applied sciences

MDPI

Article

The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities

Fátima Trindade Neves *, Manuela Aparicio and Miguel de Castro Neto

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa (UNL), Campus de Campolide, 1070-312 Lisboa, Portugal; manuela.aparicio@novaims.unl.pt (M.A.); mneto@novaims.unl.pt (M.d.C.N.)

* Correspondence: fneves@novaims.unl.pt

Abstract: In the rapidly evolving landscape of urban development, where smart cities increasingly rely on artificial intelligence (AI) solutions to address complex challenges, using AI to accurately predict real estate prices becomes a multifaceted and crucial task integral to urban planning and economic development. This paper delves into this endeavor, highlighting the transformative impact of specifically chosen contextual open data and recent advances in eXplainable AI (XAI) to improve the accuracy and transparency of real estate price predictions within smart cities. Focusing on Lisbon's dynamic housing market from 2018 to 2021, we integrate diverse open data sources into an eXtreme Gradient Boosting (XGBoost) machine learning model optimized with the Optuna hyperparameter framework to enhance its predictive precision. Our initial model achieved a Mean Absolute Error (MAE) of EUR 51,733.88, which was significantly reduced by 8.24% upon incorporating open data features. This substantial improvement underscores open data's potential to boost real estate price predictions. Additionally, we employed SHapley Additive exPlanations (SHAP) to address the transparency of our model. This approach clarifies the influence of each predictor on price estimates and fosters enhanced accountability and trust in AI-driven real estate analytics. The findings of this study emphasize the role of XAI and the value of open data in enhancing the transparency and efficacy of AI-driven urban development, explicitly demonstrating how they contribute to more accurate and insightful real estate analytics, thereby informing and improving policy decisions for the sustainable development of smart cities.

Keywords: open data; smart cities; real estate predictions; urban development; artificial intelligence; machine learning; eXplainable AI; XGBoost; Optuna; shapley additive explanations (SHAP)

Citation: Trindade Neves, F.; Aparicio, M.; de Castro Neto, M. The Impacts of Open Data and eXplainable AI on Real Estate Price Predictions in Smart Cities. *Appl. Sci.* **2024**, *14*, 2209. https://doi.org/10.3390/app14052209

Academic Editor: Luis Javier Garcia Villalba

Received: 15 January 2024
Revised: 27 February 2024
Accepted: 3 March 2024
Published: 6 March 2024

1. Introduction

In an era marked by rapid urbanization, the emergence of smart cities represents a critical evolution in how urban spaces are conceived, designed, and managed. These cities are not just innovative paradigms but also data-driven ecosystems that significantly improve the quality of life of their inhabitants while promoting sustainability and resilience [1,2]. They are at the forefront of addressing some of the most pressing challenges of our times—from complex environmental challenges like environmental degradation, resource strain, and escalating demands on infrastructure to efficient urban governance and growing digitization, leading to more efficient, data-driven, and citizen-oriented decision-making models [3,4].

Smart cities and artificial intelligence (AI) are closely intertwined [5]. AI has significant potential to address the urbanization challenges faced by cities. The prospects of smart urban technologies, including AI, range from expanding infrastructure capacity to improving decision making and supporting businesses and cities. AI systems can be integrated into various urban development areas, such as energy, mobility, public safety, healthcare,

education, and urban planning [6–8]. Integrating AI into city management promises substantial economic benefits, as smart cities are becoming hubs for job opportunities and economic development.

An advanced aspect of AI's application in urban environments is the deployment of probabilistic AI algorithms. These algorithms excel in forecasting and risk management, which is crucial for addressing urban hazards. For example, probabilistic models like the multi-head attention-based that combines convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) networks have revolutionized day-ahead wind speed forecasting [9] and may empower cities to mitigate the impact of severe weather events preemptively. Similarly, Bayesian optimization frameworks for predicting dynamic infrastructure responses to typhoons [10] offer cities the tools to enhance resilience and ensure public safety during extreme weather conditions. These examples illustrate probabilistic AI's critical role in enabling urban planners to devise more effective emergency response strategies and infrastructure resilience plans, safeguarding communities against floods, fires, and other hazards.

Integrating these advanced AI technologies into urban management underscores the necessity of responsible urban innovation. Collaborative efforts between local governments and AI developers are crucial for maximizing AI's positive impacts in cities, ensuring both beneficial and ethical applications while minimizing negative consequences [6]. AI aligns perfectly with the data-driven ethos of a smart city, making this research highly relevant in the current times.

Open data is also thoroughly related to smart cities as it plays a crucial role in their development and functioning [2,11]. Open data refers to information that is freely available and accessible to the public, allowing for its use, reuse, and redistribution. In smart cities, open data initiatives aim to promote transparency, enhance governance, and foster citizen engagement [2]. The availability of open data has led to increased research using artificial intelligence for various applications. Furthermore, open data has been recognized as a catalyst for machine learning (ML), particularly in air pollution prediction, traffic forecasting, urban design, and transport features analysis [12]. It has facilitated knowledge discovery by making data findable, accessible, interoperable, and reusable, as per the FAIR Data Principles [13]. The availability of open data allows researchers and developers to gather relevant and diverse data inputs for their machine learning models, enabling them to analyze and extract insights from large datasets [14]. However, using open data in ML applications also presents challenges, such as data quality, data format, and data integration from multiple sources [15]. The risks associated with open data, such as privacy concerns, data accuracy, and decision making based on faulty data, need to be assessed and mitigated [2]. Nevertheless, open data is crucial in enabling intelligent solutions, promoting sustainable and innovative urban environments, and propelling research in machine learning. It facilitates the creation of predictive models and decision-making tools tailored for smart city applications [15,16].

AI plays a fundamental role in sectors like engineering and construction, helping to address housing challenges and infrastructure development, thereby improving the quality and efficiency of urban services [1]. The development of smart cities significantly impacts the housing market, affecting factors such as housing supply, prices, and quality. Studies have shown that the availability of smart cities' housing infrastructure and quality are generally higher than in the surrounding regions [17]. Smart cities prioritize the development of physical infrastructure and digital technologies for urban management, resulting in improved housing resources [18]. Therefore, these cities offer more extensive and better housing infrastructure than the provinces they are located in, indicating a higher standard of living for their residents [19]. Additionally, smart cities focus on convenience and sustainability in housing, incorporating smart systems for heat supply, water supply, sewage networks, and the use of energy-efficient building materials [15,20].

The housing market dynamics in smart cities are influenced by various factors such as accessibility, diversity of amenities, and sustainability [15,21]. The concept of smart

cities emphasizes the importance of walkability and accessibility to different services in determining the price of a property [22,23]. Factors like distance to key locations and the number of commercial establishments within a certain radius can significantly impact a house's price per meter value [19]. Additionally, the housing market in major cities tends to experience more volatility and higher price fluctuations than in other regions [24]. Differences in housing market characteristics, regulations, and cultural preferences also contribute to variations in housing prices and supply across regions [25]. Overall, the housing market in smart cities is influenced by a combination of physical attributes, location-specific factors, and market dynamics, making it a complex and evolving sector.

On the other hand, the housing supply in smart cities is influenced by a complex interplay of various factors, including financing costs, construction costs, vacancy rates, and regulatory constraints [26]. Studies have shown that financing costs, such as interest rates, have a negative effect on housing starts—a crucial economic indicator, as it reflects the health and direction of the housing market and, by extension, provides insights into broader economic conditions [12]. Similarly, construction costs, including material and wage costs, can also affect the housing supply, although the results are inconclusive in some cases [19]. Vacancy rates and sales delays have been found to have a negative impact on housing starts, indicating that longer selling times can lead to fewer housing starts [20]. Regulatory constraints, such as land regulation and planning controls, can influence the housing supply, with more restrictive regulations leading to fewer housing starts [24]. The housing market in smart cities is not always perfectly competitive, and policy interventions may be necessary to maximize social welfare and ensure affordability and inclusion [19,24]. Other factors also play crucial roles in measuring and understanding housing supply, such as housing stock, demolitions and renovations, market availability, and geographical distribution. In essence, these factors highlight the complex dynamics involved in the housing supply of smart cities, and understanding these dynamics is crucial for promoting a sustainable housing market and achieving the goals of smart city initiatives.

The real estate market, characterized by its complexity and dynamism, necessitates using advanced predictive tools capable of deciphering the interplay of numerous variables. In recent years, tree-based machine learning models and artificial neural networks (ANN) have emerged as front-runners to enhance prediction accuracy for real estate prices [27–31]. Due to their capacity to capture non-linear relationships and interactions, these models consistently outperform traditional linear models [29,32,33]. However, a significant challenge arises when these models are adopted in real-world scenarios due to their lack of interpretability [34]. As black-box models, they provide limited visibility into their internal decision-making processes, causing a trade-off between accuracy and interpretability [35–40]. Recently, eXplainable Artificial Intelligence (XAI) techniques have been explored to enhance model transparency and interpretability, addressing the black-box nature of complex ML models.

Integrating fairness, transparency, and accountability in AI systems is paramount to addressing and mitigating biases within data sources and algorithms. Verhulst (2023) [41] highlighted that combating bias and discrimination involves encouraging fair and unbiased data-handling practices that ensure equal treatment for all individuals, regardless of protected characteristics such as race or gender. This approach is further supported by using synthetic data to address existing dataset biases, thus promoting better representation of populations and fostering more equitable outcomes.

Vainio-Pekka et al. (2023) [42] emphasized the critical role of Explainable AI in detecting flaws and biases within systems, thereby ensuring the transparency of these systems. XAI techniques not only aid in reducing biases in data but also facilitate a deeper understanding of the problem at hand, promoting ethical AI practices that are understandable and interpretable by humans. The necessity of integrating ethical considerations into AI development is underscored to prevent the perpetuation of societal biases and ensure equitable outcomes.

As noted by Arrieta et al. (2020) [34], the fairness discipline includes methods for bias detection within datasets, especially those concerning sensitive data that affect protected groups via variables like gender and race. Ethical concerns with black-box models arise due to their potential to unintentionally create unfair decisions by considering sensitive factors. Proposals centered on fairness aim to discover correlations between non-sensitive and sensitive variables, detect imbalanced outcomes that penalize specific subgroups, and mitigate the bias effect on model decisions.

Furthermore, the Royal Society (2019) [43] notes that data collection issues can significantly impact the performance of machine learning systems, with image recognition systems failing to work accurately for minority ethnic groups. This challenge highlights the broader ethical challenges of AI, including fairness, transparency, and privacy protection, as emphasized by Yigitcanlar et al. (2021) [6].

Collectively, these insights stress the importance of addressing biases in data and algorithms to ensure fairness, transparency, and accountability in AI systems. Via the development and implementation of XAI techniques and other bias mitigation strategies, such as the use of synthetic data, it is possible to promote ethical AI practices that are not only understandable and interpretable by humans but also ensure that AI systems do not perpetuate existing societal biases, thereby ensuring equitable outcomes for all individuals.

Like in many sectors, the importance of interpretability cannot be overstated, and in the real estate market, stakeholders, including investors, policymakers, and urban planners, often demand clear and comprehensible explanations behind predictions as they plan and make informed decisions [4]. Here, XAI techniques provide insights into the factors influencing selling prices and help stakeholders make informed decisions [27,30,32,35,44].

Developing XAI techniques conveys a promising solution for enhancing the potential use of AI in smart cities as it enables unboxing black-box AI models and explicitly describes their mechanisms, thus providing transparency, interpretability, and informed decision-making processes [45]. Currently, XAI technologies are being developed and applied in smart city projects, focusing on traffic volume prediction, population estimation, and urban analytics [7,46] by leveraging the benefits of AI while ensuring responsible and accountable use in urban governance [38].

In this study, we use a cutting-edge technique within XAI—the SHapley Additive exPlanations (SHAP) by Lundberg and Lee [47]. The SHAP framework is grounded and derived from the Shapley values [48], a concept rooted in cooperative game theory that offers a consistent and locally accurate method to interpret model predictions. In machine learning, SHAP provides a unified measure of feature importance by assigning each feature an importance value for a particular prediction. SHAP values are the SHAP method's output, quantifying each feature's contribution to a particular prediction. These values give an intuitive understanding of feature contributions toward individual predictions, thereby lifting the veil on machine learning models. The SHAP value for a feature represents the average contribution of that feature to every possible prediction. These model-agnostic values are applicable for interpreting the results from any machine learning model, and they are trendy for tree-based models, such as the eXtreme Gradient Boosting (XGBoost) we use in this study. These values can explain why a model made a specific decision and offer a consistent and locally accurate way to explain the outputs of ML models [40,49].

This study's objective is to leverage open data effectively for improving real estate price prediction. Concurrently, it aims to enhance the predictive model's transparency and interpretability by applying XAI, understanding which features are most influential, and measuring their impact on the model's price prediction for the property listings in our dataset. This dual approach seeks to bridge the gap between advanced model performance and the ability to interpret this model. Open data provides broader and more diverse data for precision analysis, while XAI will offer insights into how model predictions are made, thereby fostering trust and accountability in real estate analytics. We demonstrate that increased model interpretability can coexist with high prediction accuracy via systematic

experimentation and detailed evaluations and offer practical benefits for urban planning and smart city development.

In our three-phase approach, the proprietary data are the starting point for the analysis, which are then augmented by open data. We first train an XGBoost model with features from a proprietary dataset for predicting real estate prices and provide the rationale behind selecting XGBoost over other considered models. The second phase entails retraining this model by incorporating open data features. This method provides the ability to evaluate the consequential impact on the model's predictive accuracy and identify which features significantly affect real estate prices. We also explore the integration of Optuna—a hyperparameter optimization framework, with our XGBoost machine learning model and fine-tune it. In the third phase, we leverage SHAP values analysis to explain our model predictions effectively. This step facilitates a profound exploration of feature interactions and quantifies their relative importance. This approach is intended not only to advance the field of real estate analytics but also to contribute to more informed decision making in urban development, aligning with the needs of rapidly evolving housing markets.

Our results reveal that integrating open data features significantly enhances model predictive power, with open data features such as bank evaluations, culture, and subway distance displaying the most substantial impact on real estate prices. The findings of this study add to the understanding of the symbiotic relationship between the real estate market and urban dynamics. This novel insight offers more transparent, understandable, and accountable investment guidance to various stakeholders, assists in designing effective housing policies, and aids in formulating informed strategies for smart city development.

The remainder of this paper is structured in the following manner: Section 2 delivers a comprehensive overview of the data, methodology, and ML and XAI techniques used. Section 3 presents and discusses the results, detailing the steps and the options taken to address the research objectives. In Section 4, we engage in a discussion of the findings, exploring their implications and significance. Lastly, Section 5 concludes the paper, summarizing the key findings and contributions and suggesting directions for future research.

2. Materials and Methods

This section presents the data and the methods used in our study to predict real estate prices using ML and XAI.

2.1. Data

This sub-section provides details of the data utilized in our study. We describe the study area and the specific datasets used for analysis, detailing our selection criteria and the sources from which the data were compiled. The process of feature engineering—how we chose and refined the dataset features for optimal relevance and insight—is then discussed. We conclude with the presentation of descriptive statistics that summarize the key characteristics of our data, setting the stage for the following analyses.

2.1.1. Study Area

In aiming to leverage open data for increased and transparent real estate predictions in smart cities, the study area under scrutiny in this research focuses on Lisbon. As Portugal's capital, this city nestles on the north bank of the Tagus River and the Atlantic coastline, making it the westernmost city in continental Europe. From an administrative perspective, Lisbon is subdivided into 24 counties. According to the Portuguese Census of 2021, its resident population was 545,796 (population density 5456.32/km^2). The number of conventional dwellings was 319,640, and the number of households was 242,065 (https://censos.ine.pt/) (accessed on 11 January 2024), with an average annual growth of the House Price Index from 2011 to 2021 by about 5.25% (https://bpstat.bportugal.pt/) (accessed on 11 January 2024).

Lisbon's housing market has significantly changed over the past decade, becoming increasingly attractive to tourists and investors. These changes have resulted in a significant

increase in foreign direct investment in real estate and construction. Furthermore, the government bore the expenses for redevelopment, allowing private capital to profit from property rehabilitation. Consequently, this trend has led to widespread housing and commercial gentrification and the rise of tourism-centric urban areas, consistent with broader patterns observed across Southern Europe [23,50].

A study by Marques et al. [51] revealed that housing affordability in Lisbon has decreased in the past years, particularly affecting the middle classes and younger generations. This fact is due to local incomes not keeping pace with the escalating market prices, making housing needs persistent. A surge in international investments, especially in upscale properties and tourist accommodations, has driven up house prices, with inner-city areas of Lisbon being the most affected. Notably, in 2019, the average duration properties stayed on the market (both for rent and sale) decreased to record low levels, indicating high demand and limited supply. Traditionally, homeownership has been viewed as a wealth accumulation strategy [52]. However, in Lisbon, where the homeownership rate stands at 52%, owning a home has become less feasible for many, given the relatively modest local salaries. In response to the housing challenges, the government has introduced housing policies focused on strengthening the public housing supply in the city. These initiatives prioritize the rehabilitation of public buildings, especially those most deteriorated. Also, efforts are being made to improve public services and transportation connections outside Lisbon's central regions, making these peripheral areas more attractive to residents and ensuring they offer a comparable quality of life [53].

The spatial distribution of our dataset for the housing transactions in Lisbon from 2018 to 2021 in Figure 1 presents a distinct pattern that underscores the city's real estate dynamics and infrastructure influence. A pronounced concentration of the most expensive transactions is evident within the city center. This area of high-value real estate transactions aligns predominantly along a north/south axis. This axis mirrors the trajectory of the principal subway line, suggesting a strong correlation between transit accessibility and property values. Such a pattern underlines the importance of efficient public transportation and its role in shaping real estate prices. The subway, serving as a primary urban artery, not only facilitates mobility but also accentuates the desirability of properties in its proximity, leading to a premium on centrality. As we move away from this central axis, there is a discernible decrease in transaction values. From the city center outward, this gradient is a classic representation of urban real estate dynamics where centrality often commands a premium.

Nevertheless, new zones of real estate significance are emerging as we move toward the city's peripheries, particularly along the edges of the Tagus River. The Parque das Nações county, located on the easternmost side of the river, is witnessing a surge in real estate activity. Its contemporary urban planning, waterfront vistas, and modern amenities are driving its transformation into a sought-after residential and commercial hub. Conversely, the historic Belém county is also experiencing a renewed interest in the city's westernmost edge. Renowned for its cultural landmarks, Belém blends its rich history with modern development, making it a compelling choice for real estate investment opportunities.

This spatial distribution, from the bustling city center to the evolving river edges, exemplifies Lisbon's dynamic real estate market. The interplay between historical significance, urban infrastructure, accessibility, and waterfront development is crafting a multifaceted real estate landscape. For stakeholders, from investors to urban planners or policymakers, understanding such patterns and the factors driving them is crucial for future strategies and decision making.

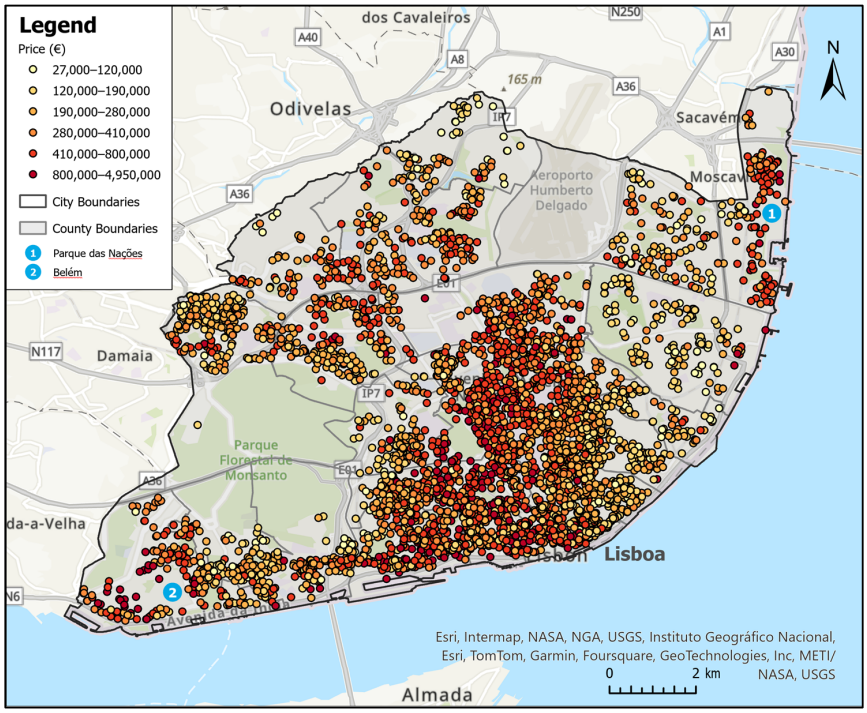


Figure 1. Spatial distribution of real estate transactions by price in Lisbon (2018–2021): This map visualizes property sales in Lisbon, differentiated by price ranges depicted in various colors, providing a clear indication of pricing hotspots and the economic landscape of Lisbon's real estate market over four years. Key counties such as Parque das Nações and Belém are marked alongside city and county boundaries for contextual reference.

2.1.2. Data Sources and Datasets

The data used in this study come from housing transactions in Lisbon that took place between 2018Q1 and 2021Q4. The original dataset under scrutiny captured 23,781 real estate transactions within Lisbon. This raw data are proprietary and were provided by Confidencial Imobiliário (https://www.confidencialimobiliario.com/ (accessed on 11 January 2024)), an independent databank, and encompass a listing collection with the properties' intrinsic attributes, such as technical specifications, amenities, and location, and the extrinsic details related to the transactions. Technical specifications include variables such as the private gross area in square meters, the energy performance certificate, and property quality status. House amenities include the number of bedrooms and whether the dwelling has a garage, pool, terrace, or patio. Concerning the location, we have information about the geographical coordinates, the postal code, and the county denomination. Transaction specifications include the selling price in Euros, the initial asking price, the transaction date, and the market placement date.

Predictive modeling aims to harness data that contribute to a robust understanding of trends and allow for precise predictions. The enrichment process is a common practice in data science to add value to existing data and to provide deeper insights by linking related information from different domains. In our study, this enrichment process combines proprietary real estate data with external open datasets for a more comprehensive analysis. Therefore, integrating contextual open data may significantly enhance the performance of machine learning models. This enrichment is essential for optimizing predictive modeling, leading to more effective solutions for smart cities [12,15,16,54]. Therefore, to improve our machine learning model performance and accuracy, we assembled several public datasets from various open data platforms covering the same period, which we used as comprehensively as possible to investigate and derive valuable insights into Lisbon's real estate dynamics during these years, understand the factors influencing property prices, and predict future trends in this ever-evolving market.

The open data gathered covers a wide range of specific urban indicators, from mobility (bus stops and proximity to subway and train stations) [44,55,56], quality of life and well-being (culture, commerce, education, health, leisure, and environment) [35,57–61], and governance (housing licensing, safety, and security) [21,62,63] to broader macroeconomic and financial indicators (inflation rate, unemployment, gross domestic product, and bank appraisals) [24,33,64,65]. These indicators play a crucial role in influencing the functionality and growth of a smart city and are instrumental in cities' assessment and evaluation. They offer valuable insights into urban development, economic performance, financial stability, and overall competitiveness [20,66,67].

Table 1 describes the primary data sources, including proprietary and open datasets, which we merge and append into one large dataset comprising 25 predictor variables to estimate real estate prices. By leveraging eXplainable AI (XAI) techniques to analyze the outputs of an XGBoost model with this enriched dataset, we aim to provide an in-depth and transparent perspective on the trends, patterns, and nuances of the housing market in Lisbon. Therefore, our approach underscores the potential of integrating publicly available datasets to enhance the accuracy and accountability of predictive analytics in property investments and urban planning initiatives.

Table 1. Real estate data variables, categories, and sources.

Variable	Description	Data Category [1]	Data Source [2]
Price	Transaction price of the dwelling (EUR) (not including taxes)	Proprietary data	Micro-SIR
BankEval	Median value of bank appraisal (EUR/m^2) by parish; monthly	Open data	INE
Bedroom	Number of bedrooms of the dwelling	Proprietary data	Micro-SIR
Bus	Number of bus stops within a 250 m radius of the dwelling	Open data	Transporlis
Commerce	Distance to the nearest shopping facilities (malls and markets) (m)	Open data	Geodados Lisboa
ConstrPermits	Number of new residential permits issued; monthly	Open data	BdP
CPI	Consumer Price Index (Inflation rate); monthly; year-on-year rate of change	Open data	BdP
Culture	Distance to the nearest cultural facilities (cinemas, museums, theaters, art galleries, and libraries) (m)	Open data	Geodados Lisboa
EPC	Energy Performance Certificate: 1—A+, 2—A, 3—B, 4—B-, 5—C, 6—D, 7—E, 8—F, 9—G	Proprietary data	Micro-SIR
GDP	Gross Domestic Product at market prices; quarterly; chained volume—year-on-year rate of change	Open data	BdP
Health	Distance to the nearest health facilities (public and private hospitals and health centers) (m)	Open data	Geodados Lisboa
Latitude	Latitude of the dwelling's postal code centroid (degrees)	Proprietary data	Micro-SIR
Longitude	Longitude of the dwelling's postal code centroid (degrees)	Proprietary data	Micro-SIR
Parks	Distance to the nearest park (m)	Open data	Geodados Lisboa
PGA	Private gross area of the dwelling (m^2)	Proprietary data	Micro-SIR
PropCond	Dwelling quality status: 0—Used, 1—New	Proprietary data	Micro-SIR
Safety	Distance to the nearest safety facilities (fire department stations and civil protection units) (m)	Open data	Geodados Lisboa
School	Number of school facilities (public and private schools and basic, secondary, and professional schools) within a 500 m radius of the dwelling	Open data	Geodados Lisboa
Security	Distance to the nearest security facilities (municipal police and public security police stations) (m)	Open data	Geodados Lisboa
Sports	Distance to the nearest sports facilities (m)	Open data	Geodados Lisboa
Subway	Distance to the nearest subway station (m)	Open data	Geodados Lisboa
Train	Distance to the nearest train station (m)	Open data	Geodados Lisboa
Trees	Number of trees within a 75 m radius	Open data	Geodados Lisboa
University	Distance to the nearest university (m)	Open data	Geodados Lisboa
UnempRate	Unemployment percentage of the active population aged between 16 and 74 years old; monthly	Open data	INE
ZIP7	Numeric representation of the 7-digit postal code in the "NNNN.NNN" format	Proprietary data	Micro-SIR

[1] Proprietary data: variables retrieved from the initial dataset (raw data); Open data: variables sourced from external, open datasets. [2] BdP—Bank of Portugal (https://bpstat.bportugal.pt/) (accessed on 3 December 2023); Geodados Lisboa—Lisbon City Council's georeferenced open data platform (https://geodados-cml.hub.arcgis.com/) (accessed on 3 December 2023); INE—Statistics Portugal (https://www.ine.pt/) (accessed on 3 December 2023); Micro-SIR—Confidencial Imobiliário (https://www.confidencialimobiliario.com/en/base-de-dados/micro-sir/) (accessed on 3 December 2023); Transporlis—Transporlis (https://www.transporlis.pt/) (accessed on 3 December 2023).

2.1.3. Descriptive Statistics

Before deploying our machine learning models, we first preprocessed the data. During this stage, we addressed missing values within the dataset, particularly for the EPC feature, which had a 12.05% missing rate. We used a simple imputation method, replacing missing values with the mode to preserve the feature's ordinal nature. Conversely, features with a high percentage of missing values, such as Garage, Pool, Terrace, and Patio, were excluded from our analysis. This decision was due to their excessively high missing rates, with the Garage at 30.14% and the Pool, Terrace, and Patio all sharing a missing rate of 34.07%. Excluding these variables prevents potential distortion in the predictive model outcomes [68].

Our approach included detecting and handling outliers in the target variable price, recognizing that outliers can significantly impact model performance and generalizability, particularly in regression models. Outliers can distort a model's generalization ability by leading it to overfit these extreme values. Therefore, we applied the Modified Z-score method to identify individual outliers in a single variable [69]. This method, relying on the median and median absolute deviation (MAD), is particularly effective for non-normally distributed data, such as the right-skewed distribution of the price feature. The Modified Z-score method, with a threshold of 3.5, was used to identify and remove 1204 outliers from the price feature that could skew our analysis.

In addition to this univariate treatment, we performed a multivariate outlier inspection by considering PGA (private gross area), the feature most highly correlated with price. We specifically targeted unusual combinations of values across these two features that are unlikely to be encountered in real-world deployment scenarios. In this process, we identified and removed observations with abnormally low prices per square meter, specifically those under EUR 1000/m^2. This analysis resulted in the exclusion of 107 observations from our dataset. By treating outliers, we refined our dataset to 22,470 records, retaining only those real estate transactions that most typically reflect the broader market trends and represent the majority of the population of our dataset.

Our study employed ArcGIS Pro (version 3.2.2) within a Geographic Information System (GIS) framework to analyze the impact of urban amenities and infrastructure accessibility on patterns of real estate transactions. The preprocessing of open data features was approached using two methods. First, we quantified the number of points of interest (POIs)—such as Trees, Bus stops, and Schools—within each property listing's 75, 250, and 500 m radii, respectively. This process involved creating spatial buffers around the properties and afterward counting the POIs within these areas to assess the density of amenities. Second, we calculated the distance from each property to the nearest POI categories, including Commerce, Culture, Health, Parks, Safety, Security, Sports, Subway, Train, and University facilities, using proximity calculations to establish potential accessibility. The outlined methodology is designed to elucidate the spatial dynamics between property transactions and the attractiveness of various urban zones, providing insights critical for real estate valuation and urban development planning.

To manage the issue of high dimensionality and the problems with one-hot encodings, like increased sparsity and multicollinearity, we converted 7-digit postal codes into a numerical format and binary encoded the property condition feature (PropCond).

Feature selection was conducted using Recursive Feature Elimination (RFE) in conjunction with feature importance scores from XGBoost models and correlation matrices. We partitioned our dataset into two segments, using 80% for training purposes and the remaining 20% for testing, in line with the standard train–test split methodology. Feature scaling was accomplished using the StandardScaler from the scikit-learn library, fitted only on the training data, and then applied to both the training and testing data to avoid data leakage.

After preprocessing, we selected 25 features for the final dataset. Table 2 presents this dataset's descriptive statistics, offering an overview of the data types, mean values, standard deviations, and a five-number summary for a succinct overview of the dataset's

distribution. It also includes measures of skewness, Fisher's kurtosis, and Pearson's correlation coefficients to highlight the features' distribution and their linear relationship with the target variable price.

Table 2. Summary statistics for the target variable and features (N = 22,470).

Feature Name	Data Type	Mean	St. Dev.	Min.	25%	50%	75%	Max.	Skew.	Kurt.	Corr. with Price
Price	Numerical	306,046.33	155,053.56	37,705	188,000	270,000	383,558	810,000	0.98	0.42	1
BankEval	Numerical	2929.44	166.91	2483	2830	2979	3040	3830	−0.43	0.56	0.10
Bedroom	Numerical	2.22	1.16	0	1	2	3	10	0.64	0.63	0.43
Bus	Numerical	7.50	3.58	0	5	7	10	23	0.73	0.88	−0.05
Commerce	Numerical	508.40	366.93	17.32	249.31	408.30	697.04	3133.37	1.73	5.23	−0.15
ConstrPermits	Numerical	2106.18	378.03	1355	1871	2082	2337	3172	0.50	0.17	0.04
CPI	Numerical	0.70	0.70	−0.70	0.10	0.60	1.00	2.70	0.80	0.63	0.03
Culture	Numerical	259.19	205.30	3.65	118.86	212.21	357.09	2539.90	2.72	16.19	−0.18
EPC	Ordinal	5.63	1.26	1	5	6	6	9	−0.83	1.48	−0.30
GDP	Numerical	1.34	5.96	−17.80	1.00	2.73	3.00	17.00	−0.70	1.75	0.01
Health	Numerical	474.15	294.26	7.37	272.74	429.62	626.59	3296.81	2.27	12.28	−0.05
Latitude	Numerical	38.73	0.02	38.69	38.72	38.73	38.74	38.80	0.62	−0.12	0.01
Longitude	Numerical	−9.15	0.03	−9.23	−9.16	−9.15	−9.13	−9.09	−0.61	−0.05	−0.07
Parks	Numerical	331.29	187.06	5.64	196.78	300.28	421.82	1295.44	1.19	1.97	−0.12
PGA	Numerical	87.29	37.90	26	60	79	107	631	1.45	4.79	0.72
PropCond	Binary	0.08	0.27	0	0	0	0	1	3.18	8.14	0.21
Safety	Numerical	870.79	525.79	17.32	502.10	769.54	1147.63	3776.78	1.31	2.32	0.01
School	Numerical	8.74	5.15	0	5	8	12	30	0.90	0.80	0.02
Security	Numerical	513.24	259.02	13.16	320.93	484.83	664.92	2073.19	0.82	1.43	−0.10
Sports	Numerical	148.41	81.91	7.36	86.37	135.91	194.09	560.71	0.85	0.72	0.09
Subway	Numerical	925.67	1123.06	3.58	319.50	520.41	988.07	6343.07	2.53	6.04	−0.10
Train	Numerical	1096.24	713.11	18.40	638.14	983.88	1357.03	5438.74	2.23	7.54	0.00
Trees	Numerical	16.73	19.43	0	1	10	27	186	1.72	4.34	0.16
University	Numerical	699.23	387.44	9.61	414.60	641.27	914.73	2921.03	1.23	3.01	−0.13
UnempRate	Numerical	6.88	0.51	5.70	6.50	6.80	7.10	8.30	0.69	0.43	−0.06
ZIP7	Numerical	1378.18	298.64	1000.00	1150.31	1300.00	1600.66	1990.62	0.64	−0.94	−0.18

2.2. Methods

This section outlines the analytical methods employed to analyze the previous dataset. The methods employed in this study contribute to the accuracy and interpretability of our model's output. We introduce XGBoost as our primary modeling technique, chosen for its efficacy in predictive modeling, followed by a description of the Optuna hyperparameter tuning strategy to maximize model performance. We then describe the performance metrics selected to assess our models' effectiveness, ensuring a multidimensional evaluation. Lastly, we detail incorporating the SHAP method to provide transparency and interpretability in our model's decision-making process. This analytical framework emphasizes the importance of XAI and open data in addressing urban challenges and the critical need for transparency in AI-driven analytics.

2.2.1. Extreme Gradient Boosting

The dynamics of real estate markets are complex and influenced by multiple interacting factors. Rather than being straightforward or directly proportional, real estate markets do not always follow a straight or predictable path. Instead of a simple cause-and-effect relationship, where a given change in one factor (like an increase in interest rates) leads to a proportional change in real estate prices, the relationship might be more complex. Small changes in one variable might significantly affect prices under certain conditions and negligible effects under others. Also, multiple factors can influence real estate prices simultaneously, and their combined effect might differ from the sum of their individual effects. For instance, the combination of low interest rates and high demand might push prices up more significantly than expected by considering each factor separately. Therefore, real estate markets exhibit non-linearity and joint effects, which machine learning algorithms adeptly handle [27,30,70].

In this study, we employ the eXtreme Gradient Boosting (XGBoost) algorithm, a highly performant and robust tree-based machine learning method developed by Chen and Guestrin [71], for predicting real estate prices [72,73]. XGBoost, an advanced implementation of Friedman's Gradient Tree Boosting algorithm [74], excels due to its unique combination of bagging and boosting techniques for ensemble learning. The bagging component enhances model stability and accuracy by training multiple models in parallel with independent sampling. The boosting aspect sequentially generates trees, with each new tree aiming to rectify the shortcomings of its predecessor, thereby incrementally improving the model's accuracy. This dual approach makes XGBoost particularly effective and widely adopted in research and industry, suitable for diverse applications, including regression and classification problems [75].

The XGBoost algorithm employs an objective function that combines a loss function for assessing prediction error and a regularization component that imposes penalties on model complexity to prevent overfitting. This combination is crucial for maintaining the balance between bias and variance. The loss function (l) measures how well the model's predictions match the actual data. For classification problems, it could be a log loss function, and for regression problems, it might be a mean squared error or another suitable metric. The regularization term (Ω) is the part of the objective function that penalizes complexity, which helps to prevent overfitting. This term is made up of two parts: one that penalizes the number of leaves in the trees (γT) and another that penalizes the magnitude of the leaf weights ($\frac{1}{2}\lambda\|w\|^2$), where γ and λ are regularization parameters.

The combination of these two elements constitutes the XGBoost objective function and is formally represented as [71]:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (1)$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$.

This composite objective function aims to find a balance between fitting the model accurately to the training data (minimizing the loss function) while keeping the model simple enough to generalize well to unseen data (minimizing the regularization term). This dual-component objective function ensures a balance between model complexity and predictive performance, which is crucial to the effectiveness of the XGBoost algorithm.

The algorithm's performance is primarily influenced by its hyperparameters, including "gamma", which controls regularization on the tree's leaves; "n_estimators", which specifies the number of trees to be built; and "max_depth", which determines the depth of each tree and is essential in controlling overfitting. These hyperparameters control model complexity, thus influencing the trade-off between bias and variance and, ultimately, the model's predictive prowess on unseen data. One can significantly enhance the model's predictive capabilities by fine-tuning these hyperparameters. The interplay of these hyperparameters with the loss reduction and regularization terms allows XGBoost to enhance its performance adaptively with each iteration, effectively handling various predictive modeling tasks. Moreover, XGBoost enables parallel tree boosting to expedite computational processes, which not only quickens the training speed but also ensures more efficient employment of hardware resources, thereby enhancing overall model training throughput and scalability [71].

2.2.2. Hyperparameter Tunning

Fine-tuning hyperparameters is a crucial step in machine learning for improving model effectiveness. Optuna (https://optuna.org/ (accessed on 11 January 2024)), developed by Akiba et al. [76], is an advanced hyperparameter optimization framework designed to automate the optimization process of machine learning models. It is renowned for its efficiency and flexibility, supporting various optimization techniques. It automatically determines the best set of hyperparameters to optimize the model's performance.

Optuna uses techniques such as Covariance Matrix Adaptation Evolution Strategy (CMA-ES) optimization, Tree-structured Parzen Estimator (TPE) optimization, Random Search, and Grid Search methods to identify the optimal hyperparameters. Optuna has been used to tune hyperparameters for tasks such as sales prediction [77], impedance forecasting [78], and multi-class classification [79], among others. By automatically adjusting the hyperparameters, Optuna can enhance machine learning models' accuracy and generalization potential, reducing the time and effort needed for manual tuning and leading to better forecast accuracy and improved performance.

By default, Optuna uses TPE as its default sampler, which is a Bayesian optimization method. TPE uses results from previous trials to inform the sampling of hyperparameters for subsequent trials. If early trials identify a promising region of the hyperparameter space, subsequent trials might focus on this region, leading to quick convergence. It introduces a novel pruning mechanism that efficiently discards unpromising trials, accelerating optimization. Optuna offers visualization tools to analyze the optimization process, providing insights into hyperparameter impacts and relationships. It has been effectively used in tuning deep neural networks [76], gradient boost models [80], AutoML systems [81], and reinforcement learning [82], among others, significantly improving model accuracy and performance [76,83].

Optuna supports parallel trials, which is an essential feature in increasing the speed of hyperparameter tuning. This parallelization allows multiple trials to be conducted simultaneously, leveraging multi-core processors and distributed computing environments. By running trials in parallel, Optuna significantly reduces the time required to find optimal hyperparameters, especially in cases where individual trials are time-consuming. This feature is particularly useful in complex machine learning tasks where hyperparameter tuning can be a bottleneck in the development process. Optuna achieves this parallelization while maintaining efficiency and accuracy in the optimization process [76].

2.2.3. Evaluation Metrics

When dealing with regression tasks such as real estate price prediction, several metrics can be used to provide a comprehensive understanding of the model's performance. These metrics are essential for assessing the performance of machine learning algorithms, as they provide a quantitative measure of their accuracy, precision, and generalization capabilities. The choice of metric can rely on the specific characteristics of the data and the business objectives [84].

When evaluating the performance of our XGBoost models, we prioritized the Mean Absolute Error (MAE) as the primary evaluation metric. This decision aligned the objective function used during the training phase and the criteria for performance assessment. Since our models were optimized using MAE as the objective function via Optuna, it stands to reason that MAE should be the primary metric for assessing the model's performance. This approach ensures that our evaluation meets the most critical objectives and constraints during the model's optimization process.

Furthermore, MAE provides a direct and interpretable measure of average prediction error, enhancing its practicality for clear and effective communication with stakeholders. This metric quantitatively expresses the average deviation of predictions from actual values without overly penalizing larger errors. This characteristic is especially crucial in our analysis, considering that the price distribution in our dataset is right-skewed. Such a skewness naturally results in outliers, which can disproportionately influence error metrics that square errors, like in Mean Squared Error (MSE) for example, where errors are squared before they are averaged, which leads to larger errors having a disproportionately more significant impact on the final metric value. By opting for MAE, we ensure a more balanced evaluation that reflects the typical prediction error, acknowledging the inherent variability within our dataset.

That said, MAE determines the average size of the errors in a set of predictions, not considering their direction, and is commonly used in statistics and machine learning for

assessing the accuracy of continuous variables. It represents the mean of the absolute differences between prediction and actual observation across the test sample, treating all individual differences with equal weight. MAE is calculated using the mathematical formula in Equation (2).

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

where

- y_i represents the actual observed values in the dataset;
- $\hat{y}_i$ denotes the predicted values generated by the model;
- n is the number of observations in the dataset;
- $|y_i - \hat{y}_i|$ represents the absolute error for each individual prediction.

While focusing exclusively on a single metric like MAE could potentially lead to a narrow view of the model's performance, incorporating a variety of metrics can offer a more holistic and nuanced evaluation. This is especially crucial as different metrics highlight different aspects of model accuracy and error behavior. To this end, in addition to MAE, we also assessed the model using three additional key metrics: MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Squared Error), and the R^2 Score (Coefficient of Determination). MAPE provides insight into the relative size of errors as a percentage, lending a sense of the error magnitude concerning the actual values. RMSE assesses the impact of larger errors as it squares them before averaging, thereby emphasizing larger deviations. The R^2 Score measures the degree to which the independent variables in a regression model account for the fluctuations observed in the dependent variable, indicating the model's goodness of fit. This combined approach, utilizing MAE, MAPE, RMSE, and R^2, ensures a comprehensive and multifaceted evaluation of the regression model, encompassing both absolute and relative error measures and overall model fit.

The Mean Absolute Percentage Error (MAPE) measures each prediction's average absolute percent error, providing a scale-independent assessment of error magnitude. It expresses accuracy as a percentage, which can be easier to interpret than other metrics. The formula of MAPE is given by Equation (3).

$$\mathrm{MAPE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{3}$$

The Root Mean Squared Error (RMSE) assesses the standard deviation of errors within predictions. It is computed by taking the square root of the mean of squared differences between actual and predicted values, and it provides error magnitude in the same units as the variable being predicted, which can be directly interpretable. This metric is sensitive to large errors and can be more informative than MAE when large residuals are undesirable or costly. The formula of RMSE is given by Equation (4).

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

The Coefficient of Determination (R^2 Score), while not an error metric per se, is a statistical indicator that measures the ratio of the variance in the target variable that is attributable to the predictor variables in a regression model. It indicates the goodness of fit and, therefore, measures how well unseen samples are likely to be predicted by the model. An R^2 value of 1 indicates that the regression predictions perfectly fit the data, whereas an R^2 value of 0 signifies that the model does not explain any variation in the response data around its mean. The formula of R^2 is given by Equation (5).

$$\mathrm{R}^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{5}$$

where $\overline{y}$ is the mean of the observed values y. It is calculated as $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

To calculate the percentage reduction in each evaluation metric (MAE, MAPE, and RMSE) from a baseline model (proprietary data) to a final model (open data), we computed the following difference ratio:

$$Percentage\ Reduction = \left(\frac{Metric\ value_{baseline} - Metric\ value_{final}}{Metric\ value_{baseline}}\right) \times 100\% \quad (6)$$

The result shows how much improvement in prediction accuracy (in terms of MAE, MAPE, and RMSE) our final model has achieved compared to the baseline model. A higher percentage indicates a more significant improvement.

To calculate the percentage gain (or improvement) in the R^2 score from the baseline model to the final model, we follow a similar approach as with the other evaluation metrics, with some adjustments for the nature of the R^2. The percentage gain can be calculated as follows:

$$Percentage\ Gain = \left(\frac{{R^2}_{final} - {R^2}_{baseline}}{{R^2}_{baseline}}\right) \times 100\% \quad (7)$$

2.2.4. Explainable Artificial Intelligence

Machine learning systems are increasingly being adopted, pushing society toward a new era of algorithmic decision making [85]. While these systems become more integral to various domains, their decisions bear a higher potential for societal impact, emphasizing the importance of trust, transparency, and understanding these complex models. Many advanced ML models operate as black boxes. Their opaqueness poses challenges as users, regardless of expertise, cannot fully understand or verify the rationale behind the system's decisions [86].

Historically, the focus of AI has shifted toward predictive power, leaving interpretability behind. Given the complexities of black box models, there is a growing emphasis and need for machine learning interpretability to make these systems more transparent and trustworthy. Explainable Artificial Intelligence (XAI), a term coined by Lent et al. [87], is an emergent field aiming to enhance the transparency of AI systems without compromising their predictive performance. XAI uses explanations to elucidate the predictions of ML models to users, serving as tools to achieve interpretability, typically relating the feature values of data to predictions in a comprehensible manner.

XAI encompasses a wide range of model-specific and model-agnostic techniques to improve the interpretability of machine learning models. Model-specific methods are designed to work with specific types of ML models. They leverage the inherent characteristics and structures of those models to provide explanations. For example, the parameters within a linear regression model or the architecture of a decision tree inherently provide interpretability. For deep neural networks, techniques like feature visualization or class activation mapping can provide insights specific to the architecture. Model-agnostic methods offer explanations for interpreting black box models independent of the ML model type, regardless of its internal architecture. They operate post hoc, meaning that after the model has made a prediction, they do not interfere with its internal workings or performance. Examples include methods that provide model-agnostic explanations, such as LIME (Local Interpretable Model-agnostic Explanations) [88], which approximates predictions with local and simpler interpretable models, and SHAP [47], which calculates the contribution of each feature to the prediction by using Shapley values, a concept initially developed in cooperative game theory [48].

The SHAP method has been widely applied across various domains for interpreting machine learning models, including finance [89], healthcare [90], and environmental sciences [91], to provide interpretable explanations of complex machine learning models. In our study, we leverage the predictive capabilities of an XGBoost ML model and enhance its interpretability using SHAP. We aim to decode the model's predictions, evaluate each

feature's significance, and provide a comprehensive view of the model's behavior and decision-making process. Using SHAP, we understand feature importance and identify the most influential features in making the predictions [47].

The SHAP method computes a value to each feature, representing its contribution to the model's prediction for a specific instance while considering all possible combinations of features. By attributing these Shapley-based values to each feature, the SHAP method enables an understanding of the relative importance of different features in influencing the model's predictions. The formula for computing SHAP values is based on Shapley values. The Shapley value formula, represented by Equation (8), calculates the contribution of each feature (or "player" in game theory) to the predictive model [47].

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (8)$$

where

- $\phi_i(v)$ is the Shapley value for feature i;
- N is the set of all features;
- S is a subset of features, with the exclusion of feature i;
- $|S|$ is the number of features within subset S;
- $|N|$ is the total number of features;
- $v(S)$ is the prediction model evaluated with the features in subset S;
- $v(S \cup \{i\})$ is the prediction model evaluated with the features in S plus feature i.

The formula calculates the average marginal contribution of feature i across all possible combinations of features. It does so by summing over all subsets of features S that do not include i, computing the difference in the model's prediction with and without feature i in each subset. The factorial terms are used for weighting these differences to ensure a fair distribution of contributions among the features.

In the context of SHAP in machine learning, this formula is adapted to calculate the contribution of each feature to the prediction of a machine learning model for a specific instance. The SHAP method applies this formula to understand how each feature value changes the prediction from what it would be if that feature were absent (or at its baseline value). The strength of this approach lies in its equitable distribution of predictive contributions among feature values. This allows for the assessment of importance ranging from local (individual sample-based) to global (overall model) explanations, leveraging the additive property of SHAP.

The SHAP values offer an intuitive understanding of feature importance by quantifying the contribution of each feature to the prediction. Positive SHAP values indicate that a feature increases the prediction, while negative values suggest the opposite. The SHAP value's magnitude represents the contribution's strength, allowing for comparing the relative importance of different features. This interpretability of SHAP values enhances the transparency and trustworthiness of machine learning models.

3. Results

This section presents the results obtained from a comprehensive three-phase analysis focused on predicting real estate prices using an XGBoost model. Initially, we provide a concise yet comprehensive comparative model analysis overview, justifying the XGBoost model selection based on empirical evidence in alignment with the study's predictive accuracy and model interpretability objectives. Next, the best-performing model, the Baseline Model—for proprietary data, is expanded into an Open Data Model by incorporating open data features and then retrained, allowing for a detailed examination of its impact on predictive precision and providing the most complete and performant solution. Subsequently, we explore SHAP values analysis to interpret the predictions of the Open Data Model, offering deep insights into the importance of features and interactions.

In this section, we present the results produced using Python (version 3.10.12), employing a variety of libraries offering an extensive array of capabilities. We utilized pandas (version 1.5.3), numpy (version 1.23.5), and scikit-learn (version 1.2.2) for data manipulation and analysis. Machine learning tasks were executed using xgboost (version 2.0.1) and tensorflow (version 2.15.0). We integrated optuna (version 3.4.0) and scipy (version 1.11.3) for optimization purposes. Visualization tasks were accomplished with the assistance of seaborn (version 0.12.2), matplotlib (version 3.7.1), plotly (version 5.15.0), and kaleido (version 0.2.1). The utility library joblib (version 1.3.2) was used for intermediate storage and loading, while platform-specific operations were managed by the google.colab library (version 1.0.0). We incorporated shap (version 0.44.0) for in-depth model explanation and interpretability. All these workflows were executed in the cloud-based runtime environment of Google Colab (https://colab.research.google.com/ (accessed on 11 January 2024)), leveraging VMs equipped with Nvidia's T4 GPU hardware acceleration and high-RAM capabilities.

3.1. Comparative Analysis of Predictive Models

Our comparative analysis aims to discern the most performant model for predicting real estate prices, a crucial step toward enhancing the accuracy and reliability of real estate analytics, particularly in the context of this study's dataset and regression task. To this end, we systematically evaluated a selection of commonly employed predictive models in real estate analytics—XGBoost [11], Random Forest Regression (RFR) [92], AdaBoost [32], and artificial neural networks (ANN) [30]—using a comprehensive suite of evaluation metrics. These metrics, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and the R^2 coefficient, were applied across training and testing datasets to measure each model's performance in terms of accuracy, consistency, and ability to generalize to new data.

Our analysis highlighted XGBoost as the standout model, particularly excelling in test dataset generalization, a key indicator of its robustness against overfitting and superior predictive accuracy. While Random Forest showed promise in training performance, XGBoost's consistent superiority across training and testing underscored its balanced approach to error reduction and precision in prediction. Conversely, despite showing potential, AdaBoost and ANN fell short of XGBoost's benchmarks, with ANN notably underperforming in predictive accuracy despite its theoretical capacity to model complex, non-linear relationships, suggesting a misalignment with the specific characteristics of our real estate dataset.

Table 3 succinctly presents our comparative findings, underlining XGBoost's optimal balance between training performance and testing applicability, particularly its lower error rates and higher R^2 values on unseen data. These results advocate for XGBoost's selection and underscore its potential in real-world applications, where generalizing from historical to future data is crucial.

Table 3. Comparative performance of predictive models on real estate proprietary data.

Dataset	Metric	XGBoost	RFR	AdaBoost	ANN
Train	MAE	40,821.22	36,166.76	73,917.53	63,803.96
	MAPE	16.66	14.34	31.27	58.37
	RMSE	55,568.64	51,184.51	95,210.09	91,295.80
	R^2	0.87	0.89	0.62	0.65
Test	MAE	51,733.88	52,678.57	74,764.45	65,366.37
	MAPE	19.58	19.74	30.43	58.31
	RMSE	72,488.87	74,482.71	97,492.31	94,369.38
	R^2	0.79	0.78	0.62	0.64

Choosing XGBoost aligns with our broader research objectives, specifically enhancing model transparency and interpretability via Explainable AI (XAI) techniques. This approach

increases the utility of XGBoost and provides stakeholders in the real estate sector with clear, understandable insights into the factors influencing price predictions. The selected model's robustness and XAI integration significantly strengthen the credibility and trustworthiness of our study's findings.

In conclusion, the systematic evaluation of predictive models supports the selection of XGBoost for its adept handling of real estate data complexities. By achieving an ideal equilibrium between minimizing overfitting risks and maximizing predictive accuracy, XGBoost stands out as the benchmark, meeting our analysis criteria. The implications of this choice extend far beyond this study, setting a foundation for innovative approaches to predictive modeling in real estate analytics.

3.2. Model Training and Optimization

Our analysis employs an XGBoost model trained on an 80–20 train–test split in our modeling process. The model undergoes training using the training dataset, and Optuna is employed for hyperparameter optimization. Within Optuna, the objective function utilizes 10-fold cross-validation to compute the model's Mean Absolute Error (MAE) with the best hyperparameters found. The data are randomized before fold-splitting, ensuring that each fold is a representative mix of the whole dataset, thus improving the robustness and reliability of the cross-validation process and effectively preventing any order-based biases during the data division process. In shuffling the data, a fixed Hitchhiker's 42 random seed was set to ensure reproducible results.

The model is trained and evaluated during each cross-validation fold, and the MAE is calculated. Optuna intelligently searches the hyperparameter space, and the objective function reports intermediate results and determines if a trial should be pruned based on performance, speeding up the optimization process. Early stopping is also used to prevent overfitting during training. This method divides the dataset into ten subsets (or folds). The model is trained on nine of these folds and tested on the remaining one. This process is repeated ten times, each time with a different fold used as the test set, and the MAE is calculated for each iteration. The average MAE across all folds is then used as a criterion for optimization, and it measures how close the model's predictions are to the actual values, with a lower MAE indicating the better performance of the model.

Optuna's objective is to identify a combination of hyperparameters that minimize this average MAE. This iterative process is optimized over 100 trials. Each trial involves training a model with a different set of hyperparameters to see which combination yields the best performance, evaluated by its MAE score. Once the best-performing model (i.e., the best set of hyperparameters) is identified in each trial, this model is then retrained on the entire training dataset. This approach validates model performance and guarantees the model's robustness. After 100 trials are completed, the overall best model is selected—the one with the best MAE performance across all trials. This model is considered the most optimal and is used for further testing and evaluation.

The model with the most optimal hyperparameter set is then used to make predictions on the test dataset to evaluate the model's performance on unseen data. Predictions are made, and the model's performance is thoroughly evaluated using the metrics MAE, MAPE, RMSE, and R^2 computed for the test set. Additionally, residuals were analyzed to diagnose potential issues in predictions.

The above predictive analytics pipeline, involving training, hyperparameter tuning, and evaluation for an XGBoost model, was thus first used to develop a Baseline Model containing the seven features derived from proprietary data. Building on the Baseline Model, we retrained the XGBoost model by integrating 18 additional features from open data sources. Like the Baseline Model, the Open Data Model underwent the same pipeline. The Baseline and Open Data Models are thus instantiations of the same XGBoost model.

Additionally, for both models, we conducted a fine-tuning process with several runs to refine the hyperparameter space further and curb overfitting by iteratively adjusting the ranges based on the previous best values found, considering the individual importance

of each hyperparameter. This process allowed the optimization algorithm to fine-tune the hyperparameters more closely. The goal was to balance model complexity with its generalization ability to unseen data, achieving a balance between performance and avoidance of overfitting.

Table 4 presents the optimal hyperparameters for the two XGBoost models' instantiations (Baseline Model and Open Data Model), outlining their best-found values and detailing the respective search ranges for these parameters for each model.

Table 4. Optimal hyperparameters of the XGBoost model (best value and search range).

Hyperparameter	Description	Type	Baseline Model	Open Data Model
n_estimators	The number of gradient boosted trees. Equivalent to the number of boosting rounds. Too many trees can lead to overfitting by capturing more noise.	Integer	1961 (1800, 2100)	2031 (2000, 2050)
max_depth	Maximum tree depth for base learners. Controls how deep each tree is allowed to grow during any boosting round. Deeper trees can capture more noise.	Integer	8 (6, 8)	11 (10, 11)
learning_rate	Step-size shrinkage is used to prevent overfitting. It scales the contribution of each tree by a factor between 0 and 1.	Float	0.007 (0.007–0.010)	0.013 (0.012, 0.013)
subsample	The proportion of data samples utilized to train each base learner in the model. Values less than 1 make the algorithm more conservative and prevent overfitting.	Float	0.74 (0.65–0.75)	0.72 (0.70, 0.74)
colsample_bytree	The fraction of features to be used for each tree. A value of 0.6 means that 60% of features are used to train each tree.	Float	0.57 (0.50–0.60)	0.60 (0.50, 0.60)
gamma	Required minimal reduction in loss to further split a leaf node in the tree. A larger gamma simplifies the model, preventing overfitting.	Float	0.27 (0.20–0.30)	0.10 (0.10, 0.15)
reg_lambda	L2 regularization term (Lasso) on weights. It is used to avoid overfitting by penalizing large weights.	Float	1.84 (1.50–2.00)	1.95 (1.90, 2.20)
reg_alpha	L1 regularization term (Ridge) on weights. It encourages sparsity in the final model representation, allowing some features to be entirely ignored.	Float	2.23 (1.80–2.30)	2.10 (1.90, 2.20)
min_child_weight	Minimum sum of instance weight (hessian) required in a child node. Higher values help prevent overfitting by avoiding learning from overly specific data patterns.	Integer	3 (3, 5)	11 (9, 11)

Both models demonstrate a balanced strategy in subsampling ratios and feature selection (colsample_bytree), carefully managing the trade-off between model complexity and overfitting. The Open Data Model's reduced gamma setting allows for more extensive tree splits, thus enabling a finer granularity in capturing data patterns. With increased L2 (reg_lambda) and moderated L1 (reg_alpha), its regulatory strategy emphasizes feature efficacy while managing model complexity.

A crucial aspect of the Open Data Model is the heightened min_child_weight, strategically set to prevent overfitting. Overfitting occurs when a model learns too much from the training data, including the noise and fluctuations, which are not generalizable. This adjustment is vital in real estate modeling, characterized by market variability across locations, property types, and economic conditions, and an overly tailored (overfitted) model fitted to a market segment, such as a particular location or property type, risks underperforming in other segments. The deliberate enhancement of min_child_weight in the Open Data

Model reflects a strategic choice to maintain the model's versatility, as the model ensures wide-ranging applicability by avoiding overemphasizing specific market segments.

Overall, the Open Data Model is engineered for precision and adaptability, aligning with the dynamic nature of real estate markets. It deftly combines detailed pattern detection with strategies to ensure reliability and generalizability, making it suitable for high-stakes, varied market predictions.

Figure 2 depicts the price prediction error plots for the Baseline and Open Data Models. It shows how well the predicted values (on the y-axis) match the true values (on the x-axis). A perfect prediction would mean all points lie precisely on the dashed line, where the predicted value equals the true value.

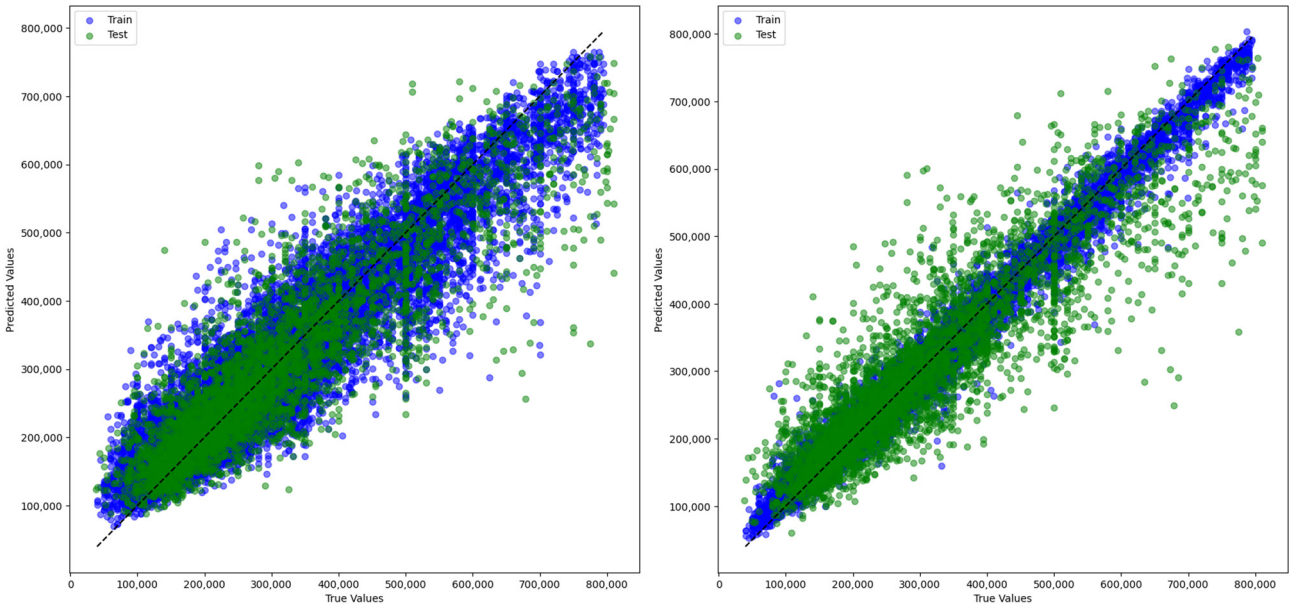
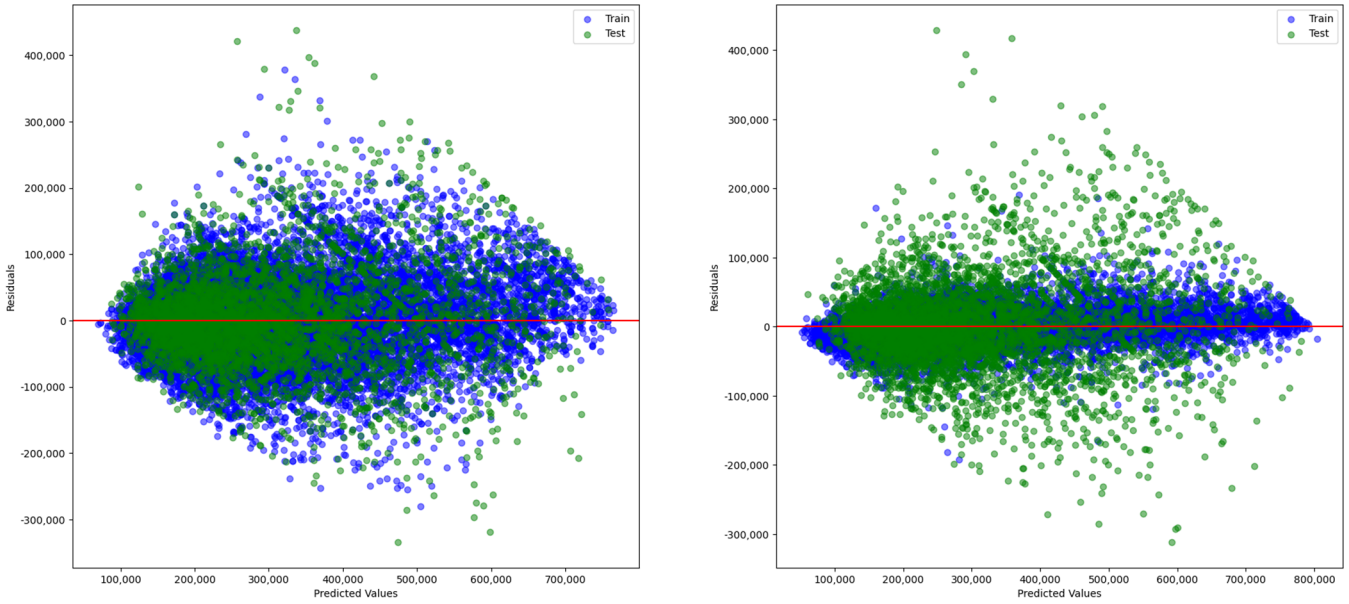


Figure 2. Comparative analysis of price prediction accuracy between XGBoost models: This figure presents two scatter plots comparing predicted versus true values of real estate prices, with the train data in blue and the test data in green. Plot (**a**) shows the performance of the Baseline Model, while plot (**b**) depicts the Open Data Model. Each plot contains points representing individual predictions. The closer the points are to the dashed line of perfect fit (where predicted values equal true values), the more accurate the predictions. The Baseline Model shows a broader dispersion of points, indicating less accuracy, while the Open Data Model's points are more concentrated around the identity line, suggesting higher accuracy in predictions.

Several insights emerge by examining the price prediction error plots for both models. The Baseline Model plot shows a broad dispersion of predictions as the true values increase, indicating increased difficulty in predicting higher-priced properties. There are distinct outliers, especially for higher true values, suggesting the model may not capture all the factors influencing higher real estate prices. In the Open Data Model plot, the tighter clustering of points around the dashed line indicates that the model predictions are generally more accurate across the full range of prices. Its test points do not exhibit as much scatter relative to the training points as in the Baseline Model, which might indicate better generalization. The improved performance in the Open Data Model suggests that incorporating open data features enhanced the model's precision. Its consistent performance across different price ranges implies a better understanding of the underlying relationships in the data.

The prediction errors concerning the predicted values can be assessed by analyzing the residual plots of the Baseline and Open Data Models (Figure 3). The price prediction error plots and the residual plots provide a comprehensive picture of each model's performance.



(**a**) (**b**)

Figure 3. Residual analysis of price prediction models: The figure displays two residual plots for evaluating the prediction errors of the two real estate price prediction models, with the train data in blue and the test data in green. Plot (**a**) corresponds to the Baseline Model and plot (**b**) to the Open Data Model. Each plot illustrates the residuals (the differences between predicted and actual prices) on the y-axis against the predicted prices on the x-axis for both train (blue dots) and test (green dots) datasets. Ideally, residuals should be randomly distributed around the horizontal line at zero (red line), indicating that the model has no systematic error. The Baseline Model exhibits a broader spread of residuals, suggesting higher variability in prediction errors, while the Open Data Model shows a tighter clustering around the zero line, implying more consistent prediction accuracy.

The residual plot for the Baseline Model shows a noticeable spread in residuals as the predicted property values increase, which could suggest a heteroscedasticity pattern, where the errors' variability increases with the prediction's magnitude. This pattern might indicate that the model's predictive accuracy is not consistent across the range of property values. The presence of residuals with large magnitudes, particularly on the higher end of the predicted values, highlights the model's limitations in accurately predicting more expensive properties.

For the Open Data Model, the residuals are more tightly clustered around the zero line, suggesting more consistent prediction errors across different values. This homoscedastic pattern indicates that the model's performance is more uniform, regardless of the property's value. The residuals are distributed more symmetrically around the horizontal axis, implying fewer systematic biases in prediction.

When comparing the two models, it is evident that the Open Data Model provides a more uniform performance with minor errors across the entire range of predicted values. These facts suggest that the Open Data Model is more robust, potentially due to the inclusion of the additional features of open data. The Baseline Model's performance is more variable, with the quality of predictions decreasing as property values increase, which could indicate missing explanatory variables or features better captured in the Open Data Model. Ultimately, the Open Data Model is more effective in predicting real estate prices across all ranges with more consistent precision. The analysis suggests that incorporating comprehensive and relevant open data features while ensuring the model captures the complexity inherent in real estate pricing is crucial for reliable predictions.

3.3. Model Evaluation

After training and hyperparameter optimization, we evaluate the performance of the Baseline and Open Data Models using various metrics for a holistic assessment—MAE, MAPE, RMSE, and R^2, computed for the test set. Subsequently, a comparative analysis of the model's performance and reliability is made by juxtaposing both models' performance metrics. We calculate and discuss the Reduction/Improvement (%) in predictive precision, clearly showing how open data contributes to our model. Via this dual approach of an individual (standalone) performance evaluation and comparative analysis, we offer a comprehensive overview of the models' capabilities and quantitatively assess the value added by the open data features.

The calculated performance metrics for the Baseline and Open Data Models in Table 5 provide quantitative insights into their predictive capabilities for real estate prices. The performance metrics calculated on the test set estimate how the models will likely perform in real-world scenarios, thereby providing an understanding of the model's value.

Table 5. Evaluation metrics of the XGBoost model for prediction of real estate prices.

Dataset	Metric	Baseline Model	Open Data Model	Reduction (%)	Gain R^2 (%)
Test	MAE	51,733.88	47,469.62	8.24	N/A
	MAPE	19.58	17.75	9.36	N/A
	RMSE	72,488.87	68,347.98	5.71	N/A
	R^2	0.79	0.81	N/A	2.96

The Baseline Model's performance on the test set, with an MAE of 51,733.88 and a MAPE of 19.58%, indicates that while it can capture some of the trends in real estate pricing, its predictions may not be reliable enough for precise financial decision making. The RMSE value of 72,488.87 further suggests that the model is particularly challenged by properties with higher prices or unique features not well-represented in the testing data. An R^2 of 0.79 shows that the model has a fair predictive capacity but also indicates that about a fifth of the variance in housing prices is left unexplained.

In contrast, the Open Data Model's performance metrics show notable improvements. The reductions in MAE and MAPE by 8.24% and 9.36%, respectively, mean that the predictions are closer to the actual selling prices with fewer outliers, which can be particularly valuable for investors and realtors seeking to price properties or forecast market trends. The RMSE reduction of 5.71% indicates fewer significant errors, which is essential for evaluating high-value properties. The increase in R^2 by 2.96% reflects a better fit to the data, suggesting that incorporating open data has provided the model with additional explanatory power to account for the price variability.

The performance gains from the Open Data Model suggest that the features derived from open data sources provide a more nuanced view of the real estate market. These improvements in the model's predictive capabilities are due to the additional data giving more detailed neighborhood characteristics, such as enhanced mobility, comprehensive quality of life, well-being assessments, and improved governance, macroeconomic, and financial indicators not accounted for in the Baseline Model.

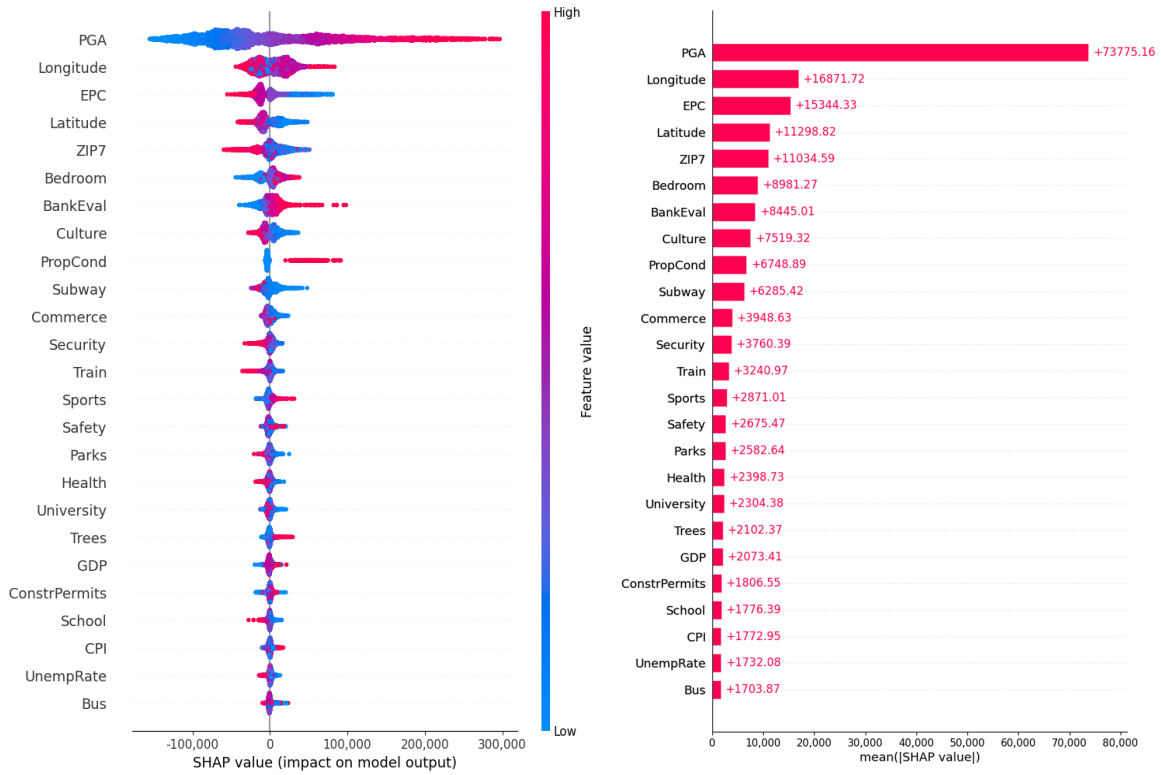
To summarize, the Open Data Model's improved accuracy makes it a more robust tool for stakeholders in the real estate market, offering more reliable predictions and insights. The gains in precision and reduced errors highlight the value of incorporating open data into predictive models. These results underscore the significance of using open data in real estate AI-driven solutions, particularly for smart cities. Integrating diverse open data sources can provide superior models with more accurate and insightful predictions. These model improvements can lead to better-informed decisions in real estate investments, policymaking, and market analysis.

3.4. Model Interpretability

The final phase of our analytical approach involves SHAP values analysis to interpret the predictions of the Open Data Model, which, in our study, yields the most comprehensive and performant solution for real estate price prediction. This analysis delves into the importance of individual features and their impact on the model's decision-making process, giving profound insights into the factors driving real estate prices.

The analysis employed SHAP's TreeExplainer to interpret the predictions of the XGBoost Open Data model by calculating SHAP values for the test set and gaining insights into feature importance and individual predictions on unseen data. These SHAP values are measures of feature importance that indicate how much each feature contributes, positively or negatively, to each individual prediction, thus quantifying the impact of each feature on the model's predictions. The TreeExplainer within the SHAP framework is specially optimized to analyze tree-based machine learning models efficiently and accurately.

During SHAP values analysis, several visual outputs are commonly generated to understand the contributions and importance of features in a machine learning model (https://shap.readthedocs.io/ (accessed on 11 January 2024)). Figure 4 shows two visually generated outputs for interpreting model predictions with SHAP values analysis.



(**a**) (**b**)

Figure 4. SHAP (SHapley Additive exPlanations) value analysis of feature impact on real estate price predictions: This figure is divided into two main parts: (**a**) on the left, the SHAP Summary plot (beeswarm) visualizes the impact of each feature on the model's prediction for individual instances. Each point represents a SHAP value for a feature and an instance. High feature values are colored in red and low in blue. Features are ordered by the sum of SHAP value magnitudes across all samples. (**b**) on the right, the SHAP Global Summary plot shows the average impact of each feature on the model's output. Features are ranked by their importance. The length of the bar represents the mean absolute SHAP value for each feature, indicating its importance. Positive SHAP values indicate a feature's increasing impact on the model's prediction, while negative values indicate a decreasing impact. Together, these plots provide insight into which features are most influential for the model's predictions and how they affect real estate prices.

The beeswarm plot in Figure 4a shows the distribution of the SHAP values for each feature across all the data points. The position on the x-axis indicates the impact of the feature on the model's output, where features pushing the prediction higher are displayed on the right, and those pushing the prediction lower are on the left. The color intensity represents the feature's value, with pink indicating higher values and blue indicating lower values. The distribution of points illustrates the variability of the impact of each feature.

The global summary plot in Figure 4b shows each feature's mean absolute SHAP values across all the data, measuring its overall importance. A higher value indicates a more significant impact on the model's predictions. The bars are color-coded and include a numeric indication of the mean absolute SHAP value for each feature, and the sign $(+/-)$ before the numbers indicates the direction of the impact. Unlike feature importances, typically described using abstract units derived from sophisticated concepts like impurities in tree algorithm nodes, mean absolute SHAP values are more straightforward and intuitive as they are presented in the same units as the target variable. In this case, they are quantified in Euros. Here, PGA is the most significant feature, impacting the predicted house price by an average of ±EUR 73,775.16. On the contrary, the Bus is the least informative feature, contributing just ±EUR 1703.87 to each house price prediction—a fact that is expected given the widespread availability of bus stops, which makes them less of an influential factor for property prices.

The SHAP analysis results in Figure 4 show that our model for predicting real estate prices reveals a nuanced approach to property valuation. Key findings include the following:

1. Primary predictors: The size of the property, indicated by the private gross area (PGA), emerges as the most significant factor, confirming that larger dwellings typically command higher prices. The property's exact location, denoted by Longitude, Latitude, and postal code, is also crucial, aligning with the well-known real estate emphasis on location;
2. Secondary factors: The condition of the property and its energy performance play important roles, although their impact might vary in different contexts. Proximity to amenities like transport, commerce, and cultural sites is significant, highlighting the value of convenience and lifestyle associated with the property;
3. Economic indicators: broader economic conditions, reflected by indicators such as GDP (Gross Domestic Product) and CPI (Consumer Price Index), are considered in the valuation but have a lesser impact compared to the direct physical and locational attributes;
4. Environmental and health attributes: features like proximity to parks and health facilities moderately affect property values, suggesting that environmental quality and access to healthcare are important to potential buyers;
5. Variability in impact: the model indicates a range of impacts for certain features, suggesting that the desirability of attributes like proximity to amenities may depend on specific market or neighborhood dynamics.

Overall, while the model places the most significant emphasis on physical characteristics and location specifics, it also integrates various factors, from accessibility to amenities to macroeconomic conditions, illustrating the multifaceted nature of real estate pricing.

4. Discussion

This section comprehensively analyzes and discusses the results obtained, elucidating their significance and implications in the broader real estate market analysis context.

4.1. Private Gross Area

From the results of the SHAP analysis, private gross area (PGA) exerts the highest average impact on property prices among the features analyzed, with a mean absolute SHAP value of 73,775.16. This prominent position underscores the critical role of a dwelling's size in price determination. This finding validates previous research by Guliker et al. (2022) [65], Ho et al. (2020) [31], Iban (2022) [93], Rampini et al. (2021) [30], Tchuente et al. (2022) [32],

and Xu et al. (2022) [94], indicating the housing area as a strong predictor for the price of a property. The presence of non-linearities revealed by a broad distribution of PGA's SHAP values in the beeswarm plot, sprawling across both sides of the baseline, reflects a complex relationship with price. Larger properties generally command higher prices, a well-known trend in real estate, but the spread of SHAP values suggests this trend is not universally linear. Smaller properties can similarly fetch high prices, contingent on additional factors.

Conversely, a larger area does not unconditionally translate to a premium in price. Indeed, the observable variability along the x-axis of the plot indicates that the influence of PGA on price is subject to inconsistencies. This pattern implies the existence of nonlinear dynamics, where the effect of PGA on the price does not consistently correlate with the area. In some instances, the increment in PGA can lead to a proportionate increase in price predictions. In contrast, the effect may be lesser or inversely related in other instances, potentially influenced by market trends, property location, and desirability.

Figure 5 represents a geospatial heatmap of SHAP values for the PGA feature from our XGBoost model for predicting real estate prices. This visualization helps understand which geographic areas have properties where the PGA significantly affects their predicted prices according to the model. Regions with no points or smaller points indicate areas where PGA is less important in predicting real estate prices or has a potentially negative impact. Each point corresponds to a property within the Lisbon area, with its location given by Latitude and Longitude coordinates. The color of a point indicates the SHAP value associated with the PGA feature for that property, with the color scale extending from −150 k to 250 k. The negative values on the color scale are essential for understanding the direction of the impact, while the absolute values are used for the size to visualize the magnitude. The size of each point reflects the absolute magnitude of the SHAP value, ensuring visibility for both positive and negative contributions.

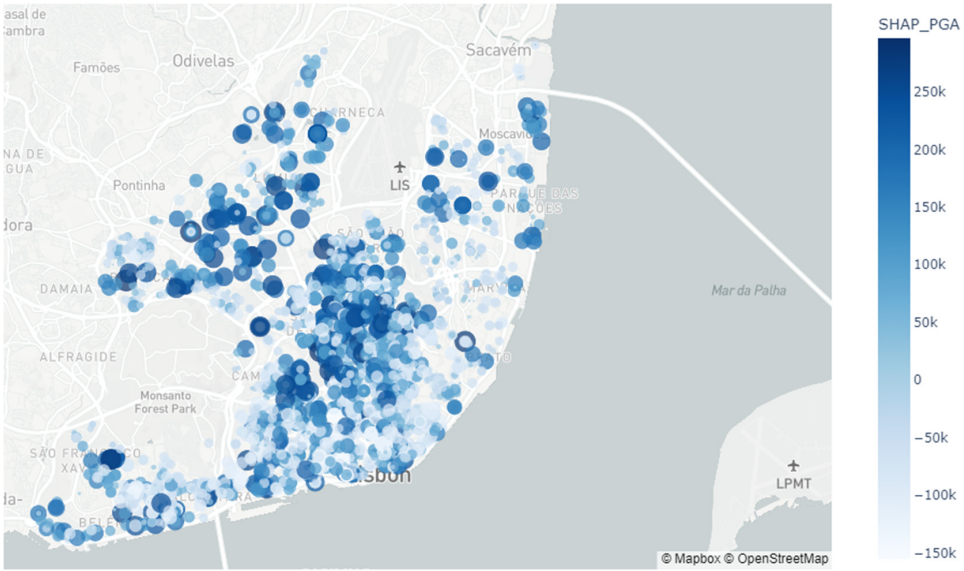


Figure 5. Geographic distribution of SHAP values for private gross area (PGA) in Lisbon: This map illustrates the influence of the private gross area feature on real estate price predictions across different locations in Lisbon. Each bubble represents a property, with its size proportional to the SHAP value assigned to the PGA feature, indicating the magnitude of impact on the prediction. The color scale ranges from light to dark blue, corresponding to the range of SHAP values from negative to positive. Higher SHAP values (darker blue) suggest a more significant positive impact of PGA on the property's predicted value, whereas lower values (lighter blue) indicate a lesser or negative impact. This visualization highlights areas where PGA is a stronger predictor of real estate prices and where it may have a reduced or inverse effect.

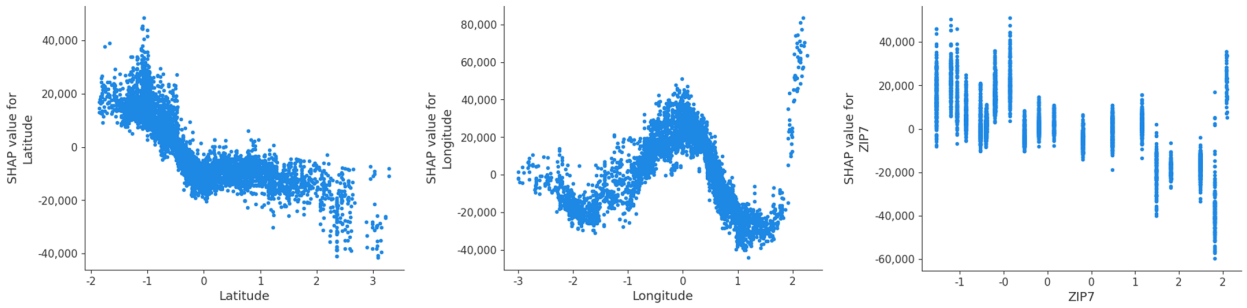
Darker blues represent higher positive SHAP values, denoting a more significant positive impact of PGA on the model's price prediction, particularly noticeable in central Lisbon, which suggests a higher space valuation in these areas. Lighter blues, conversely, indicate negative SHAP values, signaling areas where larger PGAs may not be as valued or could even decrease a property's predicted price. The varying shades of blue indicate that the impact of PGA on the model's predictions is not uniform across Lisbon. Some areas have a shallow impact (light blue), while others have a high impact (dark blue). This variability could be due to neighborhood desirability, local amenities, or zoning regulations that might make PGA more or less important in different areas.

The nuances captured by this analysis suggest that PGA's impact on price predictions is modulated by a spectrum of variables that extend beyond the mere area, underlining the importance of a multifaceted approach in price estimation models. This insight is critical for stakeholders who must consider the size of a property and how it interacts with other market and property features to drive value in the real estate market.

4.2. Geolocation

The insights gathered from the SHAP analysis reveal the significant impact of geospatial features, namely Latitude, Longitude, and ZIP7 (7-digit postal code), on real estate valuation.

Longitude emerges as a particularly influential factor, with a mean absolute SHAP value of 16,871.72. The beeswarm plot indicates its pronounced effect on property prices, suggesting that properties within certain longitudinal zones, potentially due to their location within sought-after neighborhoods, proximity to commercial hubs, or other desirable amenities, tend to command higher prices. The SHAP beeswarm plot illustrates this trend: red dots skewed to the right indicate areas with higher property values, while blue dots to the left denote lower-valued areas. Figure 6b also displays a similar pattern, where Longitude exhibits pronounced peaks and troughs, indicating that certain longitudinal zones significantly impact property prices positively or negatively. The broader range of SHAP values and clear clusters suggest a more localized and complex impact on property prices. This variation implies that east–west positioning within the city has diverse implications on property values, thus reflecting geographical desirability.



(**a**) (**b**) (**c**)

Figure 6. SHAP dependence plots for geographic predictors of real estate prices: This series of plots illustrates the relationship between the SHAP values of real estate prices and three geographic features: (**a**) Latitude, (**b**) Longitude, and (**c**) ZIP code (ZIP7). Each plot visualizes the individual contribution of these features to the predictive model's output. Plot (**a**) shows the SHAP values for Latitude, displaying a trend where certain latitudes correspond to higher or lower impacts on price predictions, suggesting a geographical preference or aversion in the real estate market. Plot (**b**) for Longitude reveals a complex, non-linear relationship with the price predictions, indicating that some longitudinal regions might be more favorable or unfavorable for property values. Lastly, plot (**c**) demonstrates the variability of the ZIP code's impact on price predictions, with some ZIP codes having a distinctly positive or negative influence, reflecting the localized real estate market trends. Collectively, these plots enable a nuanced understanding of how specific geographic locations contribute to the valuation of real estate properties as per the model's predictions.

Latitude, although less influential than Longitude with a mean absolute SHAP value of 11,298.82, still plays a significant role in the model. It indicates the north–south desirability of properties within the city. The SHAP Partial Dependence plot in Figure 6a for the Latitude shows a broadly descending trend, where the SHAP value decreases as the Latitude increases. This trend suggests a consistent north–south axis relevance in property valuation, and the impact on property prices tends to decrease as we move north or south from a certain point. The spread of the points is relatively uniform, with no apparent clusters, indicating that the impact of Latitude on property prices is more evenly distributed and possibly less localized than Longitude or ZIP7. This plot also seems to exhibit a less dramatic range of SHAP values than Longitude and ZIP7, indicating a more stable but widespread influence on property prices.

ZIP7, representing the 7-digit postal code, also shows a notable influence on property prices, with a mean SHAP value of 11,034.59. This feature reflects the desirability of different areas, capturing how local dynamics, such as school districts, local market conditions, or community features, impact property values. The SHAP values for ZIP7 underscore the intricacies of regional price variations, indicating that specific postal codes correspond to areas with differing levels of desirability.

In Figure 6c, the dependence plot for ZIP7 shows a series of vertical lines, indicating that discrete categories within the data correspond to individual postal codes. The variation in SHAP values is substantial, with some postal codes having a strong positive impact on prices and others a negative impact. This plot illustrates the granularity at which ZIP7 can affect property prices, likely reflecting very localized factors such as neighborhood desirability, school districts, or other community features.

The SHAP values for these geospatial features exhibit positive and negative impacts, highlighting the model's ability to discern location-based pricing disparities. These results collectively underscore the importance of geospatial features in property valuation models, corroborating with the studies of Rampini (2021) [30], Rico-Juan et al. [95], and Tchuente et al. (2022) [32]. They also reveal how such features can capture broad regional trends and very localized market conditions mentioned by Chen et al. (2023) [39], providing a multifaceted view of the real estate market.

In terms of real estate market analysis, these insights can be instrumental. Market segmentation can be refined using these geospatial features, allowing for targeted marketing and development strategies. For investors, these data points can highlight potential areas of growth or decline. Urban planners may also use these insights to strategically allocate resources, identify conservation regions, or plan infrastructure to enhance property values.

The SHAP analysis confirms the central role of location in determining real estate prices, with specific geographic coordinates and ZIP codes serving as proxies for many factors that influence desirability and value. Using tools like SHAP, machine learning models provide a powerful means to uncover these complex relationships, offering granular insights that can inform strategic decision making in the real estate market. However, for stakeholders in real estate, it is important to interpret these results cautiously, as correlation does not equate to causation, and overfitting remains a concern with highly granular features. Nonetheless, these results can be applied with a nuanced understanding of the market's complexity.

4.3. Energy Performance Certificate

The energy performance certificate (EPC) emerges as a significant predictor in property valuation, ranking as the third-most influential feature after PGA and Longitude, as evidenced by its substantial mean SHAP value of 15,344.33. This value underscores the significant role that energy efficiency plays in property valuation models. The SHAP beeswarm plot reinforces this, displaying a concentration of high positive SHAP values for properties with superior energy efficiency, which corresponds to lower numerical EPC ratings. This cluster of positive values on the plot illustrates a market trend where higher energy efficiency is often rewarded with higher property valuations.

However, the plot also exhibits some negative SHAP values for EPC, although these seem to be fewer and less extreme than the positive ones. These negative values represent instances where lower energy efficiency or higher numerical EPC ratings are linked to a reduction in the predicted price. This data pattern portrays a more complex real estate market where energy efficiency is valued but not consistently across all properties. Such inconsistencies may arise from varying buyer awareness or preferences concerning energy-efficient features.

The distribution and concentration of the SHAP values depicted in the plot capture the EPC's variable influence on property prices. The dense clustering of positive SHAP values suggests a clear correlation recognized by the model—higher energy efficiency commonly leads to higher property prices. In contrast, the scattered and less dense negative SHAP values hint at a less consistent pattern where properties with poorer energy performance are sometimes, but not consistently, associated with lower prices. This variability indicates that while energy efficiency is a crucial factor in property valuation, the extent to which it affects prices can vary, reflecting the dynamic nature of market perceptions and priorities related to energy efficiency in homes.

Previous studies from Guliker et al. (2022) [65], Lenaers et al. (2023) [44], and Rampini et al. (2021) [30] confirm the importance of EPC in the valuation process, and our results are aligned with these work's findings, further demonstrating that properties with higher EPC ratings tend to have a positive effect on predicted prices, reinforcing the value placed on energy efficiency in the housing market.

The significant SHAP values associated with EPC may also indicate broader trends in the real estate market. They suggest that energy efficiency is valued and could serve as a differentiator in the market, influencing buyer preferences and property values. This finding can have implications for various stakeholders: (a) sellers might be motivated to invest in energy efficiency improvements to enhance property appeal and justify higher prices; (b) buyers may consider the EPC rating a critical factor in their decision making, valuing the long-term cost savings and environmental impact; and (c) policymakers and regulators could use these insights to promote sustainability in housing via incentives or regulations.

4.4. Housing Characteristics

Several nuanced insights valuable for real estate market analysis and predictive modeling can be derived by analyzing the SHAP plots in Figure 4 concerning the housing characteristics—specifically, the Bedroom and Property Condition (PropCond) features.

The beeswarm plot reveals that the number of bedrooms (Bedroom feature) generally positively impacts the predicted property prices. The distribution of SHAP values for this feature is spread out, suggesting that additional bedrooms tend to increase the dwelling's price, but the degree of this increase varies significantly from property to property. This variability may be due to interactions with other influential features, such as the size of the property (PGA) or its location. It also hints at a possible diminishing return on value as the bedroom count increases, which would be an essential consideration for market segmentation and pricing strategies.

The analysis of SHAP values highlighting the Bedroom feature significance aligns well with Abidoye et al.'s (2018) [96] research, which established a statistically significant positive relationship between the number of bedrooms and property value. This correlation underlines the market's valuation of space and utility, particularly in how bedrooms cater to the needs of families or individuals seeking more living space. Furthermore, Bauer et al. (2023) [37]'s study reinforces this perspective by ranking bedrooms as an upper mid-range feature in their SHAP analysis. This insight suggests that bedrooms reflect a property's practical utility and significantly influence its market appeal. The consistency of these findings across different studies validates the conclusion that bedrooms are a critical, flexible indicator of a property's utility and attractiveness to buyers in the real estate market.

For stakeholders in the real estate market, these insights could be instrumental in guiding development priorities, investment choices, and targeted marketing campaigns.

In contrast, the Property Condition (PropCond) feature exhibits a more clustered distribution of positive SHAP values, especially for new properties. This clustering indicates a more consistent market expectation that new dwellings command a premium, likely due to lower anticipated maintenance costs and contemporary amenities. Whether a dwelling is new or used also affects its price, with new properties generally commanding higher prices, reflected by the average SHAP value for this feature. However, the Global Summary plot indicates that while PropCond is significant, its average impact on the model's output is less than that of the number of bedrooms, suggesting that while buyers value newness, the size and utility of the property play a more crucial role in determining price.

The consistent value attributed to the PropCond feature, particularly for newer dwellings, aligns with Morano et al. (2021) [97] findings, establishing a direct functional correlation between property condition and selling price. This insight reinforces the observed market trend where buyers are willing to pay a premium for new construction. This effect could be due to the modern design and amenities, energy efficiency, or reduced need for immediate repairs and renovations. Teoh et al.'s (2022) [98] study further supports this by identifying OverallQual, which rates a house's overall material and finish as a significant determinant of housing prices, indicating that buyers value not just the age of a property but its overall quality and condition, which are often superior in newer constructions. Additionally, Rico-Juan et al.'s (2021) [95] research further corroborates this viewpoint, highlighting that the accumulation of individual absolute Shapley values indicates that the age of a property, closely related to its condition, is a critical factor in market valuation. Real estate developers and investors can use these insights to inform construction and development decisions, ensuring that new properties meet the market demand for quality and modernity.

4.5. Governance

From the SHAP plots in Figure 4, we can analyze the impact of governance-related features such as construction permits (ConstrPermits), Safety, and Security on real estate prices, gain valuable insights into their significance within the model-based perspective and, by extension, their potential implications in the real estate market. The SHAP analysis reveals that governance features like construction permits, safety, and security are significant but not predominant factors in the model's predictions. Their impact on the real estate market can be multifaceted.

The SHAP values for construction permits (ConstrPermits) show a central concentration with a moderate spread, indicating that this feature has a variable impact on price predictions. This effect suggests that construction permits moderate influence on price predictions in some instances, while they may have a limited effect in others. This fact is corroborated by the study of Zhang et al. (2023) [99] that establishes bidirectional causal relationships between the number of building permits and the housing value index, indicating that this index responds negatively to an increase in building permits in the short term of 4–7 months but positively to an increase in building permits with a lag of 10–12 months. Also, in the work of Chen et al. (2023) [39], the number of new permits is not regarded as a significant factor in their global model results, but its SHAP values in the high-income neighborhoods sub-model reflect it as a top factor in predicting housing values. A higher number of construction permits is typically associated with growth, potentially leading to higher property values. Nevertheless, this could also signal an oversupply that may drive prices down. This dual implication suggests that stakeholders must analyze these metrics within broader economic and development contexts, ensuring an accurate valuation of their investment's potential value, as construction permits reflect market dynamism and can guide strategic development.

The safety-related SHAP values are clustered near the lower mid-range of the impact spectrum, suggesting a more uniform but lesser influence on the model's predictions. While

it is not a dominant factor, its consistent presence indicates it is a feature that the model reliably considers. These insights are aligned with the study by Bazan-Krzywoszanska et al. (2018) [63], where the authors discuss the relative importance of "district safety", ranked as one of the top ten factors in their ANN model, highlighting the importance of safety in the district as a determinant of property value.

Security displays a broader range of SHAP values, indicating a more inconsistent but occasionally significant influence on property price predictions. This pattern indicates that for specific-located properties, the presence of security services is a mid-range price determinant, as the proximity to security services might be a proxy for a low-crime environment, adding to an area's appeal.

In the 2023 study by Cellmer [21], "security" is defined as a variable in the context of points of interest (POIs) linked to public safety comprising elements such as camera surveillance, fire stations, police stations, and prisons. This study contributes to our understanding of the security feature by providing statistical data and assigning a relative importance score to this variable. However, its *p*-value of 0.678 indicates that "security" does not have a statistically significant impact on housing prices in the model used. The lack of a strong correlation between the security factor and housing prices in this research may be attributed to the complexity of combining various distinct elements into a single POI category.

Our SHAP analysis shows that safety and security values can differ significantly based on market type and location. Although our model may not prioritize these features, their impact on property prices could be substantial, especially in markets with more pronounced concerns. Real estate professionals should consider the baselines for safety and security levels in specific areas and how they could synergize with other desirable location attributes to increase property attractiveness.

4.6. Macroeconomic and Financial Indicators

The SHAP plots in Figure 4 provide a general overview of the significance of macroeconomic and financial indicators, including inflation rate, unemployment, gross domestic product, and bank appraisal in the real estate price prediction model and their implications in the broader real estate market analysis context.

The bank appraisal (BankEval), featuring a mean SHAP value of 8445.01, is highlighted as a top-10 predictor of property prices in the prediction model. This substantial value indicates that the financial sector's appraisals influence real estate market prices significantly. As an open data feature, BankEval adds a layer of transparency to the prediction process, offering a standard benchmark for market participants. The SHAP values for BankEval exhibit a wide range, with a notable red presence, which points to a general market trend where higher bank valuations are often in step with higher property market prices.

The distribution of SHAP values for BankEval reveals a complex pattern: while there is a clear tendency for higher bank evaluations to elevate property prices, the model also recognizes numerous instances where this is not the case. These exceptions, some significant outliers, suggest scenarios where bank valuations may not align with the final sale prices, possibly due to unique property features, local market conditions, or divergences in evaluation methodologies. Deppner et al.'s (2023) [100] research highlights the complex nature of appraisal errors, revealing how differences between appraised values and actual sale prices can arise due to market changes and valuation biases. This insight resonates with our findings, where BankEval is a significant but varied predictor. While bank appraisals typically reflect market trends, they are subject to inconsistencies.

The Gross Domestic Product (GDP) shows a consistent positive impact, illustrating the connection between economic growth and real estate valuation. As GDP increases, signaling economic strength, there is typically an uptick in property prices due to heightened demand and investment capability. This trend reflects the fundamental role of economic health in shaping the real estate market's trajectory.

The observed relationship in the model between property prices and the Consumer Price Index (CPI) supports the idea of real estate being an effective hedge against inflation. As CPI rises, indicating inflation, property values also tend to increase, suggesting that real estate can be a stable investment in times of currency devaluation.

The unexpected positive SHAP values associated with the Unemployment Rate (UnempRate) warrant a nuanced interpretation. This effect could reflect a non-linear relationship or interactions with other variables that the model captures. For instance, these values may reflect specific economic contexts where real estate markets are resilient to unemployment, such as areas with a high proportion of non-working wealthy residents or strong social safety nets.

The SHAP analysis reveals that macroeconomic and financial indicators correlate positively with property prices. Nevertheless, their effect can differ significantly across individual predictions, thus providing a valuable perspective on the predictive significance of macroeconomic and financial indicators within our model. While the general positive associations align with economic theory, the real-world implications of these findings should be contextualized within the broader economic landscape, considering the intricate interplay of local conditions, market sentiment, and economic policy.

Following this line of reasoning, the discussion on property prices is enriched by integrating the findings from Abidoye et al. (2019) [33], Vaidynathan et al. (2023) [101], and the SHAP analysis, each contributing unique perspectives on the determinants of real estate valuation. Abidoye's study, focusing on Hong Kong's property market, and Vaidynathan's research on the US housing market underscore the traditional economic view that GDP, CPI, and unemployment rates are key predictors of housing prices. They highlight how a strong GDP, low CPI, and unemployment can increase demand and property prices. In contrast, our SHAP analysis offers a more nuanced approach, suggesting that other factors, such as geographical location, PGA, and EPC ratings, may substantially impact a specific predictive model. This divergence suggests that while macroeconomic trends provide a general framework for understanding property market movements, specific market conditions or unique dataset features can lead to different influencing factors emerging as more significant.

Therefore, a comprehensive understanding of property prices must consider the interplay between these broader economic indicators and the more detailed localized factors insights that predictive analytics provide. This approach enables stakeholders to comprehend the dynamics in the real estate market with a broader lens, incorporating established economic principles and detailed, context-specific analysis to inform strategic planning and decision making.

4.7. Mobility

In analyzing the SHAP plots in Figure 4 of our ML model for predicting housing prices, we can derive insights about the relative importance of mobility features (Bus, Subway, and Train), their specific impact on predictions, and their implications in real estate market analysis context.

The SHAP beeswarm plot, which visualizes the distribution of each feature's impact on the model output, shows that proximity to subway stations (Subway) is a significant predictor of real estate prices, as indicated by the broader spread and higher positioning of its SHAP values. This effect suggests that properties closer to subway stations are generally valued higher, likely due to the convenience and efficiency of subway transport in urban settings. The higher mean absolute SHAP value for this feature suggests that properties near subway stations are often at a premium. This pattern reflects a broader market trend where homebuyers value the convenience of rapid transportation, which can be particularly appealing in dense urban areas where traffic congestion is a concern. The positive correlation with property prices might also indicate areas subject to gentrification or urban development. Proximity to subway stations often spurs investment and can catalyze neighborhood revitalization, increasing property values.

Our results on the influence of subway transportation on real estate values parallel the findings of Cárdenas et al. (2023) [55]. Their study reveals how transport infrastructure significantly impacts real estate values, showing a 5.2% to 10.5% increase in flat prices within a 1–1.5 km radius of future subway stations, affirming the premium placed on subway accessibility in urban property markets. These insights are particularly relevant to our study, demonstrating the tangible value added by proximity to main transport links, a factor crucial in shaping real estate trends and urban development strategies.

On the other hand, nearby bus stops (Bus) seem to have a negligible effect on pricing. The SHAP values for Bus are tightly clustered around zero, indicating that this feature does not substantially sway the model's predictions. This pattern could be attributed to the widespread availability of bus stops, making them less of a differentiating factor for property prices. The lower impact of Bus on property prices relative to Subway and Train may suggest that while bus routes are essential for basic accessibility, they do not add as much premium value to properties as access to faster and more reliable modes of transportation does.

Given the lower SHAP values, there might be more flexibility in the valuation of properties based on bus access, which could benefit more location-sensitive market segments. This finding aligns interestingly with Liu et al.'s (2018) [102] study, where the Jiangbei submarket, on the city's outskirts, showed significant influence from bus stops. This pattern suggests a nuanced landscape within the city, where bus stops' impact on real estate values might be more pronounced in less central areas like Jiangbei. Such variations highlight the importance of considering specific urban contexts when assessing the influence of transportation infrastructure on property valuation.

The proximity to train stations (Train) also positively influences property values, albeit somewhat less than subway stations. The SHAP values for the Train feature are spread out but less so than for Subway, reflecting a moderately positive effect on price. This finding validates previous research by Lenaers and De Moor (2023) [44] and implies that while train station accessibility to regional or national rail networks is a valued asset, it may not be as critical for daily commuting within the city as subway access is. The SHAP beeswarm plot also shows positive and negative impacts, and it could reflect the dual nature of train stations as hubs for opportunity and potential nuisances. While they offer connectivity, they can also bring noise, which some buyers might find undesirable.

Overall, subway proximity emerges as a premium attribute in urban real estate, overshadowing the influence of bus stops and somewhat outpacing train station access. The findings from SHAP values reinforce the importance of transportation infrastructure in real estate valuation. Properties with better access to public transportation can command higher prices, and this factor can often outweigh other considerations like property features and amenities.

Gleaned from SHAP values, these insights have implications for various real estate stakeholders. For urban planners, the data underscores the value of investing in rapid transportation infrastructure. Real estate investors and developers might focus on properties near subways for higher returns. Agents can leverage these insights in their sales strategies by emphasizing properties with advantageous transportation links.

However, while SHAP values provide a nuanced view of feature importance, they represent the model's internal reasoning and are not causal explanations. External factors such as demographic shifts, urban development policies, and cultural trends must be considered when translating these findings into market strategies, emphasizing the need for a broader and more integrated approach to understanding and leveraging these dynamics in real estate valuations.

4.8. Quality of Life and Well-Being

The SHAP summary plots in Figure 4 provide nuanced insights into the predictive power of different features within our XGBoost model for real estate pricing. Focusing on the quality of life and well-being features—Culture, Commerce, Education (Schools

and University), Health, Leisure (Sports), and Environment (Parks and Trees)—we can understand their significance in the context of the broader real estate market.

Starting with Culture, the beeswarm plot shows a broad spread of SHAP values, indicating that proximity to cultural amenities significantly impacts property values. However, the varied range of impact across the dataset suggests that not all cultural amenities are valued equally. The dense clusters at higher SHAP values point to a non-linear relationship, suggesting that specific cultural facilities may carry more weight in the valuation process.

The Commerce feature shows a moderate average SHAP value with a concentrated cluster of points. This pattern indicates that while the distance to shopping facilities is generally valued, its impact on property prices is consistent and less variable than Culture. It suggests a baseline value attributed to convenience but implies that additional proximity to commerce does not significantly increase property value at a certain point.

Education-related features (School and University) have a positive impact, yet the SHAP values spread is narrower than for Culture. This effect indicates a more uniform valuation of educational proximity, which could be tied to the consistent demand for accessibility to educational institutions from specific market segments, like families and long-term investors.

While showing a consistent positive effect, the distance to Health facilities has a mid-lower average impact on property values, as evidenced by the SHAP values. The plot reveals a positive yet plateauing effect, which suggests that while health facility proximity is important, it might be a basic expectation in specific markets, beyond which its value does not significantly increase.

Leisure (Sports) facilities exhibit a range of SHAP values, but overall, their impact on property prices is more muted than other features. The beeswarm plot suggests variability, where the significance of sports facilities can be relatively high for some properties but not others, perhaps reflecting personal preferences or the saturation of such amenities in certain areas.

The environmental features, particularly proximity to parks and the presence of Trees, show generally positive SHAP values. However, there is a concentration of points toward the lower end of the impact scale, indicating that while green spaces are beneficial and contribute to property desirability, they may be less influential than the property's size or energy performance.

The Global Summary plot reinforces the perception that quality of life and well-being features, while positively contributing to real estate prices, are not the top predictors. The color intensity in the beeswarm plot, which could correlate with the feature value, suggests that higher or lower values of these features have varying degrees of impact on price predictions.

In synthesizing these observations, we see that features associated with quality of life and well-being are indeed valued in the real estate market, but their influence is often secondary to intrinsic property characteristics such as size and location. For developers and policymakers, this underscores the importance of creating balanced environments that cater to both the physical quality of the living spaces and the well-being of the residents.

Our findings align with the study by Cellmer (2023) [21], which offers insightful observations in discussing the influence of POIs on housing prices. It reveals that a high density of POIs generally acts as a stimulant for housing prices, suggesting that a wide variety of local amenities and services can enhance the attractiveness of a neighborhood. However, the study also underscores the complexity of this relationship, noting that not all POIs contribute positively. These effects highlight the need for a nuanced understanding of how different types of POIs and their specific locations influence real estate values. The research thus challenges the simplistic notion of a direct causal relationship between POI density and housing prices, advocating for a more detailed and contextual analysis in real estate assessments.

It is important to note that SHAP values provide local explanations, and while their aggregate interpretation offers valuable insights, they should be approached with caution

when projecting broader market trends. These findings are contingent on the current market conditions and the model's training data. As the market evolves, reassessment of the impact of these features will be necessary to maintain an accurate understanding of their influence on property valuation.

4.9. Synergy of Proprietary and Open Data in Real Estate Price Prediction

The SHAP heatmap plot from Figure 7 allows us to see which features are most important across all instances and how each feature's value affects the model's predictions for each individual instance. It highlights the multifaceted nature of real estate pricing, where many factors interact in complex ways to influence the final property valuation.

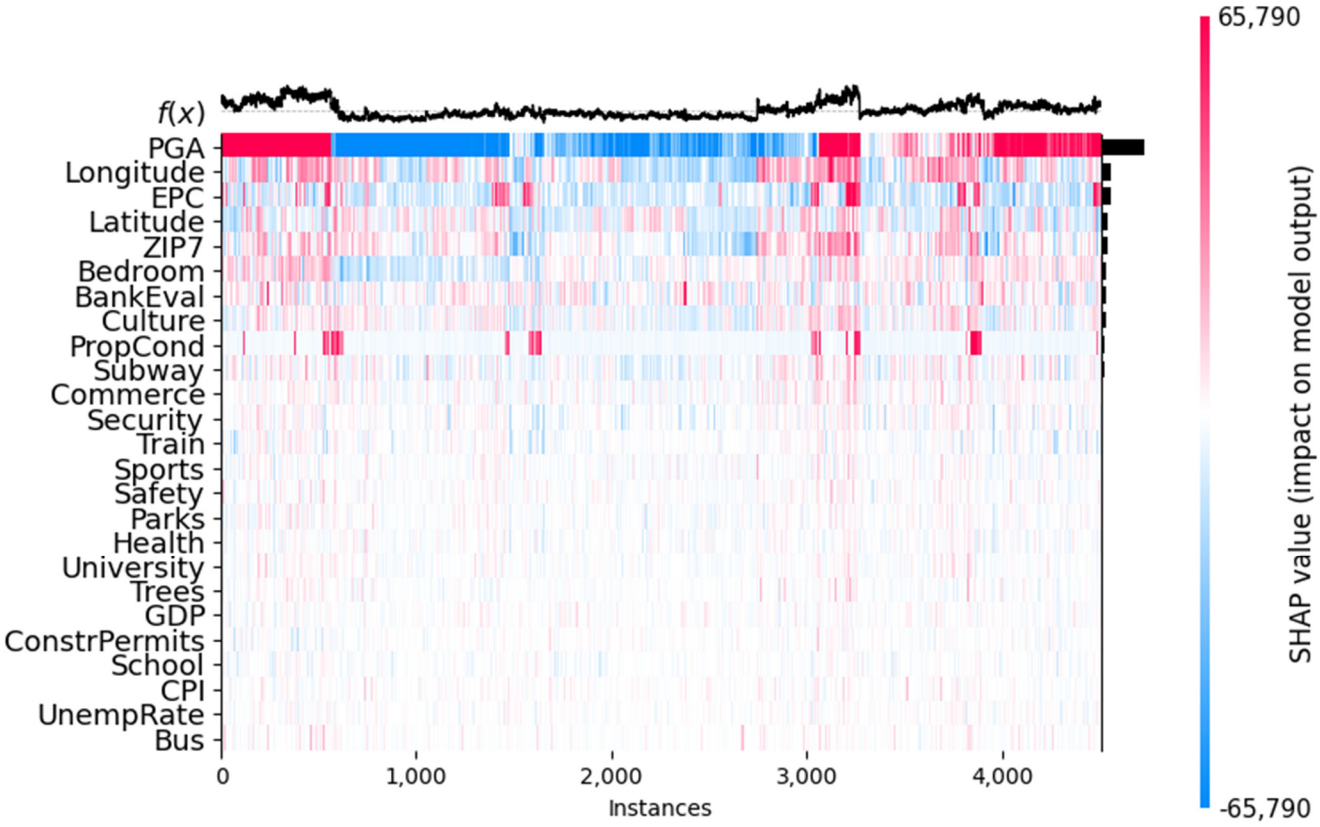


Figure 7. SHAP value heatmap for feature impact analysis in real estate price Prediction: This heatmap displays the distribution of SHAP values for each feature across 4508 instances in the test dataset. Each row represents a feature such as PGA, Longitude, or EPC, and each column corresponds to an instance. The color intensity reflects the impact of the feature on the model's prediction for that instance, with red indicating a higher positive impact and blue indicating a negative impact. The black line graph at the top illustrates the distribution of prediction outputs ($f(x)$) across all instances. Features are ordered vertically by their overall impact on model output. This visual analysis helps identify which features are most influential and how they vary across different property instances, providing insights into the predictive behavior of the model.

The SHAP heatmap visually represents the dual impact of proprietary and open data on real estate price predictions. Proprietary data, which include specific details about a property like its size (PGA), condition, and exact location (Latitude, Longitude, and ZIP7), directly and significantly impact price predictions due to the close relation to the property's inherent qualities. For example, PGA's color intensity and spread underscore its role as a primary valuation factor, reflecting the market's emphasis on property size. These factors are primary considerations for buyers and sellers, reflected in their strong influence as indicated by SHAP values in the predictive model.

Open data, conversely, offers essential contextual information encompassing socioeconomic, environmental, and infrastructural aspects of a property's surroundings and displays a more varied influence across the heatmap. This information includes accessibility to cultural amenities, public transportation, and economic indicators. The shades and distribution of colors for features like BankEval and proximity to transportation reveal their conditional importance. These features may not always exhibit a strong, direct impact like the proprietary data, but they are critical for capturing the nuanced interplay of external factors contributing to a property's market value. For instance, the impact of proximity to cultural amenities varies widely across the dataset, as seen in the SHAP heatmap, suggesting that cultural access is a significant value driver in this market.

The SHAP analysis also highlights the synergy between these two data categories. While proprietary data points to the inherent value based on property specifics, open data encapsulates external factors that can either amplify or diminish this inherent value. Stakeholders such as investors and urban planners can leverage this insight for targeted interventions; for example, improving public transportation links might increase property values in a given area, an insight that can be inferred from the model's sensitivity to these features.

Furthermore, the heatmap's patterns reveal how integrating proprietary and open data in the model enables a more detailed real estate market segmentation. The varying SHAP values across properties for open data features suggest that the importance of specific amenities can differ significantly depending on the property's location and target demographic.

The model achieves a more holistic understanding of property valuation by integrating these data types. Proprietary data provide detailed insights into the property, while open data adds breadth, reflecting the property's broader context and external influences. This combination allows for more accurate and nuanced real estate price predictions, underscoring the importance of direct property attributes and broader environmental and economic factors in the valuation process and offering actionable insights for market participants.

Integrating open data in real estate price prediction is a theme that is present within several contemporary studies, each offering unique insights into this emerging field. Alvarez et al. (2022) [103] emphasized the role of incremental learning models in adapting to dynamic urban data for real estate valuation, while Cellmer (2023) [21] focused on how points of interest derived from open data sources affect housing prices. Adje et al. (2023) [15] broadened the scope by highlighting the role of open data in smart city contexts, including real estate price estimation. Tchuente and Nyawa (2022) [32] leveraged open datasets to understand the impact of location features in French cities, emphasizing granularity in data analysis. Karamanou et al. (2022) [11] combined Open Government Data with machine learning to comprehensively view the Scottish real estate market. Tsagkis et al. (2022) [16] demonstrated the utility of open data in understanding urban growth and its influence on property prices, and Hurbean et al. (2021) [12] showed how Open Government Data, enhanced by AI technologies, leads to more accurate house price predictions.

Together, these studies underscore the growing importance of open data in offering nuanced, adaptable, and transparent approaches to real estate price prediction and urban planning. The results of our research dovetail with the emerging narrative in the field, emphasizing the critical role of XAI and open data in augmenting the precision and transparency of AI applications in urban development. Our findings align with contemporary studies and extend their insights by demonstrating the tangible benefits of these technologies in making more informed and sustainable policy decisions for smart city development.

5. Conclusions

This study made significant strides in applying AI for real estate price prediction within smart cities, particularly by leveraging proprietary and open data synergy. Our findings reveal that incorporating open data about socio-economic indicators and infrastructure,

such as public transport links and cultural facilities, plays a pivotal role in enhancing the accuracy of an XGBoost model, as evidenced by the 8.24% improvement in the MAE.

The application of SHapley Additive exPlanations (SHAP) has provided transparency into the model's decision making, elucidating the weighted significance of each variable. Proprietary data on a property's size and exact location have emerged as critical predictors of real estate values. Meanwhile, open data variables like accessibility to amenities and economic health indicators have been instrumental in capturing the contextual nuances of the property's environment.

The practical implications of these results are far-reaching for urban planners and policymakers. Integrating diverse data sources can refine urban development strategies, ensuring they are grounded in a comprehensive understanding of the factors driving real estate values. This approach can support informed decision making that promotes sustainable and equitable growth, allowing for more targeted and effective urban planning policies that cope with the evolving needs and trends of the real estate market. Furthermore, our conclusions emphasize the importance of transparency in AI-driven analytics. Tools such as SHAP heatmaps increase model accountability and serve as a bridge in communication with non-technical stakeholders, making complex AI assessments more accessible and understandable.

However, the study acknowledges certain limitations, including the model's reliance on the quality and availability of open data, which can vary significantly across different contexts. Those data limitations could be subject to speculation about potential biases inherent in both our proprietary real estate data and the open data about points of interest (POIs) and macroeconomic indicators. For instance, proprietary real estate data may be biased toward properties listed on a specific platform and might not reflect the market as a whole. Open data on POIs could exhibit a geographical bias, being more detailed in urban areas than peri-urban ones, potentially skewing the model's performance across different regions. Also, the absence of detailed socioeconomic and demographic data might limit the model's ability to accurately predict prices in diverse communities. For example, the model might not account for the impact of gentrification, changes in neighborhood demographics, or residents' economic mobility, which can significantly affect real estate values and might affect the generalizability of our findings.

Furthermore, changes in housing policy, zoning laws, and real estate regulations could affect market prices. Such changes can have significant, sometimes rapid, impacts on real estate values and market dynamics, aspects that our model may not entirely capture. Therefore, further refinement of our model may be necessary to reflect these impacts accurately.

The study also has temporal limitations. The data gathered from 2018 to 2021 provide a snapshot of the market during this period. Longer-term trends and cyclical market behaviors may not be fully captured within this timeframe. Additionally, this timeframe may not reflect the impact of rapid technological advancements and changing market dynamics on real estate prices. For instance, emerging trends such as remote work could alter the attractiveness of different locations after 2021, potentially shifting demand and prices in ways our model might not accurately foresee.

By advancing the field of XAI and highlighting the value of open data, our study contributes to the responsible and informed use of AI in developing smart cities. It calls for a collaborative approach among data scientists, urban developers, and policymakers to foster intelligent, efficient, transparent, and inclusive smart city ecosystems. Our research demonstrates the innovation potential of AI in urban planning, suggesting a future where AI-driven insights are integral to crafting policies that reflect the dynamic interplay among economic, social, and environmental factors within urban landscapes.

Integrating XAI and open data is a technical enhancement and a paradigm shift toward more informed and democratic urban governance. It is hoped that the methodologies and insights from this study will inspire further innovations in the realm of smart city planning,

encouraging a data-informed approach to creating cities that are resilient, adaptive, and attuned to the needs of their inhabitants.

The following steps in research should focus on enhancing the generalizability of our model. These future steps involve extensive testing and validation across diverse markets and regions beyond Lisbon to ensure the model's broad applicability. Tailoring the model to account for the unique characteristics, regulatory environments, and economic conditions of different markets will be essential. This process will evaluate the model's generalizability and ensure it can be adapted to reflect the complexities of various urban landscapes.

Future paths of research could explore the integration of additional data types and sources, incorporating environmental quality indices, detailed socioeconomic and demographic information, image data, real-time data streams, such as social media sentiment, immediate market changes, and traffic patterns, among others, to assess their potential to improve our model's accuracy and relevance further.

Future research may also introduce seasonality into the model to capture the temporal dynamics that influence market prices, offering a more nuanced understanding of how and why prices fluctuate throughout the year. Also, conducting longitudinal studies to capture the temporal fluctuations in market prices will offer deeper insights into how urban development policies shape urban growth, housing markets, and community development.

The future exploration of probabilistic AI models that have been applied in other predictive modeling domains presents a logical progression to build upon the foundational insights we have garnered from our current study. Probabilistic AI algorithms could complement the existing framework by providing new insights into uncertainty management and improving the predictive performance of real estate analytics, illustrating their versatility and effectiveness in real estate appraisals and offering valuable insights for urban planning and investment decisions.

Furthermore, future research should delve deeper into the ethical dimensions of AI and data utilization, focusing on safeguarding privacy, ensuring robust data security, and promoting equitable access to AI benefits across all segments of urban populations, while addressing data bias and fairness in model predictions should strengthen the commitment to social inclusivity, preventing the perpetuation of existing inequalities. This direction will help build smart cities that are intelligent, ethically responsible, and socially inclusive.

Author Contributions: Conceptualization, F.T.N., M.A. and M.d.C.N.; methodology, F.T.N., M.A. and M.d.C.N.; software, F.T.N.; validation, F.T.N., M.A. and M.d.C.N.; formal analysis, F.T.N., M.A. and M.d.C.N.; investigation, F.T.N., M.A. and M.d.C.N.; resources, F.T.N. and M.d.C.N.; data curation, F.T.N.; writing—original draft preparation, F.T.N., M.A. and M.d.C.N.; writing—review and editing, F.T.N., M.A. and M.d.C.N.; visualization, F.T.N.; supervision, M.A. and M.d.C.N.; funding acquisition, M.d.C.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by FCT—Fundação para a Ciência e Tecnologia, I.P. (Portugal), under research grant UIDB/04152/2020—Centro de Investigação em Gestão de Informação (MagIC).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy reasons.

Acknowledgments: The authors would like to express their gratitude to Confidencial Imobiliário for graciously providing the real estate transactions data that supported this work.

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

References

1. Allam, Z.; Dhunny, Z.A. On big data, artificial intelligence and smart cities. *Cities* **2019**, *89*, 80–91. [CrossRef]
2. Neves, F.T.; de Castro Neto, M.; Aparicio, M. The impacts of open data initiatives on smart cities: A framework for evaluation and monitoring. *Cities* **2020**, *106*, 102860. [CrossRef]

3. Bibri, S.E.; Alexandre, A.; Sharifi, A.; Krogstie, J. Environmentally sustainable smart cities and their converging AI, IoT, and big data technologies and solutions: An integrated approach to an extensive literature review. *Energy Inform.* **2023**, *6*, 9. [CrossRef]
4. Tekouabou, S.C.; Gherghina, Ş.C.; Kameni, E.D.; Filali, Y.; Gartoumi, K.I. AI-Based on Machine Learning Methods for Urban Real Estate Prediction: A Systematic Survey. *Arch. Comput. Methods Eng.* **2023**, *31*, 1079–1095. [CrossRef]
5. Costa, C.J.; Aparicio, M. Applications of Data Science and Artificial Intelligence. *Appl. Sci.* **2023**, *13*, 9015. [CrossRef]
6. Yigitcanlar, T.; Corchado, J.M.; Mehmood, R.; Li, R.Y.M.; Mossberger, K.; Desouza, K. Responsible urban innovation with local government artificial intelligence (AI): A conceptual framework and research agenda. *J. Open Innov. Technol. Mark. Complex.* **2021**, *7*, 71. [CrossRef]
7. Koseki, S.; Jameson, S.; Farnadi, G.; Rolnick, D.; Régis, C.; Denis, J.-L. *AI and Cities: Risks, Applications, and Governance*; UN-Habitat: Nairobi, Kenya, 2022.
8. Herath, H.; Mittal, M. Adoption of artificial intelligence in smart cities: A comprehensive review. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100076. [CrossRef]
9. Zhang, Y.-M.; Wang, H. Multi-head attention-based probabilistic CNN-BiLSTM for day-ahead wind speed forecasting. *Energy* **2023**, *278*, 127865. [CrossRef]
10. Zhang, Y.-M.; Wang, H.; Mao, J.-X.; Xu, Z.-D.; Zhang, Y.-F. Probabilistic framework with bayesian optimization for predicting typhoon-induced dynamic responses of a long-span bridge. *J. Struct. Eng.* **2021**, *147*, 04020297. [CrossRef]
11. Karamanou, A.; Kalampokis, E.; Tarabanis, K. Linked open government data to predict and explain house prices: The case of Scottish statistics portal. *Big. Data Res.* **2022**, *30*, 100355. [CrossRef]
12. Hurbean, L.; Danaiata, D.; Militaru, F.; Dodea, A.-M.; Negovan, A.-M. Open data based machine learning applications in smart cities: A systematic literature review. *Electronics* **2021**, *10*, 2997. [CrossRef]
13. Goodey, G.; Hahnel, M.; Zhou, Y.; Jiang, L.; Chandramouliswaran, I.; Hafez, A.; Paine, T.; Gregurick, S.; Simango, S.; Palma Peña, J.M.; et al. *The State of Open Data 2022*; Digital Science Report; Digital Science: London, UK, 2022.
14. Davies, T.; Walker, S.B.; Rubinstein, M.; Perini, F. *The State of Open Data: Histories and Horizons*; African Minds: Cape Town, South Africa, 2019.
15. Adje, K.D.C.; Letaifa, A.B.; Haddad, M.; Habachi, O. Smart City Based on Open Data: A Survey. *IEEE Access* **2023**, *11*, 56726–56748. [CrossRef]
16. Tsagkis, P.; Bakogiannis, E.; Nikitas, A. Analysing Urban Growth Using Machine Learning and Open Data: An Artificial Neural Network Modelled Case Study of Five Greek Cities. *Sustain. Cities Soc.* **2023**, *89*, 104337. [CrossRef]
17. Pašalić, I.N.; Ćukušić, M.; Jadrić, M. Smart city research advances in Southeast Europe. *Int. J. Inf. Manag.* **2021**, *58*, 102127. [CrossRef]
18. Radchenko, K. The economic and social impacts of smart cities: Multi-stakeholder pre-study results. *Smart Cities Reg. Dev. (SCRD) J.* **2023**, *7*, 25–38. [CrossRef]
19. Jonek-Kowalska, I. Housing Infrastructure as a Determinant of Quality of Life in Selected Polish Smart Cities. *Smart Cities* **2022**, *5*, 924–946. [CrossRef]
20. Gutman, S.; Rytova, E. Indicators for assessing the development of smart sustainable cities. In Proceedings of the International Scientific Conference on Innovations in Digital Economy, Saint-Petersburg, Russia, 24–25 October 2019; pp. 55–73.
21. Cellmer, R. Points of Interest and Housing Prices. *Real Estate Manag. Valuat.* **2023**, *31*, 69–77. [CrossRef]
22. Nijskens, R.; Lohuis, M.; Hilbers, P.; Heeringa, W. *Hot Property: The Housing Market in Major Cities*; Springer Nature: Cham, Switzerland, 2019.
23. Shin, H.-S.; Woo, A. Analyzing the effects of walkable environments on nearby commercial property values based on deep learning approaches. *Cities* **2024**, *144*, 104628. [CrossRef]
24. Garcês, P.; Pires, C.P.; Costa, J.; Jorge, S.F.; Catalão-Lopes, M.; Alventosa, A. Disentangling Housing Supply to Shift towards Smart Cities: Analysing Theoretical and Empirical Studies. *Smart Cities* **2022**, *5*, 1488–1507. [CrossRef]
25. Butryn, K.; Jasińska, E.; Kovalyshyn, O.; Preweda, E. Sustainable formation of urban development on the example of the primary real estate market in Krakow. *E3S Web Conf.* **2019**, *86*, 00010. [CrossRef]
26. Murray, C.K. A housing supply absorption rate equation. *J. Real Estate Financ. Econ.* **2022**, *64*, 228–246. [CrossRef]
27. Xu, X.; Zhang, Y. House price forecasting with neural networks. *Intell. Syst. Appl.* **2021**, *12*, 200052. [CrossRef]
28. Chollet, F. *Deep Learning with Python*; Simon and Schuster: New York, NY, USA, 2021.
29. Sagi, A.; Gal, A.; Czamanski, D.; Broitman, D. Uncovering the shape of neighborhoods: Harnessing data analytics for a smart governance of urban areas. *J. Urban Manag.* **2022**, *11*, 178–187. [CrossRef]
30. Rampini, L.; Cecconi, F.R. Artificial intelligence algorithms to predict Italian real estate market prices. *J. Prop. Invest. Financ.* **2021**, *40*, 588–611. [CrossRef]
31. Ho, W.K.; Tang, B.-S.; Wong, S.W. Predicting property prices with machine learning algorithms. *J. Prop. Res.* **2021**, *38*, 48–70. [CrossRef]
32. Tchuente, D.; Nyawa, S. Real estate price estimation in French cities using geocoding and machine learning. *Ann. Oper. Res.* **2022**, *308*, 571–608. [CrossRef]
33. Abidoye, R.B.; Chan, A.P.C.; Abidoye, F.A.; Oshodi, O.S. Predicting property price index using artificial intelligence techniques Evidence from Hong Kong. *Int. J. Hous. Mark. Anal.* **2019**, *12*, 1072–1092. [CrossRef]

34. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
35. Lenaers, I.; Boudt, K.; De Moor, L. Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques. *Int. J. Hous. Mark. Anal.* **2024**, *17*, 96–113. [CrossRef]
36. Li, Z. Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Comput. Environ. Urban Syst.* **2022**, *96*, 101845. [CrossRef]
37. Baur, K.; Rosenfelder, M.; Lutz, B. Automated real estate valuation with machine learning models using property descriptions. *Expert Syst. Appl.* **2023**, *213*, 119147. [CrossRef]
38. Javed, A.R.; Ahmed, W.; Pandya, S.; Maddikunta, P.K.R.; Alazab, M.; Gadekallu, T.R. A survey of explainable artificial intelligence for smart cities. *Electronics* **2023**, *12*, 1020. [CrossRef]
39. Chen, Y.; Jiao, J.; Farahi, A. Disparities in affecting factors of housing price: A machine learning approach to the effects of housing status, public transit, and density factors on single-family housing price. *Cities* **2023**, *140*, 104432. [CrossRef]
40. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; Lulu: Morrisville, NC, USA, 2022.
41. Verhulst, S. *Unlocking the Potential: The Call for an International Decade of Data*; UNU-CPR: New York, NY, USA, 2023. [CrossRef]
42. Vainio-Pekka, H.; Agbese, M.O.-o.; Jantunen, M.; Vakkuri, V.; Mikkonen, T.; Rousi, R.; Abrahamsson, P. The Role of Explainable AI in the Research Field of AI Ethics. *ACM Trans. Interact. Intell. Syst.* **2023**, *13*, 26. [CrossRef]
43. Royal Society. *Explainable AI: The Basics-Policy Briefing*; Royal Society: London, UK, 2019.
44. Lenaers, I.; De Moor, L. Exploring XAI techniques for enhancing model transparency and interpretability in real estate rent prediction: A comparative study. *Financ. Res. Lett.* **2023**, *58*, 104306. [CrossRef]
45. Popelka, S.; Zertuche, L.; Beroche, H. *Urban AI Guide*; Urban AI: Paris, France, 2023.
46. Liu, P.; Zhang, Y.; Biljecki, F. Explainable spatially explicit geospatial artificial intelligence in urban analytics. *Environ. Plan. B Urban Anal. City Sci.* **2023**, *2023*, 1–20. [CrossRef]
47. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
48. Shapley, L.S. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*; Harold William, K., Albert William, T., Eds.; Princeton University Press: Princeton, NJ, USA, 1953; pp. 307–318.
49. 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]
50. Estevens, A.; Cocola-Gant, A.; López-Gay, A.; Pavel, F. The role of the state in the touristification of Lisbon. *Cities* **2023**, *137*, 104275. [CrossRef]
51. Marques, T.S.; Saraiva, M.M.; Matos, F.L.d.; Maia, C.; Ribeiro, D.; Ferreira, M.; Van Heerden, S. *Property Investment and Housing Affordability in Lisbon and Porto*; Publications Office of the European Union, Joint Research Centre (JRC): Luxemburg, 2022. [CrossRef]
52. Samadani, S.; Costa, C.J. Forecasting real estate prices in Portugal: A data science approach. In Proceedings of the 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), Chaves, Portugal, 23–26 June 2021.
53. Ahrend, R.; Bétin, M.; Caldas, M.P.; Cournède, B.; Ramirez, M.D.; Pionnier, P.-A.; Sanchez-Serra, D.; Veneri, P.; Ziemann, V. *Changes in the Geography of Housing Demand after the Onset of COVID-19*; OECD: Paris, France, 2023.
54. Boeing, G.; Higgs, C.; Liu, S.; Giles-Corti, B.; Sallis, J.F.; Cerin, E.; Lowe, M.; Adlakha, D.; Hinckson, E.; Moudon, A.V. Using open data and open-source software to develop spatial indicators of urban design and transport features for achieving healthy and sustainable cities. *Lancet Glob. Health* **2022**, *10*, e907–e918. [CrossRef] [PubMed]
55. Cárdenas, J.; Gallego, J.M.; Urrutia, M.A. Announcement of the First Metro Line and its Impact on Housing Prices in Bogotá. *Case Stud. Transp. Policy* **2022**, *11*, 100941. [CrossRef]
56. Kalliola, J.; Kapociute-Dzikiene, J.; Damasevicius, R. Neural network hyperparameter optimization for prediction of real estate prices in Helsinki. *PeerJ Comput. Sci.* **2021**, *7*, e444. [CrossRef]
57. Shen, H.; Li, L.; Zhu, H.H.; Li, F. A Pricing Model for Urban Rental Housing Based on Convolutional Neural Networks and Spatial Density: A Case Study of Wuhan, China. *Isprs Int. J. Geo-Inf.* **2022**, *11*, 26. [CrossRef]
58. Büchler, S.; Niu, D.; Kinsella Thompson, A. *Predicting Urban Growth with Machine Learning*; MIT Center for Real Estate Research Paper: Cambridge, MA, USA, 2021.
59. Bouwknegt, L.; Rouwendal, J. *The Effect of Urban Trees on House Prices: Evidence from Cut-Down Trees in Amsterdam*; Tinbergen Institute: Amsterdam, The Netherlands, 2023.
60. Sisman, S.; Aydinoglu, A.C. Improving performance of mass real estate valuation through application of the dataset optimization and Spatially Constrained Multivariate Clustering Analysis. *Land Use Policy* **2022**, *119*, 106167. [CrossRef]
61. Yang, L.; Wang, B.; Zhou, J.; Wang, X. Walking accessibility and property prices. *Transp. Res. Part D Transp. Environ.* **2018**, *62*, 551–562. [CrossRef]
62. Gude, V. A multi-level modeling approach for predicting real-estate dynamics. *Int. J. Hous. Mark. Anal.* **2023**, *17*, 48–59. [CrossRef]
63. Bazan-Krzywoszanska, A.; Bereta, M. The use of urban indicators in forecasting a real estate value with the use of deep neural network. *Rep. Geod. Geoinform.* **2018**, *106*, 25–34. [CrossRef]

64. Hacıevliyagil, N.; Drachal, K.; Eksi, I.H. Predicting House Prices Using DMA Method: Evidence from Turkey. *Economies* **2022**, *10*, 64. [CrossRef]
65. Guliker, E.; Folmer, E.; van Sinderen, M. Spatial determinants of real estate appraisals in the Netherlands: A machine learning approach. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 125. [CrossRef]
66. Barcelos, M.; Barcelos, A.; Bernardini, F.; Silva, G.V. Analyzing the use of economic and financial indicators in smart cities context. In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, Delft, The Netherlands, 30 May–1 June 2018; pp. 1–2.
67. Wu, H.-N.; Yin, L.; Zhou, T.; Ye, S. City smart-growth evaluation system. In Proceedings of the 2017 IEEE International Conference on Smart Grid and Smart Cities (ICSGSC), Singapore, 23–26 July 2017; pp. 293–297.
68. García, S.; Luengo, J.; Herrera, F. *Data Preprocessing in Data Mining*; Springer: Cham, Switzerland, 2015; Volume 72.
69. Iglewicz, B.; Hoaglin, D.C. *Volume 16: How to Detect and Handle Outliers*; Quality Press: Seattle, WA, USA, 1993.
70. Monika, R. House price forecasting using machine learning methods. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2021**, *12*, 3624–3632.
71. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
72. El Mouna, L.; Silkan, H.; Haynf, Y.; Nann, M.F.; Tekouabou, S.C. A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms. *E3S Web Conf.* **2023**, *418*, 03001. [CrossRef]
73. Jha, S.B.; Babiceanu, R.F.; Pandey, V.; Jha, R.K. Housing market prediction problem using different machine learning algorithms: A case study. *arXiv* **2020**, arXiv:2006.10092.
74. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
75. Jafar, A.; Lee, M. Comparative Performance Evaluation of State-of-the-Art Hyperparameter Optimization Frameworks. *Trans. Korean Inst. Electr. Eng.* **2023**, *72*, 607–620. [CrossRef]
76. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.
77. Arai, K.; Fujikawa, I.; Nakagawa, Y.; Momozaki, T.; Ogawa, S. Modified Prophet+ Optuna Prediction Method for Sales Estimations. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 58–63. [CrossRef]
78. Lai, J.-P.; Lin, Y.-L.; Lin, H.-C.; Shih, C.-Y.; Wang, Y.-P.; Pai, P.-F. Tree-Based Machine Learning Models with Optuna in Predicting Impedance Values for Circuit Analysis. *Micromachines* **2023**, *14*, 265. [CrossRef] [PubMed]
79. Joy, J.; Selvan, M.P. A comprehensive study on the performance of different Multi-class Classification Algorithms and Hyperparameter Tuning Techniques using Optuna. In Proceedings of the 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS), Kochi, India, 23–25 June 2022; pp. 1–5.
80. Elshewey, A.M. hyOPTGB: An Efficient OPTUNA Hyperparameter Optimization Framework for Hepatitis C Virus (HCV) Disease Prediction in Egypt. *Res. Sq.* **2023**. [CrossRef]
81. Agrawal, T. Optuna and autoML. Hyperparameter Optimization. In *Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*; Springer: Cham, Switzerland, 2021; pp. 109–129.
82. Eimer, T.; Lindauer, M.; Raileanu, R. Hyperparameters in Reinforcement Learning and How To Tune Them. *arXiv* **2023**, arXiv:2306.01324.
83. Shekhar, S.; Bansode, A.; Salim, A. A comparative study of hyper-parameter optimization tools. In Proceedings of the 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, Australia, 8–10 December 2021; pp. 1–6.
84. Steurer, M.; Hill, R.J.; Pfeifer, N. Metrics for evaluating the performance of machine learning based automated valuation models. *J. Prop. Res.* **2021**, *38*, 99–129. [CrossRef]
85. De Castro Neto, M.; De Melo Cartaxo, T. *Algorithmic Cities: A Dystopic or Utopic Future?* Springer International Publishing: Cham, Switzerland, 2021; pp. 59–73.
86. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **2019**, *8*, 832. [CrossRef]
87. Van Lent, M.; Fisher, W.; Mancuso, M. An explainable artificial intelligence system for small-unit tactical behavior. In Proceedings of the National Conference on Artificial Intelligence, San Jose, CA, USA, 25–29 July 2004; pp. 900–907.
88. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
89. Tan, B.; Gan, Z.; Wu, Y. The measurement and early warning of daily financial stability index based on XGBoost and SHAP: Evidence from China. *Expert Syst. Appl.* **2023**, *227*, 120375. [CrossRef]
90. Silva-Aravena, F.; Núñez Delafuente, H.; Gutiérrez-Bahamondes, J.H.; Morales, J. A hybrid algorithm of ML and XAI to prevent breast cancer: A strategy to support decision making. *Cancers* **2023**, *15*, 2443. [CrossRef]
91. Dorosan, M.; Dailisan, D.; Valenzuela, J.F.; Monterola, C. Use of machine learning in understanding transport dynamics of land use and public transportation in a developing city. *Cities* **2024**, *144*, 104587. [CrossRef]

92. Kansal, M.; Singh, P.; Shukla, S.; Srivastava, S. A Comparative Study of Machine Learning Models for House Price Prediction and Analysis in Smart Cities. In Proceedings of the International Conference on Electronic Governance with Emerging Technologies, Poznan, Poland, 11–12 September 2023; pp. 168–184.
93. Iban, M.C. An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants. *Habitat Int.* **2022**, *128*, 102660. [CrossRef]
94. Xu, K.; Nguyen, H. Predicting housing prices and analyzing real estate market in the Chicago suburbs using Machine Learning. *arXiv* **2022**, arXiv:2210.06261. [CrossRef]
95. Rico-Juan, J.R.; de La Paz, P.T. Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Syst. Appl.* **2021**, *171*, 114590. [CrossRef]
96. Abidoye, R.B.; Chan, A.P.C. Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network. *Pac. Rim Prop. Res. J.* **2018**, *24*, 71–83. [CrossRef]
97. Morano, P.; Tajani, F.; Di Liddo, F.; Darò, M. Economic evaluation of the indoor environmental quality of buildings: The noise pollution effects on housing prices in the city of Bari (Italy). *Buildings* **2021**, *11*, 213. [CrossRef]
98. Teoh, E.Z.; Yau, W.-C.; Ong, T.S.; Connie, T. Explainable housing price prediction with determinant analysis. *Int. J. Hous. Mark. Anal.* **2022**, *16*, 1021–1045. [CrossRef]
99. Zhang, X.; Yang, E. Observation of relationship between housing value and the number of building permits in the United States using time series method. *Int. J. Hous. Mark. Anal.* 2023, *ahead-of-print*.
100. Deppner, J.; von Ahlefeldt-Dehn, B.; Beracha, E.; Schaefers, W. Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach. *J. Real Estate Financ. Econ.* **2023**, 1–38. [CrossRef]
101. Vaidynathan, D.; Kayal, P.; Maiti, M. Effects of economic factors on median list and selling prices in the US housing market. *Data Sci. Manag.* **2023**, *6*, 199–207. [CrossRef]
102. Liu, Z.; Yan, S.; Cao, J.; Jin, T.; Tang, J.; Yang, J.; Wang, Q. A Bayesian approach to residential property valuation based on built environment and house characteristics. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 1455–1464.
103. Alvarez, F.; Roman-Rangel, E.; Montiel, L.V. Incremental learning for property price estimation using location-based services and open data. *Eng. Appl. Artif. Intell.* **2022**, *107*, 104513. [CrossRef]