

Article

A Machine Learning-Based Untapped Potential Index for Regional Photovoltaic Sales in Poland

Rafał Jankowski  and Anna Zielińska *

Faculty of Management, AGH University of Krakow, al. A. Mickiewicza 30, 30-059 Krakow, Poland;
rjankowski@agh.edu.pl

* Correspondence: azielinska@agh.edu.pl

Abstract

The rapid growth of the photovoltaic (PV) market in Poland has increased the importance of data-driven sales management and strategic resource allocation in renewable energy companies. Despite dynamic market expansion, significant regional differences in PV adoption indicate the existence of underutilized sales opportunities. This study aims to develop a decision-support framework for identifying company-specific regional gaps between actual and model-predicted photovoltaic contract volumes. The analysis is based on more than 15,000 real contracts obtained from a large PV installation company and enriched with regional demographic, economic, energy-related, housing, and regulatory data. Several machine learning models were evaluated, including Random Forest, XGBoost (XGB), HistGradientBoosting, and CatBoost. To improve transparency, explainable artificial intelligence (XAI) techniques based on SHapley Additive exPlanations (SHAP) were applied. Additionally, a model-based Untapped Potential Index (UPI) was developed as an operational indicator of the company-specific gap between actual and expected regional sales performance. The results indicate that CatBoost achieved the highest predictive accuracy and enabled the identification of regions with the largest positive company-specific sales gaps. The proposed UPI supports the optimization of marketing activities, sales resource allocation, and regional expansion strategies within the analyzed company. The broader methodological framework may be extended to regional market assessment through recalibration and external validation using independent market data.

Keywords: photovoltaic market; sales management; company-specific sales gap; untapped market potential; machine learning; decision support systems; energy economics; regional development; prosumer



Academic Editor: Alina Yakymchuk

Received: 18 June 2026

Revised: 24 July 2026

Accepted: 30 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

The development of the photovoltaic (PV) market in Poland in recent years has been characterized by exceptionally high growth dynamics, positioning the country among the leading European Union (EU) Member States in terms of newly installed capacity. The Polish PV sector experienced particularly rapid expansion between 2019 and 2025. Total installed PV capacity increased from less than 2 GW in 2019 to nearly 22 GW by the end of the first quarter of 2025 [1–3]. More specifically, in 2024, installed PV capacity reached 21.157 GW and further increased to 21.994 GW by the end of the first quarter of 2025 [3]. This rapid development placed Poland fifth among EU countries in terms of annual photovoltaic capacity additions and sixth in terms of cumulative installed capacity [1,2]. The Introduction is organized around four elements. First, it presents the market context of

photovoltaic development in Poland. Second, it identifies the research gap related to the limited use of real company-level contractual data in regional PV sales analysis. Third, it defines the objective of the study. Fourth, it summarizes the methodological and practical contribution of the paper.

Installed capacity growth has remained systematic and stable in recent years. Annual PV capacity additions amounted to approximately 4.1 GW in both 2022 and 2024, while a peak increase of more than 4.6 GW was recorded in 2023. The annual increase in newly commissioned PV installations amounted to approximately 4.1 GW in both 2022 and 2024, while 2023 recorded a peak expansion of more than 4.6 GW. Forecasts published by the Institute for Renewable Energy indicate that Poland may reach 38 GW of installed photovoltaic capacity as early as 2028, confirming the sustained growth trend and the increasing importance of solar energy within the national energy mix (Figure 1) [1,3].

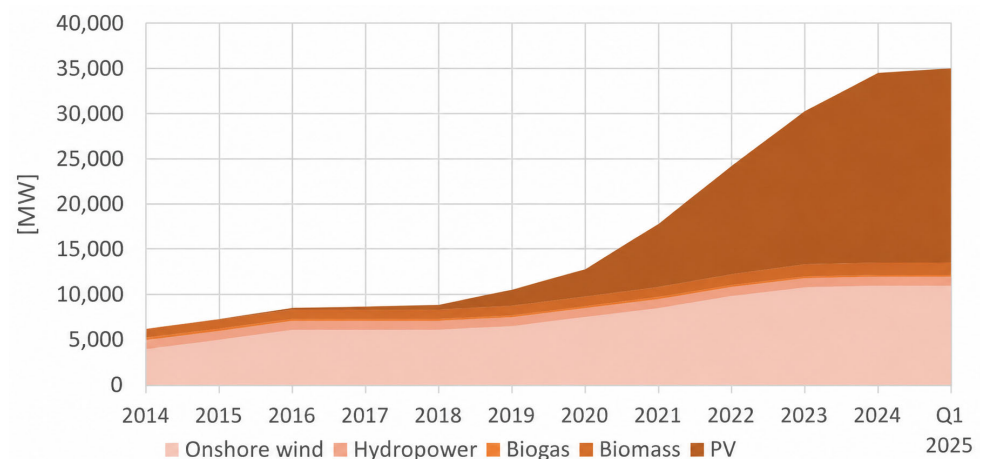


Figure 1. Installed Capacity Structure of Renewable Energy Sources in Poland [MW] [3].

The structure of the Polish photovoltaic market has undergone a significant transformation in recent years. While the prosumer micro-installation segment initially played a dominant role, a dynamic expansion of utility-scale PV farms has been observed since approximately 2023. Although micro-installations remain a key component of the market and have historically served as the primary driver of PV development, their relative share has gradually declined. At the end of 2024 and the beginning of 2025, micro-installations still accounted for approximately 59–60% of the total installed photovoltaic capacity [4,5].

In 2024, the number of active prosumers reached 1.53 million. During the same year, approximately 141.5 thousand new micro-installations, with a combined capacity of 1.43 GW, were connected to the grid, while the average capacity of a newly installed system amounted to 10.3 kW. Consequently, prosumers continue to represent a central pillar of the Polish photovoltaic market. Moreover, the growing interest in energy storage systems and electricity consumption management technologies indicates the increasing maturity of this market segment and its rising level of technological awareness [4,5].

A distinctive feature of the present study is the use of real contractual data obtained from the sales processes of a company operating in the Polish renewable energy market. The research gap addressed in this study concerns the limited use of real contractual data from installation companies to support regional sales decision-making in the photovoltaic market. While existing studies often rely on public statistics, technical potential assessments, or aggregated adoption data, company-level sales records make it possible to analyze the gap between expected and actual sales performance from the perspective of a specific market actor. Such data are rarely employed in scientific research because information regarding contracts, sales structures, and regional commercial activity is typically confidential and

constitutes a strategic asset of enterprises. In contrast to analyses based solely on publicly available datasets or aggregated statistics, the use of operational business data makes it possible to capture actual market behavior, sales seasonality, and regional differences in the sales effectiveness of photovoltaic installations. Consequently, the company dataset provides a foundation for developing a model with direct practical applicability, aimed at supporting sales decision-making processes.

To estimate the expected number of contracts, a machine learning-based approach was adopted. The analysis included ensemble models based on decision trees, namely Random Forest, XGBoost, CatBoost, and HistGradientBoosting with a Poisson loss function. The selection of these methods was motivated by the nonlinear relationships between regional characteristics and the number of concluded contracts, as well as by the need to model interactions among demographic, economic, energy-related, spatial, and regulatory variables. These models enable a transition from traditional descriptive analyses toward predictive assessments of expected sales performance at the region–month level.

Region–month observations with zero contracts were explicitly included, as they provide important information for modelling demand variability.

Furthermore, to enhance the transparency and interpretability of the results, Explainable Artificial Intelligence (XAI) techniques were employed, particularly SHAP (SHapley Additive exPlanations) values, which facilitate the interpretation of the contribution of individual variables to model predictions. The integration of real-world enterprise data, machine learning techniques, and XAI-based interpretability provides a foundation for constructing an indicator of unrealized sales potential and for formulating practical recommendations to support the management of regional PV sales strategies.

The study seeks to quantify the gap between the number of PV micro-installation contracts expected by the predictive model and the actual number of contracts achieved by the analyzed company across regions. Based on this approach, regions were identified in which company sales remained below the level expected on the basis of regional structural, economic, energy-related, and market characteristics.

An additional objective of the study is to develop a company-specific indicator of untapped regional sales opportunities. The indicator may serve as an analytical tool for identifying areas in which the number of contracts achieved by the analyzed company remains below the level expected on the basis of regional characteristics. The methodological framework may subsequently be extended to a broader regional market assessment through calibration against independent market data.

The motivation for undertaking this research stems from the substantial transformation and expansion of the Polish photovoltaic market in recent years, within which prosumers continue to play a central role. Although utility-scale PV farms have become an increasingly important driver of sectoral growth—with as much as 2.4 GW of such capacity connected to the grid in 2024—prosumer installations remain the cornerstone of the ongoing energy transition. The rapid development of this segment reflects the increasing professionalization of the market and the growing participation of institutional capital in photovoltaic investments [2,5].

Poland's position within the European Union photovoltaic market is particularly strong. In 2024, the country ranked fifth among EU Member States in terms of annual PV capacity additions, with 3.7 GW of newly installed capacity, and sixth in terms of total cumulative photovoltaic capacity. Moreover, Poland has one of the highest per capita photovoltaic capacity indicators worldwide, surpassing countries such as the United States and China. This achievement is primarily attributable to the widespread deployment of prosumer micro-installations. The contribution of photovoltaics to national electricity consumption has also been steadily increasing; in 2024, PV generation accounted for 10.6%

of total electricity consumption, while photovoltaics represented as much as 38.5% of the country's renewable energy generation mix [4,6].

1.1. Regional Heterogeneity and Underutilized Sales Opportunities

The contribution of the paper is primarily empirical and applied-methodological rather than algorithmic. The study does not propose a new machine learning algorithm or a new XAI method. Instead, it integrates real contractual data with regional socio-economic and energy-related characteristics, compares established machine learning models for predicting expected contract volumes, and operationalizes the UPI as a company-specific indicator supporting regional sales planning.

The rapid development of the photovoltaic market in Poland observed in recent years has revealed a number of significant structural and infrastructural challenges that increasingly limit the full utilization of the country's solar energy potential. One of the most critical issues is the phenomenon of curtailment, defined as the systematic reduction of electricity generation from photovoltaic installations due to constraints and congestion within the power grid. In the present study, these grid-related factors are treated as contextual constraints rather than direct model inputs, due to the sales-oriented scope of the analysis and the lack of consistent region-month data for the full study period. Although previously considered a marginal issue, curtailment has become increasingly important as installed renewable energy capacity continues to expand. It is estimated that during the first half of 2025, more than 600 GWh of electricity generation from renewable energy sources was curtailed in Poland, providing a clear indication that grid infrastructure development has not kept pace with the ongoing energy transition [3,7].

Problems associated with the limited capacity of the electricity grid constitute one of the main barriers to the further development of the photovoltaic market, particularly in the segment of large-scale installations and utility-scale PV farms. In practice, this means that in many regions, despite favorable solar irradiation conditions and the availability of suitable land for investment, the implementation of new high-capacity projects is constrained or even prevented by insufficient grid connection opportunities. The scientific literature emphasizes that infrastructural limitations, including the low flexibility of the power system and the insufficient pace of grid modernization, contribute to the emergence of so-called development bottlenecks, which restrict the effective integration of additional renewable energy capacity into the electricity system [8,9].

However, it should be emphasized that the problem of underutilized photovoltaic market potential cannot be attributed solely to technical constraints. Economic and social factors also play a significant role, including variations in household income levels, environmental awareness, building structure characteristics, and access to investment financing mechanisms. Consequently, PV adoption remains relatively low in some Polish regions despite favorable solar irradiation conditions. This phenomenon may be interpreted as an expression of underutilized market potential that is not reflected in the actual level of photovoltaic deployment and market development [9–11].

In response to the growing challenges associated with integrating renewable energy sources into the electricity system, energy storage technologies have become increasingly important. Energy storage systems, particularly battery energy storage systems (BESS), constitute a key element in enhancing system flexibility and enabling a better alignment between electricity supply and demand. Poland has experienced dynamic growth in this market segment, as evidenced by the increasing volume of contracted energy storage capacity within the capacity market framework and the growing number of related investment projects [12,13].

At the same time, support schemes for renewable energy investments increasingly promote hybrid solutions that combine photovoltaic installations with energy storage systems or thermal energy storage technologies. The introduction and subsequent evolution of support mechanisms, such as the “My Electricity” (Mój Prąd) and “Clean Air” (Czyste Powietrze) programs, toward encouraging self-consumption and energy storage solutions indicate a shift in the operational model of the photovoltaic market—from an approach focused primarily on maximizing electricity generation to one centered on the efficient utilization and management of produced energy.

Despite these developments, the issue of underutilized photovoltaic market potential remains highly relevant and warrants further investigation, particularly from a regional perspective. In this context, the development of analytical tools capable of identifying areas characterized by the largest discrepancies between the potential arising from natural and structural conditions and the actual level of market development is of considerable importance. Such analyses may provide a foundation for more effectively targeting investment activities and energy policy measures, thereby contributing to a more comprehensive utilization of Poland’s photovoltaic potential.

1.2. The Role of Economic Factors in Photovoltaic Market Development

The development of the photovoltaic market in Poland is significantly influenced by economic factors and regional conditions, which affect both the level of investment in micro-installations and the rate of technology adoption among households. Among the economic determinants, public support schemes play a particularly important role, as they substantially reduce entry barriers for prosumers and improve the economic viability of investments in photovoltaic installations.

A key instrument supporting the development of prosumer photovoltaics in Poland has been the Mój Prąd (“My Electricity”) program, implemented by the National Fund for Environmental Protection and Water Management. The primary objective of the program has been to increase electricity generation from PV micro-installations and to enhance energy self-consumption through the development of energy storage systems [14]. Successive editions of the program have demonstrated a gradual evolution of support mechanisms—from subsidies focused exclusively on photovoltaic installations toward a model promoting the integration of PV systems with energy storage technologies and energy management solutions.

In particular, under the most recent editions of the program, financial support may cover up to 50% of eligible investment costs, including expenditures related to PV installations, battery energy storage systems, and thermal energy storage facilities. At the same time, the net-billing system was introduced, replacing the previous net-metering scheme and encouraging the direct consumption of electricity generated by prosumers. This regulatory change has significantly altered the economic framework of photovoltaic investments, increasing the importance of self-consumption and efficient energy management practices [15].

In this context, the economic viability of PV systems increasingly depends on the ability of prosumers to consume a larger share of generated electricity directly or to shift electricity use through energy storage and energy management systems.

It is worth emphasizing that current support policies increasingly focus on the development of complementary infrastructure, particularly energy storage systems. In the most recent subsidy programs, the installation of either an energy storage system or a thermal energy storage unit has effectively become a standard requirement for newly supported micro-installations. This approach aims to improve the balancing of the electricity system and reduce grid congestion [12,13]. Such solutions enhance the flexibility of the energy

system while simultaneously reshaping demand within the PV market, shifting it toward more integrated and comprehensive energy solutions.

In 2023, households were eligible to apply for financial support not only for photovoltaic installations but also for technologies such as heat pumps, solar thermal collectors, battery energy storage systems, thermal energy storage systems, and energy management systems. Since August 2024, investment in either thermal energy storage or battery energy storage has become a prerequisite for receiving support. The capacity of PV installations eligible under the program ranged from 2 to 20 kW, whereas prior to August 2024 the eligible capacity range was limited to 2–10 kW [3].

Under the *Mój Prąd* program, subsidies were granted in 2024 and the first quarter of 2025 for a total of 53,509 photovoltaic installations with a combined capacity of 337,616 kW. The average installed capacity amounted to 6.4 kW. In addition to financial support mechanisms, the development of photovoltaics in Poland is also driven by ongoing technological advancements. Hybrid solutions integrating photovoltaic systems with wind energy generation and energy storage technologies are gaining increasing popularity, as they enhance the stability and reliability of renewable energy generation. Innovative concepts such as cable pooling and direct power lines are also becoming increasingly important; however, their broader implementation will require further legislative and regulatory adjustments. At the same time, Energy Management Systems (EMS) are becoming more widespread, enabling prosumers to optimize self-consumption, improve energy efficiency, and enhance the economic performance of photovoltaic investments [4,5].

In 2024 and the first quarter of 2025, the National Fund for Environmental Protection and Water Management (NFOŚiGW) disbursed a total of PLN 635 million to support investments in PV micro-installations and technologies designed to enhance the efficient utilization of electricity generated by these systems. The total value of the supported projects reached PLN 2.182 billion, indicating that subsidies provided under the *Mój Prąd* (“My Electricity”) program covered approximately 29.1% of total investment expenditures. The average investment value amounted to approximately PLN 41,000, while the average subsidy per project was approximately PLN 11,000 [3].

1.3. The Role of Regional Factors in Photovoltaic Market Development

In addition to economic determinants, regional factors are also of considerable importance, particularly geographical conditions and solar irradiation levels. Solar irradiation constitutes one of the key factors influencing the efficiency of photovoltaic installations and their utilization across different regions of Poland. The country is located within a moderate solar irradiation zone, where the annual global solar radiation typically ranges from approximately 950 to 1250 kWh/m², providing relatively favorable conditions for solar energy development compared with many other European countries [16,17]. This spatial distribution is illustrated in Figure 2, which presents the solar irradiation zones in Poland.

A more detailed analysis indicates that the performance of photovoltaic systems depends on several local climatic parameters, the most important of which are annual solar irradiation (kWh/m²), sunshine duration (the number of sunshine hours), and ambient temperature. Poland is located between 49° N and 55° N latitude, placing the country within a temperate transitional climate zone. For optimally oriented PV panels (with a tilt angle of approximately 30°), the average annual solar irradiation reaches approximately 1289 kWh/m². Cloud cover and seasonal variability are among the most important factors affecting fluctuations in PV electricity generation throughout the year.

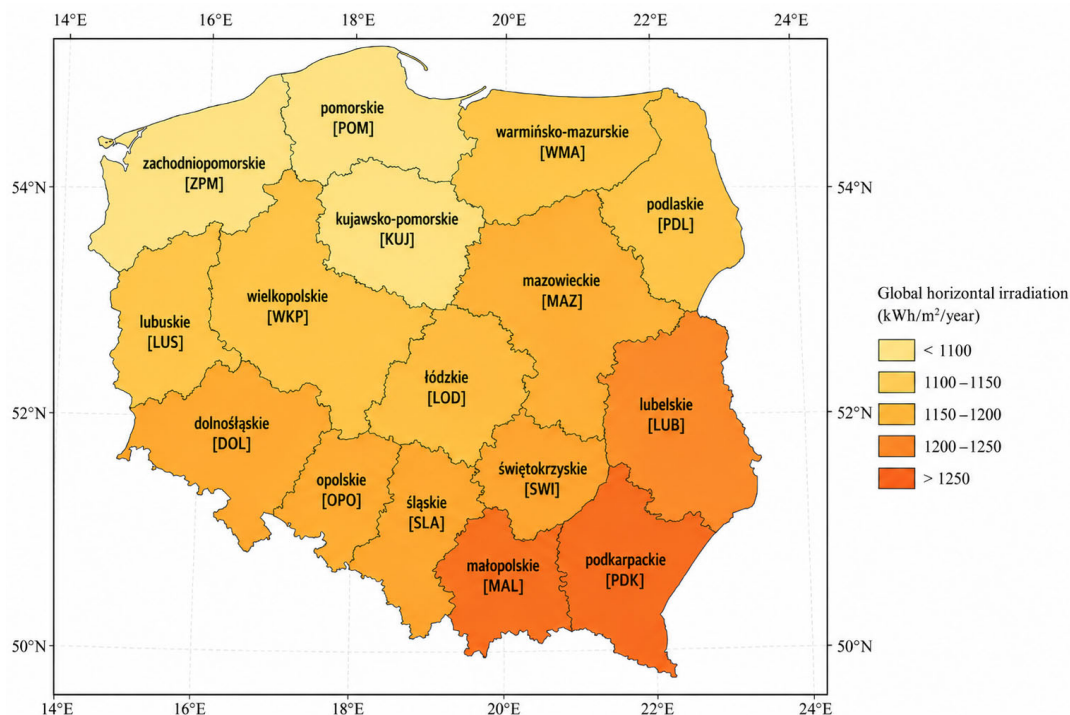


Figure 2. Solar irradiation map of Poland (kWh/m²/year) [18].

A defining characteristic of Poland's climate is the pronounced seasonality of solar irradiation, which directly translates into seasonal variations in photovoltaic electricity generation. During the summer months (May–August), solar radiation levels and the number of sunny days reach their highest values, resulting in peak electricity production. In contrast, during the winter period (November–February), solar irradiation is substantially lower, daylight hours are shorter, and cloud cover is more frequent, leading to significantly reduced energy yields from PV installations. The annual number of sunshine hours in Poland ranges from approximately 1400 to 1680 h and exhibits considerable regional variation (Figure 2). The highest values are recorded in southeastern Poland, whereas the lowest occur in coastal regions and the Carpathian Mountains [18,19].

From a quantitative perspective, approximately 70–80% of annual photovoltaic electricity generation in Poland occurs between spring and early autumn, while winter production contributes only a relatively small share of total annual output. This seasonal generation profile creates significant challenges for the electricity system, including periods of excess electricity production during summer months and potential supply shortages during winter. It should be emphasized that the number of sunshine hours should not be interpreted as the direct annual operating time of PV systems. Photovoltaic installations can generate electricity not only under direct sunshine but also under diffuse solar radiation. Therefore, the economic viability of PV systems depends not only on the number of sunshine hours, but also on annual energy yield, self-consumption levels, electricity prices, investment costs, subsidy schemes, settlement rules, and the possible use of energy storage systems. The seasonal concentration of PV generation therefore represents an operational and system-integration challenge rather than direct evidence of economic non-viability. In this study, energy storage is considered as a complementary technology for on-grid prosumer systems, supporting self-consumption and system flexibility, rather than as a component of off-grid autonomous PV systems.

Poland therefore possesses substantial potential for the further expansion of photovoltaics, both in the utility-scale solar farm segment and in the prosumer micro-installation sector. According to industry forecasts, total installed PV capacity in Poland may reach

38 GW by 2028, while the National Energy and Climate Plan (NECP) envisages a system capacity of 38 GW by 2035. The technical potential for further development is determined by the availability of suitable land, the capacity of the electricity grid, the pace of new installation deployment, and the efficiency of the technologies employed [3,19].

However, the realization of this potential requires consideration of seasonal production patterns and economic factors affecting investment profitability, particularly self-consumption, grid constraints, and the increasing role of energy storage.

Spatial analyses indicate that regional differences may translate into variations in annual electricity generation from PV installations of up to several percentage points. In practical terms, this means that installations with the same installed capacity may achieve substantially different energy yields depending on their geographical location. In southern regions of Poland, electricity generation tends to be higher and more stable, whereas in northern regions weather variability and cloud cover exert a greater influence on system performance [19].

Despite relatively favorable solar irradiation conditions across the country, a significant challenge remains the mismatch between the actual utilization of photovoltaic installations and their technical potential. In particular, the seasonal concentration of electricity generation during the summer months contributes to grid congestion and the increasing need for generation curtailment, implying that a portion of the available solar energy potential remains unused. At the same time, during winter periods, when electricity demand is generally higher, PV generation is substantially reduced. This situation highlights the importance of integrating energy storage systems and diversifying renewable energy sources within the energy mix. The PV systems discussed in this context refer primarily to on-grid prosumer micro-installations connected to the electricity distribution network. The present study does not analyze autonomous off-grid PV systems.

From the perspective of market potential assessment, it is therefore essential to consider not only average annual solar irradiation levels but also their seasonal and regional variability. Recent studies indicate that PV market analyses are frequently based on annual average values alone, which may lead to an underestimation of actual regional differences and, consequently, to inaccurate assessments of investment potential [20,21].

In summary, solar irradiation constitutes an important, yet insufficient, determinant of photovoltaic market development in Poland. Its pronounced seasonality and regional heterogeneity mean that the available potential for PV electricity generation is not always fully utilized. Incorporating these factors into analytical and predictive models is therefore essential for identifying regional differences in photovoltaic development.

1.4. Other Factors Influencing Photovoltaic Market Potential: The Number of Single-Family Houses

The structure of the housing stock in Poland clearly indicates the dominant role of single-family housing as the primary foundation for the development of photovoltaic installations. According to data from the Central Statistical Office (GUS), approximately 6.9 million residential buildings existed in Poland in 2021, of which as many as 6.3 million were single-family houses, accounting for more than 90% of all residential buildings [22]. These figures demonstrate the existence of a substantial infrastructural base for the further expansion of PV micro-installations.

The spatial distribution of these housing resources is equally important. Approximately 3.3 million single-family houses are located in rural areas, while around 1.7 million are situated in urban areas [22]. This indicates that the primary potential for the development of the prosumer market is concentrated outside highly urbanized regions, with significant implications for the spatial distribution of photovoltaic installations.

From a regional perspective, the highest concentration of single-family housing is observed in southeastern Poland and parts of central Poland. Regions such as the Lubelskie, Podkarpackie, Świętokrzyskie, and Małopolskie voivodeships are characterized by a high share of single-family housing and a dispersed settlement structure. At the same time, dynamic growth in single-family housing construction is observed in the suburban zones of major cities, including Kraków County (Małopolskie Voivodeship), Poznań County (Wielkopolskie Voivodeship), and Lublin County (Lubelskie Voivodeship), highlighting the increasing importance of suburbanization processes [22].

In this context, the level of urbanization is particularly important, as it directly influences the development potential of prosumer photovoltaics. Poland remains a moderately urbanized country, with approximately 60% of the population residing in urban areas. However, a substantial proportion of urban residents live in multi-family residential buildings [22]. This limits opportunities for the deployment of household-level PV systems, as multi-family buildings require collective investment decisions and provide relatively limited roof area per household.

In contrast, rural and suburban areas, characterized by lower levels of urbanization, offer significantly more favorable conditions for photovoltaic development. Single-family houses provide larger roof areas, greater opportunities for optimal system orientation, and increased investment autonomy for property owners. Consequently, regions with lower levels of urbanization tend to exhibit higher rates of photovoltaic technology adoption, which is reflected in the number of installations per capita.

At the same time, suburbanization processes are creating new areas with considerable photovoltaic development potential. Suburban zones combine characteristics of both urban and rural environments, featuring relatively high household incomes alongside a predominance of single-family housing. As a result, these areas represent some of the most dynamic segments of the PV market, where investments in photovoltaic installations are particularly intensive.

In highly urbanized regions, such as the metropolitan part of the Mazowieckie Voivodeship, the Śląskie Voivodeship, and the Małopolskie Voivodeship, the share of multi-family housing is substantially higher, limiting the development of prosumer micro-installations. Paradoxically, these areas may simultaneously exhibit high levels of energy demand, leading to the emergence of significant yet underutilized photovoltaic market potential within the distributed generation segment.

It is also important to emphasize that the single-family housing sector in Poland continues to expand. Such buildings account for more than 97% of all newly completed residential buildings [22]. This trend further enlarges the pool of potential prosumers, particularly in suburban areas.

In summary, the level of urbanization and the structure of residential development constitute key determinants of photovoltaic market development in Poland. Regions characterized by a high share of single-family housing and lower levels of urbanization demonstrate considerably greater potential for PV deployment. At the same time, highly urbanized regions may possess substantial but underutilized market potential resulting from structural and organizational constraints. Incorporating these factors into analytical and predictive models is essential for accurately identifying the spatial heterogeneity of photovoltaic market potential.

The results of the analysis of the distribution of photovoltaic micro-installations in Poland confirm the crucial role of residential building structure and the level of urbanization as key determinants of prosumer market development. The highest levels of installed capacity in PV micro-installations are observed in the Mazowieckie (Masovian), Wielkopolskie (Greater Poland), and Śląskie (Silesian) voivodeships (Table 1). This pattern is directly

associated with large population sizes, well-developed suburban zones, and a high share of single-family housing. These findings confirm that the availability of rooftop infrastructure and the decision-making autonomy of households are among the most important factors supporting the development of photovoltaic installations.

Table 1. Installed Photovoltaic (PV) Microinstallation Capacity by Voivodeship in Poland [GW] [3].

Voivodeship	Short Code	PV Microinstallation Capacity (GW)
mazowieckie	MAZ	1.6–1.8
wielkopolskie	WKP	1.5–1.7
śląskie	SLA	1.3–1.5
małopolskie	MAL	1.0–1.3
łódzkie	LOD	1.0–1.2
dolnośląskie	DOL	0.9–1.1
lubelskie	LUB	0.8–1.0
podkarpackie	PDK	0.8–1.0
kujawsko-pomorskie	KUJ	0.8–1.0
pomorskie	POM	0.6–0.9
zachodniopomorskie	ZPM	0.6–0.9
warmińsko-mazurskie	WMA	0.5–0.8
podlaskie	PDL	0.4–0.7
świętokrzyskie	SWI	0.4–0.7
opolskie	OPO	0.3–0.6
lubuskie	LUS	0.3–0.6

At the same time, the results indicate that the spatial distribution of PV micro-installations in Poland is driven to a greater extent by demographic and structural factors than by natural conditions, particularly solar irradiation levels. Regions characterized by a high proportion of single-family houses and moderate levels of urbanization exhibit higher rates of PV technology adoption, whereas highly urbanized areas dominated by multi-family residential buildings tend to display relatively lower installed capacities within the prosumer segment.

In this context, it is particularly noteworthy that some regions with favorable solar irradiation conditions, such as the Lubelskie and Podkarpackie voivodeships, are not among the leading regions in terms of PV micro-installation deployment. This situation indicates the existence of a discrepancy between technical potential and its actual utilization, which may be interpreted as evidence of underutilized market potential [23]. Incorporating these relationships into subsequent analytical and predictive models is essential for identifying regions where investment activities and support mechanisms should be intensified to unlock further photovoltaic market development.

2. Materials and Methods

Despite the dynamic development of the prosumer photovoltaic market in Poland and the growing number of studies addressing technical, economic, and regulatory determinants, relatively few studies have focused on identifying company-specific regional sales gaps using real-world contractual data obtained from installation companies. Existing research on photovoltaic adoption demonstrates that household investment decisions exhibit substantial spatial heterogeneity and depend not only on solar irradiation conditions but also on housing characteristics, income levels, population density, previous technology diffusion, and neighborhood effects [24,25]. At the same time, the application of machine learning methods in prosumer market analysis can support the identification of areas with

high sales potential while reducing customer acquisition costs, as confirmed by studies employing predictive models in the residential PV sector [26].

However, existing studies have primarily focused on the adoption of PV technology from statistical, spatial, or techno-economic perspectives. Comparatively less attention has been devoted to the practical use of contractual data from installation companies to identify regions where actual photovoltaic sales remain below the level predicted on the basis of regional characteristics and historical sales patterns [27]. From the perspective of an installation company, the key question is not merely where the highest number of installations is observed, but rather where a gap exists between expected and actual sales performance. Such a gap may indicate regions requiring intensified commercial, marketing, or operational activities. Therefore, regions with high positive UPI values should be treated as candidates for further managerial diagnosis rather than as regions with automatically confirmed market-wide photovoltaic potential.

The research gap addressed by this study lies in the absence of a regional sales decision-support model that integrates real contractual data from an installation company with the socio-economic, spatial, energy-related, and regulatory characteristics of regions in order to identify underutilized sales opportunities within the photovoltaic installation market. Previous studies have focused primarily on public datasets, aggregated market statistics, technical energy production parameters, or general determinants of PV technology adoption. Data derived directly from actual enterprise sales processes have been used far less frequently, despite their ability to reflect the practical conditions underlying market decisions more accurately.

The methodological value of this study lies in the integration of established machine learning and XAI methods with real contractual sales data and regional market characteristics. The study does not claim algorithmic novelty. Rather, it proposes an applied decision-support framework in which expected regional contract volumes are estimated using predictive models and then compared with actual company sales to identify regional sales gaps. In this context, the regional criteria refer to the groups of variables used to describe each voivodeship in the predictive model, including structural and demographic characteristics, housing-related prosumer potential, economic conditions, solar resource conditions, previous PV market maturity, and regulatory support periods.

The contribution of this study is mainly empirical and practical. First, it uses real operational company data that are rarely available for academic research. Second, it integrates these data with regional variables describing market conditions. Third, it applies established machine learning models to estimate expected contract volumes at the region-month level. Fourth, it operationalizes the UPI as a transparent indicator of company-specific regional sales gaps. Fifth, SHAP values are used to support global model interpretation. The proposed framework may support sales management, resource allocation within installation companies, and regional planning activities in the renewable energy sector.

2.1. Data Source and Scope of the Study

The data used in this study were obtained from a company operating in the Polish renewable energy market since 2013. The data provider is listed on the NewConnect market operated by the Warsaw Stock Exchange, indicating its formal status as a capital market participant and its established presence within the renewable energy sector. Due to the confidential nature of contractual data and their limited availability for scientific research purposes, the dataset constitutes a valuable empirical source for analyzing actual sales processes in the Polish photovoltaic installation market. It should be noted that obtaining comparable contract-level data from competing PV installation companies is highly difficult, as such information is commercially sensitive and usually not publicly dis-

closed. The dataset used in this study was obtained from a large nationwide PV installation company operating across Poland. Therefore, the data provide valuable insight into real sales processes across the national market, although they still represent the activity of a single enterprise.

The original dataset covers the period from 2023 to 2025 and contains more than 58,700 records describing the company's sales activities. These records include photovoltaic installations, energy storage systems, electric vehicle chargers and charging stations, electricity sales contracts, and complementary products such as air conditioning systems and thermal modernization services. For the purposes of this study, only records relevant to photovoltaic installations were extracted and retained for further analysis.

The temporal scope of the study was restricted to the years 2023–2025 because, during this period, the company utilized a new Customer Relationship Management (CRM) system that ensured a uniform standard for recording contractual data. Limiting the analysis to this timeframe reduces the risk of inconsistencies arising from changes in sales registration procedures.

It should be emphasized that the use of real operational data is of critical importance for the development and validation of machine learning (ML) models [28]. Unlike synthetic or highly aggregated datasets, data derived from actual business processes reflect real customer behavior, regional market conditions, sales seasonality, and practical operational constraints faced by enterprises. Consequently, the proposed model can achieve higher predictive reliability and greater practical value in business applications.

The practical significance of such datasets is particularly evident in ML-based decision-support systems. The quality, diversity, and authenticity of input data directly influence forecasting performance and the ability to identify hidden relationships among variables. As a result, a model developed using real sales data extends beyond a purely theoretical or descriptive framework and can serve as a practical tool supporting business operations [29,30].

Following the initial data selection process, the analysis was restricted to contracts related to photovoltaic micro-installations, defined as systems with an installed capacity of up to 10 kWp. The analyzed PV micro-installations should be understood as grid-connected prosumer systems; off-grid installations were not separately identified or analyzed in the dataset. This criterion was adopted to ensure the homogeneity of the analyzed market segment and to focus on the prosumer sector, which constitutes the primary application area of the proposed sales decision-support model. After applying the filtering criteria, the final dataset used for modeling consisted of 15,670 contractual observations related to photovoltaic installations meeting the specified capacity threshold.

The contractual data were subsequently enriched with a set of regional variables describing the socio-economic, demographic, spatial, energy-related, and regulatory characteristics of individual voivodeships. Public national and regional data were used during the data enrichment and preprocessing stage to characterize the broader market environment. In particular, official and publicly available information on regional PV development and installed capacity was transformed into contextual indicators of market maturity. These variables were included in the model indirectly, for example through lagged installed PV capacity per capita and the lagged number of installations per 1000 single-family houses. This approach allowed the model to account for the general level of regional PV market development while reducing the risk of information leakage from contemporaneous market outcomes.

Because the original dataset consisted of operational contractual records, it was necessary to transform the data into an aggregated format.

Prior to aggregation, the dataset was subjected to basic preprocessing. Missing values in regional variables were limited and were handled using the most recent available

observation or constant regional values. No extensive imputation procedures were required. Observations with zero contracts were retained in the dataset, as they represent meaningful information about the absence of sales activity in a given region–month. Due to the aggregated nature of the data, extreme values were not removed; instead, tree-based models were used, which are robust to outliers.

To achieve this, observations were grouped according to regional codes and time indices. For each region–month pair, the number of concluded contracts was calculated and used as the target variable in the predictive model (1). After aggregation at the region–month level, the final modelling dataset consisted of 576 observations.

$$Y_{r,t} = \text{count}(\text{contracts}_{r,t}) \quad (1)$$

where $Y_{r,t}$ denotes the actual number of PV micro-installation contracts concluded by the analyzed company in region r during month t . The term count refers to the number of retained contract records assigned to a given region–month pair and does not refer to installed capacity, sales value, or the number of sales leads.

Regional variables that remained unchanged within a given region during the analyzed period were assigned to the aggregated observations as constant values or as the first available value within each group. The variable *installed_capacity_kw* was not included as an explanatory feature in the primary contract prediction model, as it may reflect the outcome of sales activity rather than its underlying determinants. Nevertheless, this variable was retained for use in descriptive analyses and in an alternative modeling framework in which the target variable would be the total installed capacity within a given region and month. Such an approach is identified as a potential direction for future research.

The modeling process incorporated a set of variables describing the temporal dynamics of sales, regional scale and structure, economic conditions, the technical potential of photovoltaic installations, the existing maturity of the PV market, and the regulatory environment. The selection of variables was guided by the objective of predicting the expected number of contracts at the region–month level and identifying regions where actual sales performance was lower than the level expected based on underlying market characteristics:

- Temporal variables: Time index and monthly seasonality encoded using sine–cosine transformations (*time_index*, *month_sin*, *month_cos*). The *time_index* variable represents consecutive months of observation, while *month_sin* and *month_cos* were employed to preserve the cyclical nature of seasonal patterns. This encoding allows the model to account for the fact that December and January are adjacent months within the annual cycle, despite being numerically distant in a conventional monthly representation.
- Spatial and demographic variables: Regional area, population size, population density, and urbanization level (*region_code*, *area_km²*, *population*, *population_density_per_km²*, *urbanization_rate*). These variables describe the scale and structural characteristics of each region. Their inclusion is important because the number of contracts may partly reflect the size of a voivodeship, the number of inhabitants, or the degree of urbanization. Incorporating these factors reduces the risk of incorrectly interpreting larger regions as inherently more attractive sales markets.
- Variables describing prosumer housing potential: The number of single-family houses and the number of single-family houses per 1000 inhabitants (*single_family_houses*, *single_family_per_1000_population*). Particular importance was assigned to the number of single-family houses, as prosumer PV micro-installations are strongly associated with the availability of individual rooftop areas (or, alternatively, available land for ground-mounted PV systems) and the investment autonomy of property owners. The

variable *single_family_per_1000_population* was included to normalize this information relative to the size of the regional population.

- Economic variables: Gross domestic product per capita (*gdp_per_capita_pln*). GDP per capita was used as a regional indicator of economic prosperity and potential investment capacity. In the context of PV installations, this variable is particularly relevant because investments in micro-installations typically require either the availability of private capital or access to external financing, regardless of whether public support programs are available.
- Energy-related variables: Annual Global Horizontal Irradiance (GHI) (*ghi_kwh_m2_year*). Annual GHI was employed as a measure of the natural conditions supporting photovoltaic development. This variable represents the availability of solar resources within a given region and indirectly captures regional differences in technical energy-generation potential across voivodeships (or, more broadly, regions). It should be noted that GHI is used here as a static proxy for regional solar resource conditions. The model does not include regional PV electricity generation, curtailment, grid-connection availability, or power accommodation capacity, as the present framework focuses on sales performance rather than on a full technical assessment of the power system.
- Variables describing PV market maturity: Lagged installed PV capacity per capita and lagged number of installations per 1000 houses (*installed_capacity_kw_per_capita_lag1*, *installations_per_1000_houses_lag1*). These variables characterize the previous level of PV market penetration within a region. Lagged values were used to reduce the risk of information leakage between explanatory variables and the target variable. As a result, the model relies on information available prior to the prediction period rather than on data originating from the same period (month) as the dependent variable. These variables were derived from public regional market data and therefore provide an indirect representation of the overall development of the PV market in each voivodeship, although they do not replace direct information on competitors' sales or company-level regional market shares. These lagged variables should be interpreted as stock-based indicators of previous PV market development rather than as short-term sales indicators. In particular, lagged installed capacity per capita partly captures the cumulative historical diffusion of PV installations in a region and may therefore reflect market maturity or partial saturation effects.
- Regulatory variables: Periods during which support schemes were in effect (*subsidy_period*). This variable represents periods associated with public support mechanisms, including government subsidy programs such as Mój Prąd [13–15]. Its inclusion is important because subsidy schemes affect the economic attractiveness of PV investments and may alter demand patterns over time. From the perspective of the decision-support framework proposed in this study, this variable is expected to play an important role, as its effects may potentially be mitigated through appropriate commercial and operational strategies implemented by installation companies.

2.2. Conceptualization of the Company-Specific Sales-Gap

In this study, the UPI is defined as a company-specific regional sales-gap indicator rather than as a direct measure of general market potential or technical photovoltaic potential. It compares the actual number of contracts achieved by the analyzed company with the number predicted by the model. The model-predicted contract level is estimated on the basis of energy-related conditions, housing structure, the number of potential prosumers, regional economic prosperity, urbanization, previous market penetration, and

public support mechanisms. A machine learning model is employed to estimate the expected number of contracts in a given region and month, as expressed in Equation (2):

$$\widehat{Y}_{r,t} = f(X_{r,t}) \quad (2)$$

where $X_{r,t}$ denotes the vector of variables describing region (r) during period (t); $\widehat{Y}_{r,t}$ represents the predicted number of contracts; and $Y_{r,t}$ denotes the actual number of contracts.

Based on these quantities, the absolute sales gap (UP) was defined as follows (Equation (3)):

$$UP_{r,t} = \widehat{Y}_{r,t} - Y_{r,t} \quad (3)$$

where $UP_{r,t}$ denotes the untapped sales opportunity in region (r) during period (t).

To enable comparisons across regions characterized by different market sizes, a relative indicator of untapped sales opportunities was introduced, as expressed in Equation (4). The UPI was intentionally formulated as a simple and transparent indicator. Its purpose is not to introduce a mathematically complex metric, but to provide an interpretable operational measure that can be used by managers to compare regions and identify areas requiring further sales diagnosis.

$$UPI_{r,t} = \frac{\widehat{Y}_{r,t} - Y_{r,t}}{\widehat{Y}_{r,t}} \quad (4)$$

The interpretation of this indicator is as follows. Positive UPI values indicate that the actual number of contracts achieved by the analyzed company was lower than the number predicted by the model, corresponding to a positive company-specific regional sales gap. Values close to zero indicate that actual sales were close to the model-predicted level, whereas negative values indicate that the company achieved more contracts than predicted by the model. Values close to zero indicate that actual sales performance is consistent with the model-estimated market potential, whereas negative values indicate that a region achieved sales results exceeding those predicted by the model.

From the perspective of an installation company, this indicator may serve as a decision-support tool for regional expansion planning and for identifying areas where commercial and sales activities should be intensified to better exploit existing market opportunities.

In its present implementation, the UPI is a company-specific indicator of the regional sales gap because both the actual and predicted values refer to contracts concluded by the analyzed firm. Therefore, the indicator does not directly measure the total market or technical photovoltaic potential of a region. Its value may partly reflect the company's commercial strategy, local sales coverage, marketing intensity, operational capacity, and competitive position. However, the methodological framework is transferable and may be extended to a broader regional market assessment by incorporating independent data on the total number of installations or contracts, competitors' activity, and the company's corresponding regional market share. Such an extension would require recalibration and external validation before the resulting indicator could be interpreted at the market-wide regional level.

2.3. Research Methodology

The study adopted a comparative approach involving several machine learning algorithms designed to predict the number of contracts at the region-month level. The objective of this stage was not to employ a single arbitrarily selected model but rather to empirically compare the predictive performance of multiple methods and select the model that most accurately estimates the expected number of contracts [31,32]. Tashman justifies the use of out-of-sample testing and forecasting validation procedures, while Hyndman and Koehler discuss forecast accuracy assessment and the limitations of various error metrics.

These studies support the rationale for comparing alternative models using a common validation framework.

Model hyperparameters were selected using validation set performance rather than exhaustive grid search, in order to maintain consistency across models. Default or near-default configurations were used for most algorithms, with limited tuning of key parameters such as learning rate, tree depth, and number of estimators. The same validation framework was applied to all models to ensure comparability.

Only the model demonstrating the highest predictive performance was subsequently employed to construct the Untapped Potential Index (UPI).

Due to the nonlinear relationships between regional characteristics and the number of contracts, ensemble models based on decision trees were selected for analysis [33–35]. The Random Forest approach originates from the work of Breiman, while gradient boosting was introduced by Friedman. More recent evidence regarding tabular datasets further supports the suitability of tree-based and boosting models for such prediction problems. These models are particularly well suited to tabular data containing heterogeneous variables, including temporal, demographic, economic, spatial, energy-related, and regulatory features. They also enable the modeling of complex interactions among variables without requiring their complete manual specification.

The following models were included in the analysis:

- **Regional Mean Baseline**—A benchmark model based on the historical regional average. This model served as a reference point for evaluating whether more advanced machine learning approaches provide meaningful predictive improvements over a simple baseline.
- **Random Forest Regressor**—An ensemble model based on the aggregation of multiple decision trees. It was employed as a robust reference model capable of capturing nonlinear relationships while reducing the variance associated with individual trees. Breiman defines Random Forest as a combination of multiple tree predictors, supporting its characterization as a stable ensemble method that mitigates the variance of individual trees [33].
- **XGBoost Regressor**—A gradient boosting model widely used in predictive analytics involving tabular datasets. It was included due to its strong predictive performance and its ability to model complex interactions among explanatory variables.
- **CatBoost Regressor**—A gradient boosting model particularly effective for heterogeneous datasets and categorical variables. In the present study, CatBoost was treated as one of the candidate models, and its potential selection as the primary model depended exclusively on its predictive performance on the validation and test datasets. Prokhorenkova et al. demonstrated that CatBoost was specifically designed as a gradient boosting framework incorporating dedicated mechanisms for handling categorical variables [35].
- **HistGradientBoosting Regressor with Poisson loss**—A model included due to the nature of the target variable, namely the number of contracts. Since the dependent variable represents count data, a model employing a Poisson loss function provided an additional benchmark relative to conventional regression-based approaches.

The use of multiple models enabled an assessment of the robustness of the results and reduced the risk of drawing conclusions from the performance of a single algorithm. All models were compared using an identical set of input variables and the same time-based validation framework.

Model performance was evaluated using standard forecasting accuracy metrics, defined in Equations (5)–(9).

MAE (Mean Absolute Error)

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y_i - \hat{Y}_i| \quad (5)$$

The MAE metric represents the average absolute prediction error expressed in terms of the number of contracts. It provides a direct measure of the typical deviation between predicted and actual values.

RMSE (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \quad (6)$$

The RMSE metric penalizes large prediction errors more heavily than MAE and is therefore particularly useful for identifying models that occasionally produce substantial deviations from the observed values.

MAPE (Mean Absolute Percentage Error)

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \quad (7)$$

The MAPE metric expresses prediction error in percentage terms, making it particularly useful from a business perspective. It facilitates the interpretation of model accuracy relative to the magnitude of the observed values and enables comparisons across regions characterized by different sales volumes.

WAPE (Weighted Absolute Percentage Error) was used as the primary metric for evaluating predictive performance because it is more stable than the classical MAPE [36] in the presence of observations with a low or zero number of contracts. WAPE relates the sum of absolute errors to the total actual number of contracts, thereby enabling the interpretation of prediction error relative to the total sales volume. From a business perspective, this metric is particularly useful because it indicates what proportion of total sales was incorrectly estimated by the model.

$$WAPE = \frac{\sum_{i=1}^n |Y_i - \hat{Y}_i|}{\sum_{i=1}^n |Y_i|} \cdot 100\% \quad (8)$$

R² (Coefficient of Determination)

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (9)$$

The R² coefficient was used as a complementary measure indicating the proportion of variability in the number of contracts explained by the model. Chicco, Warrens, and Jurman discuss R² as a useful metric for regression evaluation [37]; however, in the present study, this metric was treated as supplementary.

Particular emphasis was placed on the MAE and WAPE metrics during model evaluation. MAE indicates the average prediction error expressed in the number of contracts, which facilitates the operational interpretation of results. WAPE, in turn, enables the assessment of error relative to the total sales volume and is less sensitive than MAPE to observations with very low numbers of contracts. RMSE and R² were used as complementary metrics, whereas MAPE was interpreted with caution due to the possible occurrence of observations with low actual contract counts.

Due to the use of nonlinear ensemble models based on decision trees, explainable artificial intelligence methods were applied to enhance the interpretability of the results [38,39]. Lundberg and Lee introduced SHAP as a unified framework for interpreting model predictions, while Lundberg's subsequent work focuses on TreeSHAP for tree-based models. In particular, SHAP values were employed to estimate the contribution of individual variables to model predictions for each region-month observation.

In this study, SHAP analysis was used as a global model-interpretability tool. Mean absolute SHAP values were calculated to compare the average magnitude of the contribution of individual variables to model predictions across the analyzed observations. The values were also compared between the FULL and REDUCED specifications to assess whether the relative global importance of the main predictors remained similar after modifying the feature set. Because no SHAP dependence plots, directional summary plots, or local explanations were included, the analysis does not determine whether higher or lower values of a variable increased or decreased individual predictions and does not explain predictions for specific regions.

The interpretation of SHAP values was conducted with caution, taking into account the regional nature of the data. The mean absolute SHAP values describe the magnitude of the predictive contribution of variables but do not indicate the direction of their relationship with the predicted number of contracts and should not be interpreted as causal effects. This is particularly important for structural variables such as population, the number of single-family houses, population density, and previous PV market penetration, which may be strongly correlated. Therefore, the SHAP results were interpreted as global measures of predictive contribution rather than as evidence of directional or causal relationships. This is particularly important for structural variables such as population size, the number of single-family houses, population density, and previous PV market penetration, which may be strongly correlated with one another. Therefore, SHAP results were interpreted as a tool for model explanation and decision support rather than as a basis for direct causal inference [40].

Due to the temporal nature of the dataset, a random split into training and testing subsets was not applied because of the potential risk of information leakage between past and future observations. Instead, a chronological split was employed. The data were ordered according to the *time_index* variable, representing consecutive months of observation from January 2023 to December 2025. In the primary validation procedure, earlier months were used for model training, subsequent months for validation, and the final observation period for testing predictive performance. The data partitioning scheme adopted in this study was as follows: Train: $t = 1, \dots, 24$; Validation: $t = 25, \dots, 30$ and Test: $t = 31, \dots, 36$.

The adopted validation strategy reflects realistic forecasting conditions, where models are trained on historical data and applied to future periods. The test set covers the most recent six months, providing an out-of-sample evaluation under current market dynamics. Consequently, the regional UPI ranking derived from this test set should be interpreted as conditional on the selected six-month evaluation window rather than as a time-invariant ranking of regions.

All analyses were conducted in Python (Python Software Foundation, Wilmington, DE, USA) using pandas, scikit-learn (Scikit-learn Developers), XGBoost (XGBoost Developers), CatBoost (Yandex, Moscow, Russia), and SHAP (SHAP Developers).

2.4. Procedure for Constructing the Untapped Potential Index (UPI)

In this section, the UPI is operationalized as a company-specific regional sales-gap indicator. The model-predicted number of contracts is derived from regional energy-related,

housing, demographic, economic, market-maturity, and regulatory variables and is subsequently compared with the actual number of contracts achieved by the analyzed company. Although the current implementation is company-specific, the same operational procedure may be adapted to broader regional market data following appropriate recalibration and external validation. After selecting the model with the highest predictive performance, the expected number of contracts was calculated for each region-month observation. The predicted values generated by the model were then compared with the actual number of contracts. The difference between these values formed the basis for determining both the absolute sales gap and the relative indicator of untapped sales opportunities.

The procedure for constructing the indicator consisted of the following stages: aggregation of contractual data to the region-month level, training of regression models on historical data, selection of the model with the highest predictive performance, calculation of $\left(\hat{Y}_{r,t}\right)$ for the test observations, determination of the difference between predicted and actual values, normalization of this difference relative to the expected value, and aggregation of the results to the regional level.

For the aggregation of monthly results to the regional level, a cumulative sales gap measure was used, as expressed in Equation (10):

$$UP_r = \sum_{t=1}^T \left(\hat{Y}_{r,t} - Y_{r,t} \right) \quad (10)$$

where UP_r denotes the cumulative sales gap in region (r), while (T) represents the number of months included in the aggregation period.

To enable comparisons across regions characterized by different market sizes, a relative indicator of untapped sales opportunities was applied, as expressed in Equation (11).

$$UPI_r = \frac{\sum_{t=1}^T \left(\hat{Y}_{r,t} - Y_{r,t} \right)}{\sum_{t=1}^T \hat{Y}_{r,t}} \quad (11)$$

The use of the aggregated form of the indicator enables comparisons between voivodeships characterized by different sales volumes and reduces the influence of individual monthly prediction deviations on the final ranking of regions.

The UPI indicator requires the comparison of two quantities: the actual number of contracts and the number of contracts expected based on regional characteristics. Consequently, before calculating the indicator, it was necessary to select a model capable of estimating the expected level of sales as accurately and consistently as possible. For this reason, several predictive models were compared in the first stage of the analysis, and the model with the highest predictive performance was subsequently used to construct the UPI indicator and develop the regional ranking.

The UPI indicator was interpreted as a predictive decision-support measure rather than as an indicator of technical photovoltaic potential or a causal effect. Its purpose was to identify regions exhibiting the largest negative difference between actual sales and the level expected based on regional characteristics and market dynamics. Based on the values of (UPI_r) , a ranking of regions according to the level of untapped sales opportunities was developed.

The interpretation of the UPI indicator is predictive and decision-oriented rather than causal. It should not be interpreted as a direct measure of technical photovoltaic potential or as evidence of causal determinants of PV adoption. Accordingly, the UPI should be interpreted as a sales-side decision-support indicator and not as a comprehensive

measure of technical PV potential, electricity generation potential, or grid accommodation capacity. Consequently, the model should not be interpreted as a causal market-saturation model, even though lagged market maturity variables partially account for previous PV penetration. A high positive value of the indicator does not automatically imply the highest technical photovoltaic potential of a region or a strategic error on the part of the company. Instead, it indicates that, given the regional characteristics and market dynamics, the model predicted a higher number of contracts than was actually achieved. From the perspective of an installation company, regions with positive UPI values may represent areas requiring more detailed sales analysis, intensified commercial activities, or better-targeted marketing campaigns.

For the aggregation of monthly results to the regional level, the average value of the UPI indicator and the cumulative sales gap over the analyzed period were used.

During the preliminary drafting stage, GenAI-assisted visualization was used only to prepare an initial visual concept of the maps. The final maps included in the manuscript were prepared using a reproducible geospatial workflow. Regional UPI values were joined with polygon boundaries of Polish voivodeships using regional codes. The maps were generated using standard geospatial and plotting libraries, with fixed legends, labelled units, and high-resolution export settings. GenAI was not used as a source of data or as an analytical tool.

3. Results

The use of multiple machine learning algorithms reduced the risk of arbitrarily selecting a single method and made it possible to assess whether the model chosen as the primary model actually outperformed both the simple baseline model and the alternative predictive models. The selection of the primary model was based primarily on the WAPE metric, as it is more stable than the classical MAPE in the presence of observations with a low actual number of contracts. At the same time, it enables the interpretation of prediction error relative to the total sales volume.

The comparison of models (Table 2) on the test dataset showed that CatBoost achieved the best predictive performance.

Table 2. Comparison of the Predictive Performance of Models on the Validation and Test Datasets.

Model	Dataset	MAE	RMSE	MAPE Nonzero%	WAPE%	R ²
Regional mean baseline	Validation	19.21	23.11	179.80	141.97	−5.34
	Test	16.28	20.76	134.84	94.49	−2.53
Random Forest	Validation	3.70	5.04	40.60	27.31	0.70
	Test	6.88	9.63	46.44	39.95	0.24
XGBoost	Validation	4.34	5.48	44.03	32.08	0.64
	Test	6.97	9.04	55.21	40.48	0.33
HistGradientBoosting Poisson	Validation	4.47	6.55	42.43	33.01	0.49
	Test	7.29	10.08	45.07	42.31	0.17
CatBoost	Validation	3.65	4.66	37.64	26.96	0.74
	Test	6.05	8.73	42.66	35.11	0.38

Table 2 presents the predictive performance of all evaluated models on the validation and test datasets. Lower values of MAE, RMSE, WAPE, and MAPE for non-zero observations indicate better predictive accuracy, whereas a higher R² value indicates a larger proportion of explained variance. WAPE was treated as the primary evaluation metric

because it is more stable than classical MAPE when the dataset contains observations with low or zero numbers of contracts. The results show that CatBoost achieved the most favorable combination of error metrics, particularly the lowest WAPE and MAE on the test set. Therefore, CatBoost was selected as the primary model for constructing the UPI.

This model achieved the lowest values of MAE, WAPE, and MAPE for non-zero observations. The WAPE value was 35.11%, compared to 39.95% for Random Forest, 40.48% for XGBoost, and 42.31% for HistGradientBoosting Poisson. The baseline model based on the regional mean achieved a WAPE of 94.49%, indicating that it was unable to accurately capture sales dynamics during the test period.

CatBoost also achieved the lowest MAE value, equal to 6.05 contracts per region-month observation. This indicates that its average absolute prediction error was lower than that of the other analyzed models.

Further calculations of the UPI indicator were performed using the CatBoost model, as it achieved the best overall predictive performance. The selection of a single primary model was necessary to avoid generating multiple potentially inconsistent rankings of regional potential. In this study, the remaining models were used for comparison and validation. Their results confirmed that the selection of CatBoost was not arbitrary but resulted from an empirical comparison of predictive performance.

After selecting CatBoost as the primary model, an additional robustness test was conducted using a reduced set of features. The purpose of this stage was not to identify a new best-performing model but to verify whether the conclusions obtained from the full model remained stable after reducing the multicollinearity of regional variables. Based on the observed VIF values for the CatBoost model, presented in Table 3, the models were recalculated in two variants (specifications): the FULL model (unchanged specification) and the REDUCED model, with the variables *single_family_houses*, *area_km²*, and *installations_per_1000_houses_lag1* removed. The removal of these variables was justified by their high multicollinearity with other predictors and their partial informational redundancy relative to *installed_capacity_kw_per_capita_lag1*.

Table 3. Variance Inflation Factor (VIF) Values for the Explanatory Variables.

Feature	VIF
population	803.889
single_family_houses	764.800
population_density_per_km ²	76.379
area_km ²	36.690
installed_capacity_kw_per_capita_lag1	29.249
installations_per_1000_houses_lag1	28.689
single_family_per_1000_population	22.418
gdp_per_capita_pln	13.583
urbanization_rate	7.465
ghi_kwh_m ² _year	3.280
time_index	1.074
month_sin	1.069
month_cos	1.005

Table 3 reports the VIF values used to diagnose multicollinearity among the explanatory variables. The highest VIF values were observed for population and the number of single-family houses, which reflect the structural relationship between demographic scale and the size of the housing stock. High VIF values were also observed for population density, area, and previous PV market penetration indicators. This confirms that several regional characteristics are strongly interrelated. For this reason, the VIF analysis was used

not to exclude the FULL model, but to motivate an additional robustness check based on a REDUCED specification.

The high multicollinearity of these variables results from the nature of regional data. Variables describing demographic scale, the number of single-family houses, population density, regional area, and previous PV market penetration are naturally interrelated. In the case of tree-based models, multicollinearity does not invalidate the predictive procedure to the same extent as in linear models, because these models do not directly estimate linear regression coefficients. However, it may affect the interpretation of feature importance, including SHAP values.

Table 4 compares model performance under the FULL and REDUCED feature specifications. The REDUCED specification was introduced to verify whether the removal of highly collinear variables substantially changes predictive performance. The results show that the CatBoost REDUCED model achieved performance close to the CatBoost FULL model, which indicates that the predictive conclusions are relatively robust to changes in the variable set. Nevertheless, CatBoost FULL retained the best overall performance in terms of the main business-oriented error metrics, particularly WAPE and MAE, and was therefore used as the primary model for the final UPI calculation.

Table 4. Comparison of Models for Different Sets of Variables.

Feature set	Model	MAE	RMSE	MAPE Nonzero%	WAPE%	R ²
FULL	CatBoost	6.05	8.73	42.66	35.11	0.38
REDUCED	CatBoost	6.28	8.96	44.76	36.43	0.34
REDUCED	XGBoost	6.58	8.70	53.16	38.20	0.38
REDUCED	Random Forest	6.83	9.55	46.35	39.63	0.25
FULL	Random Forest	6.88	9.63	46.44	39.95	0.24
FULL	XGBoost	6.97	9.04	55.21	40.48	0.33
FULL	HGB Poisson	7.29	10.08	45.07	42.31	0.17
REDUCED	HGB Poisson	7.45	10.20	46.15	43.26	0.15
FULL/REDUCED	Regional mean baseline	16.28	20.76	134.84	94.49	−2.53

Although the XGBoost model in the REDUCED specification achieved a slightly higher R² value, the CatBoost FULL model obtained the best values of MAE, WAPE, and MAPE for non-zero observations. Since WAPE and MAE have a direct business interpretation in the context of the number of contracts, CatBoost FULL was selected as the primary model.

3.1. Selection of the Untapped Potential Index (UPI)

After selecting the primary model, the regional Untapped Potential Index (UPI) was calculated. This indicator compares the number of contracts predicted by the model with the actual number of contracts in a given voivodeship.

Negative UPI values should not be interpreted as indicating a lack of regional photovoltaic potential. They indicate only that the analyzed company achieved more contracts than predicted by the model during the analyzed period. Table 5 distinguishes regions with positive company-specific sales gaps, regions where sales were close to the model-predicted level, and regions where actual sales exceeded that level.

Table 5 presents the regional UPI values calculated using the CatBoost FULL model. The table compares the actual number of contracts achieved by the analyzed company with the number of contracts predicted by the model. The absolute sales gap indicates the numerical difference between predicted and actual contracts, while the UPI expresses this difference relative to the predicted value. Positive UPI values indicate regions where the company achieved fewer contracts than expected by the model, whereas negative values

indicate regions where actual sales exceeded model-based expectations. Therefore, the classification in Table 5 should be interpreted as a company-specific sales performance assessment rather than as a direct measure of total regional PV market potential.

Table 5. UPI Values and Classification of Company-Specific Regional Sales Gaps.

Region Code *	Actual Contracts	Predicted Contracts	UP Absolute	UPI	Interpretation
ZPM	42	65.97	23.97	0.36	large positive sales gap
LOD	71	104.41	33.41	0.32	large positive sales gap
POM	173	169.06	−3.94	−0.02	sales close to predicted level
OPO	86	83.87	−2.13	−0.03	sales close to predicted level
KUJ	81	75.06	−5.94	−0.08	sales close to predicted level
DOL	184	154.98	−29.02	−0.19	sales above predicted level
LUS	44	34.29	−9.71	−0.28	sales above predicted level
WKP	176	136.41	−39.59	−0.29	sales above predicted level
SLA	149	109.16	−39.84	−0.36	sales above predicted level
SWI	43	31.22	−11.78	−0.38	sales above predicted level
WMA	85	59.63	−25.37	−0.43	sales above predicted level
MAL	189	104.46	−84.54	−0.81	sales above predicted level
MAZ	167	90.75	−76.25	−0.84	sales above predicted level
LUB	57	30.26	−26.74	−0.88	sales above predicted level
PDL	35	17.25	−17.75	−1.03	sales above predicted level
PDK	72	34.67	−37.33	−1.08	sales above predicted level

* Note: Region codes are defined in Table 1.

The interpretation presented in Table 5 refers to the sales performance of the analyzed company and should not be understood as a direct assessment of total regional photovoltaic market potential. Therefore, high positive UPI values should be treated as signals for further managerial diagnosis rather than as definitive evidence of unmet regional photovoltaic demand.

The highest positive UPI values in the CatBoost FULL model were obtained for the ZPM and LOD regions. In the ZPM region, the actual number of contracts during the test period was 42, whereas the model predicted 65.97 contracts, corresponding to a UPI value of 0.363. In the LOD region, the actual number of contracts was 71, while the model predicted 104.41 contracts, resulting in a UPI value of 0.320.

This result indicates that ZPM and LOD had the largest positive company-specific regional sales gaps during the analyzed period. From the perspective of the analyzed company, these regions may be treated as exploratory priority areas for further diagnosis of sales barriers, marketing activities, and local sales coverage within the analyzed test window, rather than as definitive long-term targets for resource allocation.

The POM, OPO, and KUJ regions were characterized by UPI values close to zero. This indicates that actual sales in these regions were close to the levels expected by the model. In contrast, regions such as PDK, PDL, LUB, MAZ, and MAL achieved negative UPI values, indicating that actual sales exceeded the levels predicted on the basis of the regional characteristics included in the model. This result may reflect a stronger market position of the company and more effective commercial activity.

The robustness test indicated that the regional UPI ranking was relatively stable across the FULL and REDUCED model specifications within the same six-month test window. This analysis evaluates sensitivity to changes in the feature set but does not establish temporal stability across different market periods. Validation across multiple rolling-origin or expanding-window test periods would be required to assess whether the ranking remains stable over time. The ZPM and LOD regions remained the two highest-ranked

regions in both the full and reduced CatBoost specifications. In the FULL model, the UPI values were 0.363 for ZPM and 0.320 for LOD. In the REDUCED model, these values were 0.402 and 0.328, respectively.

The stability of the highest-ranking positions indicates that the identification of ZPM and LOD as the regions with the largest positive company-specific sales gaps was not an artifact of multicollinearity among regional variables. Similarly, the POM, OPO, and KUJ regions remained close to their expected levels in both specifications, although their ranking positions changed slightly.

Given that two CatBoost model specifications were considered, changes in the UPI indicator were also analyzed with respect to the selected model variant. The results of the UPI stability analysis are presented in Table 6. Comparing the values obtained for the full and reduced sets of variables makes it possible to assess whether the identification of regions with positive company-specific sales gaps remains stable across alternative model specifications. Particular attention was paid to differences in UPI values and changes in the ranking positions of regions.

It should be emphasized that the presented UPI values describe the relative regional sales performance of the company analyzed after accounting for regional structural characteristics. They do not represent the overall photovoltaic potential, technical solar potential, or total market potential of individual regions. The regional distribution of the company-specific UPI_FULL values is presented in Figure 3, while the corresponding numerical values are reported in Table 6.

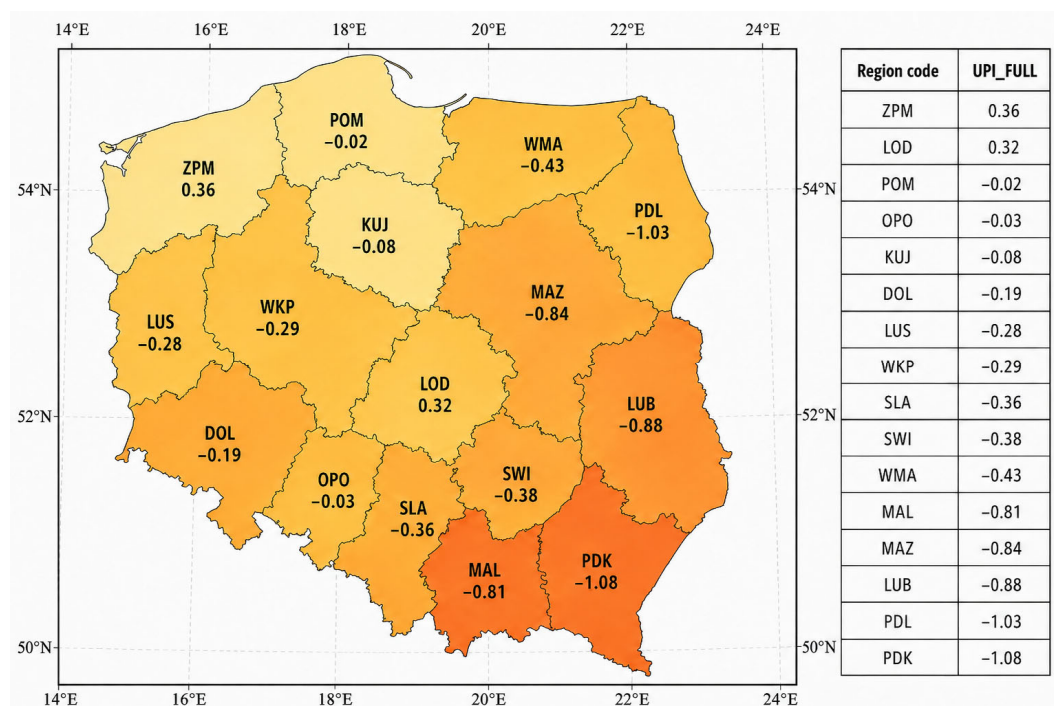


Figure 3. Regional distribution of UPI_FULL company-specific coefficient values across Polish voivodeships. Positive UPI values indicate regions where the analyzed company achieved fewer PV installation contracts than predicted by the machine learning model based on regional characteristics. Negative values indicate regions where the company exceeded the model-predicted sales level. The map represents the analyzed company's relative sales performance and should not be interpreted as a map of the general regional photovoltaic potential or technical solar potential. The ranking of Polish voivodeships according to company-specific UPI_FULL values is presented in Figure 4. The citation to Figure 4 has been added to the text.

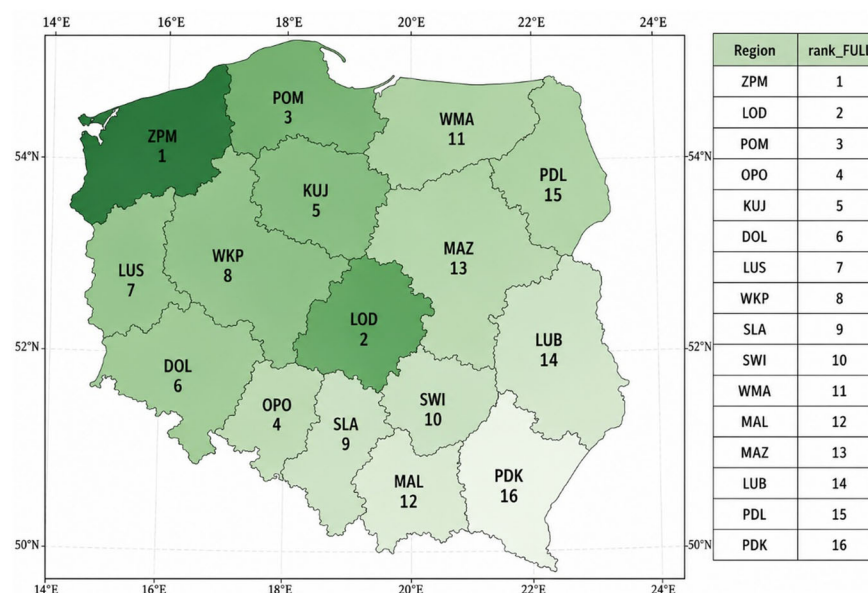


Figure 4. The ranking of regions according to the UPI values. Ranking of Polish voivodeships according to company-specific UPI_FULL values. Higher positive values indicate larger gaps between the analyzed company's actual and model-predicted numbers of PV installation contracts. Negative values indicate regions where the company exceeded the model-predicted sales level. The ranking reflects the analyzed company's regional sales performance and should not be interpreted as a ranking of overall regional photovoltaic market potential.

Table 6. Stability of the UPI Indicator with Respect to the Variable Set Specification.

Region Code *	UPI FULL	UPI REDUCED	Rank FULL	Rank REDUCED	Rank Difference	UPI Difference
ZPM	0.36	0.40	1	1	0	−0.04
LOD	0.32	0.33	2	2	0	−0.01
POM	−0.02	−0.03	3	4	−1	0.01
OPO	−0.03	0.01	4	3	1	−0.03
KUJ	−0.08	−0.06	5	5	0	−0.02
DOL	−0.19	−0.24	6	6	0	0.05
LUS	−0.28	−0.36	7	9	−2	0.07
WKP	−0.29	−0.31	8	7	1	0.02
SLA	−0.36	−0.33	9	8	1	−0.03
SWI	−0.38	−0.41	10	10	0	0.03
WMA	−0.43	−0.55	11	11	0	0.13
MAL	−0.81	−0.92	12	14	−2	0.11
MAZ	−0.84	−0.85	13	12	1	0.01
LUB	−0.88	−0.86	14	13	1	−0.02
PDL	−1.03	−1.01	15	15	0	−0.02
PDK	−1.08	−1.10	16	16	0	0.02

* Note: Region codes are defined in Table 1.

3.2. Model Interpretability

Model interpretability was analyzed using SHAP values, which make it possible to determine the relative contribution of individual variables to model predictions. The aggregated results are presented in Table 7. Comparing SHAP values for the full and reduced sets of variables makes it possible to assess whether the importance of the most influential predictors remains stable across alternative model specifications.

Table 7. Global Variable Importance Based on Mean Absolute SHAP Values.

Feature	SHAP Full	SHAP Reduced
time_index	18.55	18.66
population	5.17	5.22
installed_capacity_kw_per_capita_lag1	3.08	4.49
population_density_per_km ²	1.97	2.25
installations_per_1000_houses_lag1	1.95	—
single_family_per_1000_population	1.17	1.55
month_cos	1.16	1.22
month_sin	1.02	0.92
gdp_per_capita_pln	0.94	1.28
single_family_houses	0.83	—
urbanization_rate	0.53	0.79
subsidy_period	0.36	0.46
ghi_kwh_m ² _year	0.30	0.31
area_km ²	0.18	—

SHAP values should be interpreted as predictive contributions of variables rather than as causal effects. This is particularly important in the case of regional data, where many variables are structurally related to one another. Multicollinearity may distribute predictive importance across correlated variables or shift part of the apparent contribution from one variable to another. For this reason, the SHAP results should not be interpreted as evidence that one regional characteristic independently determines sales outcomes. Rather, they indicate which groups of variables were most informative for the predictive model.

The values reported in Table 7 are mean absolute SHAP values. Therefore, they describe the average magnitude of each variable's contribution to model predictions, but they do not indicate the direction of the relationship between a variable and the predicted number of contracts. Accordingly, the results presented in Table 7 cannot be used to conclude that an increase in a given variable increased or decreased the predicted number of contracts.

In the FULL model, the highest mean absolute SHAP value was obtained for the *time_index* variable. This indicates that *time_index* had the largest average absolute contribution to model predictions among the variables included in the analysis. However, the mean absolute SHAP value does not show whether later observations were associated with higher or lower predicted contract volumes. The result may reflect predictive information related to temporal market dynamics, but its direction and the underlying mechanisms cannot be established without additional directional or local SHAP analyses.

In this study, SHAP was used primarily as a global interpretability tool to identify the variables that contributed most strongly to model predictions. A detailed directional analysis based on SHAP dependence plots or local SHAP explanations for individual regions was beyond the main scope of the present study, which focused on the construction and validation of a company-specific regional sales-gap indicator.

The second most important variable was population, indicating that the model captured the scale effect of the regional market. Regions with larger populations provide a larger pool of potential prosumer PV customers. Population had the second-highest mean absolute SHAP value in the FULL model, indicating a relatively large average contribution to model predictions. However, the reported value does not establish whether larger populations increased or decreased the predicted number of contracts. High mean absolute SHAP values were also observed for variables describing previous PV market maturity, particularly *installed_capacity_kw_per_capita_lag1* and *installations_per_1000_houses_lag1*. These

results identify variables that were informative for the model but do not establish directional or causal relationships.

In the REDUCED model, the structure of feature importance remained consistent. The most important variables were *time_index*, *population*, *installed_capacity_kw_per_capita_lag1*, *population_density_per_km²*, *single_family_per_1000_population*, and *gdp_per_capita_pln*. After removing *installations_per_1000_houses_lag1*, the mean absolute SHAP value of *installed_capacity_kw_per_capita_lag1* increased from 3.08 to 4.49. This change indicates a redistribution of predictive attribution after modifying the feature set, but it does not establish the direction of either variable's relationship with the predicted number of contracts.

The relatively low importance of *ghi_kwh_m²_year* does not imply that solar irradiation is unimportant for the technical performance of PV installations. Rather, this result suggests that, within the regional sales model for Poland, differences in annual solar irradiation between voivodeships were less important than variables describing temporal dynamics, market scale, previous PV adoption, and socio-demographic structure.

The relatively low importance of *subsidy_period* may appear surprising, particularly given the common narrative that subsidy programs effectively stimulate the market and contribute to the adoption of green energy. The *subsidy_period* variable obtained a relatively low but non-zero SHAP importance. In the full model, the mean absolute SHAP value for this variable was 0.356, while in the reduced model it increased to 0.458. This indicates that information about subsidy periods contributed to the model's predictive performance, although its importance was lower than might intuitively be expected.

The mean absolute SHAP value of *subsidy_period* indicates that this variable contributed less to the model predictions than the highest-ranked predictors. However, the analysis does not show whether subsidy periods increased or decreased predicted contract volumes. Because *subsidy_period* and *time_index* both vary over time, their predictive attributions may partially overlap, particularly in the presence of correlated temporal information. The current analysis does not quantify the extent of this overlap or identify a causal effect of subsidy programs.

It should also be noted that the authors expected a relatively low importance value for this variable because of the specific sales strategy employed by the company that provided the data. This strategy was based on concluding conditional contracts that were activated when a support program covering specific activities was introduced. This company-specific sales mechanism may have affected how subsidy-related information was represented in the analyzed data and attributed by the model. Therefore, the relatively low mean absolute SHAP value of *subsidy_period* should not be interpreted as evidence that subsidies are unimportant for the photovoltaic market.

4. Discussion

The results show that regional differences in photovoltaic development in Poland cannot be explained solely by natural conditions such as solar irradiation. Although solar resources remain an important determinant of PV installation efficiency, the analyses conducted indicate that regional differences in photovoltaic adoption are determined to a much greater extent by socio-economic, demographic, infrastructural, and market-related factors. The *time_index* variable had the highest mean absolute SHAP value, indicating that it made the largest average absolute contribution to the model predictions. This finding identifies temporal information as highly relevant to predictive performance but does not establish the direction of the relationship or demonstrate that temporal market dynamics causally determined monthly contract volume. This result is consistent with previous studies indicating that the diffusion of renewable energy technologies in the prosumer

sector is multidimensional and depends on interactions between technical potential and the socio-economic readiness of regions [24–27].

This also means that the economic viability of PV installations should be assessed in a broader framework that includes annual yield, self-consumption, support mechanisms, energy prices, investment costs, and system-integration constraints, rather than on the basis of sunshine duration alone.

At the same time, the significance of regional variables indicates that the temporal trend alone does not fully explain the prediction mechanism. Population and the lagged variables describing previous PV penetration also had relatively high mean absolute SHAP values. This indicates that these variables contained information used by the model when generating predictions. However, the present analysis does not determine whether higher values of these variables increased or decreased predicted sales, nor does it identify the specific mechanisms underlying their predictive contribution. Potential explanations involving market maturity, technology visibility, neighborhood influences, or previous prosumer activity would require additional directional analysis and independent validation.

One of the most important findings of this study is the identification of a difference between the number of contracts expected on the basis of regional characteristics and the company's historical data and the number of contracts actually achieved by the analyzed firm. The proposed UPI enabled the identification of voivodeships in which the company's sales remained below the level predicted by the model. This result should not be interpreted as a direct measure of total regional photovoltaic market potential.

The developed UPI enables a transition from prediction alone to practical decision support for companies selling PV installations and energy storage systems. This result indicates that ZPM and LOD were the regions with the highest company-specific sales gaps during the analyzed period. However, this finding should be interpreted with caution. Lower observed sales in these regions may reflect not only untapped regional sales opportunities, but also firm-specific operational factors, such as weaker local sales coverage, a smaller number of sales representatives, lower marketing intensity, stronger local competition, limited installation capacity, or other organizational constraints.

A particularly important conclusion of the analysis concerns the role of housing structure and suburbanization processes. Variables describing the single-family housing structure contributed to model predictions, although their mean absolute SHAP values were lower than those of *time_index*, population, and the main market-maturity variables. The present SHAP analysis does not establish whether a larger housing stock increased predicted contract volumes. Nevertheless, the inclusion of these variables is substantively justified by previous literature describing the relevance of rooftop availability and household investment autonomy to prosumer PV adoption. This result is consistent with findings reported in the European literature, according to which single-family housing significantly increases the likelihood of investment in renewable energy technologies. At the same time, highly urbanized regions characterized by a predominance of multi-family housing face significant organizational and infrastructural barriers that limit the development of PV micro-installations.

Another important finding is that favorable solar irradiation conditions alone did not correspond to higher contract volumes or positive company-specific sales gaps. In the predictive model, annual solar irradiation had lower mean absolute SHAP importance than temporal dynamics, population, and previous PV market maturity. This result describes the relative predictive contribution of the variables included in the current company-specific model and should not be interpreted as a causal relationship. A broader regional market assessment would require independent market and technical data. This indicates that natural conditions alone are insufficient to stimulate PV market development without an

adequate level of household prosperity, access to financing, development of electricity infrastructure, and maturity of the installation market. This finding supports the view that the energy transition is a socio-technical process rather than a purely technological one.

From the perspective of machine learning applications in energy economics, the study demonstrates that predictive models can support company-level sales decision-making in the photovoltaic installation market. The proposed approach enables the identification of regions in which the analyzed company's actual sales fall below the level predicted on the basis of regional characteristics and historical sales patterns. Therefore, the model should be understood as a decision-support tool rather than as a causal model of photovoltaic adoption or a direct measure of technical PV potential [41]. The proposed model enables the identification of regions in which the analyzed company's actual sales fall below the level predicted on the basis of regional characteristics and historical sales patterns. Accordingly, the regional ranking may provide an exploratory input for resource allocation, regional expansion planning, and marketing optimization within the analyzed company. However, these decisions should not be based on the ranking alone and should be reassessed as new sales and operational data become available.

The study has several limitations that should be considered when interpreting the results. First, the contractual data originate from a single installation company. Although the company operates nationwide and the dataset covers sales activity across all Polish voivodeships, the results should not be interpreted as fully representative of the entire PV installation market. In particular, the model does not include contract-level data from competing companies or direct information on regional market shares. Consequently, a positive UPI value may reflect both regional market characteristics and firm-specific factors, including local sales coverage, marketing intensity, available operational capacity, commercial strategy, and competitive position. Therefore, the indicator should not be interpreted as a direct and universal measure of total regional photovoltaic market potential. At the same time, the proposed modeling approach enables the adoption of this solution by any entity operating in this market, provided that historical sales data are available. The results therefore reflect both the characteristics of the regional market and the specific business practices of the analyzed company.

Second, some regional variables in the FULL specification exhibited high multicollinearity. This is an expected phenomenon in regional datasets, where demographic scale, building stock, population density, and previous market maturity are naturally interrelated. The impact of this limitation was partially mitigated through the use of the REDUCED specification as a robustness test. Nevertheless, SHAP values should be interpreted as measures of predictive contribution rather than causal effects.

Third, the test dataset covered the final six months of the analyzed period and the validation procedure was based on a single chronological split. Consequently, the reported regional ranking is conditional on this specific test window and should not be interpreted as a stable long-term prioritization of regions. At the current stage, it should be treated as an exploratory decision-support result requiring confirmation with more recent data and additional temporal validation. Future research should therefore extend the analysis through rolling-origin or expanding-window validation to assess the stability of model performance and regional rankings across different market subperiods. Another limitation concerns the representation of long-term historical PV market saturation. Although the model includes lagged indicators of PV market maturity, such as installed capacity per capita and installations per 1000 houses, it does not include a separate long historical panel of regional installed capacity prior to the study period. As a result, the model may not fully capture saturation effects resulting from the cumulative stock of PV installations.

This limitation should be considered when interpreting temporal dynamics and regional UPI values.

Fourth, the model does not include several operational variables that are central to the interpretation of a company-specific sales-gap indicator, such as the number of leads, the number of active sales representatives, marketing expenditure, local competition, the availability of installation teams, installation capacity, or customer acquisition costs. The absence of these variables means that high positive UPI values should not be interpreted solely as evidence of untapped market demand. They may also indicate weaker company presence, lower sales effort, stronger competition, or operational constraints in a given region.

A further limitation is that the analytical framework is primarily sales-side. Although the manuscript discusses grid constraints and curtailment as important contextual factors, these variables were not included directly in the predictive model. Consequently, the model does not capture regional differences in actual PV electricity generation, curtailment levels, grid-connection constraints, or the capacity of the power system to accommodate additional PV generation. This limitation should be considered when interpreting the UPI as an indicator of sales opportunities rather than technical energy-system potential.

The study also highlights the growing importance of infrastructural constraints and power system flexibility in the context of further photovoltaic market development. The increasingly frequent occurrence of curtailment in Poland, resulting from congestion in transmission and distribution networks, may reduce the actual utilization of solar energy potential. These infrastructural constraints form part of the broader context of photovoltaic market development but are not measured by the current UPI, because regional curtailment and grid-capacity variables were not included in the predictive model. Accordingly, the present UPI should be interpreted as a company-specific regional sales-gap indicator. Nevertheless, these infrastructural variables could be incorporated into a future market-wide extension of the framework.

The article proposes a regional decision-support model for the photovoltaic installation market in Poland, based on real contractual data from an installation company and regional socio-economic, demographic, energy-related, market, and regulatory variables. The uniqueness of the study also results from the use of actual contractual data from an installation company. Such data are rarely available for scientific research purposes because companies treat information on sales, customers, contracts, and the regional effectiveness of commercial activities as strategic assets. Therefore, the possibility of developing a model based on operational data rather than solely on public or aggregated statistics constitutes an important empirical and practical contribution of this study.

The CatBoost model in the FULL specification achieved the best predictive performance, obtaining a WAPE of 35.11% and an R^2 value of 0.377 on the test dataset. This model significantly outperformed the baseline model based on the regional mean, for which WAPE reached 94.49% and R^2 assumed a negative value. This indicates that machine learning models were better able to capture the sales dynamics of PV installations than a simple extrapolation of historical regional averages.

The robustness test using the REDUCED specification confirmed the stability of the results. Despite the removal of highly collinear variables, predictive performance remained similar, and the ranking of regions according to the UPI indicator remained highly stable. The ZPM and LOD regions remained the two highest-ranked areas with the largest positive company-specific sales gaps in both the FULL and REDUCED models.

The proposed UPI indicator enables the identification of regions in which actual sales were lower than the model-predicted level during the analyzed test period. The resulting ranking may serve as an exploratory decision-support input for the allocation of sales resources, the planning of marketing campaigns, and the identification of regions requiring

further managerial diagnosis. Because the ranking is conditional on the six-month test window, it should be updated and reassessed before being used for long-term operational decisions. It should be emphasized that, despite their high predictive performance, machine learning models do not enable direct causal inference. The identified relationships should therefore be interpreted primarily as predictive relationships rather than strict cause-and-effect relationships. Future research could usefully extend the analysis by incorporating spatial econometric methods, panel data models, or causal inference techniques, which could provide a better understanding of the mechanisms underlying regional differences in the photovoltaic market.

Future research directions should also include the integration of additional infrastructural and power-system variables, such as regional PV electricity generation, curtailment levels, grid-connection availability, power accommodation capacity, regional electricity prices, and energy consumption profiles. Before the framework can be used to formulate broader regional development or policy recommendations, future studies should validate the UPI against independent regional market data and include additional variables describing grid capacity, local competition, sales coverage, and operational constraints. Incorporating these factors could improve the explanatory power of the model and enable a more comprehensive assessment of barriers to PV market development. Extending the analysis to other energy transition technologies, such as heat pumps and energy storage systems, could also facilitate the development of integrated models of regional energy transition potential.

A further research direction should involve external validation of the UPI against independent data on the total number of photovoltaic installations or contracts in individual regions, as well as data on competitors' sales activity where available. Incorporating the analyzed company's regional market share as a calibration variable or external benchmark could help distinguish broader market conditions from firm-specific effects and support a broader regional interpretation of the indicator. Future research should also consider the application of rolling-origin or expanding-window validation approaches to assess the temporal stability of model predictions and the resulting UPI rankings. Such methods would provide a more comprehensive evaluation of the robustness of the results under changing market conditions.

Future research should also incorporate longer historical time series of regional PV installed capacity, annual additions, and explicit market saturation indicators. Such variables would make it possible to distinguish more clearly between regions with genuinely lower future demand due to prior market saturation and regions where sales remain below the level expected from current regional characteristics.

Future research should also extend the explainability analysis by including directional SHAP summary plots, SHAP dependence plots, and local explanations for selected regions, particularly those with the highest positive UPI values. Such analyses would make it possible to assess the direction and regional specificity of variable contributions, which cannot be determined from the mean absolute SHAP values reported in the present study. Such an extension would allow a more detailed assessment of the direction and regional specificity of variable effects, while also helping to distinguish global model behavior from local prediction mechanisms.

In conclusion, this study demonstrates that machine learning can support company-specific regional sales decision-making in the Polish photovoltaic installation market. Among the evaluated models, CatBoost achieved the best predictive performance and was used to construct the Untapped Potential Index (UPI). The UPI enabled the identification of regions where the actual number of contracts achieved by the analyzed company was lower than the level expected on the basis of regional characteristics and historical

sales patterns. The highest positive UPI values were observed for ZPM and LOD within the six-month test window, indicating company-specific regional sales gaps that may warrant further managerial diagnosis. The resulting ranking should be interpreted as exploratory and time-window dependent rather than as a definitive basis for long-term regional prioritization. Broader market-level conclusions require validation using independent data on regional markets, infrastructure, and the company's market share.

Author Contributions: Conceptualization, A.Z. and R.J.; methodology, A.Z. and R.J.; software, A.Z. and R.J.; validation, A.Z. and R.J.; formal analysis, A.Z. and R.J.; investigation, A.Z. and R.J.; resources, A.Z. and R.J.; data curation, A.Z. and R.J.; writing—original draft preparation, A.Z. and R.J.; writing—review and editing, A.Z. and R.J.; visualization, A.Z. and R.J.; supervision, A.Z. and R.J.; project administration, A.Z. and R.J.; funding acquisition, A.Z. and R.J. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded under subvention funds for the Faculty of Management and by the “Excellence Initiative—Research University” program for the AGH University of Krakow.

Data Availability Statement: The data presented in this study are only available upon request from the corresponding author due to restrictions resulting from the signed confidentiality agreements. The data are not publicly available due to the sensitive nature of the collected information, which is protected by the non-disclosure agreement signed between the parties.

Acknowledgments: During the preparation of this manuscript, generative AI (ChatGPT, GPT-5.5; OpenAI, San Francisco, CA, USA) was used only as an auxiliary tool during the preliminary visualization drafting stage. The final figures included in the manuscript were prepared using a reproducible workflow based on the authors' data and administrative boundary files. The authors reviewed and edited all visual materials and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Łaczmńska, M. Polish Photovoltaic Market 2025—Data, Trends and Challenges of the Energy Transition. Lighthief. 2025. Available online: <https://lighthief.com/rynek-fotowoltaiki-w-polsce-2025/> (accessed on 27 July 2025).
2. Rynek Fotowoltaiki w Polsce 2025—Raport. Inżynier Budownictwa. 2025. Available online: <https://inzynierbudownictwa.pl/rynek-fotowoltaiki-w-polsce-2025-raport/> (accessed on 27 July 2025).
3. Wiśniewski, G.; Kowalak, T.; Kalinowski, K.; Łoziński, J.; Orzechowski, S.; Wiśniewski, K.; Twardawa, M.; Urbański, P.; Wycisk, A. *Photovoltaic Market in Poland 2025: Photovoltaics on the Electricity and Heat Market*; Instytut Energetyki Odnawialnej (IEO): Warsaw, Poland, 2025; Available online: <https://ieo.pl/raport-rynek-fotowoltaiki-w-polsce-2025> (accessed on 15 May 2026).
4. Smart-Grids.pl. Raport ‘Rynek Fotowoltaiki w Polsce 2025’. Podsumowanie. 2 July 2025. Available online: <https://smart-grids.pl/> (accessed on 26 May 2026).
5. Globenergia. “Fotowoltaika w Polsce—Głos Branży i Prosumentów 2025. Pobierz Raport.” Globenergia. 2025. Available online: <https://globenergia.pl/Globenergia> (accessed on 26 May 2026).
6. Inżynieria PV. Rozwój Rynku Fotowoltaiki w Polsce i na Świecie—Analiza Trendów, Prognoz i Wyzwań do 2025 Roku. Inżynieria PV. 2025. Available online: <https://inzynieriapv.pl/blog/rozwoj-ryнку-fotowoltaiki-w-polsce-i-na-swiecie-analiza-trendow-prognoz-i-wyzwan-do-2025-roku/> (accessed on 26 May 2026).
7. ESG.pl. Nadpodaż—Przyczyna Nowych Wyzwań dla Rynku Fotowoltaiki. ESG.pl. 2025. Available online: <https://esg.pl/artukul/nadpodaz-przyczyna-n2256535> (accessed on 26 May 2026).
8. Rynek Elektryczny. Moc zainstalowana OZE w Polsce. Rynek Elektryczny. 2025. Available online: <https://www.rynekelektryczny.pl/moc-zainstalowana-oze-w-polsce/> (accessed on 26 May 2026).
9. Czyżak, P.; Sikorski, M.; Wrona, A. *Co po węglu? Potencjał OZE w Polsce*; Instytut Policy Paper 06/2021; Fundacja Instytut: Warszawa, Poland, 2021. Available online: <https://instytut.pl/potencjal-oze/> (accessed on 26 May 2026).
10. Gram w Zielone. Rynek Prosumencki w Polsce Wchodzi w Fazę Testu. Gram w Zielone. 2025. Available online: <https://www.gramwzielone.pl/energia-sloneczna/20349615/rynek-prosumencki-w-polsce-wchodzi-w-faze-testu> (accessed on 26 May 2026).
11. Siuta-Tokarska, B.; Duda, J.; Thier, A.; Basiura, B. Energy Use and Sustainability Performance in Poland: An Indicator-Based Assessment in the EU Context. *Energies* **2026**, *19*, 2191. [CrossRef]

12. Globenergia. Duże Zmiany w OZE, a na ich Czele Magazyny Energii. Rok 2025 Należał do Nich. Globenergia. 2025. Available online: <https://globenergia.pl/duze-zmiany-w-oze-a-na-ich-czele-magazyny-energii-rok-2025-nalezal-do-nich/> (accessed on 26 May 2026).
13. ElektroSieć. Mój Prąd i Magazyn Energii—Kompleksowy Przewodnik po Dofinansowaniach 2025. ElektroSieć. 2025. Available online: <https://elektrosiec.pl/blog/moj-prad-i-magazyn-energii-kompleksowy-przewodnik-po-dofinansowaniach-2025/> (accessed on 26 May 2026).
14. Narodowy Fundusz Ochrony Środowiska i Gospodarki Wodnej (NFOŚiGW). *Program Priorytetowy „Mój Prąd”*. Część 3: *Program Mój Prąd na lata 2024–2027*; NFOŚiGW: Warszawa, Poland, 2024.
15. Ministerstwo Klimatu i Środowiska. “Skorzystaj z programu ‘Mój Prąd’”. 2025. Available online: <https://www.gov.pl/web/gov/skorzystaj-z-programu-moj-prad> (accessed on 15 May 2026).
16. European Commission; Joint Research Centre (JRC). *PVGIS Photovoltaic Geographical Information System*; European Commission: Ispra, Italy. Available online: https://re.jrc.ec.europa.eu/pvg_tools/en/ (accessed on 26 May 2026).
17. Instytut Meteorologii i Gospodarki Wodnej (IMGW). IMGW-PIB—Instytut Meteorologii i Gospodarki Wodnej. Available online: <https://imgw.pl/> (accessed on 26 May 2026).
18. Rudniak, J. Analiza regionalnego potencjału energii promieniowania słonecznego. *Inż. Ochr. Śr.* **2017**, *20*, 371–386.
19. Mirowski, T.; Sornek, K. Potencjał energetyki prosumenckiej w Polsce na przykładzie mikroinstalacji fotowoltaicznych w budownictwie indywidualnym. *Energ.—Społecz.—Polityka* **2015**, *2*, 73–84.
20. Angliviell de La Beaumelle, N.; Blok, K.; de Chalendar, J.A.; Clarke, L.; Hahmann, A.N.; Huster, J.; Nemet, G.F.; Suri, D.; Wild, T.B.; Azevedo, I.M.L. The Global Technical, Economic, and Feasible Potential of Renewable Electricity. *Annu. Rev. Environ. Resour.* **2023**, *48*, 419–449. [\[CrossRef\]](#)
21. Ali, A.O.; Elgohr, A.T.; El-Mahdy, M.H.; Zohir, H.M.; Emam, A.Z.; Mostafa, M.G.; Al-Razgan, M.; Kasem, H.M.; Elhadidy, M.S. Advancements in Photovoltaic Technology: A Comprehensive Review of Recent Advances and Future Prospects. *Energy Convers. Manag.* **2025**, *26*, 100952. [\[CrossRef\]](#)
22. Statistics Poland (GUS). Bank Danych Lokalnych (Local Data Bank)—Budownictwo Mieszkaniowe—Wskaźniki. Available online: <https://bdl.stat.gov.pl/bdl/start> (accessed on 26 May 2026).
23. Mazur, M.; Ingaldi, M. Strategies for Solar Energy Utilization in Businesses: A Business Model Canvas Approach. *Energies* **2025**, *18*, 6533. [\[CrossRef\]](#)
24. Graziano, M.; Gillingham, K. Spatial Patterns of Solar Photovoltaic System Adoption: The Influence of Neighbors and the Built Environment. *J. Econ. Geogr.* **2015**, *15*, 815–839. [\[CrossRef\]](#)
25. Zhang, J.; Ballas, D.; Liu, X. Neighbourhood-Level Spatial Determinants of Residential Solar Photovoltaic Adoption in the Netherlands. *Renew. Energy* **2023**, *206*, 1239–1248. [\[CrossRef\]](#)
26. Dong, C.; Zhong, Q. Evaluating Solar Photovoltaic Potential of Buildings Based on Installation Parameters and Economic Feasibility. *Sol. Energy* **2025**, *288*, 113304. [\[CrossRef\]](#)
27. Scheller, F.; Graupner, S.; Edwards, J.; Weinand, J.; Bruckner, T. Active Peer Effects in Residential Photovoltaic Adoption: Evidence on Impact Drivers among Potential and Current Adopters in Germany. *Energy Res. Soc. Sci.* **2021**, *78*, 102125. [\[CrossRef\]](#)
28. Staszewska, J.; Knop, L. Identifying Resilience Factors of Power Company Business Models. *Energies* **2025**, *18*, 992. [\[CrossRef\]](#)
29. Przybylski, W.; Kaczmarek, P.; Kłosowski, G. Practical Aspects of Implementing an Integrated System for the Sale, Installation and Monitoring of Photovoltaic Systems. *Procedia Comput. Sci.* **2024**, *246*, 656–665. [\[CrossRef\]](#)
30. Sułek, A.; Borowski, P.F. Business Models on the Energy Market in the Era of a Low-Emission Economy. *Energies* **2024**, *17*, 3235. [\[CrossRef\]](#)
31. Tashman, L.J. Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review. *Int. J. Forecast.* **2000**, *16*, 437–450. [\[CrossRef\]](#)
32. Hyndman, R.J.; Koehler, A.B. Another Look at Measures of Forecast Accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [\[CrossRef\]](#)
33. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
34. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [\[CrossRef\]](#)
35. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, QC, Canada, 3–8 December 2018; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 6639–6649. Available online: <https://dl.acm.org/doi/10.5555/3327757.3327770> (accessed on 10 June 2026).
36. Gneiting, T. Making and Evaluating Point Forecasts. *J. Am. Stat. Assoc.* **2011**, *106*, 746–762. [\[CrossRef\]](#)
37. Chicco, D.; Warrens, M.J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.

39. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [[CrossRef](#)] [[PubMed](#)]
40. Chou, Y.-L.; Moreira, C.; Bruza, P.; Ouyang, C.; Jorge, J. Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications. *Inf. Fusion* **2022**, *81*, 59–83. [[CrossRef](#)]
41. Zielińska, A.; Jankowski, R. Forecasting Installation Demand Using Machine Learning: Evidence from a Large PV Installer in Poland. *Energies* **2025**, *18*, 4998. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.