Real Estate Valuation Decision-Making System Using Machine Learning and Geospatial Data

Ahmet Yagmur
Department of Business Information Systems, Hochschule Furtwangen University, Robert-Gerwig-Platz 1, 78120 Furtwangen im Schwarzwald

Abstract

Forecasting and analyzing the real estate market is crucial globally due to the presence of investors, prospective property owners, and sellers motivated by various reasons. These motivations are universal and can arise anywhere on the globe. However, specific factors such as geographical and geo-spatial locations, geological attributes, and human foresight regarding the future value of real estate make this a challenging research area. With an abundance of scientific data available, it is indeed feasible to predict prices or establish reasonable valuation intervals. While existing studies have integrated multiple data sources, including house imagery [14, 16] and economic indicators [15], this article will focus solely on geo-spatial data. Economic conditions, property conditions, and construction years will not be examined in this discussion. Although research has acknowledged the predictive capabilities of geo-spatial data, there is a scarcity of studies that isolate its unique aspects. This research aims to fill that gap by exploring how geo-spatial features can be processed and utilized within machine learning algorithms to inform accurate real estate value predictions. Employing machine learning techniques, such as Random Forest Regressor for price prediction and XGBoost Regressor for valuation intervals, this study analyzes geo-spatial data from the 2024-2025 German real estate market. The prediction pipeline computes geo-related scores, like distances to the nearest airport and city center, before running the models. The findings intend to enhance decision-making for real estate stakeholders by offering insights into accurate, geo-spatially informed valuations. This research aspires to inspire future studies incorporating additional factors for a comprehensive understanding of the real estate market.

Keywords: Real Estate Price Prediction, Machine Learning in Real Estate, Geo-spatial Data, Spatial Data Analysis, Random Forest Regression, XGBoost Regression, German Real Estate Market, Big Data

1. Introduction

1.1 Growing Importance, Factors of Real Estate Valuation and Need of Predictive Models

Predictive modeling is essential in real estate, where accurate forecasts improve decision-making in a competitive market. Investors, sellers, and buyers face significant financial stakes, making informed choices critical for maximizing value. The real estate market's

complexity, with diverse properties and fluctuating prices, requires continuous assessment of influencing factors. Price trends establish baselines, demanding ongoing analysis over time.

Advancements in machine learning (ML) enable efficient handling of complex datasets. Integrating geographical, financial, and real estate data enhances forecasting accuracy. Key valuation factors include living size, location (proximity to amenities, neighborhood safety, cleanliness, and desirability), and the age of the property [7, 8]. ML can analyze details ranging from micro-developments to slight changes in square footage or proximity to grocery stores—factors often overlooked in traditional assessments.

As more data becomes available, advanced predictive models offer stakeholders improved tools for analysis. These innovations simplify transactions and create a more transparent, data-driven marketplace.

1.2 The Importance of Geospatial Scores

Geospatial data has become vital in predictive modelling, especially in real estate, where location heavily affects property values. With advancements in technology like remote sensing, IoT, and satellite mapping, spatial data is now easily accessible, enhancing its precision and allowing for thorough analysis of location-based factors. Proximity to amenities such as schools, shopping centers, and parks, as well as access to transit, significantly influences real estate prices. These elements are crucial for daily convenience and increase property appeal and value. Such spatial details highlight the possible worth of real estate while bridging the gap between broad economic indicators like market trends and localized factors. For instance, assessing distances to the nearest city center or airport can reveal how closely a property connects to major economic hubs. The surrounding infrastructure further enriches context, allowing for more accurate price predictions that extend beyond general economic insights. Thus, effective processing and selection of relevant geospatial data are vital in real estate valuation.

The study processes geospatial data to generate a score based on location, including proximity to airports, city centers and key locations. As stated in Rey-Blanco, Zofío and González-Arias (2024) paper the one way can be used is introducing the geospatial data is as variables to the machine learning model as features [5]. These estimated values are integrated into machine learning models such as random forest regressor for price prediction and XGBoost regressor to evaluate the valuation intervals. This approach emphasizes how geographical spatial data completes the conventional evaluation method and provides more subtle and measured opinions on real estate prices. This highlights the growing importance of spatial analytics to create accurate data-driven models that benefit real estate stakeholders.

1.3 Questions about fairness and scaling

Fairness of predictive models is a critical consideration, and there debate on machine learning models fairness and its bias in sensitive areas such as justices, commercial use and health care [6]. Especially for research commercial use where geospatial data is involved and where valuations are based on geospatial data in real estate pricing. This project has taken steps to ensure fairness by developing a data collection process that avoids potential self-inflicted biases. In particular, equal representation was maintained by collecting the same number of property registers in all 233 cities across Germany: cities with fewer records were excluded

from the analysis to maintain consistency and avoid biased results due to missing data. In addition to that, the dataset is filtered to avoid obvious outliers, some ads that may lead to false training, and misleading ads whose sole purpose in uploading is to provide free advertising to real estate professionals. These approaches provide a balanced dataset, ensuring that no single city has a disproportionate influence on the model results.

From a calibration perspective, optimizing the reliability of predictions has been a priority since the model is wanted for only a precise predictions on valuation. Furthermore, calibration in classification which means probability estimates or combination of models and probability estimates would give a meaningful information. However, that is really hard according to research that conducted Pleiss et al. (2017) [19]. Different regression models, including Random Forest, Lasso, Linear, Polynomial, and XGBoost, were tested to identify the most suitable algorithms for the task. As a matter of course, the machine learning models that uses deep bagging methods such as decision trees needs an comprehensive hyperparameter tuning phases performed on machine learning models due to the fact that it is prone to be overfitted on data [1]. Random Forest emerged as the optimal choice for price prediction due to the fact that it builds is an aggregation of decision trees, in general Random Forest is trained with the bagging method so that the forest of decision trees can indicate high precision and accurate predictions [1]. For price interval predictions, XGBoost was selected. Owing to the fact that it is an open-source machine learning library that uses group gradient boosted decision trees. These trees are constructed boosted mechanism means, sequential ensembles, errors made by previous decision trees are corrected by subsequent trees in the sequence. The grouping of trees approach contributes to improving predictive performance of XGBoost model [2]. These models underwent thorough evaluation to ensure they not only provide accurate results but also maintain computational efficiency since the machine learning models are well tuned.

The research aims to understand how location-specific factors impact property pricing in Germany, rather than solely focusing on forecasting accuracy. It evaluates factors like proximity to airports and amenities, revealing regional differences. The study identifies the most expensive city and assesses if machine learning models can ensure fair price listings.

Although the models are well-tuned to the dataset, questions remain about the completeness of the feature set. The research does not explore how additional features, such as infrastructure, historical trends, or demographic data, might improve predictive accuracy. These aspects are covered in the literature.The goal is to align machine learning capabilities with real estate valuation complexities, considering geological factors to ensure models are fair, precise, and practical for stakeholders.

1.4 Data Collection, Ethical Considerations and Regulatory Impact

The integration of geospatial data in predictive modeling for real estate valuation introduces significant ethical concerns, especially during dataset construction. This type of data may unintentionally act as proxies for sensitive attributes like socioeconomic status, age, education, and ethnicity, leading to potential indirect discrimination. Such biases risk skewing property valuations, where models might unfairly over- or under-value properties based on the embedded racial or age-related information, ultimately favoring or disadvantaging specific

To address ethical issues, transparency in data use and modeling is essential, alongside identifying and eliminating biases in data collection. This research ensured equal

representation across cities and anonymized sensitive real estate information like owner's name, race, age, profession, and marital status. Additional measures such as justice and audits are needed to assess and rectify unintended model displacement. With that said, the dataset is scraped from Germany's one of biggest real estate web site called Immowelt [10]. While there are no identity determinator data is scraped, potential data which might cause bias is also cleansed. The address is counted as semi-private data when the case the full address is recorded. That is why, the address data is collected as street, city, postal code and state. There are no apartment number recorded in the dataset, associated with a record.

From a regulatory standpoint, regulations such as the General Data Protection Regulation (GDPR) in the European Union enforce rigorous mandates regarding the utilization of personal data. GDPR highlights the importance of minimizing data, limiting its use to specific purposes, and obtaining informed consent when handling datasets with potentially identifiable information. Although geospatial data does not inherently reveal individual identities, merging it with other variables may pose privacy threats, requiring anonymization or aggregation methods to protect individual rights.

Additionally, the European Union Artificial Intelligence (EU AI) Act highlights the significance of transparency and accountability in machine learning usage, particularly in critical domains like real estate assessment. The EU AI Act will categorize AI systems based on their risk levels and enforce stringent requirements on those deemed high risk [3]. For example, models used in real estate pricing must provide clear explanations of their predictions, ensure the traceability of decision-making processes and respect the principles of fairness and non-discrimination as well as forcing itself to be as transparent as possible [3].

Integrating ethical and regulatory guidelines into machine learning workflows is essential for legal compliance and stakeholder trust. This involves documenting data preprocessing, justifying feature selection, and ensuring model interpretability. Adhering to these principles can foster a fairer real estate market and mitigate ethical and regulatory risks.

1.5 Granularity vs. Aggregation

Establishing the appropriate level of data granularity is essential in geospatial analysis for property valuation. Granularity denotes the degree of detail found in datasets, including property records and associated statistics. The appropriate granularity improves model accuracy and understanding, directly influencing stakeholders.

Excessively detailed data can lead to significant variability in model results. For instance, data on individual properties might include outliers that distort predictions. This sensitivity may diminish the model's resilience. Extremely detailed data may also elevate computational complexity and storage requirements, reducing the overall efficiency of the process.

Conversely, excessively consolidated data may hide significant insights. For example, aggregating data at the city or state level might miss the influence of local amenities. This may result in broad predictions, overlooking distinct neighborhood-specific factors that influence property values.

Achieving the correct equilibrium between granularity and aggregation is crucial for precise predictions. In this study, this was accomplished by guaranteeing equal representation among all 233 cities and omitting incomplete or inconsistent data sets.

The level of granularity is determined by the aim of the analysis. For urban price trends, a wider aggregation might be appropriate. For comprehensive property assessments, a greater level of detail, taking into account elements such as closeness to transit, educational institutions, and retail areas, is favored. In this research, mailing addresses were utilized to evaluate the properties.

Ultimately, selecting details requires balancing interpretability with model complexity. Instruments such as clustering and spatial hierarchies assist in ensuring that predictive models deliver precise insights for participants in the real estate market.

1.6 Research Question

"How can machine learning and geospatial data improve the accuracy and efficiency of real estate valuation and decision making on the real estate market?"

Hypothesis

A data science-driven valuation model will yield more accurate results compared to traditional methods with feeding the model appropriate training data embedded the geologic data inside.

Discussion

The integration of machine learning (ML) and geospatial data can transform real estate valuation. Traditional methods often rely on static models or human expertise, which struggle to account for the complexity of modern markets. Machine learning enables processing large datasets, uncovering hidden patterns, and making data-driven predictions, offering more accuracy than conventional methods. This decision-making mechanism can then be embedded in automated pipelines.

The research suggests that ML approaches offer higher accuracy by integrating multiple features and interactions, including those from geospatial data. Geospatial data adds detail, considering factors like amenities, transportation, environmental conditions, and local characteristics. These spatial variables are often overlooked in traditional models, but they significantly impact property values.

2. Nature of Data and Its Use

2.1 Types of Data

ad_id	Int	Unique ID for all records.
street	Str	The street name of the real estate.
city_code	Int	The zip code and the city name of the real estate.
price	Float	The price of the real estate.
number_of_rooms	Int	The number of rooms of the real estate.

living_area	Float	The size of living area of the real estate.
land-size	Float	The size of land area of the real estate (if exist).
URL	Str	The URL of the real estate ad.
city_id	Int	Unique city ID for all the unique city, derived for the dataset. It does not have any relation with reality.
city_score	Float	The city score calculated with statistical methods
population	Int	The population of the city that real estate located.
geo_spatial	Int	The score of amenities around the real estate
center_distance	Float	The score of closeness of the real estate to the nearest city center
airport_distance	Float	The score of closeness of the real estate to the nearest airport

Table 1: Columns and Explanations

In the context of real estate price prediction, the effectiveness of machine learning models depends heavily on the nature and quality of the data used. This study utilizes three primary types of data: structured data, and geospatial data. Each type plays a critical role in building a comprehensive and accurate predictive model.

2.2 Structured Data

Structured data refers to highly organized, numerical information that fits neatly into tables or databases. In this research, structured data includes variables such as property prices, size of living area, size of land area and number of rooms. These features are quantifiable and serve as the backbone of almost all of real estate valuation models, offering direct correlations with property value. For instance, properties with larger size of living area or larger size of land area, typically fetch higher prices than those which has smaller size of the attributes.

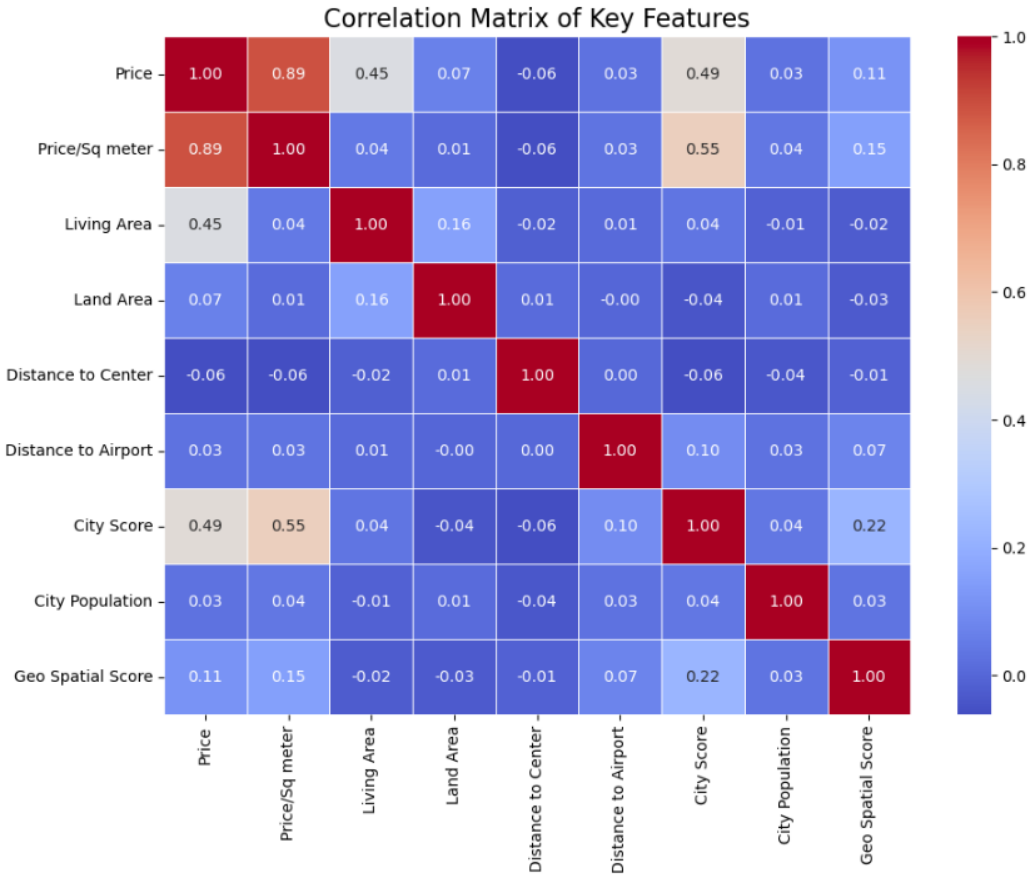

Figure 1: Correlation Matrix of Key Features

This type of data is well-suited for machine learning algorithms because it allows for straightforward feature extraction and analysis. Furthermore, structured data provides the foundational inputs for models such as Random Forest and XGBoost, which rely on numerical features directly or categorical values that turned into numerical format for training and prediction. However, the predictive power of structured data is limited without the inclusion of more contextual variables, which are addressed through geospatial and textual data.

2.3 Geospatial Data

Geospatial data is a critical component in modern real estate valuation, offering insights into the geographic and locational attributes of a property. This data includes street, zip code, city name, and some other city and state identity determination number are derived for specifically for the dataset of the real estate. There are four types of geospatial data in the dataset. They are saved as numerical scores to corresponding addresses. They are, City Score, Geospatial Score, Distance to City Center Score and Distance to Airport Score. City Score: A score that is calculated with statistical analysis on mean real estate prices for each city that be part of the dataset. Geospatial Score: Distances to key amenities (super markets, convenience stores, shopping centers, Variety Stores, Parks, restaurants), and proximity to public transportation hubs or major stations. The same stations (arrivals and departures are counted as one station). Distance to City Center Score: A score that is calculated upon the distance from the nearest city center in km. Distance to Airport Score: A score that is calculated upon

the distance from the nearest city center in km. However, the In particular cases particular cities has a small airport serving for small plane or educational and hobby purpose flights. The influence of geospatial data is intense. That is the reason expectance from the geospatial data to impact real estate values cannot be neglected.

Supermarket, Convenience Store, Variety Store	100 m	20
	300 m	15
	600 m	10
	1000 m	5
Bus Stop, Train Station	200 m	25
	500 m	15
	1000 m	5
Park, Fast Food	100 m	10
	300 m	7
	1000 m	5

Table 2: Geospatial Scoring Logic

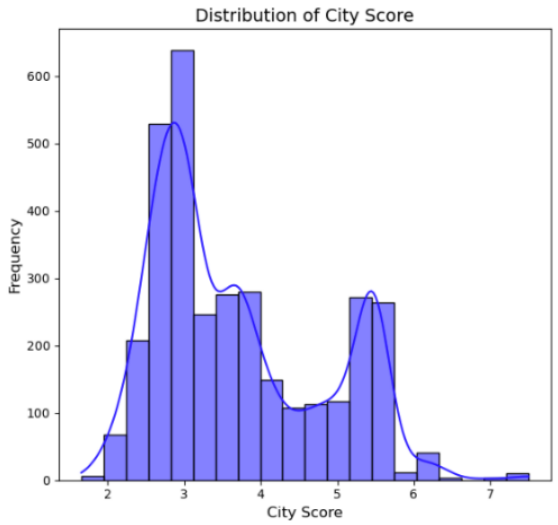

Figure 2: Distribution of City Score

That distribution in Figure 2 indicates that there is a density on city score 2.5 to 4. Since the same amount of the real estates have been taken from 233 cities, most of on scope city has a city score 2.5 to 4.

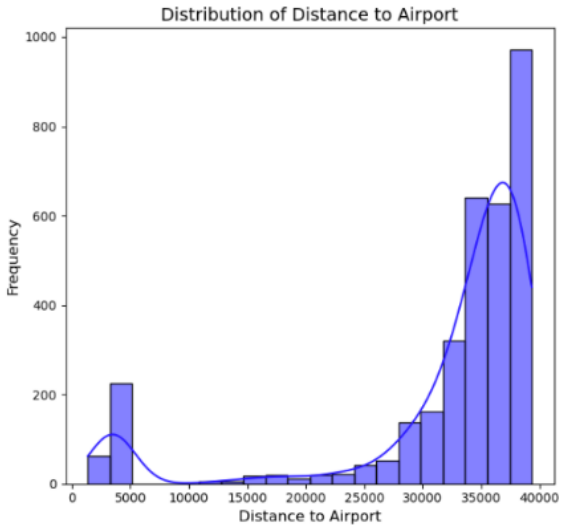

Figure 3: Distribution of Distance to Airport Score

Figure 3 indicates that, most of the real estates has a location far from airport. In contrast, there are some which are pretty close to airport. However, those are small airport for hobby and education purpose.

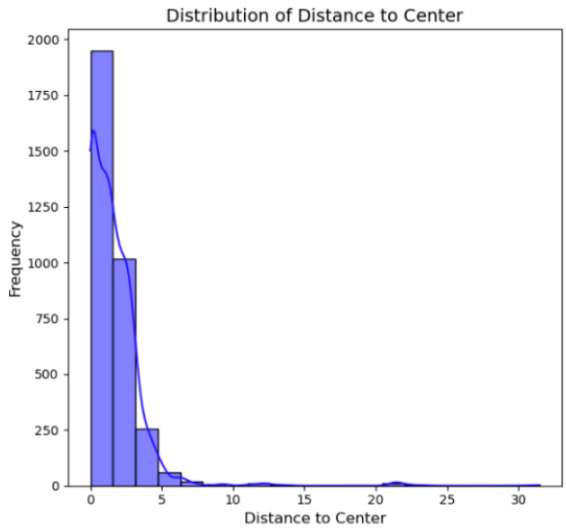

Figure 4: Distribution of Distance to Center Score

According to Figure 4 the most of the real estates located in around the center and in a close diameter. However there are minority of real estates which located rural areas

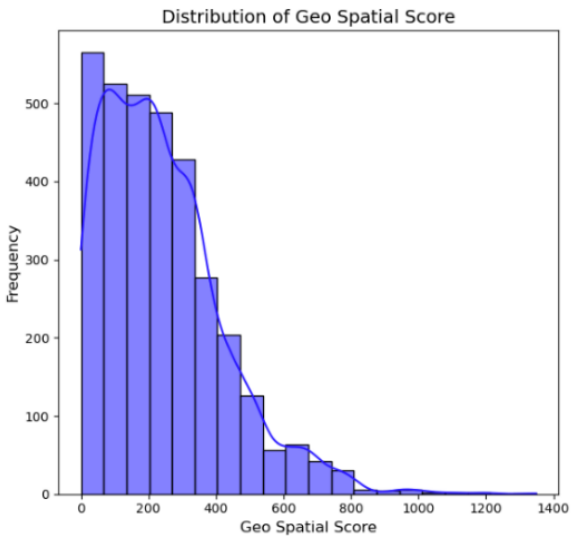

Figure 5: Distribution of Geo Spatial Score

According to distribution in Figure 5, it confidently can be said, that majority of the real estates have at least decent amount of amenities and public services provided by German Government nearby.

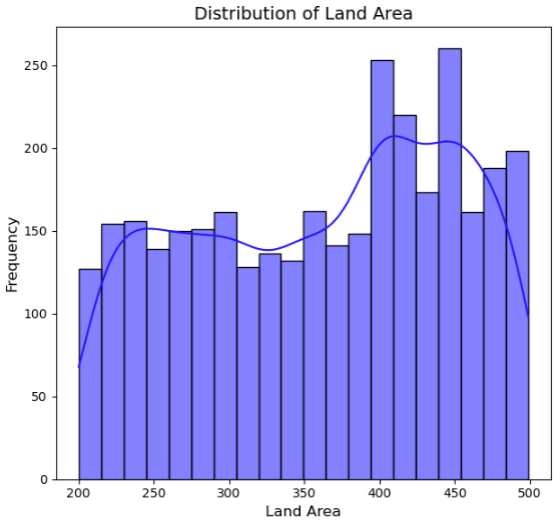

Figure 6: Distribution of Land Area

From the distribution shown in Figure 6, all of the real estates used to train the model, have an special area for home owner to use, there can be a garage or a green field.

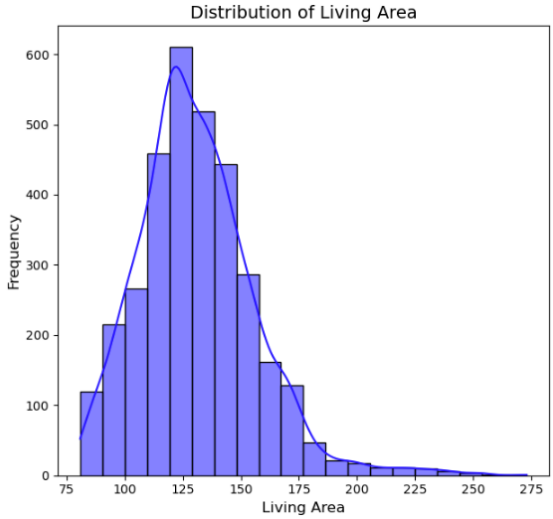

Figure 7: Distribution of Living Area

Depends on the sample visualized on the graph in Figure 7 indicates, in Germany real estates have 125 m^2 of living area size, most frequently.

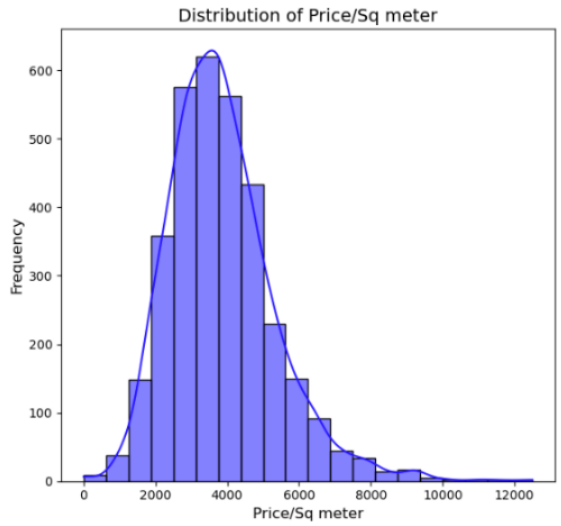

Figure 8: Distribution of Price per m^2

As shown in the Figure 8, in German real estate market, The distribution appears fairly symmetrical with a slight tail to the right (positive skew), which suggests a small number of very high-priced properties affecting the average. Due to the wide variability in price per square meter, the MAE metric remains relatively high, even after thorough data cleansing.

Moreover, geospatial data enhances the ability to create location-specific insights, such as analyzing regional price trends or identifying hotspots for real estate development. When integrated with structured data, geospatial variables significantly boost the predictive accuracy of machine learning models by capturing spatial dependencies often overlooked in traditional valuation methods.

This research aims to develop a robust and comprehensive valuation model by combining two types of data that are merged into the dataset: structured and geospatial data. Each data type complements the other to create a comprehensive data set covering both quantitative and qualitative aspects of real estate valuation, providing more accurate and actionable predictions.

2.4 Limitations

Creating accurate and efficient real estate valuation models presents several limitations. The first challenge was finding a suitable dataset that met all the project's goals. Constraints included incomplete or inconsistent data, time restrictions due to rapidly changing markets, and difficulties integrating heterogeneous data sources. Collecting data from a sufficient number of records across various regions and cities was necessary to build a diverse dataset. Since no existing dataset on the web contained all the required information, the dataset had to be created from scratch. On top of that, the unclarity of model selection is the big portion of the limitation, since choosing a machine learning model is unique for the dataset itself and the features which dataset has [11]. Addressing these limitations is critical for improving the reliability and robustness of the model via robustness of the dataset.

2.5 Missing or Inconsistent Data

A major constraint in this study is the lack of complete or consistent data. Certain values in CSV files are labeled as "Unknown" because of character encoding problems, especially with German characters that are absent in the English alphabet. Moreover, some property listings are missing essential details, like when the price is marked as "upon request," necessitating

the removal of those records. This results in missing data in the dataset, complicating the creation of a completely representative model.

Varying data formats present difficulties as well. For instance, addresses might be inputted inaccurately, or property dimensions might be given in varying units (square meters compared to square feet). These inconsistencies necessitate additional preprocessing, including cleaning and normalization, resulting in increased time and resource expenditures for data preparation.

In some cases data is filtered to handle the outliers, since the outliers data prone to potential false assessments. In order to do that, the Germany's official real estate statistics database is reviewed.

Living floor space per dwelling	Living floor space per inhabitant	Rooms	Rooms per dwelling
sq m	sq m	number	number
91.6	46.2	182,295,713	4.4
91.7	46.3	183,354,291	4.4
91.8	46.5	184,427,760	4.4
91.8	46.7	185,491,224	4.4
91.9	47.0	186,594,482	4.4
92.0	47.4	187,746,588	4.4
92.1	47.7	188,829,383	4.4
92.2	47.4	189,920,514	4.4
92.2	47.5	190,985,570	4.4

Figure 9: GENESIS-Online Database - German Real Estate Market (2015-12-31 / 2023-12-31) [17, 18]

And from there, valuable insights are gained for the filtration. According to statistical analysis of house features: Average m^2 per room is 20.95 m^2 for the year 2023. The average m^2 per room from 2015 to 2023 is respectively: 20.81, 20.84, 20.86, 20.86, 20.88, 20.90, 20.93, 20.95, 20.95. And standard deviation of m^2 per room is 0.049

Count, N: 9
Sum, Σx: 187.98
Mean, x̄: 20.886666666667
Variance, s2: 0.00245

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

Figure 10: Standard Deviation Formula

To address these issues, imputation techniques are used to fill missing values, and columns with insufficient records are dropped. Robust preprocessing methods are applied to standardize data when gaps are large. However, these solutions have trade-offs. Imputed values are based on assumptions, which can introduce bias if they do not reflect real-world scenarios.

2.6 Time Constraints

Real estate markets are dynamic, with property values fluctuating based on governmental and administrative changes, economic conditions changes particularly one half to another, policy changes, and local developments. As such, ensuring that the data remains current is a significant challenge. Delayed or outdated data can lead to predictions that are no longer accurate or relevant.

The impact of time on a property's age (construction year) and on the economic events affecting the country is inherent to the nature of time and cannot be overlooked [20]. Real-time data processing and model training are needed to keep the model up-to-date and improve prediction accuracy without downtime. However, this requires significant computational resources, a well-tuned machine learning model, and an up-to-date dataset pipeline.

This process has two parts. First, the model must be trained with the latest real estate data. Second, the input address, provided by the user, needs to be segmented for further scoring. The zip code will determine the city score. The address will then be used in geospatial functions to retrieve the longitude and latitude of the property, enabling score calculation.

To support these steps, a file storage or database system is required to retrieve the corresponding scores for the real estate addresses.

Time constraints impact model training and validation. Training a machine learning model with large, complex datasets requires significant computational effort and can be time-consuming. Regular retraining to include new data adds to this challenge.

2.7 Addressing the Limitations

To overcome these limitations, several strategies are employed. Data cleaning and normalization address a wide range of values, outliers, and missing or incorrect data. These steps ensure the dataset is as complete and uniform as possible. Real-time processing is explored to maintain data freshness and improve prediction accuracy. Advanced integration methods, such as feature engineering and machine learning pipelines, help merge data sources effectively.

While these approaches mitigate challenges, they cannot fully eliminate the limitations of the data and methodologies. Therefore, model results should be interpreted with an understanding of these constraints. Since training data is rarely perfect, tolerance is added to the model to accommodate these imperfections.

2.8 Data Use

The data used in this research serves multiple purposes, allowing for a comprehensive approach to real estate valuation. The two main aspects of data use are valuation modeling

and interaction analysis.

2.9 Valuation Modeling

The primary use of the data is to develop machine learning-based valuation models that analyze how environmental factors affect property prices. The models evaluate attributes such as proximity to key amenities, city score, and distance to city centers and airports by integrating structured and geospatial data. This approach allows for a detailed understanding of how locational and property-specific factors contribute to real estate pricing, offering a more data-driven alternative to traditional methods. On the other hand, valuation with ML model has some downsides since that is a process of automated statistical calculation algorithm, According to Nicastro (2024), [9] those reasons are training the ML model with wrong information dataset such as a dataset with huge outliers, using outdated samples, or cannot find a full disclosure dataset. If one of the scenario is the case, it cause wrong valuation on real estate pricing.

2.10 Interaction Analysis

Beyond valuation and price forecast, the data is used to study the interactions between amenities, locations, and nearby property prices or shortly term “neighborhood characteristics" [21]. This includes assessing the influence of social amenities such as schools, shopping centers, and transit stations on real estate values. For instance, high scored location (according to scoring functions used in the machine learning model) correlates with increased demand for real estate, leading to higher valuations. By capturing these interactions, the research provides meaningful insights for stakeholders ranging from real estate investors to urban planners.

3. Technologies and Tools

3.1 Libraries and Frameworks

This research leverages a range of Python libraries and tools to handle data analysis, machine learning, geospatial computations, and data cleansing tasks. Each technology plays a specific role in the pipeline, ensuring efficiency and accuracy in processing and modeling.

3.2 Geospatial Analysis Tools

Geospatial data is essential for real estate valuation and all the focus of this research on Geospatial features. To effectively analyze and incorporate geospatial data, this study used a combination of specialized tools and APIs. Geospatial platforms such as QGIS, ArcGIS, Google Maps API, and OpenStreetMap (OSM), the latter being free and open-source, are commonly used as types of Geographic Information Systems (GIS) [13].

QGIS: This open-source platform is used for geospatial analysis and mapping. It offers tools to visualize, process, and analyze spatial data, making it suitable for detailed geospatial calculations.

ArcGIS: A commercial alternative to QGIS, ArcGIS provides advanced tools for analyzing and visualizing spatial data. Its geoprocessing tools are useful for complex real estate valuation with multi-layer data.

Google Maps API: This API allows geocoding and distance calculations, helping to determine proximity to key places like city centers, transport hubs, schools, and hospitals. Its real-time capabilities enhance geospatial data accuracy.

OpenStreetMap (OSM): OSM is a free, open-source alternative that provides detailed geographic data. It can be integrated into Python workflows using libraries like osmnx for proximity calculations and network analysis.

Integration and Customization: Custom geospatial functions in Python were developed to integrate APIs like Google Maps and libraries like Geopy, ensuring efficient and accurate geospatial data processing. Specialized libraries were also created to score addresses for the machine learning pipeline. These tools and APIs work together to provide comprehensive geospatial analysis for property valuation models.

3.3 Machine Learning Models and Used Metrics

In the literature review, previous studies on identifying the optimal machine learning model for real estate valuation remain inconclusive as this area of research is still relatively new. Although no single machine learning algorithm is deemed most appropriate for that particular task, studies [11] indicate that SVM falls short when compared to RF (Random Forest) and GBM (Gradient Boosting Machine). In some cases Random Forest performs well compared to XGBoost algorithm as other researchers indicates this fact [12]

In this research two specific algorithms—Random Forest and XGBoost—were employed to address the complexities of the data and ensure optimal performance with most high accuracy possible with given dataset.

Random Forest

Antipov and Pokryshevskaya (2012) believe that, Random Forest is the one of right and proper model for “mass appraisal”. Random Forest is a versatile ensemble learning method that operates by constructing multiple decision trees during training and averaging their predictions for accuracy [23]. This model is particularly effective in handling:

Feature Importance: By evaluating the contribution of each variable to the model’s predictions, Random Forest helps identify key drivers of property value, such as proximity to amenities or size attributes.
Robustness: It minimizes the risk of overfitting by combining multiple trees and averaging their outcomes, which is especially beneficial when working with datasets containing a mix of structured and geospatial data.
Ease of Interpretation: Feature importance scores and straightforward implementation make Random Forest a reliable baseline model in real estate valuation tasks.

XGBOOST

XGBoost (Extreme Gradient Boosting) is a powerful algorithm optimized for speed and performance. It is well-suited for modeling complex interactions in real estate data [24]. Key advantages include:

Handling Complex Relationships: XGBoost captures non-linear patterns and interactions between variables, such as how the combination of proximity to amenities and property size might influence value.

Efficiency: Its computational efficiency ensures faster training and prediction, even with large datasets.
Regularization: XGBoost includes built-in regularization techniques to prevent overfitting, making it ideal for datasets with many features or potential noise.

3.4 Model Integration and Comparison

Both models were applied to the dataset to evaluate their predictive capabilities. Random Forest served as an interpretable and robust benchmark, while XGBoost provided an advanced mechanism to refine predictions for more complex relationships. By comparing the performance metrics—such as Mean Absolute Error (MAE) and R^2 scores—the strengths of each model were leveraged for accurate real estate valuation.

The combinations of Random Forest and XGBoost in a way, ensures that the research model balances interpretability, and predictive accuracy, addressing the diverse requirements of modern real estate valuation.

3.5 Evaluation and Comparison Metrics

MAE

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Figure 11: Formula of MAE

Mean Absolute Error (MAE) to assess the effectiveness of regression models [26]. By calculating the predicted value's MAE and compering the other models

MAD

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - m|$$

Figure 12: Formula of MAD

Mean absolute deviation (MAD) is a measure of the average absolute distance between each data value and the mean of a data set [27]. In the research MAD value of predicted intervals is scaled by 20% to be added into intervals as a tolerance.

Creating Tolerance for Price Prediction Range
The Mean Absolute Deviation (MAD) is calculated as 1596.41, while the scaled MAD value using min-max scaling is determined to be 319. MAD is less sensitive to skewed data distributions compared to the standard deviation, making it a suitable metric for datasets with positive skewness, such as price per square meter in this case.

The purpose of the research is not to assign a specific price to real estate properties but to provide a price range that reflects the variability and sensitivity associated with other features

of the property. The inclusion of MAD ensures a more robust tolerance in predictions, considering the variability in data.

While geo-spatial data remains an essential factor in real estate valuation, other attributes—such as the construction year of the property and economic conditions—also significantly affect the valuation process.

By utilizing MAD for tolerance, the methodology acknowledges the complexity and multi-dimensional nature of real estate valuation, focusing on building a flexible and reliable prediction range.

3.6 Percentage of Accuracy and Findings

In this research the accuracy of prediction is calculated by a range which is an interval of predicted minimum price per m^2 subtracted by scaled MAD score multiplied by the real estate's actual living size and predicted maximum price per m^2 in addition on scaled MAD score multiplied by the real estate's actual living.

The overall accuracy of the test dataset is 62.63% with the city the real estate located is the most important feature in the model.

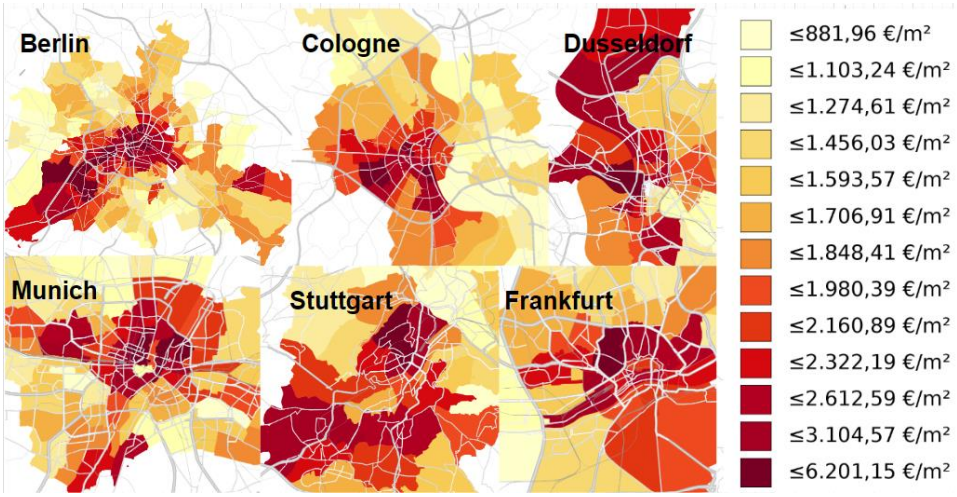

Figure 15: Graph of the Germany's Price per m^2 Data on Six of Biggest and Expensive Cities [25]

According to the analysis done by (citation), the real estate price is higher when it is close to the city center. That is why the distance to city center is a great indicator on house price per m^2. With that said, according to the XGBoost regression model, the distance to city center is not the only great indicator, but the answer of the question: Which city? Is the most important feature of the real estate market valuation. (the city rank about having most expensive real estates will be added as appendix)

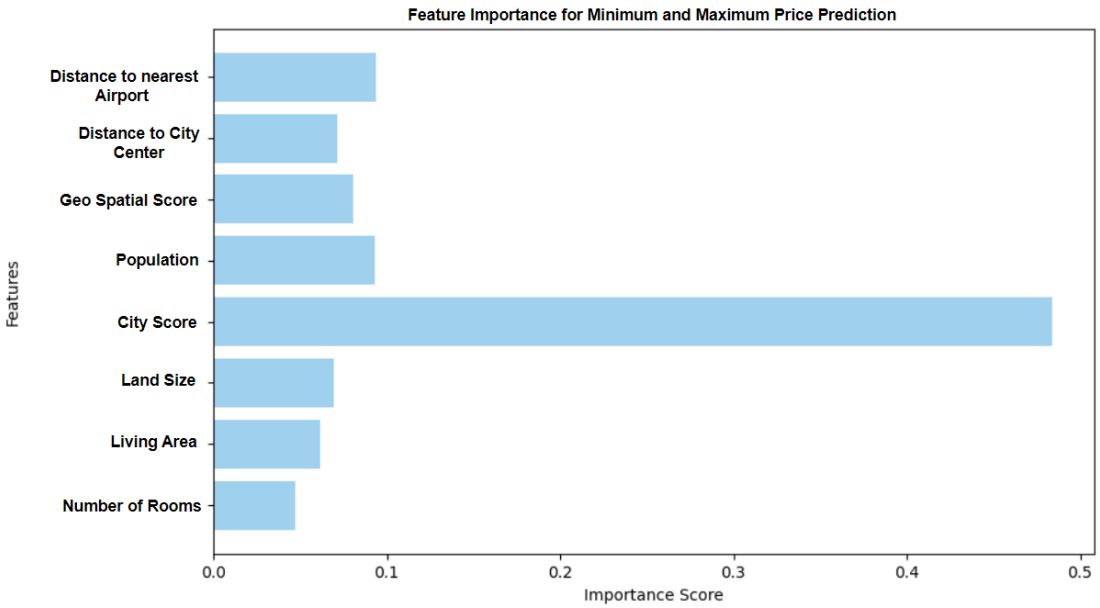

Figure 16: Feature Importance for Minimum and Maximum Price Prediction

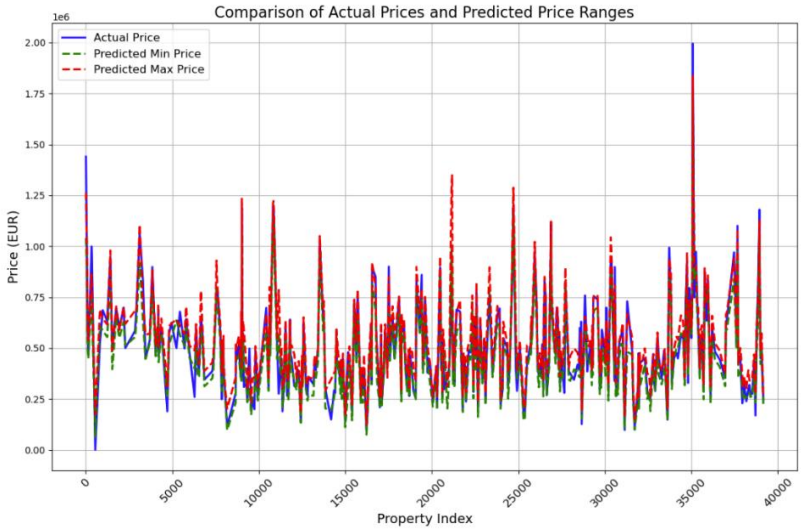

Figure 17: Comparison of Actual and Predicted Price Ranges

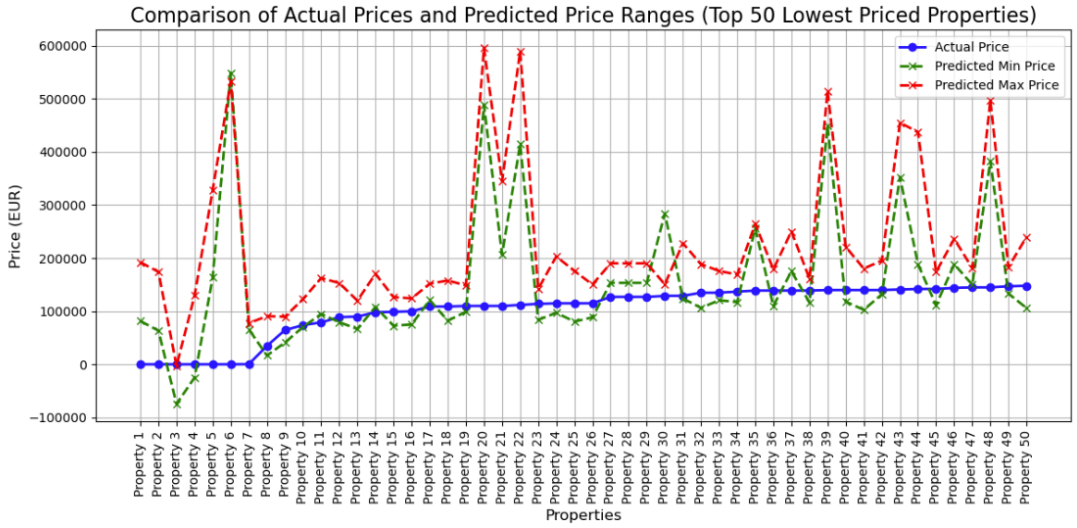

Figure 18: Comparison of Actual and Predicted Price Ranges For the top 50 Lowest Real Estate Sample

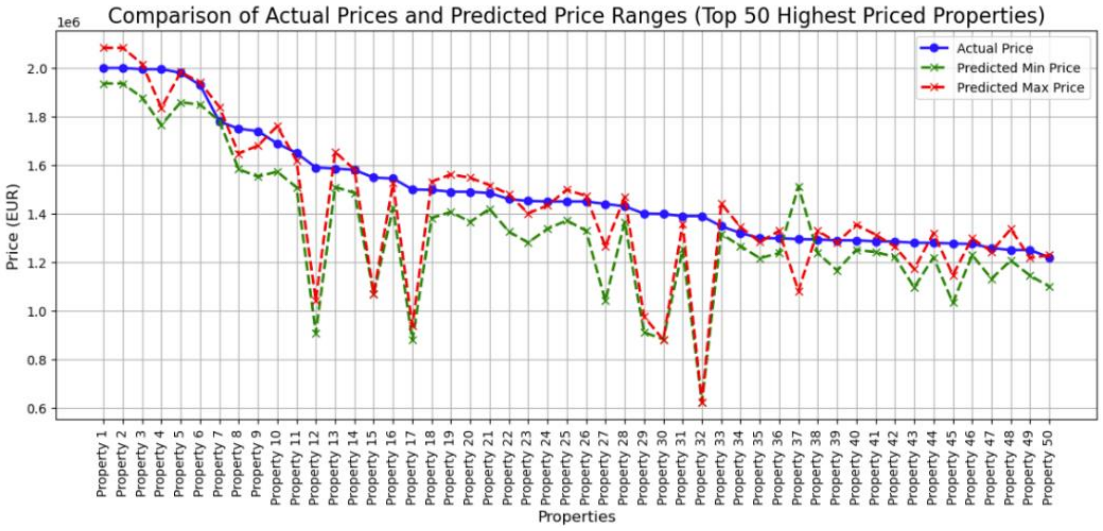

Figure 19: Comparison of Actual and Predicted Price Ranges For the top 50 Highest Real Estate Sample

4.Improving Business with Data: Solutions to Real Estate Market Challenges

4.1 Accurate Pricing

Problem statement: Financial losses may be happened as a result of incorrect real estate valuations are prevalent, stemming from the complex and diverse factors influencing real estate prices.

Solution: Utilizing machine learning models such as Random Forest and XGBoost for price prediction. These models integrate geospatial data like proximity to city centers, airports, and other amenities, significantly enhancing valuation accuracy.

The research demonstrates that geospatial scores, such as distance to key locations and city scores, provide a robust foundation for understanding regional price variations. This approach reduces errors in valuation, thereby minimizing financial losses, as well as maximizing the profitability.

4.2 Enhancing Negotiation Transparency

Problem statement: A lack of transparent information often complicates negotiations between buyers and sellers, resulting in mistrust or suboptimal outcomes.

Solution: Data-driven insights can serve as a mediator by providing objective price intervals and regional comparisons.

The use of valuation interval models (e.g., XGBoost for maximum and minimum price predictions) ensures that buyers and sellers have access to clear, data-supported information. This transparency fosters more efficient and fair negotiations.

4.3 Demand Forecasting and Creating Valuation Criteria

Problem Statement: Identifying high-demand regions is challenging due to market trends and varying buyer preferences. Creating logical valuation criteria depends on key features, such as the city and living area size, which are essential for determining the optimum price.

Solution: Predictive modeling with geospatial data helps identify demand hotspots. Areas with high proximity to schools, parks, and shopping centers often see increased interest, though the city of the real estate is a more significant factor.

Valuation criteria are determined using feature importance, highlighting key indicators for real estate valuation. This research demonstrates how machine learning can analyze price trends and score amenities to identify demand clusters. This information can guide future real estate owners and investors toward high-potential regions.

5. Conclusion

In summary, the research identified three main price drivers in the German real estate market. The most important factor is the city where the real estate is located. The second most important factor is the city’s total population. The third factor is the distance to the nearest airport. These values are geographically derived.

While living space and land area are also important, as they provide property details, two real estates with identical features can still have different prices if they are located in different areas.

Acknowledgement

Prof. Dr. Leonard Noriega, Zeynep Isik.

Abbreviations and Glossary

AI: Artificial Intelligence

API: Application Programming Interface

CSV: Comma-separated values

EU AI: European Union Artificial Intelligence

EU: European Union

GBM: Gradient Boosting Machine

GDPR: General Data Protection Regulation

GIS: Geographic Information System

JSON: JavaScript Object Notation

MAD: Mean Absolute Deviation

MAE: Mean Absolute Error

ML: Machine Learning

OS: Operating System

OSM: OpenStreetMap

R^2: A metric to assess the ML algorithm performance

RF: Random Forest

SVM: Support Vector Machine

XGBoost: eXtreme Gradient Boosting

Zip Code: An identifier for cities' districts

References

[1] Donges, N., Whitfield, B., Pierre, S., (26/11/2024). Random Forest: A Complete Guide for Machine learning – All you need to know about the random forest model in machine learning. Builtin.

[2] Kavlakoglu, E., Russi, E. (2024). What is XGBoost? IBM. Available at: https://www.ibm.com/topics/XGBOOST. Retrieved date: (03/12/2024)

[3] EU Artificial Intelligence Act (13/07/2024).Transparency Obligations for Providers and Deployers of Certain AI Systems. Official Journal of the European Union (chapter 4, article 50). Available at: http://data.europa.eu/eli/reg/2024/1689/oj. Retrieved date: 03/12/2024.

[5] Rey-Blanco, D., Zofío, J. L., & González-Arias, J. (2024). Improving hedonic housing price models by integrating optimal accessibility indices into regression and random forest analyses. Expert Systems with Applications, 235, 121059.

[6] White-House. (2016). Big data: A report on algorithmic systems, opportunity, and civil rights. Technical report.

[7] Cassidy, G. (06/10/2024). 7 Neighborhood Factors That Impact Home Values. Homes. Retrieved from: Available at: https://www.homes.com/blog/how-neighborhoods-affect-home-values/. Retrieved date: 11/12/2024

[8] Pfeiffer, M. (2022).10 Shocking Factors That Can Affect Your Property Value and Tax Assessment. Mortgage Mark. Available at: https://mortgagemark.com/mortgage-resource-library/factors-that-can-affect-your-property-value/. Retrieved date: 11/12/2024

[9] Nicastro, S. (04/04/2024). The 5 Most Accurate Home Value Estimators. Clever. Available at: https://listwithclever.com/real-estate-blog/home-value-estimate-websites/. Retrieved date: 11/12/2024

[10] Immowelt. A German Real Estate Website. Available at: https://www.immowelt.de/. Retrieved date: 11/12/2024

[11] Ho, W., K., O. Tang, B. S., & Wong, S., W. (2020). Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1), 48–70.

[12] Mete, M. O., & Yomralioglu, T. (2022). Mass valuation of Real Estate Using GIS-based nominal valuation and machine learning methods. *European Real Estate Society, ERES*, 1-7.

[13] Gold, C. M. (2006). What is GIS and What is Not?. *Transactions in GIS*, *10*(4), 505-519.

[14] Vaddi, S. S., Yousif, A., Baraheem, S., Shen, J., & Nguyen, T. V. (2022). House Price Prediction via Visual Cues and Estate Attributes. In International Symposium on Visual Computing (pp. 91-103). Cham Springer Nature Switzerland.

[15] Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the US real house price index. *Economic Modelling*, *45*, 259-267.

[16] Kang, Y., Zhang, F., Peng, W., Gao, S., Rao, J., Duarte, F., & Ratti, C. (2021). Understanding house price appreciation using multi-source big geo-data and machine learning. Land use policy, 111, 104919.

[17] Federal Statistical Office, Wiesbaden (2024). GENESIS-Online database. Database Table: Stock of residential buildings and dwellings. Code: 31231-0001.

[18] Statistisches Bundesamt (Destatis). (2024). Society and Environment. Housing. Available at: https://www.destatis.de/EN/Themes/Society-Environment/Housing/_node.html#sprg481706. Retrieved date: 26/11/2024

[19] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. Advances in neural information processing systems, 30.

[20] Qureshi, A. Mushailov, L. Herrera, P. Hale, P. McDaniel, R. (2022). "A Framework for Predicting the Optimal Price and Time to Sell a Home". SMU Data Science Review. Vol. 6: No. 2, Article 16.

[21] Lee, Chun-Chang & Chang, & Lin, Hui-Yu. (2012). The Impact of Neighborhood Characteristics on Housing Prices-An Application of Hierarchical Linear Modeling.

International Journal of Management and Sustainability. 1. 31-44. 10.18488/journal.11/2012.1.2/11.2.31.44.

[22] Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: an application of Random forest for valuation and a CART-based approach for model diagnostics. Expert Systems with Applications, 39(2), 1772-1778.

[23] Hong, Jengei & Choi, Heeyoul "Henry & Kim, Woo-Sung. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. International Journal of Strategic Property Management. 24. 1-13. 10.3846/ijspm.2020.11544.

[24] Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

[25] Hein, S., (2016). Price maps for Germany. Value Marktdaten. Available at: https://www.value-marktdaten.de/en/2016/03/14/price-maps-for-germany/. Retrieved date: 26/11/2024

[26] Ahmed, M., W. (2023). Understanding Mean Absolute Error (MAE) in Regression: A Practical Guide. Medium. Available at: https://medium.com/@m.waqar.ahmed/understanding-mean-absolute-error-mae-in-regression-a-practical-guide-26e80ebb97df. Retrieved date: 26/11/2024

[27] Thomas, S. (2021). Mean Absolute Deviation (MAD) - Meaning & Formula. Outlier. Available at: https://articles.outlier.org/mean-absolute-deviation-meaning. Retrieved date: 26/11/2024