Application of Machine Learning to Identify Factors Affecting Apartment Prices in Cau Giay Ward, Ha Noi City, During the 2023-2025 Period

Ngoc Thi Cam BUI[1)], Thu Hang DANG [2)], Then Thi BUI[2)], Tuyet Minh DANG[3*)]
[1)] University of Economics and Business - Vietnam National University
[2)] Hanoi University of Natural Resources and Environment, Hanoi, Vietnam
[3)] Thuyloi University, Hanoi, Vietnam

Corresponding author: dtminh@tlu.edu.vn; ORCID: 0000-0001-8379-1087

doi: 10.29227/IM-2026-01-01-011

ABSTRACT

Apartment prices are influenced by various factors, particularly within the context of highly volatile and non-transparent real estate markets. This study proposes a novel approach to identify and quantify the factors affecting apartment prices, aimed at enhancing transparency and reliability in valuation practices. The research employs several machine learning models, including Extra Trees Regressor (ETR), Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost), applied to a dataset of 246 apartment transaction samples in Cau Giay Ward, Hanoi City, during the 2023-2025 period, incorporating 12 influential variables. The results indicate that the ETR model outperforms other models across four out of five performance evaluation metrics. Subsequently, the study assesses the significance of these factors as quantified by the SHAP (SHapley Additive exPlanations) technique applied to the ETR model. The primary factors identified as having the greatest impact on apartment prices include: location, transaction year, developer reputation, financial factors, number of rooms, amenities, floor area, corner unit status, floor level, and market conditions. Notably, the introduction and utilization of spatial coordinate factors clarify a strong price differentiation trend along the East-West axis in the area. This research provides a new quantitative explanatory framework, contributing to increased objectivity and reliability while effectively supporting apartment management and valuation.

KEYWORDS: *machine learning models, apartment prices, land management, Cau Giay Ward*

1 INTRODUCTION

The apartment market is experiencing significant price fluctuations. According to statistics from the Ministry of Construction (2025) [1], apartment prices in Hanoi in the third quarter of 2025 increased by 33% year-on-year, making accurate valuation an urgent requirement for investors, regulatory bodies, and consumers. The market value of apartments is driven by multi-dimensional factors, including geographical location, transaction timing, physical attributes, developer prestige, and various socio-economic determinants. However, the relationship between these factors and the selling price is often complex, non-linear, and characterized by inter-variable interactions [2, 3]. This serves as the primary rationale for the application of Machine Learning (ML) methods, which are more effective at processing these intricate relationships compared to traditional hedonic price

models or conventional statistical frameworks [4, 5]. Empirical comparisons between traditional regression analysis and machine learning have demonstrated that ML offers superior performance in handling spatial dependencies and complex non-linear correlations [6].

In the era of rapidly advancing Big Data and Artificial Intelligence, machine learning (ML) methods are increasingly demonstrating superior efficiency in analyzing structured data, processing complex relationships, and forecasting market trends. This study prioritizes the use of algorithms such as Random Forest, Gradient Boosting, XGBoost, and Extra Trees Regressor due to their high accuracy and effective handling of mixed data types [7], which typically exhibit greater stability than deep learning models like ANN or SVM on medium-sized datasets. These tree-based models offer distinct advantages in processing categorical variables (e.g., developer prestige, legal status) and do not require normality assumptions regarding data distribution, unlike traditional hedonic regression models. The success of machine learning models in apartment valuation has been empirically validated across diverse markets, including Tokyo (Japan) [8], the Czech Republic [9], and Jordan [10]. Notably, recent studies [11, 12] and comparative analyses in Prishtina [13] have indicated that Ensemble Learning techniques not only yield high predictive accuracy but also possess the capability to disentangle complex determinants such as transport infrastructure, building density, and socio-economic fluctuations. This reinforces the premise that transitioning from conventional models to data-driven approaches is essential for enhancing transparency within the real estate market.

Globally, the application of machine learning (ML) in real estate valuation has emerged as a mainstream trend across various countries with diverse market characteristics, ranging from hyper-dense metropolises like Tokyo [8] to emerging markets such as Jordan [10] or Prishtina [13]. In recent years, leveraging Artificial Intelligence (AI) to address property valuation challenges has become a prominent academic movement worldwide. Landmark international studies [14, 15] have demonstrated the superior efficacy of ML techniques in disentangling the determinants of housing prices, with a particular emphasis on the role of infrastructure amenities and integrated services. These approaches reinforce the premise that machine learning models serve not only as forecasting tools but also as powerful instruments for quantifying the intangible values of apartments. Furthermore, extensive research in developed markets [4] has proven that ML algorithms can significantly reduce prediction errors by automating the extraction of non-linear features from Big Data. Numerous studies have confirmed that Ensemble Learning algorithms, such as Random Forest and Gradient Boosting, consistently yield more accurate and stable predictive results than traditional methods when processing actual transaction data [10, 16]. Beyond physical attributes, international scholarship also highlights the critical importance of spatial factors, public transportation, and population density in the formation of apartment market value [11, 12].

In Vietnam, domestic scholarship has begun to establish a theoretical framework comprising six primary groups of factors governing apartment values. Specifically, geographical location and the surrounding environment are consistently identified as top priorities [17, 18], followed by physical attributes such as floor area and internal design [19, 20], as well as developer prestige and general market fluctuations [21]. However, a significant limitation persists: these approaches primarily rely on traditional linear regression methods, which struggle to fully capture the non-linear relationships and complex inter-variable interactions within this unique market. Particularly, given the extreme price volatility in core areas like Cau Giay Ward during the 2023–2025 period, accurately quantifying the impact of each factor and demystifying the 'black box' nature of AI algorithms remains a substantial research gap. Therefore, this study aims to propose a comprehensive evaluation solution using machine learning integrated with the SHAP (SHapley Additive exPlanations) post-hoc explanation technique. The findings not only suggest a novel quantitative approach to enhancing price transparency but also enrich the theoretical framework for applying Artificial Intelligence in land management and real estate practices in Vietnam.

The research data was compiled from apartment transactions in Cau Giay Ward, Hanoi City, during the 2023–2025 period. This area was selected as a representative case study due to its high

density of apartment projects, diverse market segments (ranging from resettlement housing to luxury apartments), and price appreciation rates that are emblematic of the inner-city apartment segment in Hanoi during this timeframe. The core of this study addresses two primary research questions: (1) Which of the four developed and compared algorithms (RF, GB, XGBoost, and ETR) achieves the optimal predictive performance; and (2) How can the SHAP technique be applied to quantify and interpret the impact of each input factor on the predicted prices, thereby demystifying the "black box" nature of the models.

2 RESEARCH METHODOLOGY

2.1 Research Subject and Scope

2.1.1 Research Subject

The subjects of this study are apartment prices and the determinants affecting apartment values within Cau Giay Ward, Hanoi City. Through a comprehensive literature review, and by adopting and adapting frameworks from analogous studies [22-24], the authors utilized apartment price data alongside 12 variables tailored to the characteristics of the study area. These variables pertain to location, developer prestige, amenities and services, financial factors, market conditions, year of construction, floor area, number of rooms, corner unit status, floor level, and spatial coordinates (X and Y coordinates), as detailed below.

Table 1. Description of variables in the machine learning model

Tab. 1. Opis zmiennych w modelu uczenia maszynowego

No.	Variable Name	Description	Measurement / Evaluation Method/ Data Source
1	Location (Lo)	Distance from the apartment to urban amenities (schools, hospitals, markets, main roads, etc.)	Euclidean distance Cadastral map
2	Developer Prestige (DP)	Reflects the developer's capacity, experience, and the quality of construction materials and equipment.	Official lists from the Department of Construction (DoC) and the Department of Planning and Investment (DPI); developer websites; and homebuyer surveys.
3	Amenities (Am)	Physical features and services within or surrounding the building that enhance the quality of life and property value (e.g., security, sanitation, etc.)	Developer websites, valuation certificates, and surveys of real estate agencies and homebuyers.
4	Floor Area (FA)	The living area of the apartment unit.	Valuation certificates, and surveys of real estate agencies and homebuyers
5	Number of Rooms (NoR)	The total number of rooms within the apartment unit.	Valuation certificates, and surveys of real estate agencies and homebuyers
6	Floor Level (FL)	The specific floor level of the apartment unit	Valuation certificates, surveys of real estate agencies, and homebuyers
7	Corner Unit (CU)	Indicates whether the apartment is a corner unit or a standard unit	Valuation certificates, and surveys of real estate agencies and homebuyers.
8	Market Factors (MF)	Socio-economic and legal factors affecting the supply and demand of apartments	Surveys of real estate agencies and homebuyers.
9	Financial Factors (FF)	Factors directly affecting purchasing power and market transaction values (e.g., income, savings, mortgage rates, housing incentives, and solvency).	Surveys of real estate agencies, commercial banks, and homebuyers

10	Construction Year (CY)	The year the building was completed and put into use.	Developer websites and surveys of real estate agencies.
11	X-Coordinate (Coor_ X)	The X-coordinate of the building	Cadastral maps
12	Coordinate Y (Coor_ Y)	The Y-coordinate of the building	Cadastral maps
13	Apartment Price (AP)	The selling price per square meter (m^2) of the apartment unit.	Valuation certificates and survey questionnaires of real estate agencies and homebuyers.

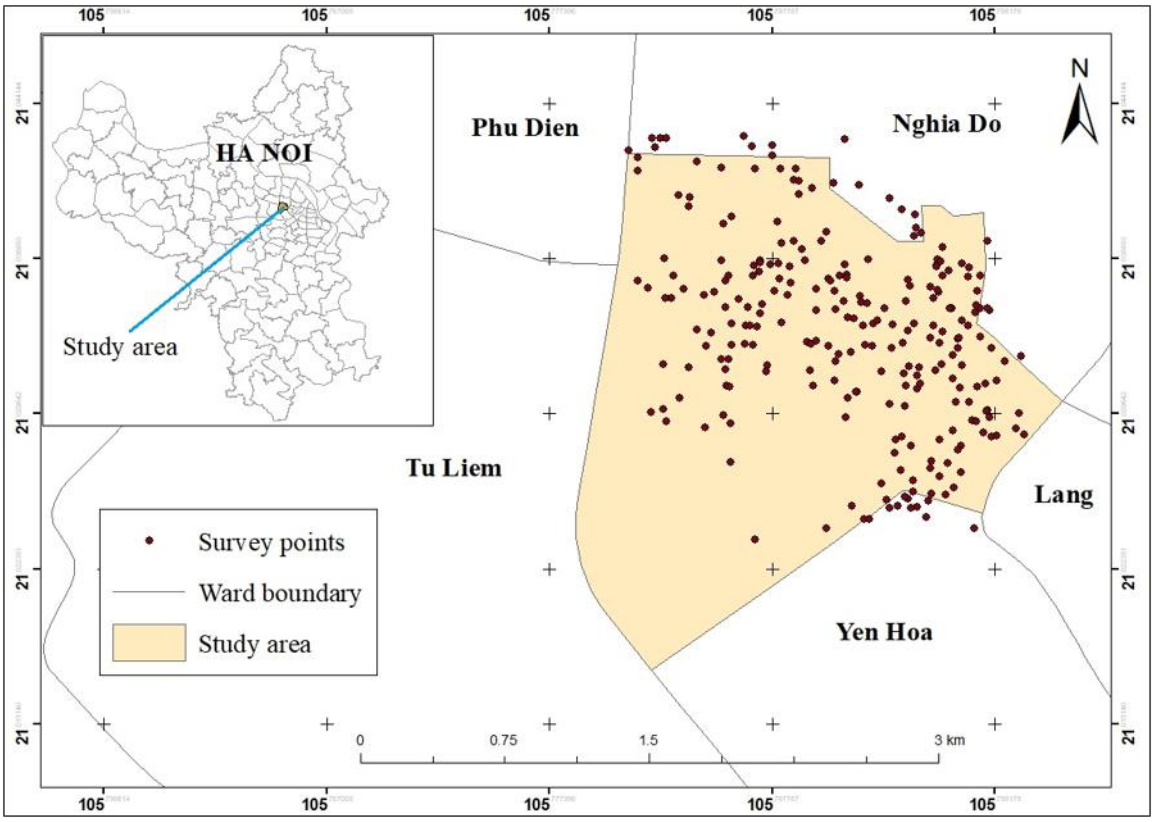

Figure 1. Spatial distribution of apartment price samples in Cau Giay Ward

Rys. 1. Przestrzenne rozmieszczenie próbek cen mieszkań w dzielnicy Cau Giay

To develop the machine learning model for apartment valuation, the dataset was compiled from cadastral maps, geospatial databases, and 246 apartment price samples surveyed in Cau Giay Ward, Hanoi. The collected apartment prices represent actual market transactions under normal conditions, primarily sourced from valuation certificates integrated with direct survey data from real estate trading floors and residents. The survey methodology was designed along major street networks. The spatial distribution of the collected samples is illustrated in Figure 1.

The pre-processed dataset was randomly split into two subsets: a training set comprising 80% (equivalent to 197 samples) and a test set comprising 20% (equivalent to 49 samples). To mitigate the risk of overfitting due to the moderate sample size, the training process integrated cross-validation techniques during the model hyperparameter tuning phase.

2.1.2 Scope of Research

The study area is Cau Giay Ward, one of the most vibrant apartment markets and a representative of Hanoi's core urban districts. This area features synchronized transportation infrastructure and a wide range of project segments with prices that fluctuate significantly based on location and amenities, making it highly suitable for testing advanced valuation models. The geographical scope is defined according to the current administrative boundaries of the ward as per Resolution 1681/NQ-UBTVQH15 (2025) to ensure consistency in policy and administrative jurisdiction. The research focuses on apartment price data from the 2023–

2025 period. This was a pivotal timeframe as the market underwent a strong upward cycle driven by the post-pandemic economic recovery and major legislative changes, notably the 2024 Land Law. In particular, administrative adjustments and urban planning shifts during this period directly influenced investor sentiment and transaction values, resulting in a high-complexity dataset ideal for applying and evaluating the performance of machine learning models.

2.2 Data Collection Methods

2.2.1 Secondary Data Collection

The study conducted a selective collection and inheritance of relevant data and documents concerning socio-economic development and real estate market trends. These were sourced from state agencies, including the Department of Natural Resources and Environment (DONRE), the Land Registration Office, the Hanoi Land Fund Development Center, and official internet resources. Additionally, cadastral maps of Cau Giay Ward, Hanoi, were collected at a scale of 1:500 in .dgn format. These maps were utilized to extract spatial information, such as geographic locations and proximity to apartment amenities, for integration into the machine learning model. The cadastral data was processed using GIS software to extract the centroid coordinates of the buildings and calculate Euclidean distances to utility infrastructures, ensuring consistent measurement units for the model.

2.2.2 Primary Data Collection

The study investigated 246 apartment transactions occurring between 2023 and 2025. The survey subjects included real estate agencies, building management boards within Cau Giay Ward, industry experts, and residents. The survey focused on the 12 variables identified in Table 1, encompassing information on location, transaction year, corner unit status, floor level, floor area (m2), and amenities such as distance to the city center, hospitals, schools, and entertainment hubs, as well as developer reputation. To ensure data integrity, each collected sample was cross-referenced between information from real estate trading floors and valuation certificates (where available). This process aimed to eliminate biases caused by psychological factors or speculative market misinformation.

2.3 Data Processing Methods

Spatial data was analyzed and processed using ArcGIS software. The machine learning models were developed and trained using the Python programming language and the Scikit-learn library. After extracting spatial features from ArcGIS, they were linked to primary attribute data via a unique building identifier (ID) to create a unified dataset prior to executing the machine learning algorithms.

2.3.1 Cartographic Data Processing

For the surveyed dataset, issues such as missing, biased, or inconsistent data were identified and resolved. Specifically, categorical variables were converted into numerical format using Label Encoding or One-hot Encoding techniques. Price outliers were controlled using the Boxplot method to ensure the model remained unaffected by anomalous transactions. Furthermore, data normalization was required, particularly for cartographic

documents. The cadastral maps, initially collected in .dgn format, were converted into a File Geodatabase (*.gdb) format. Within ArcGIS, topology rules were validated and the VN-2000 coordinate system was established. Land-use purpose labels and project types were reviewed and standardized to accurately define the boundaries of apartment areas and infrastructure utility layers for calculating distance variables. Additionally, other attribute fields were extracted from relevant documents and spatial analysis results.

2.3.2 Processing Survey Data for Machine Learning Models

Step 1: Data Pre-processing

The data pre-processing phase was conducted using the Python programming language, leveraging specialized libraries such as Pandas (for data management and manipulation) and NumPy for numerical array processing. This phase included the following steps:

(1) Data Cleaning: Removing duplicate records and eliminating missing or unreliable values within the dataset.

(2) Outlier Treatment: An apartment price sample is identified as an outlier if it falls outside the Interquartile Range (IQR). Since apartment prices are heterogeneous, samples were grouped by location and road type. The IQR was calculated independently for each group to ensure data objectivity and precision. The formula for calculating the IQR is as follows:

$$(Q1 - 1.5 \times IQR) < x < (Q3 + 1.5 \times IQR) \quad (1)$$

Where Q1 and Q3 are the 25^{th} and 75^{th} percentiles, respectively, and IQR = Q3 - Q1.

(3) Data Splitting: Finally, the dataset was randomly partitioned into training and testing sets using an 80:20 ratio, consistent with methodologies established in previous studies [22].

Step 2: Model Training

The machine learning models were developed and trained using the Python programming language, leveraging the standard scikit-learn library along with other specialized packages. This study utilized four popular regression algorithms for automated mass appraisal (AMA):

Random Forest (RF): An ensemble learning method that constructs multiple independent decision trees from random subsets of data and features, then averages their results for prediction [25].

Extra Trees Regressor (ETR): An extension of RF that introduces additional randomness by selecting split points at random and utilizing the entire training set to reduce variance, albeit at the potential cost of a slight increase in bias [26].

Gradient Boosting (GB): A sequential ensemble method where each new model is trained to correct the residual errors of the previous one. The model is optimized by following the negative gradient of the loss function (Gradient Descent), hence the term "Gradient Boosting" [19].

Extreme Gradient Boosting (XGBoost): An advanced implementation of GB optimized for both computational efficiency and predictive performance, offering faster processing speeds and superior accuracy [27].

Model performance is highly sensitive to hyperparameter configuration. To ensure objective evaluation and reliability, this study employed GridSearchCV integrated with 5-

fold Cross-Validation. This combination facilitates the identification of optimal hyperparameters, such as the number of trees (n_estimators) and maximum tree depth (max_depth), while the 5-fold cross-validation ensures that the performance metrics remain robust and independent of any single data split.

Step 3: Model Evaluation and Comparison

The performance of the models was evaluated using standard metrics for mass appraisal accuracy, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R2) [22, 27]:

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

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

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (5)$$

Where n is the number of samples, yi is the actual market transaction price, y^_i is the predicted value, and y¯_i is the mean value of the actual market transaction prices.

Among these metrics, an R2 value closer to 1 indicates that the model effectively explains the variability of the data. Conversely, lower values of MAE, RMSE, and MAPE signify lower prediction errors. Notably, MAPE (%) provides an intuitive perspective on the average error rate relative to the actual transaction values.

Step 4: Assessing Feature Importance Using SHAP Technique

The application of machine learning significantly enhances the efficiency of apartment valuation. However, the "black-box" nature of these models often obscures their internal mechanisms from appraisers, potentially undermining confidence in the output results. SHAP (SHapley Additive exPlanations) is a robust technique for model interpretation that elucidates how specific features influence model predictions. Developed by [28] and rooted in cooperative game theory, SHAP assigns each feature an importance value representing its contribution to the final prediction. The SHAP technique can interpret models at both the global level to determine the overall significance of each factor across the entire large-scale dataset and the local level, to explain specific predictions for individual data samples. This study utilizes SHAP for global interpretation to identify the roles of various factors affecting apartment prices within the valuation model. Specifically, the SHAP Summary Plot is employed to rank the importance of the 12 input features, while SHAP Dependence Plots are used to analyze the marginal impact and trends of the most influential variables on apartment prices.

3 RESULTS AND DISCUSSION

3.1 Performance Evaluation and Comparison of Machine Learning Models

The dataset was partitioned into an 80% training set and a 20% testing set, with the latter comprising 49 samples. A performance comparison between these two subsets was conducted to detect potential overfitting. The predictive performance of the models on the test set was evaluated and compared based on the optimized hyperparameters. The four performance metrics for the machine learning models, calculated according to Equations (2), (3), (4), and (5), are summarized in the table below:

Table 2. Performance metrics of the four machine learning models

Tab. 2. Miary wydajności czterech modeli uczenia maszynowego

Model	RMSE	MAE	R^2	MAPE
RF	11.22	9.07	0.77	14.60%
ETR	10.17	7.45	0.83	11.00%
GB	10.89	8.39	0.76	13.00%
XGBoost	11.39	9.20	0.77	14.50%

Following the optimization process, the best-performing hyperparameters applied to the models are summarized in Table 3:

Table 3. Optimized hyperparameter configurations for the machine learning models

Tabela 3. Zoptymalizowane konfiguracje hiperparametrów dla modeli uczenia maszynowego

Model	Hyperparameter	Value
RF	n_estimators	[100,150,200,250,300]
	max_depth	[12,13,14,15, None]
ETR	n_estimators	[100, 200, 300]
	max_depth	[6,7,8,9, None]
GB	n_estimators	[50,100,150,200,250]
	max_depth	[3,4,5,6,7, None]
	learning_rate	[0.1, 0.2, 1.0, 0.05, 0.5]
	alpha	0.01
XGBoost	n_estimators	[100,150,200,250]
	max_depth	[4,5,6,7, None]
	learning_rate	[0.1, 0.01, 0.05, 0.5]

As shown in Table 2, all models experienced a slight decrease in performance on the test set compared to the training set (indicated by a lower R2); however, this remains within an acceptable range as the test R2 values all exceeded 0.76. This demonstrates that the models possess strong generalization capabilities and are not subject to underfitting. The results indicate that the ETR model achieved the superior performance across most evaluation metrics. Specifically, ETR yielded the lowest RMSE (10.17) and the highest R2 (0.83). Regarding the mean absolute error, the ETR model also recorded the lowest MAE (7.45), while RF (9.07) and XGBoost (9.20) exhibited higher values. Similarly, ETR's MAPE was the lowest at 11.00%, representing the best overall predictive accuracy. Consequently, ETR was identified as the most accurate method in this study and was selected for further model interpretation using SHAP.

3.2 Identifying the Importance of Factors Affecting Apartment Prices

Model interpretation and the analysis of factors influencing apartment prices were conducted using the SHAP (SHapley Additive exPlanations) technique. This is a state-of-the-art method for interpreting machine learning models, widely recognized in recent literature [23]. Rooted in cooperative game theory, the SHAP technique assigns an importance value to each feature, representing its specific contribution to the model's prediction [28]. In this study, SHAP was utilized to determine the roles and relative importance of the input variables (the category of influencing factors) with respect to the output label (apartment price). The interpretation results are visualized through SHAP Bar Plots and SHAP Beeswarm Plots. To further elucidate the impact of the 12 input variables and the accuracy of the machine learning model in predicting apartment prices, the SHAP technique's identification of factor importance is presented in Figure 2 and Figure 3.

The SHAP Bar Plot and SHAP Beeswarm Plot illustrate the magnitude of influence for each input variable.

In Figure 3, the horizontal axis represents the SHAP value for each sample, with each point reflecting the specific impact of a variable on the predictive outcome. The color of the points, ranging from red (high value) to blue (low value), reveals the distribution of SHAP values and their directional influence. It is evident that Location is the most critical factor, with SHAP values spanning both positive and negative domains. Blue points (representing proximity to residential areas) are primarily concentrated on the positive SHAP side, indicating that these areas command the highest prices. This aligns with market reality, where proximity to established residential hubs facilitates infrastructure and service development, meeting essential living and community growth needs.

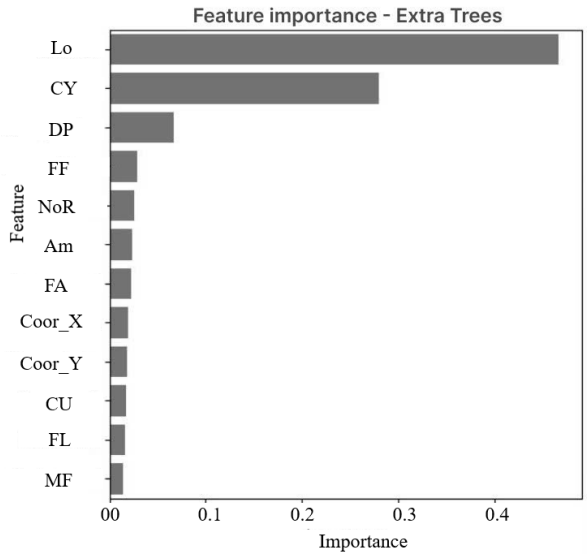

Figure 2. SHAP Bar Plot illustrating the global importance of factors influencing apartment prices

Rys. 2. Wykres słupkowy SHAP ilustrujący globalne znaczenie czynników wpływających na ceny mieszkań.

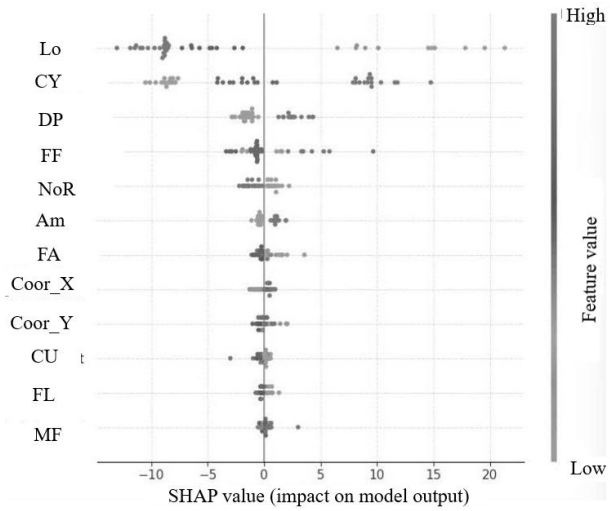

Figure 3. SHAP Beeswarm Plot representing the distribution and direction of factors influencing apartment prices

Rys. 3. Wykres rojowy SHAP (SHAP Beeswarm Plot) przedstawiający rozkład i kierunek wpływu czynników oddziałujących na ceny mieszkań

The Year of Construction also exerts a strong influence on the predictions. Low-value points (blue), corresponding to the year 2023, are distributed on the negative SHAP side, indicating lower prices compared to 2025. This likely reflects a rising price trend for newer apartments that offer modern amenities. The transaction year clearly captures market volatility over time.

The Finance factor significantly impacts apartment pricing. In practice, financial variables including personal income, developer incentives, interest rates, and access to financial products have a profound effect on housing costs.

Finally, Developer Reputation emerges as another major determinant of apartment prices. Currently, apartments developed by reputable entities, such as Vingroup or Sun Group, consistently command higher transaction prices than those from less established developers. These reputation rankings are officially evaluated and published annually by the Vietnam Real Estate Association (VNREA).

Factors such as the number of rooms, corner unit status, floor level, floor area, and amenities including proximity to schools, city centers, hospitals, transportation, and workplaces also influence apartment prices, albeit to a lesser extent.

Notably, the X-coordinate exerts a more significant impact than the Y-coordinate, reflecting a more pronounced differentiation in real estate values within Cau Giay Ward along the East-West axis compared to the North-South axis. This observation aligns with previous research [29] conducted on a broader scale across Hanoi. In practice, the development of radial traffic axes and the Urban Railway (Metro) lines along East-West corridors has catalyzed more intense price fluctuations compared to areas developing along the vertical axis (for instance, price volatility is more significant from Ring Road 3 toward the To Lich River than along the North-South direction from Hoang Quoc Viet to Tran Duy Hung).

The analysis in this study is consistent with prior research [21, 30, 31], identifying location, transaction year, developer reputation, and corner unit status as major determinants of apartment prices. Furthermore, this study incorporates spatial distance, X-coordinates, and Y-coordinates to evaluate price trend fluctuations. While the amenity group shows an impact,

it is not yet prominent, potentially due to a lack of highly granular data for each specific sample. However, this research has not yet accounted for environmental and legal factors, which can significantly influence real estate values in practice. Consequently, future studies should integrate these criteria to enhance accuracy and ensure a multi-dimensional perspective for the valuation model.

4 CONCLUSION

This study proposed and tested a machine learning-based framework to identify factors influencing apartment prices, with a pilot implementation in Cau Giay Ward yielding positive initial results. By employing the SHAP technique, the study effectively addressed the "black-box" nature of machine learning models, providing greater transparency and reliability for valuation tasks compared to traditional methods. Among the four Ensemble models tested, the Extra Trees Regressor (ETR) demonstrated superior performance, achieving an R2 of 0.83 and an RMSE of 10.17, significantly outperforming the RF and XGBoost models (R2 = 0.77). Through the application of SHAP to the ETR model, the study quantified and ranked the importance of 12 input variables, identifying location, transaction year, developer reputation, and financial factors as pivotal determinants. Notably, the integration of spatial elements (X, Y coordinates) enabled the identification of a distinct East-West price fluctuation trend within the study area. With a MAPE of 11%, the model proves highly capable of supporting regulatory authorities in developing Automated Mass Appraisal (AMA) systems. This provides a robust scientific foundation for the rapid and transparent determination of starting prices, property taxes, and real estate fees.

Despite achieving positive results, this study is subject to certain limitations regarding sample size, primarily due to administrative boundary adjustments and challenges encountered during field surveys. Moving forward, the research group plans to expand the scope of data collection and test more advanced machine learning architectures. These efforts aim to enhance the model's accuracy and generalizability across diverse geographic areas, thereby contributing to the refinement of both theoretical frameworks and practical applications in real estate price management in Vietnam.

5 REFERENCES

[1] Construction, M.o., Notice No. 523/TC-BXD dated October 30, 2025, on the disclosure of information on housing and the real estate market in the third quarter of 2025. 2025.

[2] Bourassa, S.C., E. Cantoni, and M. Hoesli, Spatial dependence, housing submarkets, and house price prediction. The Journal of Real Estate Finance and Economics, 2007. 35(2): p. 143-160.

[3] Malpezzi, S., Hedonic pricing models: a selective and applied review. Housing economics and public policy, 2003. 1: p. 67-89.

[4] Mullainathan, S. and J. Spiess, Machine learning: an applied econometric approach. Journal of Economic Perspectives, 2017. 31(2): p. 87-1 06.

[5] Varian, H.R., Big data: New tricks for econometrics. Journal of economic perspectives, 2014. 28(2): p. 3-28.

[6] Yoshida, T., D. Murakami, and H. Seya, Spatial Prediction of Apartment Rent using Regression-Based and Machine Learning-Based Approaches with a Large Dataset: Spatial Prediction of Apartment Rent using Regression-Based and Machine Learning-Based Approaches with a Large Dataset. The Journal of Real Estate Finance and Economics, 2024. 69(1): p. 1-28.

[7] Kim, W.-s., I. Choi, and M. Lee, Explainable AI-based mass appraisal: Insights from machine learning applications in Korea's residential property market. International Journal of Strategic Property Management, 2025. 29(5): p. 350-376.

[8] Peng, T.-C. and C.-C. Wang, The application of machine learning approaches on real-time apartment pr ices in the Tokyo metropolitan area. Social Science Japan Journal, 2022. 25(1): p. 3-28.

[9] Nikodym, J., Application of machine learning methods for estimating apartment prices in the Czech Republic. 2019.

[10] Thalji, N. and M. Alwadi, Predicting Apartment Prices in Jordan Using Ensemble Machine Learning Algorithms to Support Investment and Urban Planning Decisions. International Journal of Advances in Soft Computing and its Applications, 2026. 18(1): p. 321-338.

[11] Chen, Y., J. Jiao, and A. Farahi, Disparities in affecting factors of housing price: A machine learning approach to the effects of housing status, public transit, and density factors on single-family housing price. Cities, 2023. 140: p. 104432.

[12] Noorcahya, D., et al., Prediction of Apartment Price Considering Socio Economic and Crime Rates Factors in DKI Jakarta. International Journal of Advances in Data and Information Systems, 2023. 4(2): p. 145-154.

[13] Hoxha, V., Comparative analysis of machine learning models in predicting housing prices: a case study of Prishtina's real estate market. International Journal of Housing Markets and Analysis, 2025. 18(3): p. 694-711.

[14] Gadziński, J. and A. Radzimski, The first rapid tram line in Poland: How has it affected travel behaviours, housing choices and satisfaction, and apartment prices? Journal of Transport Geography, 2016. 54: p. 451-463.

[15] Hiller, N. and O.W. Lerbs, Aging and urban house prices. Regional Science and Urban Economics, 2016. 60: p. 276-291.

[16] Chu, T., Machine learning-based predictive modelling of house prices. 2025.

[17] Ngoc, N.M., N.H. Tien, and V.M. Hieu, Factors affecting the selling price of luxury apartments in Vietnam. A quantitative analysis. International journal of business and globalisation, 2023. 10(Y): p. 1-19.

[18] To MInh Thu, et al., Analysis of factors affecting apartment selling prices in Bắc Ninh Province. Journal of Banking Science and Training, 2022. 238: p. 62-73.

[19] Liem, L.D., Analysis of factors affecting apartment prices in Ho Chi Minh City. Master's thesis. University of Economics Ho Chi Minh City, 2019.

[20] Cuong, N.H., Factors affecting the selling prices of high-end apartments: A study in Hanoi. Doctoral dissertation. National Economics University, 2020.

[21] Khanh, D.V., Study on factors affecting the selling prices of Grade A apartments in Nam Tu Liem District, Hanoi. Journal of Natural Resources and Environment Science, 2023. 47: p. 51-62.

[22] Tu, B.N., T.Q. Binh, and B.T.C. Ngoc. Analyzing Mass Appraisal of Urban Residential Land with Machine Learning-A Case Study in Hanoi, Vietnam. in International Conference on Geoinformatics for Spatial-Infrastructure Development in Earth & Allied Sciences. 2023. Springer.

[23] Bilgilioğlu, S.S. and H.M. Yılmaz, Comparison of different machine learning models for mass appraisal of real estate. Survey review, 2023. 55(388): p. 32-43.

[24] Jafary, P., et al., Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors. Cities, 2024. 151: p. 105115.

[25] Breiman, L., Random forests. Machine learning, 2001. 45(1): p. 5-32.

[26] Geurts, P., D. Ernst, and L. Wehenkel, Extremely randomized trees. Machine learning, 2006. 63(1): p. 3-42.

[27] Iban, M.C., An explainable model for the mass appraisal of residences: The application of tree-based Machine Learning algorithms and interpretation of value determinants. Habitat International, 2022. 128: p. 102660.

[28] Lundberg, S.M. and S.-I. Lee, A unified approach to interpreting model predictions. Advances in neural information processing systems, 2017. 30.

[29] Bui, T., A study of factors influencing the price of apartments: Evidence from Vietnam. Management Science Letters, 2020. 10(10): p. 2287-2292.

[30] Bui Ngoc Tu, Man Quang Huy, and L.P. Thuy, Comparison of machine learning models in mass land valuation: A case study in Cầu Giấy District, Hanoi. Journal of Soil Science, 2025. 80.

[31] Hoang Thi Huong, Vu Thi Thuong, and B.L. Anh, Factors affecting apartment prices in Bac Tu Liem District, Hanoi. Journal of Economics and Forecasting. Retrieved from https://kinhtevadubao.vn/yeu-to-anh-huong-den-gia-can-ho-chung-cu-tren-dia-ban-quan-bac-tu-liem-tp-ha-noi-30212.html on December 24, 2025., 2024.