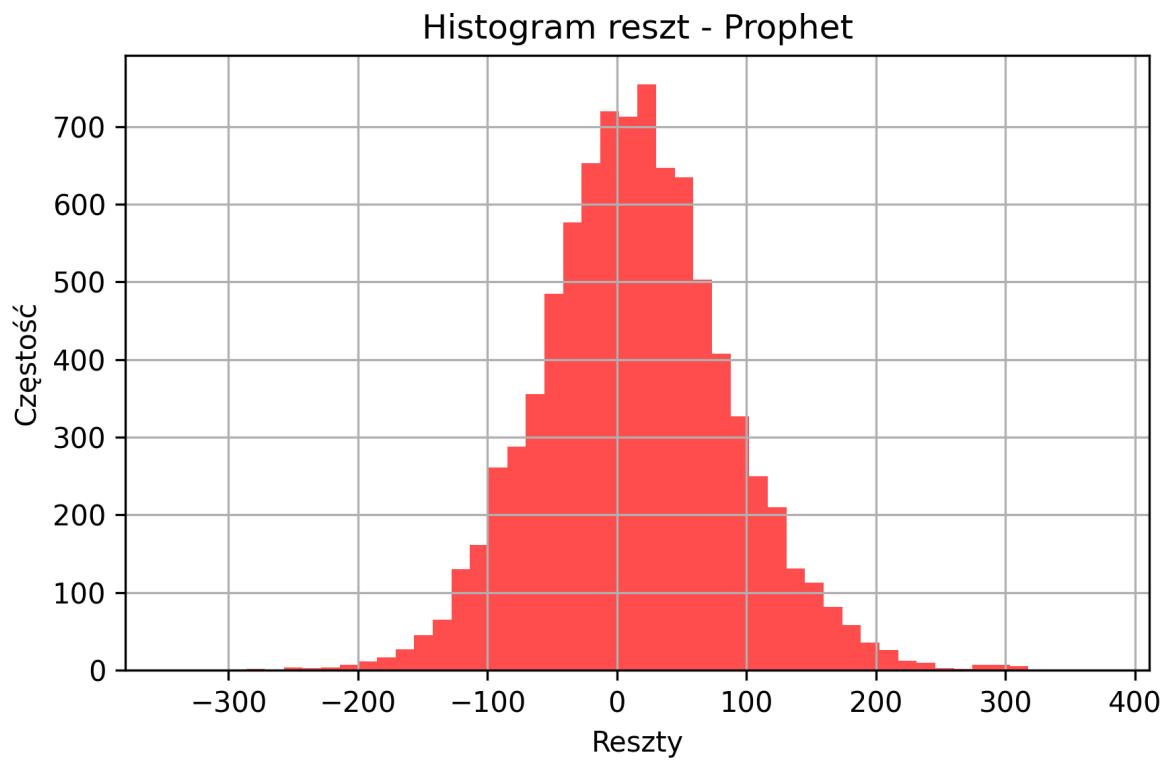
Rysunek 33. Porównanie rzeczywistych i przewidywanych wartości cen energii dla modelu Prophet w okresie niestabilnym. Opracowanie własne.

Wykres predykcji modelu Prophet w okresie niestabilnym jest podobny do wykresu regresji Ridge 30. Widać, że model przeszacowuje ceny energii elektrycznej w okresach dużej zmienności, co prowadzi do dużych błędów prognozowania.

Poniżej przedstawiono wykresy błędów prognoz oraz histogram reszt dla modelu Prophet w okresie niestabilnym.



Rysunek 34. Błędy prognoz dla modelu Prophet w okresie niestabilnym. Opracowanie własne.



Rysunek 35. Histogram reszt dla modelu Prophet w okresie niestabilnym. Opracowanie własne.

Histogram dla modelu Prophet w okresie niestabilnym jest bardziej przesunięty w kierunku wartości dodatnich niż histogram dla regresji Ridge 32. Histogram wykazuje większe odchylenia od rozkładu

normalnego, i nie posiada szczytu w zerze. Natomiast ogon histogramu po prawej stronie jest krótszy. To jest prawdopodobnie powodem, że model Prophet ma lepsze wyniki od regresji Ridge.

MLP

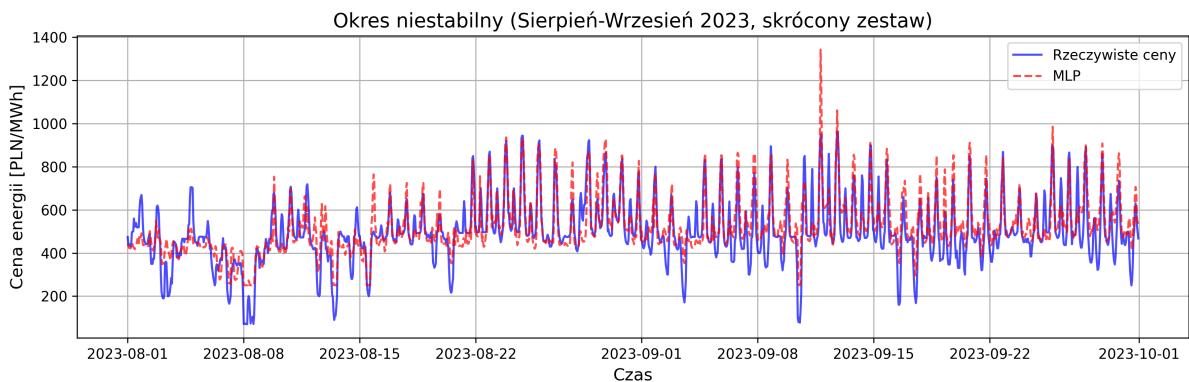
Dla uzyskania najlepszych wyników dla modelu MLP w okresie niestabilnym, liczbę epok zwiększo do 1000, a rozmiar partii zmniejszono do 64. Wartości hiperparametrów pozostały takie same jak w przypadku okresu stabilnego. Szybkość uczenia została zwiększena do 0.001, co pozwoliło na szybsze dostosowywanie wag. Po raz pierwszy wyniki skróconego zestawu parametrów przewyższyły wyniki pełnego zbioru danych we wszystkich metrykach. Sugeruje to, że model MLP jest bardziej podatny na nadmierne dopasowanie do danych treningowych w przypadku pełnego zbioru danych. Wartości metryk dla pełnego i skróconego zbioru danych przedstawiono w tabeli 14.

Tabela 14. Wyniki modelu MLP dla pełnego i skróconego zbioru danych w okresie niestabilnym (2023).

Zbiór danych	MAE	RMSE	MAPE (%)	sMAPE (%)	R^2
Pełny	98.85	126.89	1361.22	22.21	0.50
Skrócony	66.08	88.46	1702.13	17.38	0.76

Skrócony zestaw parametrów wyjaśnił 76% zmienności cen energii elektrycznej, co jest znacznie lepszym wynikiem niż w przypadku pełnego zbioru danych. Jest to jednak wciąż gorszy wynik niż w przypadku wcześniejszych modeli. Wartości absolutne MAE i sMAPE są przykładowo o 8.21 PLN/MWh i 1.03% gorsze niż w przypadku modelu Prophet.

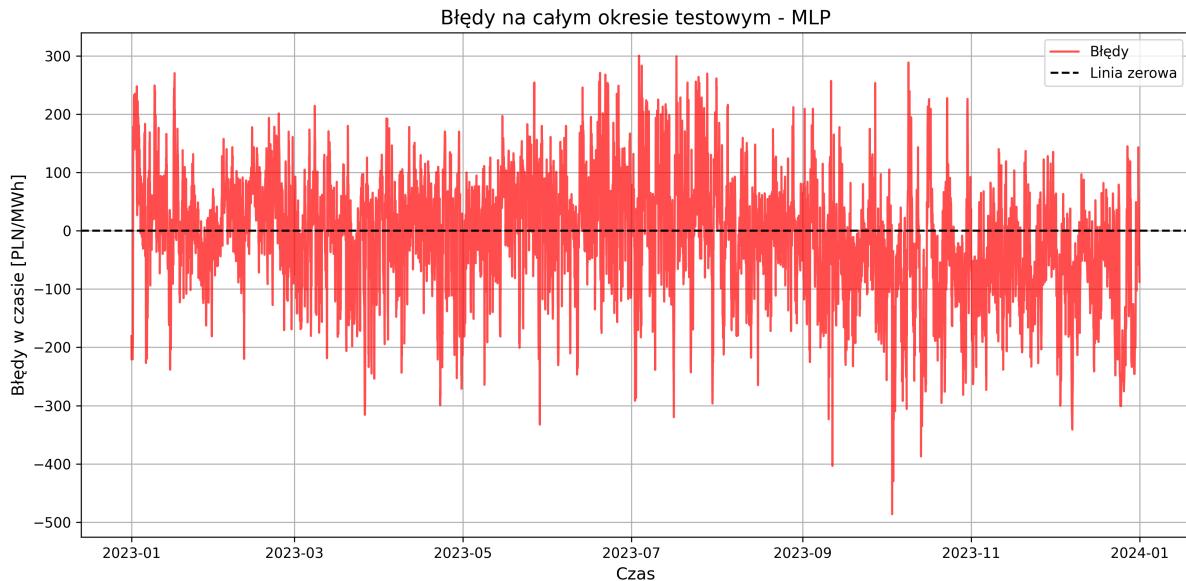
Wykres rzeczywistych i przewidywanych wartości dla modelu MLP w okresie niestabilnym dla danych skróconych przedstawiono na rysunku 36.



Rysunek 36. Porównanie rzeczywistych i przewidywanych wartości cen energii dla modelu MLP w okresie niestabilnym. Opracowanie własne.

Wykres potwierdza gorsze wyniki modelu MLP w porównaniu do modelu Prophet. Prawie wszystkie szczyty oscylacji cenowej są przeszacowane, a model MLP nie jest w stanie dokładnie

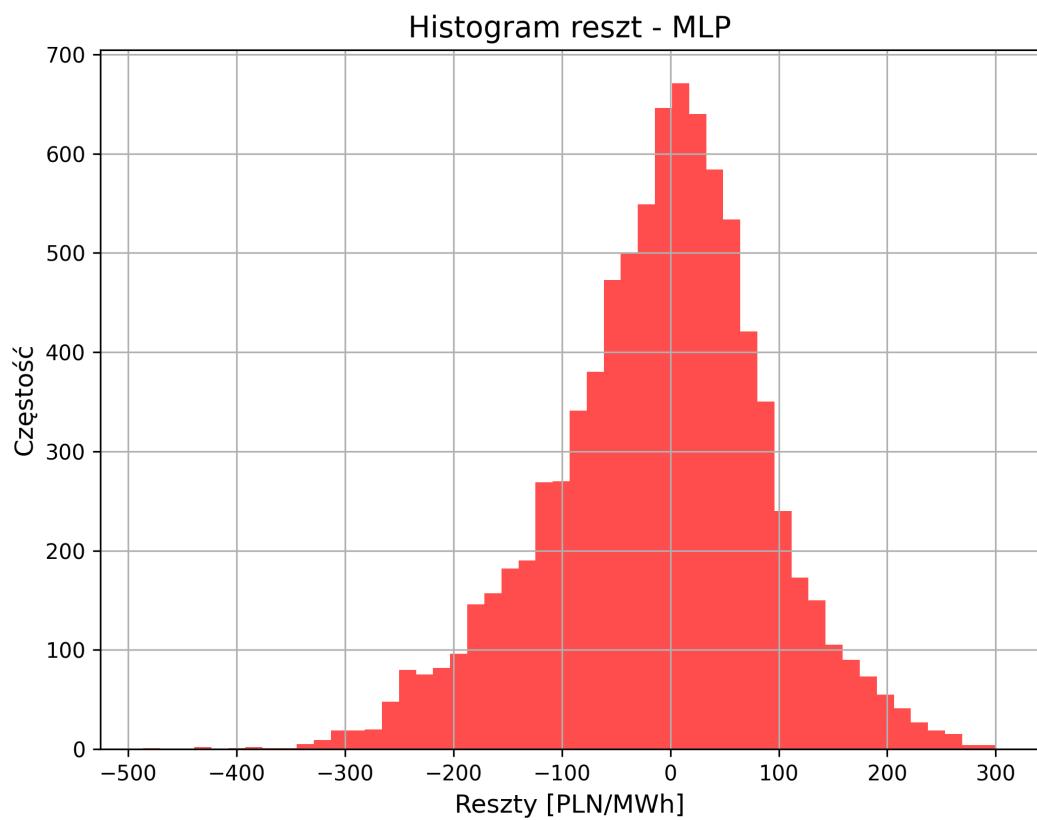
przewidzieć poziomu wzrostów ani spadków cen energii elektrycznej. Wartości prognoz są znacznie bardziej rozproszone niż w przypadku modelu Prophet.



Rysunek 37. Błędy prognoz dla modelu MLP w okresie niestabilnym. Opracowanie własne.

Wykres błędów prognoz dla modelu MLP w okresie niestabilnym 37 również potwierdza gorsze wyniki modelu MLP. Wartości absolutne błędów dochodzą do poziomu 400 PLN/MWh.

Histogram reszt 38 dla modelu MLP ma szczyt w okolicy zera, ale ma dłuższe ogony po obu stronach. Histogram wykazuje większe odchylenia od rozkładu normalnego. Widoczne są reszty o dużych wartościach, szczególnie po lewej stronie histogramu.



Rysunek 38. Histogram reszt dla modelu MLP w okresie niestabilnym. Opracowanie własne.

Rozdział 7

Analiza wyników i podsumowanie

Niniejszy rozdział stanowi podsumowanie przeprowadzonych badań dotyczących przygotowania zbioru danych wejściowych do opracowania modelu predykcji cen energii elektrycznej na Rynku Dnia Następnego (RDN) w Polsce oraz analizę wyników uzyskanych za pomocą wybranych modeli prognostycznych.

7.1 Analiza wyników

Główym celem pracy było stworzenie kompleksowego zbioru danych, który obejmowałby szeroki zakres czynników mogących wpływać na ceny energii elektrycznej w Polsce, uwzględniając zarówno okresy stabilności, jak i wysokiej zmienności rynkowej. Cel ten został zrealizowany poprzez zgromadzenie danych historycznych z lat 2016-2023, obejmujących m.in. ceny energii, dane dotyczące zapotrzebowania i generacji, ceny paliw, emisje CO_2 oraz warunki atmosferyczne. Stworzony zbiór danych charakteryzuje się granulacją godzinową i został poddany obróbce, w tym obsłudze brakujących wartości oraz ujednoliceniu rozdzielczości czasowej poszczególnych zmiennych.

Przeprowadzona analiza korelacji pozwoliła na identyfikację zmiennych najsiłniej powiązanych z ceną energii elektrycznej, co stało się podstawą do stworzenia również skróconej wersji zbioru danych. Dane zostały podzielone na dwa główne okresy: stabilny (2016-2019) oraz niestabilny (2020-2023), co umożliwiło ocenę wpływu różnych warunków rynkowych na skuteczność modeli.

Do weryfikacji użyteczności przygotowanego zbioru danych wykorzystano cztery modele progностyczne: regresję liniową, regresję Ridge, model Prophet oraz wielowarstwowy perceptron (MLP).

Okres stabilny (testowany na danych z 2019 roku)

- Modele regresji liniowej i Ridge osiągnęły zbliżone, dobre wyniki, przy czym regresja Ridge z optymalnym parametrem $\alpha = 500.0$ dla pełnego zbioru danych wykazała nieznaczną przewagę, co sugeruje występowanie współliniowości między zmiennymi. Różnice w metrykach pomiędzy pełnym a skróconym zbiorem danych były niewielkie, wskazując, że dodatkowe zmienne wniosły pewną informację, jednak skrócony zbiór nadal zawierał kluczowe prediktory.

- Model Prophet, skonfigurowany z trybem addytywnym i dostosowaniem hiperparametrów (najlepsze dla `changepoint_prior_scale=0.100`, `seasonality_prior_scale=20.0`, `holidays_prior_scale=0.1`), osiągnął wyniki porównywalne do regresji Ridge, jednak przy znacznie dłuższym czasie obliczeń. Pełny zbiór danych również w tym przypadku dał lepsze rezultaty niż zbiór skrócony.
- Model MLP, mimo testowania różnych architektur (najlepsza: pięć warstw ukrytych 64-64-32-16-8 neuronów, aktywacja ReLU, regularizacja L2, Dropout, optymalizator Adam z niską szybkością uczenia), uzyskał wyniki gorsze niż modele statystyczne dla pełnego zbioru danych. Wyniki dla zbioru skróconego były znacznie słabsze, co wskazuje na trudności modelu MLP w generalizacji.

Okres niestabilny (testowany na danych z 2023 roku)

- W przypadku modeli regresji liniowej i Ridge, wyniki dla pełnego i skróconego zbioru danych były bardzo zbliżone, a regularizacja L2 w regresji Ridge (optymalne $\alpha = 0.1$) nie przyniosła znaczących korzyści. Wartości metryk MAE i RMSE były znacznie wyższe niż w okresie stabilnym, co odzwierciedla duże trudności w prognozowaniu cen w warunkach wysokiej zmienności. Metryka MAPE okazała się problematyczna z powodu występowania wartości bliskich zeru. Mimo wysokich błędów absolutnych, współczynnik R^2 nie uległ znacznemu pogorszeniu, co sugeruje, że modele nadal wyjaśniały znaczną część zmienności cen.
- Model Prophet wykazał nieznaczną poprawę w stosunku do modeli regresji, szczególnie pod względem MAE i RMSE, przy optymalnych hiperparametrach (`changepoint_prior_scale=0.001`, `seasonality_prior_scale=50.0`, `holidays_prior_scale=0.1`). Niska wartość `changepoint_prior_scale` wskazuje na preferencję modelu do stabilniejszych trendów w tym burzliwym okresie.
- Model MLP dla okresu niestabilnego, przy zwiększonej liczbie epok (1000) i zmniejszonym rozmiarze partii (64) oraz wyższej szybkości uczenia (0.001), uzyskał lepsze wyniki dla skróconego zbioru danych niż dla pełnego. Sugeruje to większą podatność na przeuczenie oraz zakłócenia przy dużej liczbie zmiennych w warunkach wysokiej zmienności. Mimo to, wyniki MLP wciąż były gorsze niż modele Prophet i regresji. Model miał tendencję do przeszacowywania szczytów oscylacji cenowych.

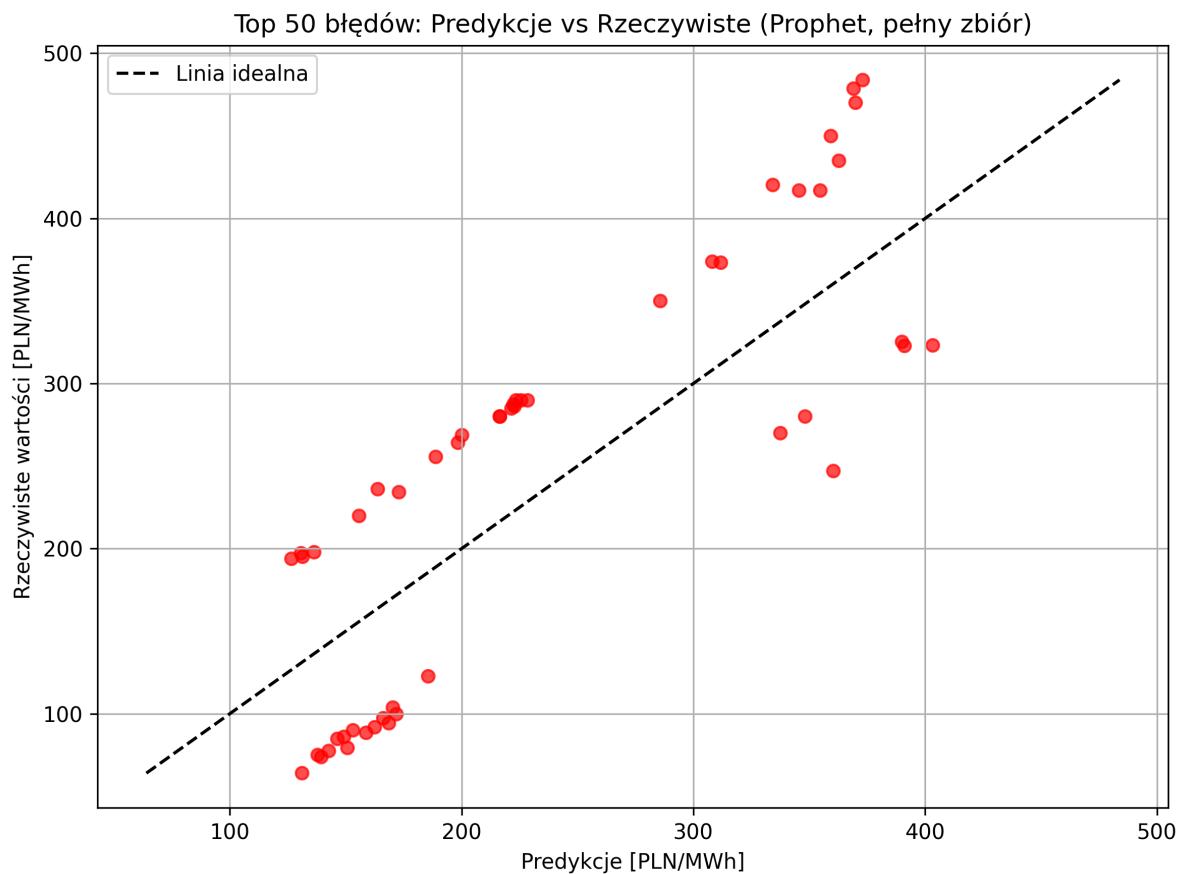
Analiza wykazała, że przygotowany zbiór danych, szczególnie w pełnej wersji, dostarcza wartościowych informacji dla modeli prognostycznych. Modele statystyczne, zwłaszcza Prophet, okazały się stosunkowo solidne i efektywne obliczeniowo w obu analizowanych okresach. Model MLP, mimo swojej elastyczności, nie wykazał przewagi nad prostszymi metodami w badanym kontekście i okazał się bardziej wrażliwy na dobór zestawu zmiennych oraz warunki rynkowe. Model MLP okazał się być dużo bardziej wymagający z punktu widzenia czasu obliczeń. Uzyskanie wyników jednego zestawu danych zajmowało 5 razy więcej czasu niż w przypadku modelu Prophet.

Dobrane dane wejściowe są odpowiednie do prognozowania cen energii elektrycznej. Szczególnie istotnymi zmiennymi według modeli statystycznych okazały się:

- `fixing_i_price_mean24` oraz `fixing_i_price_mean48`, czyli średnie ceny z poprzednich dni
- `RB_price`, czyli cena z rynku bilansującego
- `fixing_i_price_lag24`, czyli cena z poprzedniego dnia
- `lt_price` oraz `cz_price`, czyli ceny z rynków sąsiednich krajów
- `fixing_i_price_lag168`, czyli cena z tygodnia wstecz
- `non emissive_sources_percentage`, czyli udział źródeł nieemisyjnych w miksie energetycznym
- `Load`, czyli zapotrzebowanie na energię
- `wind_onshore`, czyli generacja z wiatru lądowego

Wszystkie z wymienionych zmiennych wpływały w sposób liniowy na prognozowane ceny i bezpośrednio odzwierciedlały ich zmiany. Warto zauważyć, że w przypadku modelu Prophet, zmienne te również miały znaczenie, jednak ich wpływ był bardziej złożony i zależny od sezonowości oraz trendów.

Niestety, w przypadku zmienności rynku, dobrane zmienne nie były w stanie wytłumaczyć wszystkich skoków cenowych. Niemniej jednak błędy prognoz nie były spowodowane jedynie brakiem istotnych zmiennych, ale także dużą niestabilnością rynku oraz nieprzewidywalnością zdarzeń zewnętrznych. Przykładowo, wykres 39 przedstawia największe błędy prognoz modelu Prophet w okresie stabilnym, które były spowodowane nieprzewidywalnymi skokami cenowymi. Błędy mają charakter losowy, nie mają wyraźnego trendu ani sezonowości, co sugeruje, że są wynikiem nieprzewidywalnych zdarzeń zewnętrznych.



Rysunek 39. Największe błędy prognoz modelu Prophet z okresu stabilnego. Opracowanie własne.

7.2 Podsumowanie

Niniejsza praca magisterska koncentrowała się na przygotowaniu i analizie obszernego zbioru danych na potrzeby prognozowania cen energii elektrycznej na Rynku Dnia Następnego w Polsce. Głównym celem było zgromadzenie danych obejmujących lata 2016-2023, które odzwierciedlałyby zarówno okresy stabilności cenowej, jak i momenty wysokiej zmienności spowodowane czynnikami ekonomicznymi, geopolitycznymi i pogodowymi. Udało się stworzyć kompleksową bazę danych zawierającą wiele zmiennych, wpływ których na ceny energii potwierdzają wyniki wybranych modeli, analiza korelacji oraz literatura analizująca rynki zagraniczne.

Wyniki badań wskazują, że przygotowany zbiór danych jest wartościowy i pozwala na budowę modeli o zadowalającej jakości predykcji, szczególnie w okresie stabilnym. Modele statystyczne, takie jak regresja Ridge, okazały się być solidnym i efektywnym podejściem. Model Prophet również dostarczył konkurencyjnych prognoz, zwłaszcza w okresie niestabilnym, podkreślając swoją zdolność do adaptacji do zmieniających się wzorców sezonowych i trendów przy odpowiednim doborze hiperparametrów. Model MLP, mimo potencjału do modelowania złożonych nieliniowości, nie wykazał

Rozdział 7. Analiza wyników i podsumowanie

w tym badaniu jednoznacznej przewagi, a jego wyniki były silnie zależne od doboru cech i warunków rynkowych, co sugeruje konieczność dalszych badań nad optymalizacją jego architektury i procesu uczenia dla tego konkretnego zadania.

Praca potwierdziła trudności związane z prognozowaniem cen energii w warunkach wysokiej zmienności rynkowej, jakie miały miejsce w latach 2020-2023. Wszystkie modele wykazywały wyższe błędy prognoz w tym okresie, co podkreśla wyzwania związane z wpływem nieoczekiwanych czynników zewnętrznych i nieliniowością danych.

Opracowany zbiór danych oraz przeprowadzone analizy mogą stanowić cenne źródło informacji dla uczestników rynku energii w podejmowaniu decyzji handlowych i zarządzaniu ryzykiem. Dalsze kierunki badań mogłyby obejmować eksplorację bardziej zaawansowanych metod uczenia maszynowego, w tym podejść hybrydowych łączących różne techniki modelowania, a także analizę wpływu dodatkowych czynników, takich jak nastroje rynkowe czy zmiany regulacyjne. Istotne byłoby również zbadanie metod radzenia sobie z ekstremalnymi wartościami cen i skokami, które stanowią największe wyzwanie dla modeli prognostycznych.

Bibliografia

- [1] (EIA), U. E. I. A., *Europe Brent Spot Price FOB (Dollars per Barrel)*, <https://www.eia.gov/dnav/pet/hist/RBRTED.htm>, Accessed: 2025-05-02, 2025.
- [2] Busby, J. W. i in., „Cascading risks: Understanding the 2021 winter blackout in Texas”, *Energy Research & Social Science*, s. 102106, 2021, ISSN: 2214-6296. DOI: <https://doi.org/10.1016/j.erss.2021.102106>. adr.: <https://www.sciencedirect.com/science/article/pii/S2214629621001997>.
- [3] ENTSO-E, *Transparencja rynku energii elektrycznej*, <https://newtransparency.entsoe.eu/market/energyPrices>, Accessed: 2025-05-02, 2025.
- [4] Foundation, I., *Energy Transition Insights*, <https://energy.instrat.pl/>, Accessed: 2025-05-02, 2025.
- [5] Kalisz, M. i Mantiuk, A., „Uczenie ciągłe w modelowaniu cen energii elektrycznej”, Politechnika Warszawska, Mechatroniki, 2025.
- [6] Lago, J., Marcjasz, G., De Schutter, B. i Weron, R., „Forecasting day-ahead electricity prices: A review of state-of-the-art algorithms, best practices and an open-access benchmark”, *Applied Energy*, t. 293, s. 116983, 2021, ISSN: 0306-2619. DOI: <https://doi.org/10.1016/j.apenergy.2021.116983>. adr.: <https://www.sciencedirect.com/science/article/pii/S0306261921004529>.
- [7] Laitsos, V., Vontzos, G., Paraschoudis, P., Tsampasis, E., Bargiotas, D. i Tsoukalas, L. H., „The State of the Art Electricity Load and Price Forecasting for the Modern Wholesale Electricity Market”, *Energies*, t. 17, nr. 22, 2024, ISSN: 1996-1073. DOI: [10.3390/en17225797](https://doi.org/10.3390/en17225797). adr.: <https://www.mdpi.com/1996-1073/17/22/5797>.
- [8] Lipiecki, A., Uniejewski, B. i Weron, R., „Postprocessing of point predictions for probabilistic forecasting of day-ahead electricity prices: The benefits of using isotonic distributional regression”, *Energy Economics*, t. 139, s. 107934, 2024, ISSN: 0140-9883. DOI: <https://doi.org/10.1016/j.eneco.2024.107934>. adr.: <https://www.sciencedirect.com/science/article/pii/S014098832400642X>.
- [9] Maciejowska, K., Uniejewski, B. i Weron, R., *Forecasting Electricity Prices*, 2022. arXiv: 2204.11735 [q-fin.ST]. adr.: <https://arxiv.org/abs/2204.11735>.

Bibliografia

- [10] Marcjasz, G., „Forecasting Electricity Prices Using Deep Neural Networks: A Robust Hyper-Parameter Selection Scheme”, *Energies*, t. 13, nr. 18, 2020, ISSN: 1996-1073. DOI: 10.3390/en13184605. adr.: <https://www.mdpi.com/1996-1073/13/18/4605>.
- [11] Openmeteo, *Prognozy pogody*, <https://open-meteo.com>, Accessed: 2025-04-09, 2025.
- [12] Poggi, A., Di Persio, L. i Ehrhardt, M., „Electricity Price Forecasting via Statistical and Deep Learning Approaches: The German Case”, *AppliedMath*, t. 3, nr. 2, s. 316–342, 2023, ISSN: 2673-9909. DOI: 10.3390/appliedmath3020018. adr.: <https://www.mdpi.com/2673-9909/3/2/18>.
- [13] Polski, N. B., *Roczne kursy walut - tabela A*, <https://nbp.pl/statystyka-i-sprawozdawczosc/kursy/archiwum-tabela-a-csv-xls/>, Accessed: 2025-05-02, 2025.
- [14] PSE, *Dane systemowe*, <https://www.pse.pl/dane-systemowe>, Accessed: 2025-05-02, 2024.
- [15] PSE, *Monitorowanie rynku bilansującego*, <https://www.pse.pl/obszary-dzialalnosci/rynek-energii/monitorowanie-rynku-bilansujacego>, Accessed: 2025-05-02, 2025.
- [16] PSE, *Raporty historyczne*, <https://raporty.pse.pl>, Accessed: 2025-05-02, 2025.
- [17] Tan, Z., Zhang, J., Wang, J. i Xu, J., „Day-ahead electricity price forecasting using wavelet transform combined with ARIMA and GARCH models”, *Applied Energy*, t. 87, nr. 11, s. 3606–3610, 2010, ISSN: 0306-2619. DOI: <https://doi.org/10.1016/j.apenergy.2010.05.012>. adr.: <https://www.sciencedirect.com/science/article/pii/S0306261910001807>.
- [18] Taylor, S. i Letham, B., „Forecasting at scale”, wrz. 2017. DOI: 10.7287/peerj.preprints.3190v2.
- [19] Uniejewski, B., Nowotarski, J. i Weron, R., „Automated Variable Selection and Shrinkage for Day-Ahead Electricity Price Forecasting”, *Energies*, t. 9, nr. 8, 2016, ISSN: 1996-1073. DOI: 10.3390/en9080621. adr.: <https://www.mdpi.com/1996-1073/9/8/621>.
- [20] Weron, R., „Electricity price forecasting: A review of the state-of-the-art with a look into the future”, *International Journal of Forecasting*, t. 30, nr. 4, s. 1030–1081, 2014, ISSN: 0169-2070. DOI: <https://doi.org/10.1016/j.ijforecast.2014.08.008>. adr.: <https://www.sciencedirect.com/science/article/pii/S0169207014001083>.
- [21] Zhang, J., Tan, Z. i Yang, S., „Day-ahead electricity price forecasting by a new hybrid method”, *Computers & Industrial Engineering*, t. 63, nr. 3, s. 695–701, 2012, Energy Management and Economics, ISSN: 0360-8352. DOI: <https://doi.org/10.1016/j.cie.2012.03.016>. adr.: <https://www.sciencedirect.com/science/article/pii/S036083521200085X>.
- [22] Ziel, F., Steinert, R. i Husmann, S., „Efficient modeling and forecasting of electricity spot prices”, *Energy Economics*, t. 47, s. 98–111, 2015, ISSN: 0140-9883. DOI: <https://doi.org/10.1016/j.eneco.2014.10.012>. adr.: <https://www.sciencedirect.com/science/article/pii/S0140988314002576>.

Wykaz skrótów i symboli

MLP Wielowarstwowy Perceptron 18

OZE Odnawialne Źródła Energii 12

PSE Polskie Sieci Elektroenergetyczne 19

RB Rynek Bilansujący 34

RDN Rynek Dnia Następnego 11

Spis rysunków

1	Ceny energii elektrycznej na RDN w latach 2016-2023. Opracowanie własne na podstawie danych instrat.	20
2	Histogram rozkładu zmiennej fixing_i_price. Opracowanie własne na podstawie danych instrat.	21
3	Zmienna temperatura w czasie (2016-2023). Opracowanie własne na podstawie danych open-meteo.	23
4	Zmienna prędkość wiatru w czasie (2016-2023) Opracowanie własne na podstawie danych open-meteo.	23
5	Zmienna promieniowania słonecznego w czasie (2016-2023). Opracowanie własne na podstawie danych open-meteo.	24
6	Zmienna zachmurzenia w czasie (2016-2023). Opracowanie własne na podstawie danych open-meteo.	24
7	Zmienna temperatura w czasie (2022). Opracowanie własne na podstawie danych open-meteo.	25
8	Zmienna prędkości wiatru w czasie (2022). Opracowanie własne na podstawie danych open-meteo.	25
9	Zmienna promieniowania słonecznego w czasie (2022). Opracowanie własne na podstawie danych open-meteo.	26
10	Zmienna zachmurzenia w czasie (2022). Opracowanie własne na podstawie danych open-meteo.	26
11	Zmienna produkcji energii z różnych źródeł w czasie. Opracowanie własne na podstawie danych PSE.	27
12	Porównanie produkcji energii z emisjnych i bezemisjnych źródeł w 2023 roku. Opracowanie własne na podstawie danych PSE.	28
13	Ceny paliw kopalnych i emisji CO ₂ w latach w ujęciu miesięcznym. Źródło: Opracowanie własne na podstawie danych rynkowych.	31
14	Straty mocy w wyniku awarii i w sieci w ujęciu miesięcznym. Źródło: Opracowanie własne na podstawie danych z PSE.	32

15	Zapotrzebowanie na energię i wolumen handlu w latach 2016-2023 (w ujęciu miesięcznym). Źródło: Opracowanie własne na podstawie danych z PSE.	33
16	Średnie miesięczne odchylenie ceny na rynku bilansującym (<i>rb_price</i>) od ceny na Rynku Dnia Następnego (<i>fixing_i_price</i>) w latach 2016-2023 (w PLN/MWh). Źródło: Opracowanie własne na podstawie danych z PSE.	35
17	Korelacja dni tygodnia z cenami energii na RDN. Opracowanie własne.	36
18	Wykres korelacji zmiennych objaśniających względem zmiennej docelowej <i>fixing_i_price</i> . Opracowanie własne.	42
19	Mapa cieplna korelacji zmiennych objaśniających w skróconym zbiorze danych. Opracowanie własne.	43
20	Podział szeregów czasowych cen energii na zbiory treningowe i testowe. Opracowanie własne.	45
21	Porównanie rzeczywistych i przewidywanych wartości cen energii dla regresji Ridge w okresie stabilnym. Opracowanie własne.	60
22	Błędy prognoz dla regresji Ridge w okresie stabilnym. Opracowanie własne.	60
23	Histogram reszt dla regresji Ridge w okresie stabilnym. Opracowanie własne.	61
24	Porównanie rzeczywistych i przewidywanych wartości cen energii dla modelu Prophet w okresie stabilnym. Opracowane własne	62
25	Błędy prognoz dla modelu Prophet w okresie stabilnym.	63
26	Histogram reszt dla modelu Prophet w okresie stabilnym. Opracowanie własne.	64
27	Porównanie rzeczywistych i przewidywanych wartości cen energii dla modelu MLP w okresie stabilnym. Opracowanie własne.	65
28	Błędy prognoz dla modelu MLP w okresie stabilnym. Opracowanie własne.	66
29	Histogram reszt dla modelu MLP w okresie stabilnym. Opracowanie własne.	67
30	Prognozy modelu Ridge w porównaniu do rzeczywistych cen energii w okresie niestabilnym (2023). Opracowanie własne.	69
31	Błędy prognoz dla modelu Ridge w okresie niestabilnym (2023). Opracowanie własne.	69
32	Histogram reszt dla modelu Ridge w okresie niestabilnym (2023). Opracowanie własne.	70
33	Porównanie rzeczywistych i przewidywanych wartości cen energii dla modelu Prophet w okresie niestabilnym. Opracowanie własne.	71
34	Błędy prognoz dla modelu Prophet w okresie niestabilnym. Opracowanie własne.	72
35	Histogram reszt dla modelu Prophet w okresie niestabilnym. Opracowanie własne.	72
36	Porównanie rzeczywistych i przewidywanych wartości cen energii dla modelu MLP w okresie niestabilnym. Opracowanie własne.	73
37	Błędy prognoz dla modelu MLP w okresie niestabilnym. Opracowanie własne.	74
38	Histogram reszt dla modelu MLP w okresie niestabilnym. Opracowanie własne.	75
39	Największe błędy prognoz modelu Prophet z okresu stabilnego. Opracowanie własne.	79

Spis tabel

1	Podstawowe statystyki zmiennej fixing_i_price	20
2	Średni bilans wymiany energii z sąsiadami w latach 2016-2023. Opracowanie własne na podstawie danych PSE.	29
3	Statystyki cen energii (PLN/MWh) na rynkach sąsiednich w okresie 2016-2023. Opracowanie własne na podstawie danych ENTSO-E.	30
4	Opis zmiennych dotyczących cen paliw kopalnych i emisji CO ₂	31
5	Średni kurs wymiany PLN/USD i PLN/EUR w latach 2016-2023. Źródło: Opracowanie własne na podstawie danych NBP.	37
6	Porównanie statystyk opisowych cen energii w okresach spokojnym (2016-2019) i niespokojnym (2020-2023).	44
7	Wyniki regresji liniowej i grzebietowej dla pełnego zbioru danych w okresie stabilnym (2019).	58
8	Wyniki regresji liniowej i Ridge dla skróconego zbioru danych w okresie stabilnym (2019) wraz z różnicami względem pełnego zbioru.	59
9	Wyniki regresji liniowej i Ridge z logarytmizacją dla okresu stabilnego (2019).	59
10	Wyniki modelu Prophet dla pełnego i skróconego zbioru danych w okresie stabilnym (2019).	62
11	Wyniki modelu MLP dla pełnego i skróconego zbioru danych w okresie stabilnym (2019).	65
12	Wyniki metryk dla regresji liniowej i Ridge w okresie niestabilnym (2023). Opracowanie własne.	68
13	Wyniki modelu Prophet dla pełnego i skróconego zbioru danych w okresie niestabilnym (2023).	71
14	Wyniki modelu MLP dla pełnego i skróconego zbioru danych w okresie niestabilnym (2023).	73