TRADITIONAL OR ADVANCED MACHINE LEARNING APPROACHES: WHICH ONE IS BETTER FOR HOUSING PRICE PREDICTION AND UNCERTAINTY RISK REDUCTION?

Long Phi Tran [*], Hoang Duc Le [**],
Ta Thu Phuong [***], Dung Chi Nguyen [****]

[*] *Corresponding author*, School of Banking and Finance, National Economics University, Hanoi, Vietnam
Contact details: School of Banking and Finance, National Economics University, 207 Giai Phong Street, Hanoi, Vietnam
[**] School of Banking and Finance, National Economics University, Hanoi, Vietnam
[***] Faculty of Business Management, National Economics University, Hanoi, Vietnam
[****] Smart Invest Securities Company, Hanoi, Vietnam

How to cite this paper: Tran, L. P., Le, H. D., Phuong, T. T., & Nguyen, D. C. (2025). Traditional or advanced machine learning approaches: Which one is better for housing price prediction and uncertainty risk reduction? *Risk Governance and Control: Financial Markets & Institutions, 15*(1), 27–36. https://doi.org/10.22495/rgcv15i1p3

Copyright © 2025 The Authors

This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/

ISSN Online: 2077-4303
ISSN Print: 2077-429X

Received: 06.03.2024
Accepted: 02.01.2025

JEL Classification: G17
DOI: 10.22495/rgcv15i1p3

Abstract

Predicting housing prices is particularly of interest to many scholars and policymakers. However, housing prices are highly volatile and difficult to predict. This study used both traditional and advanced machine learning (ML) approaches to address the issue of housing price prediction. This study involves and compares the predictive power between advanced ML models, including random forest, gradient boosting, k-nearest neighbors (KNN), bagged classification and regression trees (CART), and traditional ML models based on linear regression and its modifications. Notably, in this study, we employed both performance metrics, including the mean absolute error (MAE), root mean square error (RMSE), coefficient of determination (R^2), and k-fold cross-validation (CV) procedure in order to investigate the predictive performance of each model. Empirically, based on a dataset comprising 78,704 real estate sales in Hanoi, Vietnam, we find that advanced ML approaches outperform traditional approaches. Specifically, advanced ML models enhance the accuracy of house price prediction and the decision-making process related to housing buying and selling activities. Our findings also reveal that among advanced ML algorithms, the random forest algorithm performs better than the other models in predicting housing prices.

Keywords: Housing Price, Price Prediction, Machine Learning Models, Traditional Statistical Models, Performance Evaluation Framework

Authors' individual contribution: Conceptualization — L.P.T. and H.D.L.; Methodology — T.T.P. and D.C.N.; Formal Analysis — L.P.T. and D.C.N.; Writing — Original Draft — L.P.T. and D.C.N.; Writing — Review & Editing — L.P.T.; Supervision — H.D.L.; Funding Acquisition — H.D.L.

Declaration of conflicting interests: The Authors declare that there is no conflict of interest.

Acknowledgements: This study was funded by the Vietnamese Ministry of Education and Training under Grant B2022.KHA.07.

1. INTRODUCTION

Since the first quarter of 2020, many countries around the world have faced a range of considerable problems as a result of the ongoing rapid spread of coronavirus as well as earlier global risk factors (e.g., the prolonged USA-China trade war, currency fluctuations, long-term interest movements, Brexit). At the end of February 2020, global stock markets suffered the worst week, with almost $6 trillion wiped from their market value since the global financial crisis of 2008. Many experts have warned that these extreme events could lead to another global financial crisis. As a consequence of market shocks, investors prefer to keep safe assets and avoid risky assets, known as the "flight to quality" phenomenon in financial markets. Accordingly, real estate investment has been widely accepted as a potential alternative to traditional assets (e.g., stocks, bonds, and currencies), especially in times of high uncertainty. Investing in real estate is the safest long-term investment option. This is preferred to other popular investments. Unlike stocks and bonds, real estate investment is not volatile. As an effective investment asset, real estate has recently become increasingly appealing to many investors worldwide. In the past, investors typically did not consider real estate when constructing their investment portfolios.

It is generally known that the housing market is an important part of the real estate market. Therefore, the housing market is affected by a large number of factors such as economic growth, interest rates, demographics (e.g., age, population, gender, income, and migration), and government policies (e.g., tax credits, deductions, subsidies). In addition to these demand-side factors, this market can also be determined by available supply (Kim & Park, 2005; Muellbauer & Murphy, 2008). Expectations of financial returns from investments in the housing sector can put upward pressure on housing prices by increasing the demand for housing (Case et al., 2004). Meanwhile, the short-term adjustment capacity of the housing supply is limited. In addition to macro factors, there are some micro factors, such as location, area, direction, district, age of house, and number of floors, which are strongly correlated with housing prices (Chen et al., 2017; Truong et al., 2020).

It is undoubtedly true that house prices are important to many individuals and organizations (e.g., real estate owners, real estate developers, financial institutions, legislators, and the general public) (Schulz & Werwatz, 2004). Therefore, housing price prediction has attracted much attention across various academic fields, including economics, finance, politics, policy decision-making, and computer science (Selim, 2009; Abidoye & Chan, 2018; Hong et al., 2020). However, housing prices are highly volatile and difficult to predict. To date, various methods have been applied to predict housing prices. Most of these methods estimate changes in housing prices based on information from the house price index (HPI). However, these approaches have some disadvantages in forecasting the price of a single house. These disadvantages are attributed to the overwhelming evidence of inefficient information on HPI as well as the nonlinear and time-dependent characteristics of housing price data (Xu et al., 2012). Therefore, it is crucial to develop more accurate housing price prediction models.

In recent years, owing to the growing trend towards artificial intelligence (AI) and its methods, such as machine learning (ML), asset price prediction based on AI applications has become increasingly popular. The ML approach has been widely applied in various fields to facilitate the extraction of knowledge and prediction of future events. It is worth noting that several first authors have proposed traditional ML approaches to predict housing prices (Selim, 2009; Yoo & Wagner, 2012). However, these models rarely focus on the performance of the individual models. More importantly, they ignore less popular yet complex models. To address this shortcoming, advanced ML approaches have been introduced, including artificial neural networks (ANN) (Abidoye et al., 2019; Xu & Zhang, 2023; Khrais & Shidwan, 2023; Xu & Zhang, 2024), random forests (Park & Bae, 2015; Levantesi & Piscopo, 2020; Soltani et al., 2022; Adetunji et al., 2022; Rey-Blanco et al., 2024), and deep learning models (Yu et al., 2018). Many studies, including Truong et al. (2020) and Manasa et al. (2020), present strong evidence that these advanced ML models outperform traditional approaches. They demonstrated that advanced ML approaches have the ability to capture nonlinear relationships, enhance predictive accuracy, and reduce variation. However, several other studies argue against the advantages of these advanced approaches (Wang et al., 2021). This debate provides us with a strong motivation to compare the predictive performance of traditional and advanced ML approaches.

Although various methods have been proposed for dealing with house price prediction problems, it is well known that the predictive performance of a forecasting model is heavily dependent on the available datasets. Researchers commonly adopt one or more performance metrics to evaluate the predictive performance of a model. Therefore, they are able to determine, using actual data, the model with the highest prediction capacity.

For the above-mentioned reasons, in this study, this paper aims to compare traditional and advanced ML models when they are used to predict housing prices. We employ both traditional and advanced ML models to shed light on the difference between them in solving the issues of housing price prediction. Empirically, we conducted this research in Hanoi, Vietnam, with a dataset comprising 78,704 real estate sales in 2022. Hanoi is the second-largest city and official capital of Vietnam. The housing market in this city has vibrated in recent years. Studying Hanoi is interesting because Hanoi has characteristics of rapidly developing cities in the world. For example, industrialization and urbanization in Hanoi are very fast over the past few decades, making the prediction of real estate prices become more and more difficult. Recently, the Vietnamese government has had new policies related to land use and foreign ownership to attract investment from both domestic and foreign investors. The quick development of the infrastructure in Hanoi also plays a crucial role in shaping real estate prices in the city. As a result, we believe that examining real estate prices in Hanoi can provide valuable lessons and insights applicable to other emerging markets and rapidly developing cities worldwide.

However, there is a lack of holistic scientific research on forecasting Hanoi's housing prices. One of the reasons for this may arise from the problem of the absence of effective and accurate price housing prediction methods. Therefore, in this study, the problem of housing price prediction is analyzed using several traditional ML techniques, including linear regression and its modifications (e.g., generalized linear model and generalized additive model), and advanced ML models including random forest, gradient boosting, k-nearest neighbors (KNN), and bagged classification and regression trees (CART). Moreover, prediction model performance measures, including the coefficient of determination (R^2), mean square error (MSE), mean absolute error (MAE), and cross-validation (CV) technique, are used to select the models that perform better than the other models in forecasting housing prices. With higher R^2 scores and lower MSE and MAE values of advanced ML approaches, our empirical results confirm that these advanced models outperform traditional models in forecasting housing prices in Hanoi.

Overall, our study contributes to the literature in several ways. First, we add more evidence to the literature by comparing advanced ML models with traditional hedonic price models when predicting price houses (Xu & Zhang, 2023; Khrais & Shidwan, 2023; Soltani et al., 2022; Rey-Blanco et al., 2024). Our results show that the advanced ML approaches that we applied in this study to predict housing prices in Hanoi seem to outperform the other methods used in previous studies. Second, to the best of our knowledge, this study is the first to apply both traditional and advanced ML approaches to forecasting housing prices in the Hanoi real estate market. Third, holistic research comparing various techniques in housing price prediction is lacking. Our study provides a broad investigation of various types of ML algorithms and compares their predictive power. From the above contributions, the models suggested in this study can significantly improve the price prediction capacity of housing markets in Hanoi. Therefore, we firmly believe that we can provide valuable recommendations to policymakers, investors, and portfolio managers in the global housing market, and real estate in general.

The remainder of this paper is organized as follows. Section 2 provides an overview of the existing literature. Section 3 describes the methodology of this study. Our data and results are explained in Section 4. Section 5 concludes the study and discusses the implications of our research.

2. LITERATURE REVIEW

This section provides an overview of prior research on predicting house prices. In this context, our primary emphasis is on two types of ML models: traditional and advanced models.

In the existing literature, the first strand focuses on traditional statistical models. Over the past few decades, a variety of hedonic-based methods have been employed to uncover the intricate relationship between house prices and housing characteristics of housing (Selim, 2009; Malpezzi, 2003). Notably, Hort (1998) developed hedonic-based regression techniques to assess how market fundamentals impact housing price dynamics. The value of various features of residential properties is crucial in influencing the quality of life within communities, highlighting the significant role that residential housing plays in this regard. To attain this objective, researchers have frequently delineated hedonic price models in empirical studies (Nellis & Longbottom, 1981; Meese & Wallace, 2003).

Bin (2004) uses a semiparametric regression approach to estimate a hedonic pricing function. The predictive performance of this method was compared with that of typical parametric models. The utilization of geographic information system (GIS) data has facilitated the incorporation of location-based characteristics pertaining to residential properties. The results suggest that the semiparametric regression method consistently achieves superior performance compared to the parametric models in terms of predicting prices, both within the sample used for estimation and in predicting prices for new observations. This implies that the semiparametric technique is promising for accurately assessing and forecasting housing sales prices.

Fan et al. (2006) used the decision tree model, a prominent statistical pattern recognition technique, to examine the correlation between home features and housing prices. This study shows the effectiveness of this approach by examining the resale public housing market in Singapore. Kestens et al. (2006) included household-level data in hedonic regressions to examine the range of implicit pricing associated with various characteristics, such as income, household type, education, age, and prior tenure status of purchasers. Two methodologies were implemented to achieve this objective: one set of models included expansion terms, whereas the other utilized spatially weighted regressions.

Recent research has shifted its focus towards comparing the predictive capabilities of hedonic-based methodologies with those of ML algorithms. Kauko et al. (2002) investigated the application of neural network modeling in the Finnish housing market. The results demonstrated the capacity to discern diverse aspects of housing submarket patterns through the examination of the datasets. Additionally, the study demonstrated the classification capabilities of two neural network techniques: self-organizing maps and learning vector quantization. Fan et al. (2006) put forth a number of tree-based methodologies that can be utilized as significant instruments for statistical pattern detection to examine the relationship between home features and house prices. Sarip et al. (2016) proposed a novel prediction model based on fuzzy neural networks, which draws upon the principles of hedonic price theory. The objective of their study was to determine suitable price levels for newly developed real estate sites. These studies provided empirical evidence of the model's strong potential for approximating functions and its usefulness for predicting real estate prices. Selim (2009) compared the predictive capabilities of hedonic and ANN models. The findings indicate that ANN models have the potential to be a more advanced option for forecasting housing values in Turkey. Kusan et al. (2010) proposed a fuzzy logic model in separate investigations to forecast the selling prices of residential properties. Zhou et al. (2018) proposed a novel hybrid algorithm that combined fuzzy linear regression with a fuzzy cognitive map. This method was developed to address the challenges associated with housing market forecasting and optimization.

With the advent of the forthcoming era of big data, an increasing number of individuals have become involved in data analyses and mining. Among the diverse data analysis methods, ML has garnered increasing attention. The ML approach has been widely utilized in diverse domains, including business, finance, accounting, economics, and statistics, facilitating the extraction of knowledge and the prediction of future events. With recent growth and development of the real estate industry, ML has emerged as a crucial tool for accurately forecasting property prices. Scholars have increasingly utilized ML methods to estimate property prices, continually showing improved prediction abilities (Zulkifley et al., 2020; Thamarai & Nakarvizhi, 2020; Begum et al., 2022). The methods discussed in this study include ANN (Nguyen & Cripss, 2001; Abidoye et al., 2019), random forests (Park & Bae, 2015; Levantesi & Piscopo, 2020), and deep learning models (Yu et al., 2018). Empirical research suggests that ML models can harness the advantages of many methods, resulting in enhanced predictive accuracy (Truong et al., 2020; Manasa et al., 2020). Several ML strategies, such as bagging, boosting, and stacking, have been employed (Tang et al., 2018; Afonso et al., 2019). According to Pugliese et al. (2021), bagging is effective in reducing the variation. On the other hand, boosting is specifically designed to address bias. Stacking is an approach that combines multiple algorithms to minimize both bias and variance.

In recent years, there has been growing interest among researchers in conducting comparative evaluations of ML algorithms to uncover forecasting models that exhibit higher performance. In the building field, Gerek (2014) introduced adaptive neuro-fuzzy (ANFIS) methodologies to estimate the selling prices of residential properties. The findings of this study demonstrated that ANFIS models using grid partitioning exhibited greater effectiveness than models utilizing sub-clustering. Chen et al. (2017) proposed the use of particle swarm optimization (PSO) and support vector machines (SVM) as the foundation for real estate price forecasting models. The results of their study demonstrated the superior predictive accuracy of the SVM model compared to both grid and genetic algorithms.

Gu et al. (2011) and Liu and Liu (2019) proposed the utilization of a hybrid methodology that integrates genetic algorithms and support vector machines (GSVM), as a preferred strategy for the prediction of housing prices. The studies provided evidence that the forecasting accuracy of GSVM exceeded that of conventional approaches. Nevertheless, there has been a scarcity of research focused on the advancement of a more refined housing price prediction model using comprehensive assessments of diverse ML algorithms. This research addresses this disparity by conducting a comparative analysis of ML algorithms and developing a refined housing price forecast approach that offers enhanced accuracy for the real estate industry.

Recent advancements in ML and real estate price prediction have significantly expanded the scope of this field. Nguyen and Cripps (2001) explored deep learning techniques to predict housing prices in metropolitan areas, demonstrating substantial improvements in accuracy compared to traditional methods. Their study highlighted the capacity of neural networks to capture complex, non-linear relationships in housing data, often missed by conventional regression models. This aligns with Adetunji et al. (2022), who reported that random forest models outperform other ML algorithms in terms of predictive performance for housing prices in emerging markets.

Moreover, Khrais and Shidwan (2023) revisited the role of macroeconomic indicators in real estate price prediction. They employed a hybrid model combining economic factors with ML algorithms, finding that variables like gross domestic product (*GDP*) growth and *interest rates* substantially improve prediction accuracy. These recent studies underscore the continuous evolution of predictive models and their increasing sophistication, making them indispensable tools for real estate market analysis.

Overall, the literature suggests that both traditional and advanced ML models can be used to predict real estate prices. The advantage of the latter is that it can employ more information given the same set of data. As a result, we propose the following hypothesis:

H1: Advanced machine learning models are better at predicting real estate prices than traditional models.

3. RESEARCH METHODOLOGY

In this section, we present the methodology of two groups of ML models: traditional and advanced models.

3.1. Traditional machine learning models

Linear regression is simple to understand and has been widely used for many years. Therefore, this model is one of the most fundamental and well-known algorithms used in ML. In this study, we develop a multivariate linear regression model to forecast housing prices. Therefore, we can find the best possible line that fits the training set for the linear regression model and then predict the unknown house prices.

We also employed generalized linear models for the same historical housing data to predict future housing prices. Generalized linear models are a class of regression models that can be used to model a wide range of relationships between a response variable and one or more predictor variables. These models are slightly different from the linear regression ML algorithm because they can cover the nonlinear relationships between influential factors and housing prices by using a different underlying statistical distribution.

The main advantage of applying traditional ML algorithms is that they are suitable for manual data, and we can include legislative variables in the housing price prediction procedure.

3.2. Advanced machine learning models

This subsection briefly introduces several advanced ML algorithms used in this study.

3.2.1. Random forest

Introduced by Breiman (2001), the random forest has become a popular ML method. Because this algorithm integrates the predictions of numerous decision trees to provide a final prediction that is more accurate, random forest has been proven to be an effective tool for prediction.

3.2.2. Extreme gradient boosting

Extreme gradient boosting (XG Boost) is a scalable tree-boosting ML system. This model provides parallel tree boosting and is the leading ML library for regression, classification, and ranking problems. The XG Boost package is accessible as an open-source version. The system has made a big difference and has gained widespread recognition in a number of ML and data-mining problems.

3.2.3. Stochastic gradient boosting

According to Friedman (2002), gradient boosting constructs additive regression models by sequentially fitting a simple parameterized function (base learner) to current "pseudo-" residuals by least squares at each iteration. A big insight into bagging ensembles and random forests from the subsamples of the training dataset. This variation in boosting is known as stochastic gradient boosting. The main advantage of this method is that it can be used to reduce the correlation between trees in a sequence in gradient-boosting models.

3.2.4. K-nearest neighbors

KNN is one of the most basic ML algorithms based on supervised learning classifiers. This model uses proximity to make classifications or predictions regarding the grouping of individual data points.

3.2.5. Bagged trees

Bagged trees are special cases of random forests. A shortcoming of decision trees is that they are high-variance estimators. A small number of additional training observations can significantly alter the prediction performance of a learned tree. Therefore, bagging or bootstrap aggregation, which is a general-purpose procedure, has been proposed to reduce the variance of a statistical learning method.

3.3. Alternative methods for prediction

It is important to describe alternative methods that could be suitable for conducting housing price prediction research. One such method is SVM, known for its effectiveness in high-dimensional spaces and its ability to handle non-linear relationships through kernel functions. Another alternative is ANN, which is particularly powerful in capturing complex patterns and interactions within the data, making them suitable for dynamic markets like real estate.

Additionally, Bayesian regression offers a probabilistic approach to prediction, allowing for the incorporation of prior knowledge and the quantification of uncertainty in predictions. This can be particularly useful in markets with significant volatility. Decision trees and ensemble methods, such as gradient boosting machines (GBM) and XG Boost, are also viable alternatives due to their robustness and high predictive accuracy. These models can handle large datasets and complex interactions, making them suitable for real estate price prediction.

Each of these alternative methods has its strengths and can be selected based on the specific characteristics of the dataset and the research objectives. By incorporating a description of these methods, the methodology section will provide a comprehensive overview of potential approaches, enhancing the robustness and credibility of the research.

3.4. Model performance metrics

In the following, we present the performance measures, including the MAE, MSE, and R^2, which are used to evaluate the predictive performance of the prediction models.

3.4.1. Mean absolute error

The MAE represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of residuals. (e_t) in the dataset.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t| \tag{1}$$

The smaller the MAE, the better the model predictions aligned with the actual data. If the MAE of a prediction model is equal to zero, then it would mean a perfect prediction; however, in most cases, achieving such perfection is impossible.

In our study, the actual value is the listing price value and the predicted value is the predicted price value of the house property. Both were continuous variables.

3.4.2. Root mean square error

The MSE has been widely employed to assess the predictive performance of a model in relation to a numerical outcome (Willmott & Matsuura, 2005). This metric is derived from the residuals of the model, which represent the difference between the observed and expected values of the model. The MSE is computed by squaring the residuals and summing them.

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

However, the weakness of the MSE is that it loses the units of the original variable because it is the sum of the squared residuals. Therefore, to maintain the units of the data, the root mean square error (RMSE) was calculated by extracting the square root of the MSE:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \tag{3}$$

Similar to MAE, the lower the RMSE, the better the ability of the model to predict accurately.

3.4.3. Coefficient of determination

In addition to the RMSE, the R^2, a measure that provides information about the goodness of fit of a model, has also been widely used. This coefficient represents the percentage of variability in data that the model is capable of explaining (Helland, 1987). For example, the R^2 of a regression model is 0.75, which indicates that this model can explain 75% of the variability in the dependent variable. The simplest method to calculate the R^2 is based on calculating the correlation coefficient between the observed value and the predicted value and then squaring it:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \quad (4)$$

Contrary to MAE and RMSE, the higher the R^2, the more accurate the model's ability to predict.

In this study, we employed all three metrics to evaluate the accuracy of the predictive models.

3.5. K-fold cross-validation

In addition to performance metrics, this subsection provides a specific technique of CV to evaluate the performance of ML algorithms in predicting housing prices. This is a k-fold CV. This technique aims to minimize the disadvantages of the hold-out method. K-fold CV randomly divides the dataset into k-folds, trains the model on k – 1 folds, and evaluates it on the remaining fold. This procedure was repeated k times, resulting in k-test error estimates that were averaged to obtain the k-fold CV estimate.

4. EMPIRICAL APPLICATIONS

4.1. Data and variables

Our data comprised 78,704 house sales records. All data are retrieved from a reputable website about real estate transactions offering prices in Hanoi (https://alonhadat.com.vn/). Table 1 lists and describes all the variables used in this study. We collect the data from January to December 2022.

Table 1. Variable description

Field	*Description*	*Data type*
Orient	Orientation of the house	Character
Dining	Number of dining rooms	Integer
Road	Width of the road in front of the house	Numeric
Kitchen	Number of kitchens	Integer
Type	Is the house located in an alley?	Binary
Regulation	Is the house legally owned?	Binary
Width	Width of the house	Numeric
Length	Length of the house	Numeric
Bedroom	Number of bedrooms	Integer
Parking	Availability of car parking	Binary
Floor	Number of floors of the house	Integer
Bathroom	Number of bathrooms	Integer
Address	Address of the house	Integer
Agency	Involvement of a broker or not	Binary
Dist	Which inner district in Hanoi it belongs to	Character

The variables in this study were captured from real estate transactions in Hanoi during the year 2022. This timeframe provides a comprehensive view of the housing market within a single year, allowing for accurate and relevant predictions.

We select this set of variables based on data availability. The non-numeric variables are represented as dummy variables. In addition, the process of data normalization was applied to the numeric variables using the following formula:

$$X' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (5)$$

House areas, especially in major cities in Vietnam, rarely exceed 1000 square meters. As a result, houses with areas larger than this threshold were considered outliers and excluded from the model.

Currently, Hanoi consists of 12 districts: Hoan Kiem, Tay Ho, Ba Dinh, Cau Giay, Dong Da, Hai Ba Trung, Long Bien, Thanh Xuan, Nam Tu Liem, Hoang Mai, Bac Tu Liem and Ha Dong. Table 2 shows the median price of houses available for purchase by district in Hanoi. This table shows significant fluctuations in house prices among the various districts. We can clearly see a notably higher price in central districts, such as Hoan Kiem and Ba Dinh. Meanwhile, in the same district, there is a considerable difference in the median housing prices between properties situated on major roads and those located on alleyways or side streets. This difference was most clearly observed in Hoan Kiem, followed by Tay Ho.

Table 2. Median price (in billion VND) per house by district in Hanoi

District	*Median price in billion VND per house*	
	Houses in the street	*Houses in the alley*
Hoan Kiem	37.75	7.50
Tay Ho	15.20	4.70
Ba Dinh	14.50	4.60
Cau Giay	12.35	4.30
Dong Da	10.20	4.25
Hai Ba Trung	9.50	3.60
Long Bien	7.50	3.50
Thanh Xuan	6.55	3.30
Nam Tu Liem	6.30	3.20
Hoang Mai	4.50	3.00
Bac Tu Liem	4.50	2.90
Ha Dong	3.90	2.25

Table 3 illustrates the median area of houses in 12 districts in Hanoi. It can be seen that Hoan Kiem district has the highest median area of houses available for sale. Specifically, 50% of the houses situated on the streets within this district have an area of over 83 square meters. Nevertheless, when the house was inside an alley, the aforementioned number was reduced to 46 square meters.

Table 3. The median area of houses (in square meters) by district in Hanoi

District	*The median is of houses in square meters*	
	Houses in the street	*Houses in the alley*
Hoan Kiem	83	50
Tay Ho	75	48
Ba Dinh	75	46
Cau Giay	60	45
Dong Da	60	44
Hai Ba Trung	57	42
Long Bien	56	42
Thanh Xuan	55	41
Nam Tu Liem	55	40
Hoang Mai	54	40
Bac Tu Liem	50	40
Ha Dong	45	36

4.2. Results and discussions

This subsection of our study provides and explains the experimental results of the ML models used in the study for house price prediction.

Tables 4 and 5 summarize the outcomes of the prediction metrics on the training and test sets, respectively, using the 5-fold CV method. As reported in these tables, the best results were obtained using the random forest model for both training and validation sets. This algorithm showed the highest R^2 score and lowest RMSE and MAE values among the ML algorithms. Additionally, the random forest model is the most stable prediction model because it has the lowest standard deviation (SD). Conversely, linear regression showed the lowest predictive performance with the highest RMSE and MAE, as well as the lowest R^2. In addition to the random forest model, there are four more advanced ML models (XG Boost, stochastic gradient boosting, KNN, and bagged CART) that exhibit superior predictive power than regression methods. This finding aligns with several prior studies that concluded that advanced ML techniques outperform traditional ML techniques.

Table 4. Performance comparison in training set

Machine learning approach	*Algorithm*	*Avg. R^2*	*Avg. MAE*	*Avg. RMSE*	*SD. R^2*	*SD. MAE*	*SD. RMSE*
Random forest	rf	**0.825**	**2.992**	**7.194**	**0.011**	0.073	0.306
Extreme gradient boosting	xgbLinear	0.813	3.057	7.469	0.013	**0.061**	**0.292**
Stochastic gradient boosting	gbm	0.797	3.408	7.742	0.016	0.082	0.354
K-nearest neighbors	KNN	0.783	3.258	8.019	0.015	0.073	0.35
Bagged CART	treebag	0.727	4.388	8.998	0.017	0.121	0.383
Linear regression	lm	**0.704**	**4.574**	**9.376**	**0.019**	**0.107**	**0.364**
Generalized linear model	glmStepAIC	0.704	4.575	9.376	0.019	0.107	0.364
Generalized additive model using splines	gamSpline	0.693	4.436	9.533	0.023	0.084	0.452
Generalized additive model using splines	bam	0.626	4.409	11.468	0.134	0.135	5.317
Generalized additive model using splines	gam	0.599	4.425	12.2	0.144	0.15	5.711

Table 5. Performance comparison in validation set

Machine learning approach	*Algorithm*	*Avg. R^2*	*Avg. MAE*	*Avg. RMSE*	*SD. R^2*	*SD. MAE*	*SD. RMSE*
Random forest	rf	**0.821**	**3.003**	**7.198**	**0.012**	0.085	0.358
Extreme gradient boosting	xgbLinear	0.809	3.102	7.449	0.014	**0.074**	**0.294**
Stochastic gradient boosting	gbm	0.793	3.414	7.754	0.017	0.093	0.365
K-nearest neighbors	KNN	0.779	3.267	8.121	0.019	0.085	0.371
Bagged CART	treebag	0.720	4.394	9.003	0.017	0.129	0.398
Linear regression	lm	**0.701**	**4.585**	**9.424**	**0.023**	**0.115**	**0.375**
Generalized linear model	glmStepAIC	0.701	4.585	9.424	0.023	0.115	0.375
Generalized additive model using splines	gamSpline	0.675	4.447	9.565	0.031	0.096	0.491
Generalized additive model using splines	bam	0.620	4.416	11.484	0.145	0.154	5.324
Generalized additive model using splines	gam	0.578	4.441	12.342	0.156	0.15	5.725

Our results are similar to the previous finding. For example, we find that hedonic traditional pricing models are inferior to advanced ML models. This result is similar to the result of Abidoye and Chan (2018), who compare the hedonic pricing model with the ANN technique in property valuation. Khrais and Shidwan (2023) also conclude that advanced ML models can predict house prices better than conventional modeling approaches in the Middle East region. Xu and Zhang (2023) find similar results when forecasting the retail property price index in China. Overall, the empirical findings support *H1* and demonstrate that advanced ML models have superior predictive power compared with statistics-based ML, such as linear regression.

5. CONCLUSION

This study attempts to identify the most suitable model for predicting house prices in Hanoi, Vietnam. Applying ten ML models to the housing price data of Hanoi, we find that advanced ML approaches such as random forest, XG Boost, and stochastic gradient boosting outperform traditional models in forecasting future house prices. Furthermore, the empirical results reveal that the random forest algorithm is the best model for predicting housing prices in Hanoi. All of our results in this study are robust when we employ alternative evaluation metrics, including performance metrics (R^2, MAE, and RMSE), CV techniques (k-fold CV), and data visualization tests.

Our findings highly recommend policymakers, investors, and real estate agents in Hanoi, Vietnam to use ML techniques to predict the reference price of houses in particular and real estate in general. This will be useful for solving the problem of highly asymmetric information in the real estate market. Additionally, technology development can create new forecasting techniques that are better than the old ones. Stakeholders of the real estate market should update the new ML techniques frequently and apply the best ones in predicting housing prices.

Our study is not without limitations. Our data is sourced from only one website (https://alonhadat.com.vn/) and covers only Hanoi city. Although this website is a commonly used website and Hanoi is the capital of Vietnam, collecting more data from other websites and having more information about house prices in other cities may be useful in predicting house prices. In future studies, we will investigate house price prediction using a bigger set of data consisting of information from many websites and covering many cities in Vietnam.

While this study focuses on Hanoi's real estate market, the implications of our findings extend far beyond this specific locale. Hanoi, the capital city of Vietnam, is experiencing rapid urbanization and economic growth, making it a representative case study for other emerging markets globally. The city's dynamic real estate market, characterized by increasing foreign investment, population growth, and evolving government policies, mirrors similar trends seen in other fast-developing urban areas worldwide.

The methodologies employed in this research, including advanced ML models such as random forest and gradient boosting, have universal applicability. These techniques are robust and versatile, capable of handling diverse datasets and extracting meaningful insights in various contexts. By demonstrating the effectiveness of these models in Hanoi, we provide a framework that can be adapted and applied to other cities undergoing similar transformations.

Moreover, insights from Hanoi's market can inform broader strategies for urban development and real estate investment. Policymakers, investors, and developers in other regions can leverage these findings to enhance their understanding of market dynamics, improve predictive accuracy, and make more informed decisions. Thus, while our study is rooted in Hanoi, its methodologies and conclusions offer valuable lessons for global urban markets facing rapid growth and change.

This study makes significant contributions to the existing literature on housing price prediction by demonstrating the superior performance of advanced ML models compared to traditional statistical approaches. Our research provides empirical evidence that models such as random forest, gradient boosting, and KNN offer higher predictive accuracy, which is critical for stakeholders in the real estate market. These findings contribute to the growing body of literature that supports the adoption of advanced analytical techniques in economic and market forecasting.

One of the primary contributions of this study is the comprehensive evaluation of different ML models using a robust dataset comprising 78,704 real estate transactions from Hanoi, Vietnam. By employing various performance metrics, including MAE, RMSE, and R^2, we provide a detailed comparison of the model's predictive capabilities. This methodological rigor enhances the validity of our findings and offers a valuable reference for future research in similar contexts.

The implications of our study are manifold. For policymakers, the improved predictive accuracy of advanced ML models can inform better decision-making processes related to housing policies and market regulations. Accurate price predictions help in planning and implementing effective housing policies, thereby stabilizing the market and preventing potential bubbles. For investors and developers, the ability to forecast housing prices with greater precision can lead to more informed investment decisions, optimizing returns and managing risks more effectively.

Furthermore, our research underscores the importance of integrating advanced ML techniques into real estate market analysis. Traditional models, while useful, often fail to capture the complexities and nonlinearities inherent in housing data. By contrast, advanced models can handle large datasets with multiple variables and interactions, providing deeper insights and more reliable predictions.

In conclusion, this study not only advances the methodological approaches in housing price prediction but also offers practical implications for various stakeholders in the real estate market. The integration of advanced ML models represents a significant step forward in enhancing the accuracy and reliability of market forecasts, ultimately contributing to the stability and efficiency of the housing market. This work lays the groundwork for future studies to explore and refine these techniques further, expanding their applicability to different regions and market conditions.

REFERENCES

Abidoye, R. B., & Chan, A. P. C. (2018). Improving property valuation accuracy: A comparison of hedonic pricing model and artificial neural network. *Pacific Rim Property Research Journal, 24*(1), 71–83. https://doi.org/10.1080/14445921.2018.1436306

Abidoye, R. B., Chan, A. P. C., Abidoye, F. A., & Oshodi, O. S. (2019). Predicting property price index using artificial intelligence techniques: Evidence from Hong Kong. *International Journal of Housing Markets and Analysis, 12*(6), 1072–1092. https://doi.org/10.1108/IJHMA-11-2018-0095

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science, 199*, 806–813. https://doi.org/10.1016/j.procs.2022.01.100

Afonso, M., Blok, P. M., Polder, G., van der Wolf, J. M., & Kamp, J. (2019). Blackleg detection in potato plants using convolutional neural networks. *IFAC-PapersOnLine, 52*(30), 6–11. https://doi.org/10.1016/j.ifacol.2019.12.481

Begum, A., Kheya, N. J., & Rahman, Z. (2022). Housing price prediction with machine learning. *International Journal of Innovative Technology and Exploring Engineering, 11*(3), 42–46. https://doi.org/10.35940/ijitee.C9741.0111322

Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics, 13*(1), 68–84. https://doi.org/10.1016/j.jhe.2004.01.001

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Case, B., Clapp, J., Dubin, R., & Rodriguez, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. *The Journal of Real Estate Finance and Economics, 29*, 167–191. https://doi.org/10.1023/B:REAL.0000035309.60607.53

Gerek, I. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. *Automation in Construction, 41*, 33–39. https://doi.org/10.1016/j.autcon.2014.02.002

Chen, J.-H., Ong, C. F., Zheng, L., & Hsu, S.-C. (2017). Forecasting spatial dynamics of the housing market using support vector machine. *International Journal of Strategic Property Management, 21*(3), 273–283. https://doi.org/10.3846/1648715X.2016.1259190

Fan, G.-Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies, 43*(12), 2301–2315. https://doi.org/10.1080/00420980600990928

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gu, J., Zhu, M., & Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Applications, 38*(4), 3383–3386. https://doi.org/10.1016/j.eswa.2010.08.123

Helland, I. S. (1987). On the interpretation and use of R^2 in regression analysis. *Biometrics, 43*(1), 61–69. https://doi.org/10.2307/2531949

Hong, J., Choi, H., & Kim, W.-s. (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*(3), 140–152. https://doi.org/10.3846/ijspm.2020.11544

Hort, K. (1998). The determinants of urban house price fluctuations in Sweden 1968–1994. *Journal of Housing Economics, 7*(2), 93–120. https://doi.org/10.1006/jhec.1998.0225

Kauko, T., Hooimeijer, P., & Hakfoort, J. (2002). Capturing housing market segmentation: An alternative approach based on neural network modelling. *Housing Studies, 17*(6), 875–894. https://doi.org/10.1080/02673030215999

Kestens, Y., Thériault, M., & Des Rosiers, F. (2006). Heterogeneity in hedonic modelling of house prices: Looking at buyers' household profiles. *Journal of Geographical Systems, 8*, 61–96. https://doi.org/10.1007/s10109-005-0011-8

Khrais, L. T., & Shidwan, O. S. (2023). The role of neural network for estimating real estate prices value in post COVID-19: A case of the middle east market. *International Journal of Electrical and Computer Engineering, 13*(4), 4516–4525. https://doi.org/10.11591/ijece.v13i4.pp4516-4525

Kim, K., & Park, J. (2005). Segmentation of the housing market and its determinants: Seoul and its neighbouring new towns in Korea. *Australian Geographer, 36*(2), 221–232. https://doi.org/10.1080/00049180500150019

Kusan, H., Aytekin, O., & Özdemir, I. (2010). The use of fuzzy logic in predicting house selling price. *Expert Systems with Applications, 37*(3), 1808–1813. https://doi.org/10.1016/j.eswa.2009.07.031

Levantesi, S., & Piscopo, G. (2020). The importance of economic variables on London real estate market: A random forest approach. *Risks, 8*(4), Article 112. https://doi.org/10.3390/risks8040112

Liu, R., & Liu, L. (2019). Predicting housing price in China based on long short-term memory incorporating modified genetic algorithm. *Soft Computing, 23*, 11829–11838. https://doi.org/10.1007/s00500-018-03739-w

Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In T. O'Sullivan & K. Gibb (Eds.), *Housing economics and public policy* (pp. 67–89). Wiley. https://doi.org/10.1002/9780470690680.ch5

Manasa, J., Gupta, R., & Narahari, N. S. (2020). Machine learning based predicting house prices using regression techniques. In *2020 2nd International conference on innovative mechanisms for industry applications (ICIMIA)* (pp. 624–630). IEEE. https://doi.org/10.1109/ICIMIA48430.2020.9074952

Meese, R., & Wallace, N. (2003). House price dynamics and market fundamentals: The Parisian housing market. *Urban Studies, 40*(5–6), 1027–1045. https://doi.org/10.1080/0042098032000074308

Muellbauer, J., & Murphy, A. (2008). Housing markets and the economy: The assessment. *Oxford Review of Economic Policy, 24*(1), 1–33. https://doi.org/10.1093/oxrep/grn011

Nellis, J. G., & Longbottom, J. A. (1981). An empirical analysis of the determination of house prices in the United Kingdom. *Urban Studies, 18*(1), 9–21. https://doi.org/10.1080/00420988120080021

Nguyen, N., & Cripps, A. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of Real Estate Research, 22*(3), 313–336. https://doi.org/10.1080/10835547.2001.12091068

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications, 42*(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Science and Management, 4*, 19–29. https://doi.org/10.1016/j.dsm.2021.12.002

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*, Article 121059. https://doi.org/10.1016/j.eswa.2023.121059

Sarip, A. G., Hafez, M. B., & Daud, N. (2016). Application of fuzzy regression model for real estate price prediction. *Malaysian Journal of Computer Science, 29*(1), 15–27. https://doi.org/10.22452/mjcs.vol29no1.2

Schulz, R., & Werwatz, A. (2004). A state space model for Berlin house prices: Estimation and economic interpretation. *The Journal of Real Estate Finance and Economics, 28*, 37–57. https://doi.org/10.1023/A:1026373523075

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2, Part 2), 2843–2852. https://doi.org/10.1016/j.eswa.2008.01.044

Soltani, A., Heydari, M., Aghaei, F., & Pettit, C. J. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. *Cities, 131*, Article 103941. https://doi.org/10.1016/j.cities.2022.103941

Tang, F., Xiao, C., Wang, F., & Zhou, J. (2018). Predictive modeling in urgent care: A comparative study of machine learning approaches. *JAMIA Open, 1*(1), 87–98. https://doi.org/10.1093/jamiaopen/ooy011

Thamarai, M., & Malarvizhi, S. P. (2020). House price prediction modeling using machine learning. *International Journal of Information Engineering and Electronic Business, 12*(2), 15–20. https://doi.org/10.5815/ijieeb.2020.02.03

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science, 174*, 433–442. https://doi.org/10.1016/j.procs.2020.06.111

Wang, Q., Jiao, W., Wang, P., & Zhang, Y. (2021). A tutorial on deep learning-based data analytics in manufacturing through a welding case study. *Journal of Manufacturing Processes, 63*, 2–13. https://doi.org/10.1016/j.jmapro.2020.04.044

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79–82. https://doi.org/10.3354/cr030079

Xu, P., Han, Y., & Yang, J. (2012). U.S. monetary policy surprises and mortgage rates. *Real Estate Economics, 40*(3), 461–507. https://doi.org/10.1111/j.1540-6229.2011.00325.x

Xu, X., & Zhang, Y. (2023). Retail property price index forecasting through neural networks. *Journal of Real Estate Portfolio Management, 29*(1), 1–28. https://doi.org/10.1080/10835547.2022.2110668

Xu, X., & Zhang, Y. (2024). Office property price index forecasting using neural networks. *Journal of Financial Management of Property and Construction, 29*(1), 52–82. https://doi.org/10.1108/JFMPC-08-2022-0041

Yoo, S., Im, J., & Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning, 107*(3), 293–306. https://doi.org/10.1016/j.landurbplan.2012.06.009

Yu, L., Jiao, C., Xin, H., Wang, Y., & Wang, K. (2018). Prediction on housing price based on deep learning. *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering, 12*(2), 90–99. https://web.archive.org/web/20201224050108/https://zenodo.org/record/1315879/files/10008599.pdf

Zhou, J., Zhang, H., Gu, Y., & Pantelous, A. A. (2018). Affordable levels of house prices using fuzzy linear regression analysis: The case of Shanghai. *Soft Computing, 22*, 5407–5418. https://doi.org/10.1007/s00500-018-3090-4

Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House price prediction using a machine learning model: A survey of literature. *International Journal of Modern Education and Computer Science, 12*(6), 46–54. https://doi.org/10.5815/ijmecs.2020.06.04