International Journal of
Geo-Information

Article

Location, Location, Location: The Power of Neighborhoods for Apartment Price Predictions Based on Transaction Data

Christopher Kmen [1,2,*], Gerhard Navratil [1] and Ioannis Giannopoulos [1]

[1] Geoinformation Group, TU Wien, 1040 Vienna, Austria; gerhard.navratil@geo.tuwien.ac.at (G.N.); igiannopoulos@geo.tuwien.ac.at (I.G.)
[2] OTTO Immobilien GmbH, 1010 Vienna, Austria
[*] Correspondence: christopher.kmen@tuwien.ac.at

Abstract: Land and real estate have long been regarded as stable investments, with property prices steadily rising, underscoring the need for accurate predictive models to capture the varying rates of price growth across different locations. This study leverages a decade-long dataset of 83,527 apartment transactions in Vienna, Austria, to train machine learning models using XGBoost. Unlike most prior research, the extended time span of the dataset enables predictions for multiple future years, providing a more robust long-term prediction. The primary objective is to examine how spatial factors can enhance real estate price predictions. In addition to transaction data, sociodemographic and geographic variables were collected to characterize the neighborhoods surrounding each apartment. Ten models, each varying in the number of input years, were trained to predict the price per square meter. The model performance was assessed using the mean absolute percentage error (MAPE), offering insights into their predictive accuracy for both short-term and long-term predictions. This study underscores the importance of distinguishing between newly built and existing apartments in real estate price modeling. By splitting the dataset prior to training, predictive models focusing solely on newly built properties achieved an average reduction of about 6% in MAPE. The best-performing models achieved an average MAPE of 15% for one-year-ahead predictions and maintained a MAPE below 20% for predictions up to three years ahead, demonstrating the effectiveness of leveraging spatial features to enhance real estate price prediction accuracy.

Keywords: real estate price prediction; apartment; transaction data; time-related; machine learning algorithms; XGBoost

check for updates

Citation: Kmen, C.; Navratil, G.; Giannopoulos, I. Location, Location, Location: The Power of Neighborhoods for Apartment Price Predictions Based on Transaction Data. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 425. https://doi.org/10.3390/ijgi13120425

Academic Editor: Wolfgang Kainz

Received: 1 October 2024
Revised: 11 November 2024
Accepted: 14 November 2024
Published: 27 November 2024

1. Introduction

Real estate property, a crucial asset in our society, serves not only for housing but also as a significant investment aimed at monetary appreciation. In Austria, the increase in prices for single-family houses and apartments has significantly outpaced other economic indicators such as inflation or GDP over the past decade [1], highlighting the critical need for focused research in this area. The Residential Property Index (RPI), a key indicator of economic and cyclical trends in Europe, tracks both new and used residential properties within Austria. It showed an increase of 12.3% in the overall value of real estate and an increase of 15.5% for apartments in 2021 compared to the previous year [2]. This annual fluctuation, influenced by factors such as demand, availability, and attractiveness compared to other investments, underscores the dynamic nature of the real estate market. Current research often focuses highly on the internal characteristics of properties, such as room and bathroom counts. However, the variability in housing prices within a city, even among similar properties, suggests that spatial features significantly influence price formation. This study aims to address this by investigating the potential of developing predictive models using a broad array of spatial features derived through feature engineering and urban computing.

In housing price research, it is crucial to distinguish between the data sources used for modeling. In Austria, where property rights are only legally recognized when registered in the land registry, each property transaction is thoroughly documented. This makes the land registry a public and highly trusted source, with its data being reliable due to rigorous pre-registration checks. On the contrary, online listings from real estate agents or private individuals serve as another data source, although these often lack location information [3] and show significant price variations. A study by *willhaben* (willhaben.at is an online marketplace in Austria that also lists one of the largest offers for real estate) and IMMOunited (IMMOUnited is an Austrian data provider for transaction data within Austria) analyzed these discrepancies, finding up to a 21% difference between the listed offer prices on *willhaben* and actual transaction prices in 2020 [4]. This disparity highlights that actual transaction data, with fewer inaccuracies than initial offers, provide a more accurate reflection of the market dynamics.

Most current research on real estate in connection with purchase prices focuses mainly on the prediction of houses. Little can be found in the area of apartments, especially in Central Europe. For example, in Germany, access to land registry is regulated by Section 12 (1) of the German land registry code (GBO), stating that *"there must be a legitimate interest on the part of the interested party"* (Grundbuchordnung §12(1): "Die Einsicht des Grundbuchs ist jedem gestattet, der ein berechtigtes Interesse darlegt. Das gleiche gilt von Urkunden, auf die im Grundbuch zur Ergänzung einer Eintragung Bezug genommen ist, sowie von den noch nicht erledigten Eintragungsanträgen."). Unfortunately, research does not count as one of these legitimate interests. In Austria, where the transaction details are publicly available (Grundbuchsumstellungsgesetz §5(2): "Die Einsicht in das Hauptbuch, die Urkundensammlung und die Hilfsverzeichnisse ist durch die Ausfertigung von Abschriften zu gewähren ..."), access is connected to administrative fees. The collection of large datasets, therefore, is associated with high costs. In this article, we directly address this issue to provide a baseline for this topic.

While actual transaction data are critical in real estate price prediction, many studies overlook key classification distinctions that are crucial in the industry. Real estate agents typically categorize properties using specific criteria, such as distinguishing between newly built and existing apartments based on first occupancy rather than construction year. This classification means that an apartment can be considered "newly built" if it has not been used, regardless of the actual construction date. However, the prevalent research often follows different criteria, relying mainly on the year of construction to classify properties as new or existing. This discrepancy is evident in various studies [5–10]. In addition, other significant factors, such as whether a property is purchased as an investment or for owner occupation, or whether an apartment is a penthouse versus on a regular floor (every floor of a building below the top floor is referred to as a regular floor, regardless of whether the top floor is developed or not), are often overlooked in research.

This article explores whether combining apartment transactions using these industry-based classifications and urban characteristics provides a robust foundation for predicting future property prices, focusing specifically on apartments in Vienna. Unlike house transactions, apartment purchases typically include the usable area in the contract, which is crucial to predicting the price per square meter (sqm). The study also examines how the variation in the training time span affects the prediction quality and emphasizes the spatial parameters that define a neighborhood, including infrastructure, social elements, and local leisure activities. Additionally, it considers sociodemographic factors, such as education level and income, which may influence price dynamics. During model development, the following questions are addressed:

- What spatial metrics can we use for feature engineering?
- What is the prediction accuracy for one year into the future for different timespans of the input data?
- What is the optimal training time span of the input data to predict future real estate values?

The purpose of this research is to develop and evaluate predictive models for apartment prices, with a particular emphasis on the influence of spatial and socio-demographic factors. Drawing on a decade-long dataset, it addresses gaps in the existing literature, which often focus on internal property characteristics while neglecting the critical role of spatial dynamics and industry-based classifications. By integrating advanced spatial metrics such as the shortest paths to POIs and isochrones, the study quantifies neighborhood accessibility and its impact on real estate prices.

To further enhance the accuracy of the model, the research distinguishes between newly built and existing apartments, aligning with industry practices. Using the XGBoost algorithm and rigorous hyperparameter tuning, it explores how different training time spans influence prediction quality. Ultimately, the study aims to improve the robustness of long-term price forecasts by leveraging spatial metrics and comprehensive data to capture the complex dynamics of neighborhood characteristics.

The remainder of the article is organized into five sections. The discussion of the relevant literature and the analysis of their approaches is shown in Section 2. After identifying the gaps in modeling and data usage, the framework for the experiment is developed in Section 3. It contains descriptions of the data sources used, their structure, and their potential peculiarities. The spatial relations are then modeled for usage as input for XGBoost. Finally, the implementation of the prediction and the quality estimation strategy are presented. Section 4 explains the different experiments and lists the predicted qualities achieved. A discussion of the results is presented in Section 5. Section 6 summarizes the findings and presents ideas for future extensions.

2. Related Work

Approaches for real estate price prediction can be divided into three main types:

- The first type is based on the hedonic pricing model (HPM) [11]. According to [8], this model typically only considers structural characteristics. Ref. [12] lists features such as the general condition of the house, heating system, bathroom, garage, etc., and location amenities. The difference between the various models often lies less in the input features than in the method chosen for how spatial heterogeneity is modeled. In recent years, HPMs have often been questioned as to whether they are still appropriate since they are usually outperformed by artificial neural networks in most valuation scores [5,13]. Nevertheless, HMPs are still used and actively addressed in research [14,15]. In the European region, they are used either as a stand-alone solution or as a support for evaluation experts for property valuation. A disadvantage of the hedonic pricing model is that the quality of the estimate is highly dependent on the availability of data on internal characteristics.
- The second type includes the most recent research that applies machine learning (ML) and deep learning (DL) techniques (e.g., [16,17]). The focus is on a large number of internal features, which also overlap in different research papers. Usually, spatial features, such as distances to certain points of interest or the number of amenities in the neighborhood, complement these internal features. In addition, other sources of information, such as photographs, can help extract information about the environment [18] or the objects themselves [8]. Building on these advancements, Geographic Information Systems (GISs) and Automated Valuation Models (AVMs) have emerged as crucial tools in enhancing real estate price prediction models. GISs, with their ability to integrate and analyze spatial data, provide a deeper understanding of how locational attributes like proximity to amenities and landscape features impact property values. For instance, GIS-based models have successfully quantified the influence of viewsheds and spatial patterns on housing prices, highlighting the importance of spatial context in real estate appraisal [19]. Additionally, GIS combined with techniques like Geographically Weighted Regression (GWR) has demonstrated superior predictive accuracy over traditional methods [20]. Meanwhile, AVMs, powered by machine learning algorithms, complement these spatial analyses by efficiently processing

large datasets to deliver consistent and accurate valuations. These models outperform traditional hedonic approaches by capturing complex patterns in both structured and unstructured data [21]. The increasing availability of big data further enhances AVM capabilities, making them indispensable for both individual property assessments and mass appraisals [22].

- The third group focuses on the economic aspects of real estate. Economically oriented models are usually created for a larger volume and are therefore more focused on the investment and economic development of real estate [23]. The main difference is that the focus is on the development of a property value rather than on price prediction. Real estate price indices are closely related to this type of research. There is partial overlap with the hedonic price model, as the latter also incorporates economic factors [24] and also includes or generates the aforementioned indices [25]. Economic models are not used exclusively for this research but are also included in the work to a smaller extent. Ref. [26], for example, includes the economic effects of neighborhood characteristics on housing values.

Research on housing price prediction often uses house transaction data [27] or rental information [28,29]. For apartment price assessments, offer prices are typically utilized, with regional variations in data usage. Most models based on rental data focus on Asia, while in North America, more sophisticated models rely on house sales data. This is partly because transaction data are more accessible in these regions. In contrast, in Europe, where transaction data are not publicly available, offer data are commonly used instead, leading to less research in this area. It should also be noted the methodologies to estimate house and apartment prices differ.

For apartment prices, the price per square meter is usually used, whether for rental [28] or purchase, unlike house prices, which are often expressed as a total sum. This distinction arises because the usable area for houses is seldom available, and additional factors, such as the value of the accompanying land, play a significant role [17]. The use of sqm price for apartment transactions is further supported by studies like [30], who predicted apartment prices in St. Petersburg using various methods. The data used for this work comprised around 3000 data points of transactions for two-room-apartments collected in Spring 2010 in St. Petersburg.

As research in this field is very active, the most relevant works are considered to identify current trends. A variety of ML algorithms are applied in the field of house price prediction. For example, ref. [27] used XGBoost for house price prediction and achieved good results with it. Despite the results, the paper points out that the prediction models may vary from place to place. Ref. [17] used a comparatively large dataset over a 5-year period in their work. The aim of this work was to compare different ML techniques against each other. XGBoost, CatBoost, Random Forest, Lasso, Voting Regressor, and others are used. The work used transaction data for houses in Florida and concluded that XG-Boost provides excellent and comparatively the best results. The work by [8] focuses on features describing the space around the object and the object itself. Multiple linear regression and gradient boosting methods are used. Twenty thousand datasets from the Boston area are examined. The focus of the work is on the collection of (especially) visual features, which are included using DL techniques. Ref. [31] used transactional data to investigate various machine learning and deep learning techniques for predicting real estate prices, concluding that support vector machines performed best. Among the models applied, their least squares support vector regression (LSSVR) achieved exceptionally low MAPE values of 1.679% and 0.228%. However, the dataset is poorly described, with the authors stating only that most of the data are residential and spans a three-year period, leaving the exact type of real estate and other data specifics unclear. Ref. [32] used rental data in their work. They explored the ability of the integration of ML techniques with the hedonic price model to map spatial patterns. Different spatial characteristics within a 15-min walking distance, sub-district, and nearest accessibility are considered. This approach proves effective in the context of the study and can be applied to other cities.

Most work does not consider time as a parameter. For online offer data, this is usually not necessary, as one or more points in time are usually selected to reflect the status quo of the bid at that point in time. However, when transaction data are used, the transaction date is relevant and should be included in the analysis [23]. However, long-term data are usually not available [33]. The time spans in real estate research often cover only a few months or a few years (for example, [33]), with longer periods being rare. Ref. [33] emphasized temporal dynamics by using 200,122 house transactions from 2011 to 2015. The models in this research were trained over the entire duration and tested on unseen data from the final two years as a benchmark to evaluate the various methods presented in the article, achieving MAPEs of 8–12% with a time-aware latent hierarchical model. By adopting this methodology, they managed to maintain a relatively stable MAPE throughout the entire prediction period. This highlights the importance of incorporating temporal context, a focus that aligns with our study's emphasis on long-term forecasting.

Another example of a study utilizing a long time span is provided by [16]. They developed a prediction model using 7407 apartment transactions in Ljubljana from 2008 to 2013. The temporal aspect was incorporated as a feature without differentiating the results based on time. Although interior apartment characteristics were considered, the study also placed emphasis on spatial factors, in the form of the distance to POIs. In addition, the study used data on apartment buildings, including details such as floor count, number of apartments per building, and renovation history. Their study highlighted a temporal trend: a market peak in 2008 followed by a 28% decline. Despite this, the observed price fluctuations were smaller compared to Vienna. Their Random Forest model achieved a MAPE of 7.04% in training and 7.27% in testing, both within the same period. The temporal consistency of their data probably contributed to these low errors.

A notable example of a study that uses an extended time span is [34], which analyzed data over a period of approximately eight years. However, comprehensive investigations of the impact of time span on prediction accuracy are still rare.

3. Framework

The framework discussed here is based on real transaction data of apartments, as well as urban computing measures representing spatial concepts. These concepts are used to engineer features that are later utilized by XGBoost. Further important decisions for the framework are the selection of data sources used for the prediction and the method to assess the prediction quality.

The model was intentionally designed to exclude external factors, such as the COVID-19 pandemic, policies, and interest rates, to focus solely on the impact of structural and local socio-demographic changes. This approach ensures that the analysis remains clear and is not influenced by broader, unpredictable events or non-spatial variables that may not enhance the model's capacity to capture spatial variations. However, demographic changes at the district level were included to reflect localized socio-economic dynamics.

3.1. Data Collection and Feature Engineering

The data used for the models are apartment transactions for Vienna between 2010 and 2020 as extracted from the land registry. There are approximately 145,300 data points available that were collected from the Austrian land registry during this period. These data points constitute the ground truth, as they represent real-world apartment transactions.

The analysis employs a comprehensive set of features beginning with transaction data, which includes primarily the sale price of the apartment, the usable area, and the date of the transaction. Sociodemographic data and the Austrian Trade Index (ATX) are also included to provide an economic context. Furthermore, leveraging the geographical component of the transaction data, the study delves into crucial spatial features that characterize the neighborhood, as well as urban computing measures.

3.1.1. Features from Transaction Information

The features that can be directly extracted from the transactions are

- Price;
- Usable area;
- Location;
- Classification as a penthouse or as regular floor;
- Transaction date.

In addition to the directly extracted features, two additional features could be generated by analyzing the transactions: if newly built and if an investment. An apartment is considered newly built if it is bought from a developer, even if the construction was completed years ago due to the time it takes to find a buyer. This information can be extracted on the basis of the seller's name. An apartment is considered an investment if there is no value-added tax (VAT) indicated in a transaction. This is the reason why investment apartments can have lower prices than regular ones. Only a few papers address these features (e.g., [35]), but this classification can have a significant impact on pricing, regardless of the type of property.

Since only data points with a complete set of information can be used as training and test data, a total of about 87,000 data points remain after inspection. The most common reason for excluding a data point was missing information about the usable area.

The location information is supplied as an address in the transaction information. Addresses were translated to geographical coordinates using the Vienna City address register (Adressen Standorte Wien, available under https://www.data.gv.at, accessed on 30 June 2023). The latitude and longitude of the building entrance, instead of the object center, are joined. Although several distance features are collected, the use of the entrance to the building is the most realistic approach.

The sqm price, which is used as a label, is determined from the price and the usable area, a common metric used for apartments. Using the sqm price standardizes comparisons between apartments of varying sizes and facilitates the tracking of price fluctuations over time. It also provides the ability to define price ranges for small areas with a reasonable margin. The price information itself can be discarded as redundant. The concept of usable area is mostly shaped by Austrian legislation and is defined in the Condominium Act of Austria (§ 2 Abs. 7 WEG).

Vienna is subdivided into 23 districts. As can be seen in Figure 1, more data points are available in the inner districts. The farther away from the city center, the housing tendency shifts from apartments to houses. Due to the constant growth of the city, more and more apartments are being built in the outer districts, but this transition takes time.

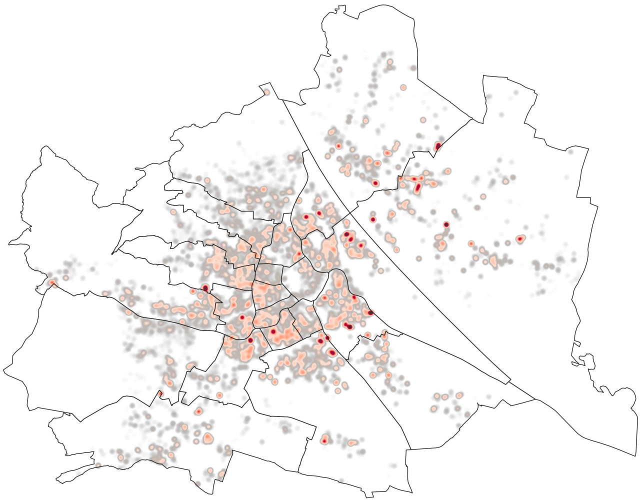


Figure 1. Distribution of apartment transactions with valid data points over Vienna in the form of a heatmap. The heatmap contains transaction data for the entire ten-year time span.

Temporal reference is also necessary. Since this work focuses heavily on time dependence, a closer look must be taken at the distribution of the data over time. As can be seen in Table 1, data availability increases over time. There are two main reasons for this. First, the number of apartment sales increases. This is supported by the general increase in transactions. Second, the availability of the parameter "usable area" is also increasing. From 2010 to 2020, availability increased from about 45% to 60%. This is a sign of an ever-improving documentation.

Table 1. Distribution of valid transaction data points by year.

Year	Data Points Count	Percentage
2010	5072	6.07%
2011	5472	6.55%
2012	5801	6.95%
2013	5754	6.89%
2014	5908	7.07%
2015	7431	8.90%
2016	7894	9.45%
2017	8377	10.03%
2018	9363	11.21%
2019	11,392	13.64%
2020	11,063	13.24%

3.1.2. Socio-Demographic and Economic Data

The socio-demographic and economic variables included in the model were selected to capture key factors influencing trends in the housing market. The average income provides insight into the purchasing power of residents, while the percentage of unemployment reflects economic stability in the area. The percentage of residents without Austrian citizenship highlights the demographic composition, which can influence demand and housing preferences. Lastly, the average age is considered to account for age-related housing needs and market dynamics.

The socio-demographic features are collected from the statistical yearbook of Vienna (Statistische Jahrbuch der Stadt Wien available under https://www.wien.gv.at, accessed on 30 June 2023). Various information about the city of Vienna is available in these annual publications. All the features listed below are collected by year and district:

- Average income;
- Percentage of unemployment;
- Aercentage of residents without Austrian citizenship;
- Average age.

In addition to these features, nine educational features are collected from Statistik Austria, the official statistics authority of Austria [36]. The features include the shares of the different educational levels of the population aged 25–64 years by district in Vienna.

The economic situation is described by the monthly median of the ATX. The information is provided by finanzen.at (available under: https://www.finanzen.at, accessed on 29 June 2023). Next, based on the assumption that there can be a delayed reaction of the real estate market on the trade index, a second feature was introduced using a 6-month-old ATX value (i.e., data points from January 2010 use the ATX median from July 2009).

3.1.3. Urban Computing Features

UrbanCore. Geographic relevant features were extracted using the UrbanCore framework [37]. For this work, the area of Vienna is divided into rectangular cells to generate statistics on infrastructure and points of interest (POIs) within these cells. The cells have a size of 99.995 m by 66.575 m. The extent of the cell is the result of the dimension of the total cell coverage (bounding box of transaction points) divided by a fixed number of cells. The UrbanCore framework provides the means to extract spatial features based on raster cells novel to real estate price predictions. This gives us the opportunity to examine a broad spectrum of spatial aspects. Thirty-two features are calculated for each of the cells (see Table 2).

Table 2. Statistics on the features extracted from cells within Vienna's bounding box, as defined by UrbanCore framework.

Feature Description	Mean	Std	Min	0.25	0.5	0.75	Max
Mean road segment length	25.85	29.05	0	0	21.72	42.82	259.77
Min road segment length	17.62	27.84	0	0	2.56	22.93	259.77
Max road segment length	36.95	38.38	0	0	30.40	68.772	303.50
Variance road seg. len.	209.81	447.64	0	0	0	246.50	11,193.26
Number of road segments	1.80	2.62	0	0	1	3	51
Mean road seg. orientation	23.45	25.35	0	0	14.80	44.87	89.99
Var. of orientation on roads	188.90	375.82	0	0	0	162.14	2008.68
Mean speed of roads	16.61	19.13	0	0	0	30	100
Min speed of roads	15.69	18.24	0	0	0	30	100
Max speed of roads	17.36	20.17	0	0	0	30	100
Variance of speed on road	7.61	29.82	0	0	0	0	900
Number of one ways	0.73	1.65	0	0	0	1	32
N._of roadcar_junctions	0.02	0.15	0	0	0	0	3
N._of pathroadcar_junc.	0.43	0.98	0	0	0	0	18
Number of junctions	1.53	2.99	0	0	0	2	75
Mean num. of ways in junc.	1.46	1.622	0	0	0	3	7
Var. num. of ways in junc.	0.035	0.11	0	0	0	0	4.58
Number of left POIs	1.64	11.38	0	0	0	0	172
Number of right POIs	0.73	2.82	0	0	0	0	46
Number of POI	2.37	13.21	0	0	0	0	205
Mean opening angle of junc.	24.36	29.90	0	0	0	56.96	197.92
Min opening angle of junc.	18.66	27.26	0	0	0	45.78	182.10
Max opening angle of junc.	29.44	36.67	0	0	0	60.13	524.28
Var. opening angle of junc.	100.84	305.23	0	0	0	0.50	13,633.72
Number of 3 ways	1.17	2.28	0	0	0	1	60
Number of 4 ways	0.34	1.07	0	0	0	0	27
Number of 5 ways	0.01	0.13	0	0	0	0	5
Number of 6 or above ways	0.00	0.06	0	0	0	0	3
Mean opening angle of 3 ways	24.97	30.86	0	0	0	59.63	150.53
Mean opening angle of 4 way	4.86	17.23	0	0	0	0	199.24
Mean opening angle of 5 way	1.013	10.26	0	0	0	0	236.42
Mean opening angle of 6 or above way	0.26	6.36	0	0	0	0	524.28

Building structure model. The building structure model (BKM (Baukörper Model der Stadt Wien available under: https://www.data.gv.at, accessed on 25 June 2023)) consists of all structural measures in Vienna. This model is coordinate-based and inch-perfect. It is maintained by the MA 41 ("The surveying department MA 41 of the city of Vienna"). The BKM is used for four features calculated within the cell structure of the UrbanCore. First, the built-up area within a cell is calculated as an absolute value for two BKMs ten years apart—2011 and 2021. These two values are the first two features of the model. The third feature is called the Building Space Ratio (BSR) and is the ratio of building to non-built-up areas within a cell. The last feature compares both BKMs and their change in a built-up area is calculated and outputted as a relative value. This feature is called the Development Index (DI).

$$BKM_{x1}Cell = \sum BKM_{x1} \text{ per Cell} \tag{1}$$

$$\text{BuildSpaceRatio} = 1 - \frac{BKM_{21}Cell}{\text{Area of Cell}} \tag{2}$$

$$\text{GrowthRate} = \frac{BKM_{21} - BKM_{11}}{\text{Area of Cell}} \tag{3}$$

Location features. The features described up to this point were average values for specified areas. However, an apartment is located at a specific address, and distances to specific POIs are relevant for the apartment's user [38]. Therefore, additional location features were defined, generated, and collected. These location features are calculated using regularly used POIs and the location from the transaction data. Some POIs, such as schools or doctors' offices, are used more frequently than others, such as museums. A variety of public transport POIs, such as bus, tram and train, are also used. Each year, Wiener Linien, the primary operator of Vienna's public transport network, conducts an analysis of the modal split [39]. The results consistently show that approximately one-third of the population uses public transport.

On top of the different public transport POIs, subway stations are considered in two different ways. As it is one of the most important transport networks in Vienna, it needs special attention. On the one hand, a single central point is projected onto the center of each station. On the other hand, several points are projected onto the individual entrances to the stations.

The next section outlines the preparation of the network to minimize calculation errors and introduces two location calculation methods (isochrones and shortest path), alongside the features used. Due to differing administrative rules and the unavailability of key data features outside Vienna, data points near the administrative border were excluded from the dataset to prevent biased estimates and ensure analysis accuracy. This leaves 83,527 data points for training and testing. Detailed figures on data collection can be found in the Table 3.

Table 3. The table provides an overview of the distribution of apartment transaction data in Vienna. It categorizes the data into different groups based on usability.

Data Points	Count	Percentage
Not usable	4083	2.8%
Limited usability (missing labels)	54,377	37.4%
Data too close to the city limits	3293	2.3%
Usable data	83,527	57.5%
Total	**145,280**	**100%**

Network preparation. As stated in Section 2, walking distance is a valid approach for location features to describe the neighborhood. The disadvantage of this approach is that the margin of error must be kept small in order to obtain accurate and comparable distances. This is especially true when calculating distances between points that are very close to each other. First, an OSMnx [40] graph for Vienna was downloaded. OSMnx only provides nodes at junctions. When a data point or POI is created on the raw OSMnx graph it maps these points on existing nodes, so the error can be quite large. To create a sufficient network for a better distance calculation, the following procedure is used:

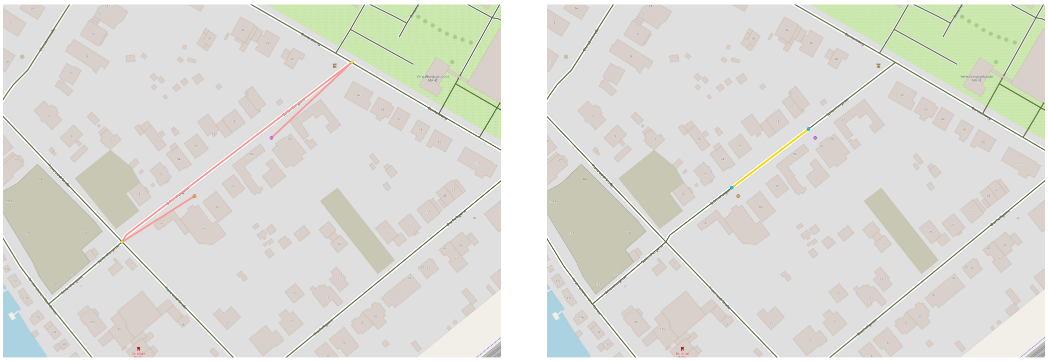
1. The downloaded OSMnx graph is decomposed into vertices and nodes. The vertices of the network, the data points, and the POIs are loaded with their geometry into PostgreSQL.
2. The nearest vertices are searched for each data point and POI, and the points are projected onto them. The nearest vertex is obtained from the shortest imaginary orthogonal line from the vertex through the given point.
3. The vertices and projected points are moved to QGIS. The vertices are split at the location of projected points via SAGA—*"Split line by Point"* (since this step resulted in

the formation of several unwanted artifacts, QGIS provided a visual component to clean them up).

4. The network is checked for errors (duplicated vertices) and then transferred back to PostgreSQL, where the network is recreated with additional new nodes at every vertex end. An ID is generated for each node, and the IDs are joined as endpoint IDs to the vertices.
5. Finally, the IDs of the closest nodes are joined to the original points, and the manipulated network is handed over to a Python script for further computation.

For the procedure, QGIS 3.16.7 with SAGA, PgAdmin 4 with PostgreSQL 14, and PostGIS 3.1.4 were used.

Figure 2 shows the difference between the calculation methods. Let us assume that the orange points are house entrances and the purple points are doctors. Since OSMnx only creates nodes at intersections (yellow dots), the distance is distorted in the calculation. On the right in Figure 2, the new nodes created by projecting them onto the segment have been used, allowing a more accurate calculation of the shortest path.



(**a**) Shortest path with OSMnx and NetworkX (**b**) Shortest path with network manipulation

Figure 2. (**a**): Calculating the shortest path with OSMnx (version 1.1.2) and NetworkX (version 2.6.3), based on the nodes created at junctions. (**b**): Calculating the shortest path with network manipulation. Basemap provided by OpenStreetMap ©.

Isochrones. In preparation, the data points are checked for duplicate locations. It is not uncommon for several data points to have the same address, especially over a ten-year period. Furthermore, it should be taken into account that a new building could hold up to hundreds of apartments.

Three different walking distances were chosen for the isochrones: 150, 300, and 1000 m. These distances were chosen to reflect the immediate neighborhood. Most people are not willing to walk more than 1 km to reach their destination. Ref. [41] showed that the average walking distance is 0.7 miles (1.13 km) and the median is 0.5 miles (0.81 km), regardless of the destination. Thus, 1 km gives a good approximation for the walking neighborhood. Isochrones for these distances are calculated with a convex hull to create spatial links of the POIs within their area. Each type of POI is summarized as a total number of occurrences within each isochrone.

Section 3.1.3 describes two types of subway features—single and multi-point—to represent stations either by their entries or as a single point to avoid bias from multiple entries within an isochrone. These features are categorized based on the station's high-priority connections, such as links to other subway lines or trains, reflecting their proven influence on real estate prices in other cities [34]. Transaction data are also summarized to enhance analysis. All transactions (except those that do not include location information) are summarized by year of occurrence. This leads to a total of thirty-five features for each isochrone.

Shortest path. The same POIs are used for the shortest path. There is no need for the shortest path to the next transaction since this calculation is prone to show the distance of POIs that are actually used by people living there. The subway stations are again used as both single and multiple points but without classification. The calculation is carried out using NetworkX's shortest path algorithm. Shortest paths were calculated to a maximum distance of 5000 m to reduce calculation time. A distance of 5000 m was used to ensure that almost all data points have the shortest path to each POI and it is a more than sufficient distance to be considered a neighborhood [42].

3.2. Features and Their Nature of Representation

The basic model contains 200 features. All features are listed in Table 4.

Table 4. Features used for the base models, along with their counts, descriptions, and forms of integration.

Feature	Count	Description	Form of Integration
Transaction	5	Usable area, investor, penthouse, newly built, and transaction date	individual
ATX	2	Monthly median and 6 months shifted in the past	by year and month
Socio-Demographic	4	Unemployed, average income, share of foreigners, and average age	by year and ZIP code
Education	9	Different levels of education: including compulsory school, secondary school, apprenticeship, among others	by year and district
Transaction count per building	1	-	by location
Urban Core	32	Variation of segment, junction, POI, etc., statistical values based on a cell grid	by location
BKM	4	Two features at fixed points in time (2011 and 2021). Two features derived from both BKMs, based on Urban Core grid cell	by location
Shortest path	15	POIs: doctor, drugstore, kindergarten, museum, parks, police station, low-priority public transport, schools, subway single point, subway multi-point, touristic attractions, train station, university, playground, restaurant/bar.	by location; individual
Isochrone	105	Shortest path POIs with finer distinctions. Separate features for bus and tram, restaurants and bars, and subway features. Includes 11 Transaction features	by location; individual
ZIP	23	One-hot-encoding; based on the 23 districts of Vienna	by ZIP code

Isochrone and Shortest Path features are emphasized in the model due to their ability to provide detailed insights into accessibility and connectivity, which are crucial for understanding spatial influences on property values. These features offer a granular view of the spatial context surrounding the apartments, capturing how easily various POIs can be reached. In addition, POIs are readily available and can be processed automatically, facilitating the creation of a diverse and comprehensive feature space. This automated processing supports the development of a robust model by incorporating significant spatial features known to impact real estate prices.

3.3. Modeling Algorithm

As mentioned above, the objective of this work is to predict the sqm price of apartments. The scope is to solve this as a regression problem. The Extreme Gradient Boosting (XGBoost) algorithm was chosen for this purpose. XGBoost is considered an efficient technique for regression problems with the aim of building predictive models [7,9,17]. The algorithm works by using multiple weak hypotheses and combining them into a single superior hypothesis ("Hypotheses Boosting Problem") [43]. In other words, several weak stems of a tree are combined into one strong model. Gradient Boosted Trees offer greater model capacity than Random Forest, allowing them to capture complex relationships and intricate decision boundaries. Among the various implementations of Gradient Boosted Trees, XGBoost was selected for its scalability and high efficiency [44].

The model was implemented using Python 3.9, and further packages were used, including XGBoost 1.3.3 and scikit-learn 0.24.2. The data were split into 90% for training and 10% for testing. A 5-fold validation with randomized search and fifty iterations to tune hyperparameters was used. A five-fold cross-validation with a randomized search over fifty iterations was used to fine-tune the model hyperparameters. The search space for each parameter is detailed in Table 5.

Table 5. Search space for hyperparameters of XGBoost models.

Hyperparameter	Search Space	Description
No. of estimators	{200, 300, 500, 800, 750, 1000, 1500, 2000}	Number of boosting rounds
Max depth	{9, 10, 11, 12, 13, 14, 15, 16}	Maximum depth of trees
Learning rate	{0.001, 0.01, 0.025, 0.05, 0.075}	Learning rate of model
Min. child weight	{4, 5, 6, 7, 8, 9}	Minimum sum of instance weight in a child
Gamma	{0.1, 0.2, 0.3, 0.4}	Minimum loss reduction for further partitioning
Subsample	{0.5, 0.6, 0.7, 0.8, 0.9}	Ratio of training data used per tree
Colsample bytree	{0.6, 0.7, 0.8, 0.9, 1.0}	Ratio of columns subsampled per tree
objective	{'reg:squarederror', 'reg:tweedie'}	Objective function for optimization
Booster	{'gbtree', 'gblinear'}	Type of boosting method used
Eval metric	{'mape'}	Metric for validation data evaluation
Eta	{0.2, 0.3, 0.4, 0.5, 0.6}	Controls the learning rate

Ten different models were developed, differing in the number of years of training input and hence data points. A more detailed look is given in Section 4.1. All models use the mean absolute percentage error for validation, solve "reg:tweedie" as the objective, and use "gbtree" as the booster. The remaining fine-tuned parameters for each model can be found in Table 6.

Table 6. Hyperparameters of models: subsample (SS), n_estimators (NE), min_child_weight (MCW), max_depth (MAD), learning_rate (LR), gamma (GA), eta (ET), colsample_bytree (COL).

No. Years Input	SS	NE	MCW	MAD	LR	GA	ET	COL
1	0.7	800	7	11	0.025	0.2	0.4	1.0
2	0.9	500	8	14	0.025	0.1	0.3	0.7
3	0.7	750	6	13	0.025	0.1	0.3	0.7
4	0.9	1500	8	12	0.025	0.1	0.3	0.7
5	0.8	800	8	14	0.010	0.2	0.3	0.6
6	0.9	500	5	14	0.050	0.2	0.3	0.9
7	0.9	800	8	14	0.025	0.2	0.4	0.8
8	0.6	800	5	15	0.025	0.3	0.3	0.6
9	0.8	1500	8	14	0.025	0.1	0.4	0.9
10	0.9	800	6	12	0.050	0.1	0.4	0.8

3.4. Evaluation

Several measurement methods are used in the scientific literature on real estate prediction models. In addition to the R^2 and the mean square error (MSE) and the mean absolute error (MAE) are also used in regression problems. The MAE reflects the average deviation from the real value. The mean absolute percentage error (MAPE) is chosen for the evaluation of the model performance. MAPE is calculated as follows:

$$\text{MAPE} = \frac{1}{n}\sum\frac{|y_{i,t} - y_{i,p}|}{y_{i,t}} \tag{4}$$

where y_i, t is the actual value of the transaction and y_i, p is the predicted value.

MAPE is chosen because of its advantages in real scenarios. Price ranges are part of the daily real estate business and, therefore, are easy to interpret. Although the R^2 error seems to be an appropriate option, it is not due to the nonlinearity of the data. Therefore, the R^2 value cannot be used correctly here. Something similar can be said about the MSE. Since the data are not adjusted for outliers and outliers are penalized higher in the MSE, it is not a suitable choice [17].

4. Experiments and Results

A number of experiments were carried out on the basis of the data presented in the last section. This section contains the concepts for the experiments and the results obtained from them.

4.1. Input Variation

A series of 10 models were developed, each differentiated by the period of years of input used for training. The first model utilized transaction data from a single year, while the subsequent models incrementally incorporated more years of data, culminating in the tenth model, which was trained on a decade's worth of data up to the year 2019. Given the extent of the available transaction data, the model trained on 10 years of data can produce only one prediction value. Each model is tasked with the same objective: to employ one or several years of data to predict sqm prices for the following years. A full sequence of this process is termed a cycle. The modeling cycle is depicted in the provided Figure 3, which illustrates the progression from the first model with a single input year to the tenth model.

Models with a large number of input years (8–10 years of input) must be considered with caution, as only a few predictions can be made, and thus only a few comparisons are available. As the number of input years used as training data increases, the number of prediction years that can be used for comparison decreases because the study period is limited to 2020.

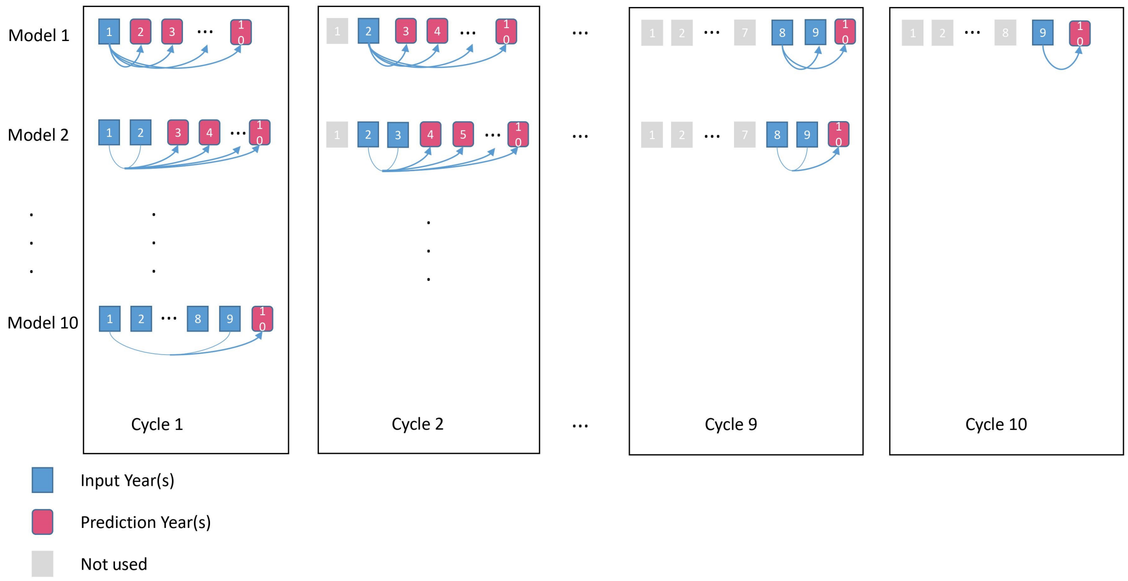


Figure 3. Training cycles of predictive models for housing prices. Illustrates the progression of 10 different models, each using a different number of input years for training. Model 1 starts with data from a single year, incrementally increasing to Model 10, which is trained on 10 years of data. Each cycle represents a complete sequence of predictions for subsequent years following the training period.

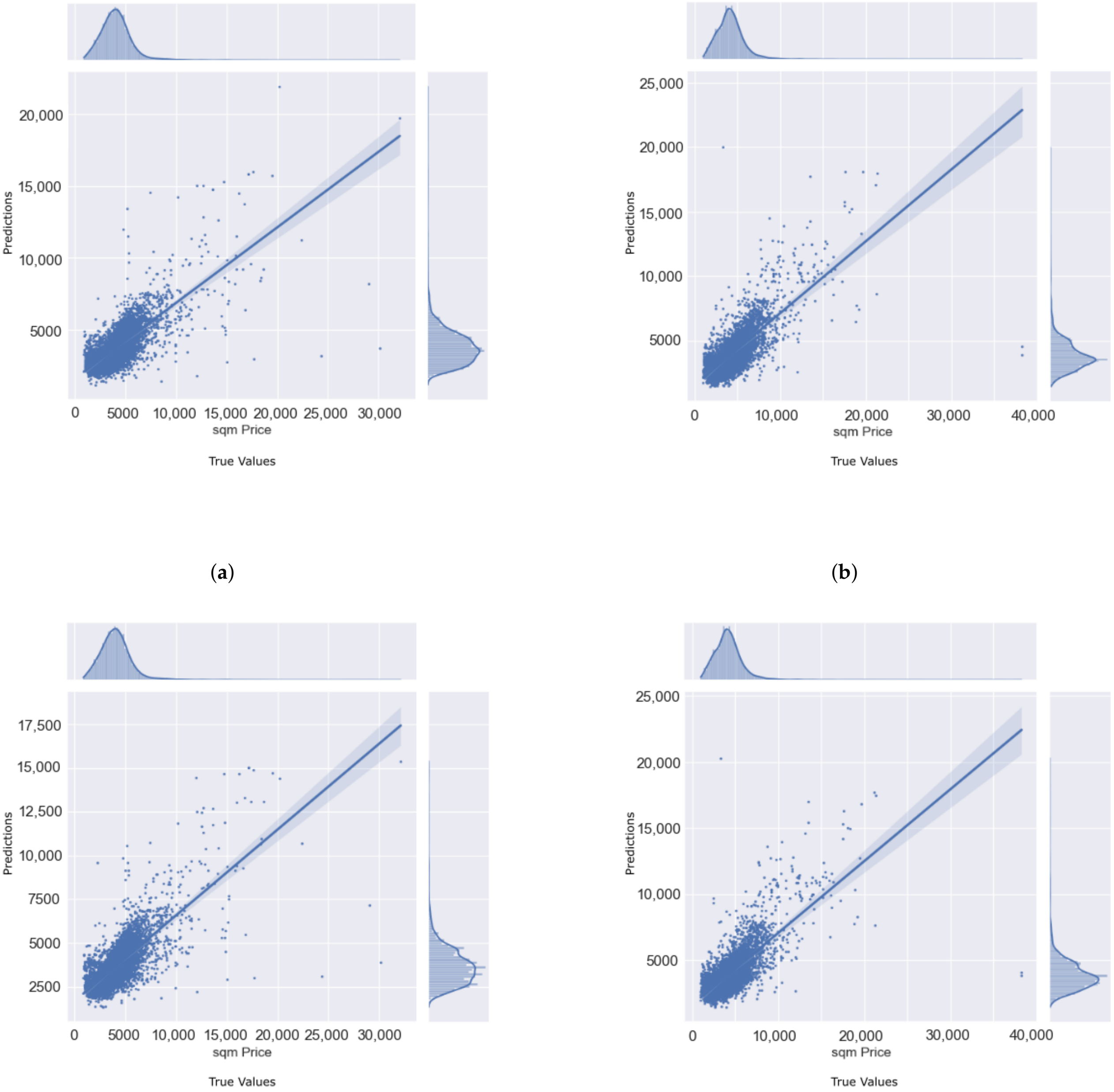
4.2. Results for Predicting One Year in the Future

In the first step, the models are compared in terms of their ability to predict a subsequent year, regardless of how many input years were used for training. The results show that the accuracy of the prediction depends on the quality and quantity of the input data. In general, most models achieve a value of around 21% or even 20% at least once in their cycle. The only exceptions are models with 9 or 10 input years. The errors of these models are usually 2–3% higher. Table 7 shows the MAPE for the top scores for the prediction of one year in the future.

Table 7. One-Year Prediction: This table showcases the optimal MAPEs obtained by each model when predicting one year ahead. The models are arranged in ascending order based on the number of training years they utilized. Additionally, alongside the number of training years, the table indicates the specific input years used for training.

MAPE	Number of Training Year (s)	Training Year (s)	Prediction Year
21.18%	1	2016	2017
21.08%	2	2013–2014	2015
21.08%	3	2015–2017	2018
20.34%	4	2013–2016	2017
20.67%	5	2012–2016	2017
21.00%	6	2011–2016	2017
20.66%	7	2011–2017	2018
20.94%	8	2010–2017	2018
22.46%	9	2010–2018	2019
24.91%	10	2010–2019	2020

The error distribution of the top models can be observed in Figure 4. Each data point is assigned coordinates that represent the predicted and the actual price achieved. It can be seen that most data points are in the range of up to 10,000 EUR/m^2 and that there are relatively few outliers, which, however, have a large influence on the MAPE.



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

(**c**) (**d**)

Figure 4. The figure illustrates the regression plots for the top-performing models, arranged in ascending order based on their MAPE. Each plot provides a comparison between the true and predicted values. What distinguishes each plot is the number of training years involved and the year being predicted. (**a**) MAPE: 20.34%; Training years: 2013–2016; Prediction year: 2017. (**b**) MAPE: 20.66%; Training years: 2011–2017; Prediction year: 2018. (**c**) MAPE: 20.67%; Training years: 2012–2016; Prediction year: 2017. (**d**) MAPE: 20.94%; Training years: 2010–2017; Prediction year: 2018.

In the second step, all predictions for the subsequent years of all models were compared. Table 8 summarizes the results and shows that the best results were achieved with 4 to 5 years of input data.

Table 8. Mean MAPE over all subsequent year predictions of each model.

Input Years	Mean MAPE
1	23.34%
2	22.86%
3	22.54%
4	22.11%
5	22.10%
6	22.34%
7	22.19%
8	22.56%
9	23.77%
10	24.68%

It should be mentioned that the predictions for 2020 could not achieve an error of less than 24%. This needs to be investigated further after a data update to rule out a qualitative issue. If a quality check does not change the outcome, this leads to the assumption that there was a massive change in the price development. Aspects of the Coronavirus pandemic also need to be considered, such as the lockdown from March until the 1st of May 2020.

4.3. Results for Predicting Multiple Years into the Future

The results of multiple subsequent years are more difficult to interpret. The main reason is again the time limitation. It is difficult to tell from an average error whether models with many input years are actually better. These models have better average values than those with more verifiable results due to the fact that they have fewer years for the verifiable forecast. Nevertheless, one statement can be made about the rate at which the error increases. As can be seen from Figure 5, the error of models with 5–9 input years grows slower in the experimental series from 2010 onward.

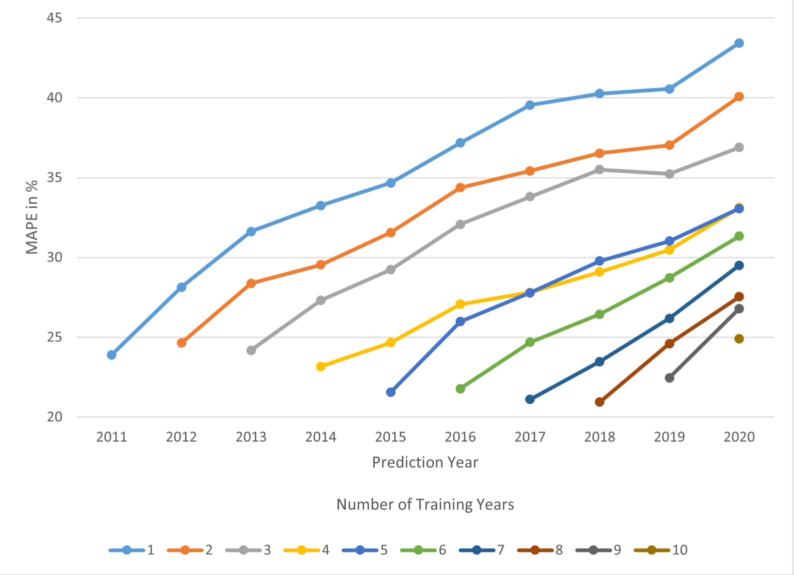


Figure 5. The figure depicts the development of error across various models over a span of multiple years. The diversity among the models lies in the number of years used for training. The graphical representation enables comparison of each model's ability not only to predict outcomes for a specific year but also to predict over a particular range of future years.

The increase in MAPE varies slightly with different starting years for training, yet it can be noted that the error grows slower with 5 and 6 input years and subsequently also with 4 input years (4 input years show good results, especially in the starting years 2012 and 2013).

4.4. Noise Elimination Experiments

With around 200 features, it is likely that some features generate unnecessary noise. Different groups of features were excluded from the training to identify redundant information. A series of experiments was conducted with the 4-year input model, as it shows the best prediction results. Figure 6 shows the change in MAPE for these experiments.

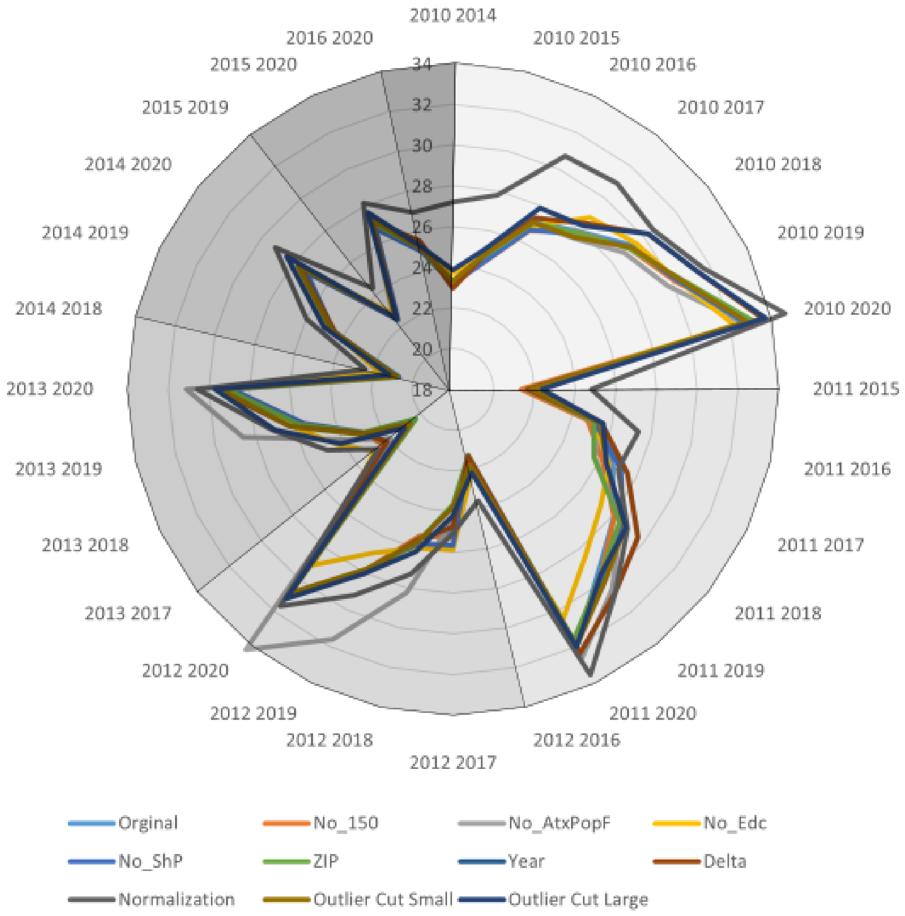


Figure 6. The figure presents the experiments conducted on the 4-year input model to mitigate noise generated by features. Each segment within the figure represents a distinct cycle of training and prediction. For instance, the initial segment illustrates training conducted with data from 2010 to 2013, followed by predictions for the years 2014 to 2020. The subsequent segment draws on training data from 2011 to 2014 and so forth. The colored lines within each segment show the MAPE achieved during the respective training and prediction cycle for the feature selection.

The following experiments were performed in the order defined by the list:

- **Original**: All features used.
- **No 150**: Smallest isochrones of 150 m are excluded.
- **No ATXPopF**: ATX and socio-demographic features are dropped.
- **No Edc**: No education feature.
- **No ShP**: Shortest path excluded.
- **ZIP**: Using ZIP Codes instead of One-Hot-Encoding.
- **Year**: Using the year of the transaction date in 20xx format instead of the year and month delta combination.
- **Delta**: Using the year delta without the month.
- **Normalization**: Testing normalization of features with standard scaler library of sklearn.
- **OutlierCutSmall**: Cut outliers only if two or more features are outliers (Tukey method).
- **OutlierCutLarge**: Cut outliers if one or more features are outliers (Tukey method).

The only notable improvements with respect to MAPE were achieved by using the ZIP code variant instead of one-hot coding. Therefore, all other models were also tested with this modification. All models except for the 2-year input performed better with the ZIP code instead of the one-hot encoding.

4.5. Data Subset Experiments

As mentioned in the Section 1, the classification of an apartment as newly built and/or as an investment property significantly influences its pricing structure. Therefore, these classifications were included as features in our earlier models. To investigate whether these features, when used to subdivide the data prior to training, testing, and prediction, could affect the error behavior of the models, the data were divided into two distinct groups: the first group included only newly built apartments, and the second included only existing apartments. All investment properties were intentionally excluded to promote price stability.

Three different predictive models for our experiment were used: 1-year, 4-year, and 5-year models. For this experimental setup, the most promising models from the original framework were selected. The rationale was to determine whether these input variations could be further enhanced through data subsetting, while simultaneously maintaining computational efficiency. Beyond excluding newly built and investment features, the same set of features, as outlined in Section 3, was maintained in all models. With the use of only the existing apartments, the models' performance slightly declined. On average, they experienced a decrease in MAPE of approximately 1–2% across all models. The best models performed around 4% worse than their original counterparts. On the contrary, the exclusive use of data from newly built properties resulted in substantial performance enhancement. On average, these models outperformed their predecessors by 6%. The best model achieved a striking 13.5% MAPE for the following year, exceeding the original model by almost 7%. Some models managed to maintain a MAPE below 20% even three years into the future.

Table 9 provides a direct comparison of MAPE for the model using 4 years of input. The model selected for this comparison is the highest-performing model. The table clearly demonstrates the superior performance of this model when only newly built apartments are applied and therefore more homogeneous data. The MAPE remains significantly lower, indicating the model's ability to provide more accurate predictions, even several years into the future.

Table 9. Comparison of the 4-year input base model to the 4-year data subset input models.

Input Years	Prediction Year	Base Model	Newly Built	Stock Only
2013–2016	2017	20.34%	13.52%	25.53%
2013–2016	2018	22.87%	16.65%	27.05%
2013–2016	2019	25.77%	19.57%	27.11%
2013–2016	2020	29.14%	23.39%	28.55%

Figure 7 underscores the superior performance of the model that exclusively uses the newly built apartment data. It presents the MAPE over the longest possible observation period using the 4-year input model.

To demonstrate that all models experienced a decrease in MAPE, Table 10 compares the three aforementioned models to the original framework for one-, two-, and three-year-ahead predictions, using 2018 as the prediction year. The results show that all variants achieve a reduction in MAPE when trained exclusively on the homogeneous dataset of newly built apartments.

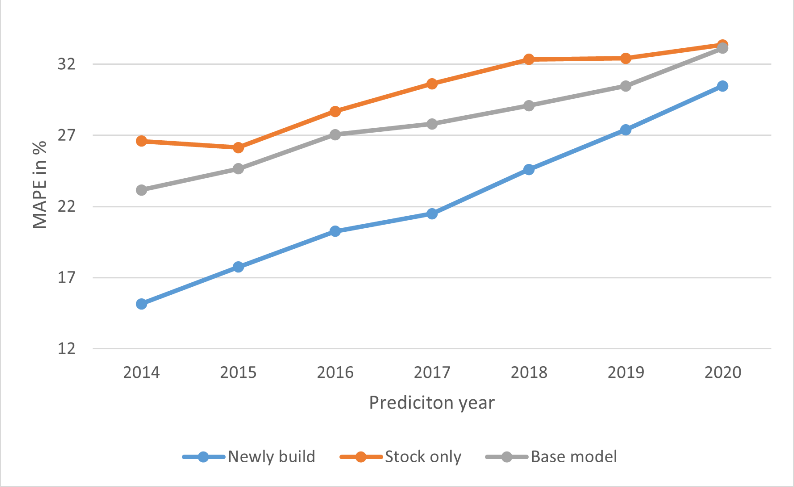


Figure 7. Development of the MAPE of the models for a longer observation period.

Table 10. Prediction model comparison for one-, two-, and three-year-ahead Forecasts.

Model	One Year Ahead	Two Year Ahead	Three Year Ahead
1-Year Base Model	21.92%	22.41%	26.24%
1-Year Newly Built	14.46%	15.79%	21.12%
4-Year Base Model	20.79%	22.87%	25.64%
4-Year Newly Built	14.08%	16.65%	19.66%
5-Year Base Model	21.42%	23.27%	27.09%
5-Year Newly Built	14.42%	17.22%	21.58%

Furthermore, Table 11 presents the average performance of each input variant, aggregated across all models and training data configurations. This table highlights the impact of the training data on model performance by summarizing the average MAPE for each variant, providing a comprehensive overview of their predictive accuracy.

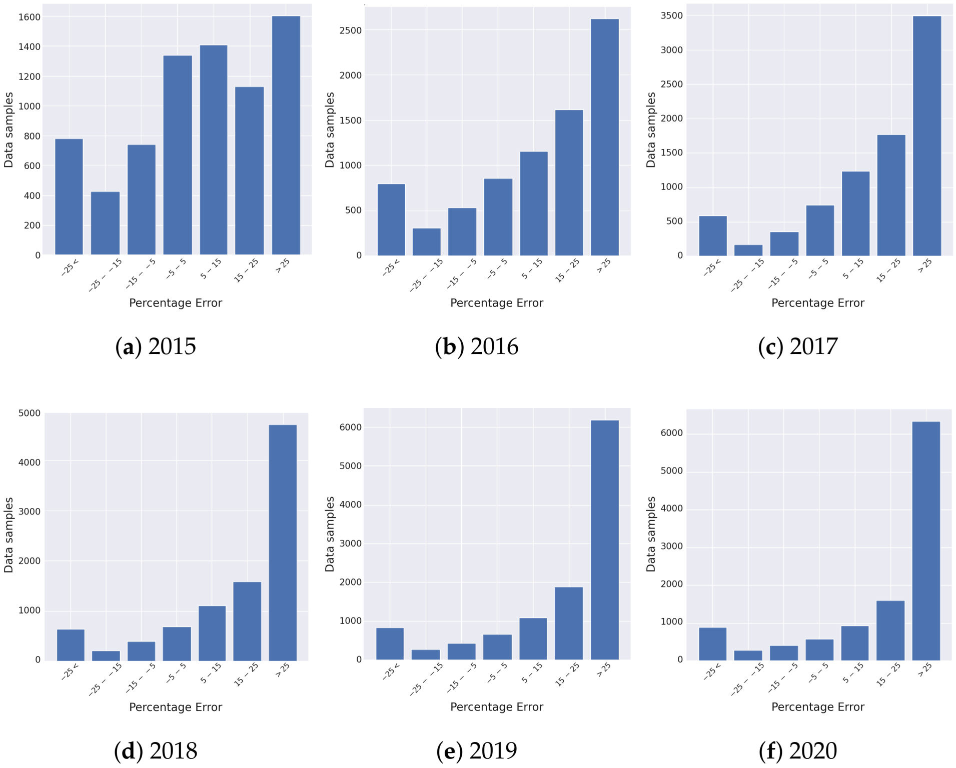
Table 11. Comparison of average MAPE and MAPE decrease for different models.

Model	Number of Variations	Average MAPE	MAPE Decrease
1-Year Base Model	56	29.29%	–
1-Year Newly Built	56	23.47%	5.82%
4-Year Base Model	28	26.18%	–
4-Year Newly Built	28	20.25%	5.93%
5-Year Base Model	21	26.40%	–
5-Year Newly Built	21	20.36%	6.04%

This sequence of experiments highlights the critical role that these classifications play in the prediction of real estate prices. Furthermore, it underscores how pre-subsetting data can impact the predictive accuracy of machine learning models. Utilizing more homogeneous data allowed the model to predict much more accurately with minimal cost to the models that used the heterogeneous data.

4.6. Error Development over Several Subsequent Years

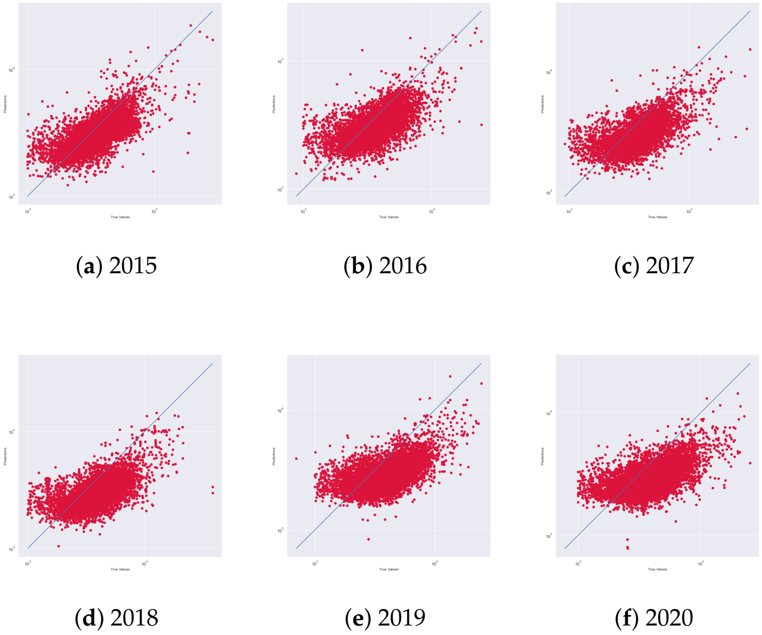
To examine the development of the error over a longer observation period, the errors of the individual predictions of the data points were divided into groups, starting with an underestimation of more than 25% up to an overestimation of more than 25% in increments of tens. The resulting seven groups can be well compared in terms of their trends. Figure 8 shows how the error shifts when predicting several years. Figure 9 shows the same distribution with a fixed axis so that the shape change of the deviations of the predictions can be observed.



(**a**) 2015 (**b**) 2016 (**c**) 2017

(**d**) 2018 (**e**) 2019 (**f**) 2020

Figure 8. The figure illustrates the categorization of labels into error classes based on their MAPE. The model used was trained using a 4-year input from 2011 to 2014. It was employed to predict the years that followed. When looking at the errors in these predictions, a clear trend emerges: the further into the future the model predicts, the more pronounced the left skew in the error distribution.



(**a**) 2015 (**b**) 2016 (**c**) 2017

(**d**) 2018 (**e**) 2019 (**f**) 2020

Figure 9. This figure shows the real versus predicted labels in the form of a point cloud, using the same 4-year input model and task as described in Figure 8. A fixed-axes approach is employed to provide a different perspective on the data shift. Through this method, it becomes apparent that the labels tend to gravitate increasingly toward the lower part of the diagonal line as the predictions progress.

The increasing left skew in the error histograms is observed in all models. The probable cause is that the model cannot use the time parameter (transaction date) appropriately. This shift to the left implies that the model is increasing the single values of its predictions too quickly. The following versions of the time parameter have been tested (all time parameters are derived from the transaction date):

- Year
- Year Delta (2010 ≡ 0; 2011 ≡ 1,. . .)
- Year + Month
- Year + Month Delta (January 2010 ≡ 0.01, February 2010 ≡ 0.02, January 2011 ≡ 1.01, . . .)

Initially, the *year + month delta* was used and confirmed as the best option for the noise-canceling experiments from Section 4.4.

5. Discussion

To better understand the nature of the prediction errors, a detailed analysis was performed on a specific model. This model, which used four years of training data to predict housing prices for 2014, was selected due to its superior performance. The focus of the analysis was on examining the deviations between the predicted and actual prices. Outliers were identified and analyzed to determine the reasons behind these discrepancies, helping to refine the accuracy and reliability of the model in future predictions.

MAPE defines a range for the expected deviation. Deviations far outside this range are outliers, and it is important to see their pattern and try to explain their existence. Extreme underestimations do not occur as frequently as overestimations. In general, this means that the model usually overestimates. These extreme outliers were investigated in more detail. It is hard to give a clear threshold for the MAPE of outliers because it varies a lot for each model. For the model used, it looks as if the extreme underestimation starts at −50% and the extreme overestimation starts at +100%. The upper threshold was chosen higher after the predictions tended to be overestimated, as can also be seen in the left slope of the figures in Figure 8. These thresholds are valid for about 4.5% of the predictions for the 2014 data set.

To explain significant prediction errors, additional data were collected for specific points. These included evaluating images of the buildings and reviewing transaction details, such as the inclusion of inventory or garage spaces. Large errors were often observed in luxury homes that have high sqm prices due to their rarity and scattered distribution in the dataset, making accurate predictions challenging for the model. On the contrary, extreme overestimation cases typically occurred when the model predicted reasonable prices based on the location and appearance of the building. However, the buyer and seller information did not clarify why some transactions closed at unexpectedly low prices. Some factors influencing low prices, such as agreements with existing tenants during building renovations or distress sales, are acknowledged to be difficult to detect and often unavailable publicly. These outliers, especially distress sales that require details on the seller's financial status, underscore inherent limitations in the dataset, preventing complete accuracy in some estimations. As long as MAPE ranges from −50% to +100%, fully accounting for discrepancies between model predictions and actual prices remains challenging. Enhancing the model with additional features might mitigate some inaccuracies. For example, a property in the 9th district sold in 2014 was underestimated by over 50%, suggesting inadequacies in the model or a lack of relevant features. Table 12 presents a comparative analysis of this property, called Object A, with another, Object B, in the same district. Sold one month apart, both properties were new constructions and investments with identical sociodemographic characteristics; the main distinguishing factor was their location within the district, with Object A in the west and Object B in the east.

Some assumptions can be made about this. First, the spatial features could have introduced a bias in the model, since this is the main distinguishing feature between the two properties. Second, there could be a local bias, if, in the vicinity of property A, apartments mainly in cheaper segments are sold, which pushes the price of the prediction down. There may also be additional features missing to make a better prediction.

Table 12. Comparison of prediction results of the 4-year model on two objects in the 9th district.

Property	Area (sqm)	Real sqm Price	Predicted sqm Price	Error	Date
Object A	40.47	6022.24	2744.50	−54%	September 14
Object A	41.38	6000.00	2743.46	−54%	September 14
Object A	43.82	6002.74	2757.11	−54%	September 14
Object A	43.84	6000.00	2757.11	−54%	September 14
Object B	77.93	5587.04	4614.81	−17%	October 14
Object B	59.89	4503.83	4526.91	1%	October 14
Object B	86.96	7205.02	4557.47	−37%	October 14
Object B	59.53	5528.39	4406.12	−20%	October 14

It is important to consider that even within the same property, selling prices can vary for apartments that are nearly identical in external features. According to the data presented in Table 13, the model accurately predicted the square meter price for several apartments within this property, with an average error of just 1% in 26 transactions. However, individual deviations from actual transaction prices ranged from −25% to +26%. This variability is challenging to explain using the current set of features, as apartments of similar size and amenities would ostensibly be priced comparably. Enhancing the model with additional parameters could potentially reduce these discrepancies.

Table 13. Performance of 4-year model on multiple apartments of similar size within the same building.

Area (sqm)	Real sqm Price	Predicted sqm Price	Error
58	2763.20	3206.96	16%
59	3286.44	3207.88	−2%
59	3112.04	3462.52	11%
59	3103.39	3213.00	4%
59	3103.39	3213.00	4%

The fact that the model makes worse predictions with fewer years than with more years seems logical. A larger number of datasets enriches knowledge about the importance of different features. The fact that the models perform worse after a certain number of input years suggests that the underlying relative importance of features changes over time. This cannot be adequately accounted for with XGBoost and may be the reason why the models deteriorate rapidly as they predict further into the future. Additionally, the situation in the later years needs to be analyzed in more detail. As mentioned in Section 4.2, it is possible that the stability of the market or the quality of the data is lower than in earlier years. In addition, it would be useful to have a method that automatically detects when the market situation has changed in a specific area. This would allow us to eliminate older data, which do not fit the new situation, from the learning process.

To put the results of this work in context, the results are compared to other studies. The most similar work was published by [30]. The models tested achieved MAPEs ranging from 14.86% to 20.53%. Further refinement through data subsets yielded MAPEs of 9.8% for apartments smaller than 61.5 sqm and 19.4% for larger ones. Restricting predictions to specific districts improved results further, achieving MAPEs of 12.9% and 23.6%. The size of the dataset and the fact that the test error is created with unseen data from the same time period explain to some extent the low MAPE. In our work, we focus on predicting the future prices of apartments, having a more general approach, without distinguishing between small and large apartments.

The setup by [16] contributed to strong performance, with the Random Forest method achieving an MAPE of 7.04% in training and 7.27% in testing. Testing was conducted within the same time frame as training, avoiding predictions beyond the observed period. Additionally, price fluctuations during the observed period were relatively low, with a mean price of EUR 2415 per sqm and a standard deviation of EUR 575. In comparison, the mean price in Vienna for the same period was EUR 2885 per sqm, with a notably higher standard deviation of EUR 1504.

An example of research that acknowledges the importance of time in interpreting results is the work by [33]. The results, broken down by each quarter for the two-year prediction period, revealed a variation in MAPE between 8–12% for their best-performing method. The utilization of houses rather than apartments, coupled with limited information about the features that are used, makes this paper non-comparable to our research. However, it warrants mention only to acknowledge the presence of time-aware real estate research with good results. In Ref. [31], the final MAPE values of the two best models are 1.676% (LSSVR1) and 0.228% (LSSVR2). These values are among the best results found in the field of real estate price prediction. However, the paper does not fully document the data used. They only state that most of the data are residential and were taken from a 3-year time period. The fact that about 2.54% of the data are commercial makes them even harder to compare. Additionally ,it is not clear to what the presented MAPE refers to and what was actually predicted. The article also contains a table comparing their results with the MAPE values of other studies. Their results are the best in the selection. However, a closer look at the papers in the table reveals that the MAPEs are based on highly different databases and different settings. This leaves only mainly [30] and [16] for comparison.

Although it is difficult to make comparisons with other works, the results can be put into a general context. Ref. [45] categorizes MAPE values as follows:

- Less than 10%: Highly accurate prediction;
- 11–20%: Good prediction;
- 21–50%: Reasonable prediction;
- above 51%: Inaccurate prediction.

The results obtained in this article fall into the category of *Good Prediction* or *Reasonable Prediction*. Notably, models that exclusively rely on "newly built" data achieve the "Good Prediction" category, maintaining this higher accuracy even when tasked with predicting several years into the future.

The process of buying real estate is closely connected with certain variables that are complex to define or track. At the heart of this complexity is an emotional component that will always be present for both individuals and companies buying or selling real estate. This aspect should not be neglected in corporate transactions either, as there is always a human being in the background who brings not only rational considerations but also emotional aspects into the decision-making process. Other non-quantifiable factors such as time constraints or financial pressure also have a significant impact on the desire to buy or sell a property. These factors can strongly influence real estate prices and lead to significant fluctuations. Even if, by and large, a common consensus on pricing emerges, it is precisely these variables that can lead to some differences. Therefore, given these non-descriptive factors, the accuracy of any model can only reach a certain level of accuracy when trying to predict real prices.

The models developed in this article introduce a new perspective on real estate by leveraging actual transaction data and employing classifications that are widely used in the industry, such as "newly built" or "stock". This approach more accurately reflects real-world conditions and significantly reduces the prediction error. Furthermore, the study demonstrates that predictive models can deliver reliable results even with minimal information about internal apartment characteristics, highlighting the critical role of location factors. By incorporating a vast spatial component through urban computing, this research distinguishes itself from previous studies, offering a novel methodology for real estate price prediction.

The findings also underscore the importance of temporal context in real estate. In rapidly evolving markets, particularly in Western regions, models must not only capture current trends but also anticipate swift changes in the temporal framework. While the study focused on spatial and socio-demographic factors, it is essential to acknowledge the potential influence of external factors such as macroeconomic policies, interest rates, and demographic shifts. These variables could impact real estate markets in ways that spatial data alone cannot fully capture. The exceptional nature of the COVID-19 pandemic presented an unprecedented challenge to real estate markets, causing fluctuations that fall outside the typical patterns captured by our model. Although the model was deliberately trained using only pre-COVID-19 data to focus on structural and local socio-demographic factors, the inclusion of 2020 prediction data allows for an evaluation of how severe external events can disrupt standard market dynamics. This approach demonstrates the model's limitations when faced with extraordinary circumstances, such as a global pandemic, and underscores the need for adaptive frameworks. By examining the model's performance during such an exceptional period, this study raises critical questions for future research. Developing adaptive models capable of integrating external shocks, including macroeconomic crises or public health emergencies, could improve robustness and predictive accuracy. Such advancements would ensure that valuation tools remain relevant and reliable, even in the face of unforeseen global events, further enhancing their real-world applicability.

Each model's predictive accuracy, measured in MAPE, was evaluated for both one and several years ahead predictions. The analysis focused on three specific configurations, the 1-year, 4-year, and 5-year setups, which proved to be the most effective input structures. The most significant improvements were observed in models trained exclusively on newly built apartment data. Notably, the 4-year newly built model consistently outperformed its base counterpart, achieving MAPEs of 14.08%, 16.65%, and 19.66% for one-, two-, and three-year predictions, respectively. In contrast, the corresponding base model recorded MAPEs of 20.79%, 22.87%, and 25.64% over the same periods. The reduction in MAPE highlights the crucial role of dataset homogeneity in improving model performance. Training exclusively on data from newly built properties yielded more accurate predictions, likely due to the reduced variance within this subset. While the base model failed to achieve an error rate below 20%, the results from the subset models demonstrate that integrating spatial features with carefully curated datasets significantly improves predictive accuracy. This approach not only improves short-term predictions but also ensures reliable long-term predictions, maintaining a MAPE mostly below 20% for predictions up to three years ahead.

6. Conclusions and Future Work

XG-Boost has proven to be an effective algorithm to predict housing prices with *Good to Reasonable* prediction accuracy. A key advantage of this method is the speed with which models can be set up, allowing for multiple model computations and extensive experimental iterations within a short period. However, making comparisons with studies from different markets, particularly in other countries, presents challenges due to the scarcity of long-term research on owner-occupied housing, especially in Western nations. The presented work establishes a baseline that needs to be improved in the future by adding additional data points and improving the quality of the data points. However, also the model needs improvement, e.g., by the automatic identification of breaks between zones of stable price development in both space and time, because an average error of 25% over longer periods of comparison and an error of 20% for subsequent years are not sufficient to keep up with the market. Despite these limitations, the models trained on newly built properties achieved significantly lower MAPE values, consistently below 20% for predictions up to three years ahead. This highlights the potential of leveraging homogeneous datasets to improve predictive accuracy and reliability. Also, the introduction of additional features such as buyer and seller classification or the marking of sales subsidized by the state may lead to great improvements. In addition, attributes of the buildings where the apartments are located could be added as a feature. Further analysis of spatial features, particularly those related to key POIs such as hospitals, could improve the model's understanding of neighborhood dynamics and accessibility, both of which play a critical role in price formation.

Understanding the impact of individual factors on housing prices is crucial. This is provided, e.g., by the analysis of feature importance. However, combinations of features may also need to be considered. Expanding the analytical approach to include deep learning methods could provide significant advances, as these techniques have recently shown considerable success in predicting house prices. Despite the differences between houses and owner-occupied apartments, the overarching similarities in the housing market suggest that these methods could be broadly applicable and effective.

Finally, the implementation of an AVM based on the proposed approach offers promising potential for practical applications. An AVM could provide automated and accurate price estimates, assisting real estate professionals, policymakers, and investors in decision-making processes. Future research should prioritize refining the current model to improve its applicability in real-world AVM systems. Before deployment, these models must undergo rigorous testing in experimental environments, involving real estate experts to evaluate their accuracy and practical utility. Plans are already underway to conduct such experiments, which will help ensure the robustness and adaptability of the model in a real-world context.

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

Funding: Christopher Kmen reports financial support was provided by OTTO Immobilien GmbH.

Data Availability Statement: Data not available due to legal and commercial restrictions.

Acknowledgments: The authors would like to express their sincere gratitude to OTTO Immobilien GmbH for their financial support and for providing access to the comprehensive real estate transaction database, which was invaluable to the success of this research. Open Access Funding by TU Wien.

Conflicts of Interest: Christopher Kmen reports a relationship with OTTO Immobilien GmbH that includes employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Pressrelease, S. Kaufpreise von Häusern und Wohnungen Stiegen im Jahr 2021 µm 12.3%. 2022. Available online: https://www.statistik.at/fileadmin/announcement/2022/05/20220324HPI2021.pdf (accessed on 19 April 2024).
2. Oesterreichische Nationalbank. Residential Property Price Index. 2023. Available online: https://www.oenb.at/en/Statistics/Standardized-Tables/Prices--Competitiveness/Sectoral-Price-Development/residential-property-price-index.html (accessed on 19 April 2024).
3. Rabiei-Dastjerdi, H.; McArdle, G.; Matthews, S.A.; Keenan, P. Gap analysis in decision support systems for real-estate in the era of the digital earth. *Int. J. Digit. Earth* **2021**, *14*, 121–138. [CrossRef]
4. OTS. Willhaben und IMMOunited Untersuchen Preisschere bei Wohnimmobilien 2020. 2020. Available online: https://www.ots.at/presseaussendung/OTS_20210628_OTS0054/willhaben-immounited-untersuchen-preisschere-bei-wohnimmobilien-2020 (accessed on 19 April 2024).
5. Limsombunc, V.; Gan, C.; Lee, M. House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *Am. J. Appl. Sci.* **2004**, *1*, 193–201. [CrossRef]
6. Lim, W.T.; Wang, L.; Wang, Y.; Chang, Q. Housing price prediction using neural networks. In Proceedings of the 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2016, Changsha, China, 13–15 August 2016; pp. 518–522. [CrossRef]
7. Koch, A.; Peremyslova, M.K.; Lemanowicz, L. Zestimate Bazinga: Predicting Selling Price for Condos in Downtown Vancouver. 2018. Available online: https://api.semanticscholar.org/CorpusID:145034093 (accessed on 10 November 2024).
8. Kang, Y.; Zhang, F.; Peng, W.; Gao, S.; Rao, J.; Duarte, F.; Ratti, C. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy* **2021**, *111*, 104919. [CrossRef]
9. Zeng, L. Research on Batch Evaluation of Real Estate Price Based on XGBoost. *Bcp Bus. Manag.* **2021**, *16*, 326–333. [CrossRef]

10. Srirutchataboon, G.; Prasertthum, S.; Chuangsuwanich, E.; Pratanwanich, P.N.; Ratanamahatana, C. Stacking Ensemble Learning for Housing Price Prediction: A Case Study in Thailand. In Proceedings of the KST 2021–2021 13th International Conference Knowledge and Smart Technology, Bangsaen, Chonburi, Thailand, 21–24 January 2021; pp. 73–77. [CrossRef]
11. Rosen, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *J. Political Econ.* **1974**, *82*, 34–55. [CrossRef]
12. Brunauer, W.; Lang, S.; Umlauf, N. Modelling house prices using multilevel structured additive regression. *Stat. Model.* **2013**, *13*, 95–123. [CrossRef]
13. Abidoye, R.B.; Chan, A.P. Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network. *Pac. Rim Prop. Res. J.* **2018**, *24*, 71–83. [CrossRef]
14. Zhan, W.; Hu, Y.; Zeng, W.; Fang, X.; Kang, X.; Li, D. Total Least Squares Estimation in Hedonic House Price Models. *Isprs Int. J. Geo-Inf.* **2024**, *13*. [CrossRef]
15. 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]
16. Čeh, M.; Kilibarda, M.; Lisec, A.; Bajat, B. Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *Isprs Int. J. Geo-Inf.* **2018**, *7*, 168. [CrossRef]
17. 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.
18. Law, S.; Jeszenszky, P.; Yano, K. Examining geographical generalisation of machine learning models in urban analytics through street frontage classification and house price regression. In Proceedings of the GIScience 2021 Short Paper Proceedings, Poznań, Poland, 27–30 September 2021. [CrossRef]
19. Cavailhès, J.; Brossard, T.; Foltête, J.C.; Hilal, M.; Joly, D.; Tourneux, F.P.; Tritz, C.; Wavresky, P. GIS-based hedonic pricing of landscape. *Environ. Resour. Econ.* **2009**, *44*, 571–590. [CrossRef]
20. Yiorkas, C.; Dimopoulos, T. Implementing GIS in real estate price prediction and mass valuation: The case study of Nicosia District. In Proceedings of the Fifth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2017), Paphos, Cyprus, 20–23 March 2017; Volume 10444, pp. 112–128.
21. Valier, A. Who performs better? AVMs vs hedonic models. *J. Prop. Investig. Financ.* **2020**, *38*, 213–225. [CrossRef]
22. Kok, N.; Koponen, E.L.; Martínez-Barbosa, C.A. Big data in real estate? From manual appraisal to automated valuation. *J. Portf. Manag.* **2017**, *43*, 202–211. [CrossRef]
23. Li, L.; Chu, K.H. Prediction of real estate price variation based on economic parameters. In Proceedings of the 2017 IEEE International Conference on Applied System Innovation: Applied System Innovation for Modern Technology, ICASI 2017, Sapporo, Japan, 13–17 May 2017. [CrossRef]
24. Helbich, M.; Brunauer, W.; Vaz, E.; Nijkamp, P. Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria. *Urban Stud.* **2014**, *51*, 390–411. [CrossRef]
25. Helbich, M.; Jochem, A.; Mücke, W.; Höfle, B. Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning. *Comput. Environ. Urban Syst.* **2013**, *39*, 81–92. [CrossRef]
26. De Nadai, M.; Lepri, B. The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–3 October 2018; pp. 323–330. [CrossRef]
27. Avanijaa, J.; Gurram, S.; Reddy, K.M. Prediction of house price using xgboost regression algorithm. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 2151–2155.
28. Peng, Z.; Inoue, R. Specifying multi-scale spatial heterogeneity in the rental housing market: The case of the Tokyo metropolitan area. In Proceedings of the GIScience 2021 Short Paper Proceedings, Poznań, Poland, 27–30 September 2021.
29. Shen, H.; Li, L.; Zhu, 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*, 53. [CrossRef]
30. Antipov, E.A.; Pokryshevskaya, E.B. Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Syst. Appl.* **2012**, *39*, 1772–1778. [CrossRef]
31. Pai, P.F.; Wang, W.C. Using machine learning models and actual transaction data for predicting real estate prices. *Appl. Sci.* **2020**, *10*, 5832. [CrossRef]
32. Hu, L.; He, S.; Han, Z.; Xiao, H.; Su, S.; Weng, M.; Cai, Z. Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy* **2019**, *82*, 657–673. [CrossRef]
33. Tan, F.; Cheng, C.; Wei, Z. Time-aware latent hierarchical model for predicting house prices. In Proceedings of the IEEE International Conference on Data Mining, ICDM, New Orleans, LA, USA, 18–21 November 2017; pp. 1111–1116. [CrossRef]
34. Trojanek, R.; Gluszak, M. Spatial and time effect of subway on property prices. *J. Hous. Built Environ.* **2018**, *33*, 359–384. [CrossRef]
35. Chaiwuttisak, P. Latent topic analysis of the post property for sales to predict a selling price of second-hand condominiums. In *Proceedings of the Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2021; Volume 2050, p. 012005.
36. StatistikAustria. Abgestimmte Erwerbsstatistik. 2023. Available online: https://www.statistik.at/en/about-us/surveys/register-based-census/register-based-labour-market-statistics (accessed on 19 April 2024)

37. Fogliaroni, P.; Bucher, D.; Jankovic, N.; Giannopoulos, I. Intersections of our world. In *Proceedings of the Leibniz International Proceedings in Informatics, LIPIcs*; Schloss Dagstuhl—Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing: Wadern, Germany, 2018; Volume 114. [CrossRef]
38. Huang, Y.; Lieske, S.N.; Wang, S.; Liu, Y. How does heterogeneity in dwelling type preferences relate to housing and built environment characteristics? *Int. J. Digit. Earth* **2023**, *16*, 93–112. [CrossRef]
39. WienerLinien. Modal Split for Vienna. 2022. Available online: https://www.wienzufuss.at/2022/03/30/wienerinnen-und-wiener-sind-klimafreundlich-unterwegs-44-aller-wege-werden-mit-dem-rad-oder-zu-fuss-erledigt/ (accessed on 19 April 2024).
40. Boeing, G. OSMnx 1.9.2 Documentation, 2016–2024. Available online: https://osmnx.readthedocs.io/en/stable/ (accessed on 19 April 2024).
41. Yang, Y.; Diez-Roux, A.V. Walking distance by trip purpose and population subgroups. *Am. J. Prev. Med.* **2012**, *43*, 11–19. [CrossRef]
42. Liang, X.; Liu, Y.; Qiu, T.; Jing, Y.; Fang, F. The effects of locational factors on the housing prices of residential communities: The case of Ningbo, China. *Habitat Int.* **2018**, *81*, 1–11. [CrossRef]
43. Kearns, M. Thoughts on hypothesis boosting. *Unpubl. Manuscr.* **1988**, *45*, 105.
44. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
45. Lewis, C. *Industrial and Business Forecasting Methods: A Practical Guide to Exponential Smoothing and Curve Fitting*; Butterworth Scientific: Oxford, UK, 1982.

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