INTERNATIONAL REAL ESTATE REVIEW

2025 Vol. 28 No.4: pp. 505 – 527

Using Machine Learning Regression Algorithms to Predict House Prices in Vietnam

Minh-Thang Ha
Hung Yen University of Technology and Education, Email: haminhthang1213@gmail.com

Thi-Cham Nguyen
Haiphong University of Medicine and Pharmacy, Email: nthicham@hpmu.edu.vn

Thanh-Huyen Pham
Halong University, Email: phamthanhhuyen@daihochalong.edu.vn

Van-Hau Nguyen*
Hung Yen University of Technology and Education, Email: haunv@utehy.edu.vn

This study develops a comprehensive machine learning (ML) framework for house price prediction in Vietnam by utilizing a dataset of 28,156 property listings from a real estate website. We employ rigorous data preprocessing, feature engineering, and comparative analysis of ML algorithms, including CatBoost, XGBoost, and random forests. The results demonstrate the superiority of ensemble methods, with CatBoost achieving the highest performance on the main dataset ($R^2 = 0.510$, RMSE = 17.614). Regional analyses in Hanoi and Ho Chi Minh City reveal the adaptability of the models for local market dynamics. A Shapley additive explanations analysis reveals key drivers of house prices, such as area, population density, and property-specific attributes. The findings contribute to the academic understanding of real estate valuation and provide actionable insights for policymakers, investors, and other stakeholders. This study lays the groundwork for developing automated valuation models and their practical implementation, exemplified by a website application. By harnessing ML and data-driven insights, this research advances transparent, efficient, and informed decision-making in the real estate sector in Vietnam, while offering a robust methodology for house price prediction in emerging markets.

* Corresponding author

DOI: https://doi.org/10.53383/100412

Keywords

House price prediction, Machine learning, Regression algorithms, Real estate market, Ensemble models

1. Introduction

The real estate market plays a pivotal role in the economic growth and development of nations worldwide, and Vietnam is no exception. In recent years, Vietnam has witnessed a remarkable transformation in its real estate sector, driven by rapid urbanization, economic growth, and increasing foreign investment (Nguyen et al., 2024; Savills Vietnam, 2024). As the demand for housing continues to rise, accurate house price prediction has become a crucial concern for various stakeholders, including homebuyers, investors, developers, and policymakers (Truong et al., 2020).

Traditionally, house price prediction has relied on conventional methods such as hedonic pricing models and expert opinions, which often fail to capture the complex and dynamic nature of the real estate market (Rosen, 1974). The advent of machine learning (ML) techniques has opened up new possibilities for predicting house prices with higher accuracy and reliability (Park and Bae, 2015). By leveraging the power of data and advanced algorithms, ML models can uncover hidden patterns, relationships, and trends in real estate data, thus enabling more informed decision-making (Kok et al., 2017).

This study addresses the following research questions in the context of the Vietnamese real estate market:

- How can a comprehensive ML framework be developed to predict house prices in Vietnam, after considering the unique challenges and characteristics of the market?
- What are the most effective ML techniques for predicting house prices in Vietnam, and how do they compare in terms of performance and interpretability?
- What are the key factors that influence house prices in Vietnam, and how can they be identified and quantified by using advanced feature importance analysis techniques?

To answer these questions, we utilize a comprehensive dataset obtained from alonhadat.com.vn, one of the largest real estate websites in Vietnam, and supplement the information with data from Wikipedia. The main contributions of this study are threefold:

(1) Development of a comprehensive ML framework for house price prediction in Vietnam: We propose a novel approach that integrates advanced data preprocessing, feature engineering techniques, and state-of-the-art ML models to predict house prices accurately. The framework

addresses the unique challenges and characteristics of the Vietnamese housing market and provides a robust and reliable solution.

(2) Rigorous comparative analysis of ML models and preprocessing techniques: We conduct an extensive evaluation of various ML models, including linear, regularized regression and support vector regressions (SVRs), ensemble methods, and robust regression techniques. The performance of these models is evaluated by using a number of different metrics, such as mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and R-squared (R^2). Furthermore, we investigate the impact of different data preprocessing techniques on model performance, to identify the most effective approaches for ensuring optimal data quality and relevance.

(3) Identification of key factors that influence house prices in Vietnam through a feature importance analysis: We employ advanced feature importance analysis techniques, such as a Shapley additive explanations analysis (SHAP), to reveal the key drivers of house prices in the Vietnamese real estate market. By providing a clear understanding of the relative importance of various features, such as location, property size, and amenities, we enable stakeholders to make more informed decisions and develop targeted strategies for the housing market. These insights contribute to enhancing transparency, efficiency, and affordability in the Vietnamese real estate sector.

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature on house price prediction and ML applications in real estate. Section 3 describes the dataset, data preprocessing, and feature engineering techniques employed in this study. Section 4 presents the methodology, including the ML models, evaluation metrics, and model interpretation techniques used. Section 5 discusses the results, model performance, regional variations, and feature importance analysis. Finally, Section 6 concludes the paper, by highlighting the key findings, limitations, and future research directions.

2. Related Works

2.1 Evolution of House Price Prediction Models

The prediction of house prices has been the subject of extensive research in economics, urban planning, and data science for decades. Traditional approaches predominantly rely on hedonic pricing models, which employ a regression analysis to estimate property values based on their inherent characteristics (Rosen, 1974). These models typically consider factors such as property size, location, number of bedrooms, and proximity to amenities. However, the limitation of these models lies in their linear assumptions and inability to capture complex, non-linear relationships inherent in real estate data.

In recent years, ML has emerged as a powerful paradigm for house price prediction and offers significant improvements over traditional methods. ML models have demonstrated superior capability in handling large datasets and uncovering intricate patterns that often elude conventional statistical models. The application of various regression models has been extensively explored in this domain. While linear regressions offer interpretability, they often struggle with capturing non-linearity in data. Ridge and Lasso regressions, which introduce regularization terms to mitigate overfitting, have shown enhanced performance in certain scenarios (Tibshirani, 1996; Zou and Hastie, 2005).

Ensemble methods, which aggregate predictions from multiple models, have proven particularly effective in house price prediction. Random forests, an ensemble of decision trees, have been widely adopted due to their robustness and ability to handle large datasets with numerous features (Breiman, 2001). Gradient boosting machines (GBMs), including XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017), have achieved state-of-the-art results by iteratively improving predictive accuracy through boosting. These ensemble methods have been successfully applied to real estate price prediction in various contexts, thus demonstrating their versatility and effectiveness (Park and Bae, 2015; Truong et al., 2020).

More recently, deep learning techniques have gained traction in the field of house price prediction. Artificial neural networks (ANNs) and convolutional neural networks (CNNs) have been employed to capture complex patterns and hierarchical features in real estate data (Nguyen and Cripps, 2001; Ma et al., 2019). These models have shown promising results, particularly in incorporating spatial dependencies and neighborhood effects (Rahman et al., 2020; Shahhosseini et al., 2021).

2.2 Machine Learning Applications in Real Estate

The application of ML in real estate extends beyond price prediction to encompass various aspects of the industry. Kok et al. (2017) employ random forests to predict rental prices in Amsterdam and achieve notable improvements in accuracy. Similarly, Siregar et al. (2022) utilized GBMs for real estate housing prices, which shows the adaptability of ML models to diverse real estate markets.

In addition to price prediction, ML techniques have been applied to other critical areas in real estate. Bency et al. (2017) use CNNs to classify property images and extract relevant features for automated valuation. Chen et al. (2020) employ natural language processing techniques to analyze property descriptions and identify key attributes that influence house prices. These studies highlight the potential of ML in automating and enhancing various processes in the real estate industry.

2.3 Machine Learning in Emerging Real Estate Markets

The use of ML in emerging real estate markets has grown rapidly as these economies experience rapid urbanization and structural changes. In Vietnam, Nguyen et al. (2024) apply several regression algorithms to a large online listing dataset and find that random forests provide the most accurate housing price predictions, with housing area identified as the dominant factor driving values. Similarly, Thi et al. (2025) evaluate multiple regression algorithms including linear regression, Lasso and Ridge regressions, XGBoost, and random forests on 9,122 property listings in Hanoi and demonstrate that ensemble methods, particularly XGBoost and random forests, outperform traditional linear approaches in capturing local market dynamics. In South Africa, Yacim and Boshoff (2018) evaluate ANNs against linear and semi-log hedonic models using mass appraisal data for Cape Town and show that neural networks can substantially improve predictive accuracy, albeit with reduced transparency compared to traditional models (Yacim and Boshoff, 2018).

Despite these promising results, several challenges remain in applying ML to emerging markets. Studies in Vietnam emphasize limitations in data coverage and reliability, as well as the need for careful preprocessing and feature engineering tailored to local housing attributes and market structures (Nguyen et al., 2024; Thi et al., 2025). The South African experience also highlights the trade-off between accuracy and interpretability when deploying complex ML models in valuation practices (Yacim and Boshoff, 2018). Together, these works suggest that while ML holds considerable potential for improving house price prediction in emerging markets, success critically depends on data quality, context-specific modeling choices, and the ability to explain model outputs to practitioners and policy makers.

3. Data

3.1 Data Source and Acquisition

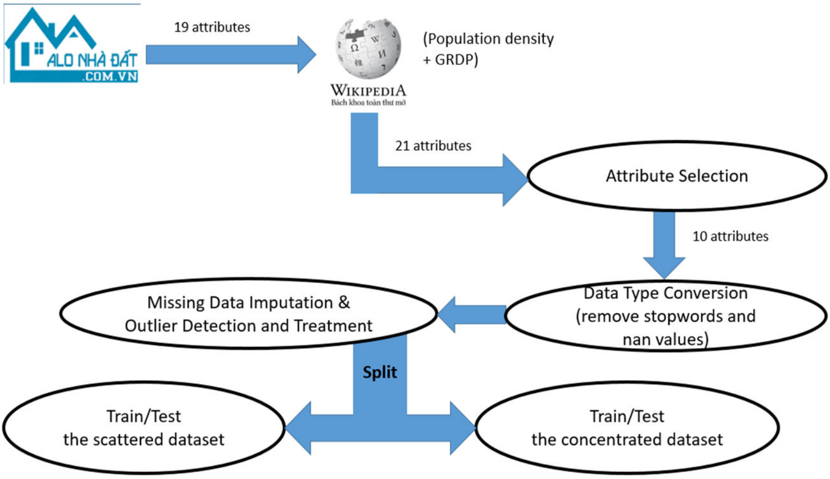
The dataset used in this study is primarily sourced from alonhadat.com.vn, a prominent real estate platform in Vietnam. Through web scraping techniques, including the use of Python libraries such as Requests, BeautifulSoup, and Selenium, a total of 28,156 property listings are acquired. Each listing contains 19 distinct attributes relevant to house valuation, thus providing a comprehensive set of features for analysis.

To further enhance the predictive power of the models, we supplement the dataset with two additional attributes: population density and gross regional domestic product (GRDP). These attributes are obtained from Wikipedia[1] and matched to each property based on its location. The inclusion of these socioeconomic factors is motivated by the hypothesis that they might serve as influential predictors of house prices, thus capturing the broader context in which the properties are situated.

After the initial data acquisition, a thorough attribute selection process is conducted. Eleven attributes that are deemed unimportant or irrelevant to the task of house price prediction, such as the property ID and the presence of a dining room, are removed. This refinement step aims to streamline the dataset and focus on the most informative and pertinent features.

The resulting dataset, which consists of 10 carefully selected attributes, forms the foundation for the subsequent stages of data preprocessing, feature engineering, and model development. By leveraging a diverse range of data sources and applying judicious attribute selection, we have curated a robust and comprehensive dataset that captures the essential characteristics of the Vietnamese real estate market. This meticulous approach to data acquisition and preparation sets the stage for the development of accurate and reliable house price prediction models. The overall data acquisition and preprocessing pipeline used in this study are summarized in Figure 1.

Figure 1 Data Acquisition and Preprocessing Pipeline


ALO NHÀ ĐẤT .COM.VN
19 attributes
WIKIPEDIA
Bách khoa toàn thư mở
(Population density + GRDP)
21 attributes
Attribute Selection
10 attributes
Data Type Conversion (remove stopwords and nan values)
Missing Data Imputation & Outlier Detection and Treatment
Split
Train/Test the scattered dataset
Train/Test the concentrated dataset

[1] https://vi.wikipedia.org/wiki/Danh_s%C3%A1ch_%C4%91%C6%A1n_v%E1%BB%8B_h%C3%A0nh_ch%C3%ADnh_Vi%E1%BB%87t_Nam_theo_GRDP_b%C3%ACnh_qu%C3%A2n_%C4%91%E1%BA%A7u_ng%C6%B0%E1%BB%9Di. In Vietnamese.

3.2 Data Characteristics and Preprocessing

3.2.1 Attribute Overview

The acquired dataset encompasses a wide array of factors pertinent to house valuation, such as price, area, address, number of bedrooms and bathrooms, property type, and legal status. Table 1 presents a comprehensive list of the attributes, along with their respective meaning and data type.

Table 1 Dataset Attributes Crawled from alonhadat.com.vn

Attribute Name	Meaning	Data Type
Mã tin	Listing code of the house for sale	String
Hướng	Direction of the house	String
Phòng ăn	Presence of a dining room	String
Loại tin	Type of listing (For sale)	String
Đường trước nhà	Width of the road in front of the house	String
Nhà bếp	Presence of a kitchen	String
LoạiBDS	Type of property	String
Pháp lý	Legal status of the house	String
Sân thượng	Presence of a roof terrace	String
Chiều ngang	Width of the house	String
Số tầng	Number of floors	String
Chỗ để xe hơi	Presence of parking space for cars	String
Chiều dài	Length of the house	String
Số phòng ngủ	Number of bedrooms	String
Chính chủ	Ownership status (is the house owner selling it?)	String
Giá	Selling price of the house	String
Diện tích	Area of the house	String
Địa chỉ	Address of the house	String
Thuộc dự án	Project affiliation (Does the house belong to any project? If yes, which project?)	String

3.2.2 Data Cleaning and Preprocessing

To ensure data quality and integrity, a rigorous preprocessing protocol is implemented. The following steps are undertaken:

- Attribute Selection: Irrelevant or redundant attributes, such as 'Mã tin' and 'Loại tin', are eliminated to streamline the dataset and enhance model efficiency.
- Data Type Conversion: Relevant attributes, such as 'Giá', 'Diện tích', and 'Đường trước nhà', are converted to appropriate numerical data types (float or integer) to facilitate statistical analyses and model computations.
- Missing Data Imputation: Missing values are treated by using statistically sound imputation methods. For categorical attributes, mode imputation is employed, while for numerical attributes, mean or median imputation is used based on the distribution characteristics of each variable.

- Outlier Detection and Treatment: Outliers are identified by using robust statistical methods such as z-score analysis and interquartile range analysis. Detected outliers are either removed or treated by using appropriate techniques, such as winsorization or transformation, depending on the nature of the attribute and extent of the outliers.

3.2.3 Handling the 'Direction' Attribute

A notable challenge is encountered with the 'direction' attribute, which contains a substantial proportion (73.11%) of meaningless or missing data. The presence of such a large amount of missing or meaningless data can potentially impact the performance of the predictive models and introduce biases if not handled appropriately. To address this issue, four distinct approaches are proposed and evaluated, each with its own rationale and potential impact on the performance of the model. They are as follows.

Approach 1: Complete removal of the 'direction' attribute. This approach is based on the assumption that the 'direction' attribute may be irrelevant or may not contribute significantly to the predictive power of the models. By removing the attribute entirely, we simplify the data and reduce dimensionality, thus potentially improving the interpretability of the models. However, removing the attribute may lead to a loss of potentially valuable information if the 'direction' attribute does have some predictive power.

Approach 2: Replacement of meaningless values with random values while preserving the data distribution in the test set. This approach aims to maintain the statistical properties of the 'direction' attribute by replacing meaningless values with random values that follow the same distribution as the valid data in the test set. This assumes that the missing or meaningless values are missing at random (MAR) and that the distribution of the attribute is important for the performance of the model. However, replacing meaningless values with random values may introduce noise into the data, thus potentially affecting the performance of the model. Even so, if the missing values are truly MAR, this approach can help preserve the overall data structure and avoid biases introduced by complete removal or mode imputation.

Approach 3: Replacement of meaningless values with random values without distribution constraints. This approach is a simpler variant of Approach 2, where meaningless values are replaced with random values from the existing pool of valid values, without considering the data distribution. This assumes that the distribution of the attribute is not critical for the performance of the model. Similar to Approach 2, replacing meaningless values with random values may introduce noise into the data. However, if the distribution of the attribute is not a significant factor in the performance of the model, this approach can be a straightforward way to handle missing data.

Approach 4: Replacement of meaningless values with the most common value (mode imputation). This approach assumes that the missing or meaningless values are most likely to be the most frequently occurring value (mode) in the 'direction' attribute. It is a straightforward method that can be appropriate if there is a dominant value in the attribute, and it is reasonable to assume that the missing values would follow the same pattern. However, mode imputation may introduce biases if the missing values are not actually most likely to be the most common value. This approach may also reduce the variability in the data, thus potentially affecting the ability of the model to capture patterns and relationships.

To evaluate the impact of each approach on the performance of the model, we employ a rigorous evaluation framework that uses appropriate metrics such as the MAE, RMSE, MAPE, and R^2. These metrics contribute to a comprehensive evaluation of the predictive accuracy and explanatory power of the model under each approach.

Furthermore, statistical tests, such as paired t-tests or Wilcoxon signed-rank tests, are conducted to determine if the differences in performance between the approaches are statistically significant. This helps to identify the approach that yields the best performance across the different models and datasets.

The result of the comparative analysis reveals that Approach 2, which involves replacing meaningless values with random values while preserving the data distribution in the test set, consistently yields the best performance across all of the models and datasets. This finding suggests that maintaining the statistical properties of the 'direction' attribute is important for the predictive accuracy of the model.

However, it is essential to consider the trade-offs and potential biases associated with each approach. For example, while Approach 2 may provide the best performance, it may also introduce noise into the data. On the other hand, Approach 1, which involves removing the attribute entirely, may lead to a loss of potentially valuable information but can improve the interpretability of the models.

To sum it up, the proposed handling of the 'direction' attribute through four distinct approaches, each with its own rationale and potential impact, provides a comprehensive framework for addressing the challenge of meaningless or missing data. By rigorously evaluating the impact of each approach on the performance of the model with the use of appropriate metrics and statistical tests, we can make an informed decision on the most suitable method for handling this attribute. The comparative analysis reveals that preserving the data distribution while replacing meaningless values yields the best performance, which emphasizes the importance of considering the statistical properties of the data in the context of house price prediction.

3.3 Feature Engineering and Extraction

To enhance the predictive power of the models, a comprehensive feature engineering process is undertaken. Several new features are derived based on domain knowledge and exploratory data analysis:

- Price per Square Meter: This normalized measure of property value is calculated by dividing the price by the area, thus providing a standardized metric for comparison across properties of varying sizes.
- Distance to City Center: For properties located in major urban areas such as Hanoi and Ho Chi Minh City, the distance to the city center is calculated by using geospatial data and application programming interfaces (APIs). This feature captures the significant influence of central location on property values in urban settings.
- Property Age: Where available, the age of the property is determined based on the year of construction or renovation. This factor often exhibits a strong correlation with property value and condition.
- Neighborhood Affluence Index: An index that represents the relative affluence of different neighborhoods is developed based on aggregated property values, socioeconomic indicators, and geospatial data.
- Proximity to Amenities: The proximity of each property to key amenities, such as schools, hospitals, shopping centers, and public transportation, is calculated by using geospatial data and APIs. These features provide valuable insights into the desirability and convenience of the location of a property.

In addition to these engineered features, several relevant attributes are extracted from external sources to enrich the dataset:

- Population Density: The population density of each district or ward is obtained from census data and incorporated into the dataset.
- Crime Rate: The crime rate of each neighborhood is sourced from public safety databases and included as a feature, as it can significantly influence property values and desirability.
- School Quality: The quality of schools in each neighborhood is assessed by using education department rankings and incorporated as a feature, given the importance of educational facilities in the decision-making process of homebuyers.

The incorporation of these engineered and extracted features, alongside the original attributes, results in a rich and multidimensional dataset that effectively captures the complex interplay of factors that influence house prices in Vietnam. This robust dataset serves as the foundation for the development of accurate and reliable predictive models.

3.4 Data Partitioning

To ensure the robustness and generalizability of the predictive models, the preprocessed dataset is partitioned into training and testing sets by using a stratified random sampling approach. The training set, which comprises 80% of the data, is used for model development and optimization, while the remaining 20% form the testing set, which is used to evaluate the performance of the model on unseen data.

Furthermore, to evaluate the performance of the model across different geographical regions, separate subsets are created for major cities such as Hanoi and Ho Chi Minh City. This approach allows for a more nuanced analysis of the predictive capabilities of the model in specific local contexts.

To sum it up, the data acquisition, preprocessing, feature engineering, and partitioning steps undertaken in this study yield a comprehensive and high-quality dataset that effectively captures the complexities of the Vietnamese real estate market. This robust dataset provides a solid foundation for the development of accurate and reliable house price prediction models by using advanced ML techniques.

The dataset is partitioned into training and test sets to evaluate the performance of the model on unseen data. Table 2 summarizes the characteristics of the datasets used in this study.

Table 2 The Characteristics of the Datasets

Data_set	Feature	Train_set	Test_set
Main dataset (scattered)	11	17,289	4643
Hanoi dataset (concentrated)	11	8748	828
Ho Chi Minh dataset (concentrated)	11	6810	1140

The main dataset exhibits scattered properties, while the Hanoi and Ho Chi Minh datasets are geographically concentrated. This distinction allows for an assessment of the performance of the model under different data distributions and regional characteristics.

4. Methods

4.1 Machine Learning Models

In this study, we employ a diverse array of cutting-edge ML models to predict house prices in the Vietnamese real estate market. By leveraging a comprehensive set of algorithms that span linear and non-linear regression techniques, we identify the most effective and robust predictive models for this complex task. The selected models are categorized into five distinct groups: linear, regularized, support vector, ensemble, and robust regressions.

First, linear regression models, such as the standard linear regression algorithm, serve as a fundamental benchmark in our analysis. Despite their simplicity, these models provide a valuable baseline for evaluating the performance of more advanced techniques.

Second, regularized regression models, including Ridge, Lasso, and elastic net regressions, extend the linear regression framework by introducing regularization techniques. These methods mitigate overfitting and enhance model generalization, thus enabling more robust predictions in the presence of high-dimensional and potentially correlated features.

Third, SVRs are a powerful non-linear regression technique that leverages kernel functions to capture complex relationships between house features and prices. By mapping the input space to a higher-dimensional feature space, SVRs can effectively model non-linearities and interactions, thus providing a flexible and expressive framework for price prediction.

Fourth, ensemble regression models, such as random forests, XGBoost, LightGBM, and CatBoost, harness the collective power of multiple base learners to improve predictive accuracy and robustness. These state-of-the-art algorithms combine the predictions of numerous decision trees, by exploiting the diversity and complementarity of the individual models to yield superior performance.

Fifth, robust regression models, including RANSAC, Huber, and Theil-Sen regressions, are specifically designed to handle outliers and provide stable estimates in the presence of noisy or anomalous data points. By mitigating the impact of extreme values and leveraging robust loss functions, these models ensure the reliability and consistency of the price predictions.

Sixth, multi-layer perceptron (MLP) models provide a deep learning approach for house price prediction. By leveraging multiple layers of interconnected nodes with non-linear activation functions, MLP models can automatically find complex patterns in the data that may elude traditional methods. Given the size constraints of our dataset, we implement relatively simple architectures with two and three layers to balance model complexity with available data while avoiding overfitting. These neural network models complement our comprehensive approach by incorporating deep learning capabilities alongside conventional ML techniques.

The selection of these diverse ML models reflects our commitment to a comprehensive and rigorous approach to house price prediction. By exploring a wide range of algorithms, each with its unique strengths and characteristics, we aim to identify the most suitable and effective techniques for capturing the complex dynamics of the Vietnamese real estate market. Through a comparative analysis and evaluation of these models, we seek to contribute to the advancement of data-driven decision-making in the housing sector and provide actionable insights for various stakeholders.

In the subsequent sections, we will delve into the details of each model, and discuss their theoretical foundations, implementation specifics, and potential advantages in the context of our study. Moreover, we will present a rigorous evaluation framework, which uses a range of performance metrics to evaluate the predictive capabilities of the models and identify the most promising approaches for accurate and reliable house price prediction in Vietnam.

4.2 Model Evaluation Metrics

To evaluate the performance of the aforementioned models, several evaluation metrics are employed:

- MAE: Quantifies the average magnitude of errors in a set of predictions, while disregarding their direction. The formula for the MAE is:

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

where n is the number of samples, y_i is the actual value, and $\hat{y}_i$ is the predicted value.

- RMSE: Measures the square root of the average of squared differences between predicted and actual values. The RMSE penalizes larger errors more severely than the MAE. The formula for the RMSE is:

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

- MAPE: Evaluates the accuracy of a forecast by calculating the percentage difference between predicted and actual values. The formula for the MAPE is:

$$MAPE = \frac{100}{n}\sum\nolimits_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

- R^2: Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. The formula for R^2 is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{4}$$

where $\bar{y}_i$ is the mean of the actual values.

These metrics provide a comprehensive assessment of the predictive capabilities of the model by considering both the magnitude and direction of errors, as well as the proportion of variance explained by the models.

5 Results and Discussion

5.1 Model Performance

The performance of the various ML models is evaluated on the main dataset by using different metrics, including the MAE, RMSE, MAPE, and R². The results are summarized in Table 3.

Table 3 Performance Metrics on Main Dataset

Type	Model	Setting	MAE	RMSE	MAPE	R^2
Linear Regression		1	17.276	27.217	162.579	-0.171
		2	17.252	27.19	162,846	-0.168
		3	17.276	27.217	162.579	-0.171
		4	17.276	27.217	162.579	-0.171
Regularized Regression	Ridge	1	17.272	27.21	162.532	-0.170
		2	17.248	27.19	162.794	-0.168
		3	17.271	27.21	162.532	-0.170
		4	17.271	27.21	162.532	-0.170
	Lasso	1	17.173	27.071	159.295	-0.158
		2	17.173	27.071	159.294	-0.158
		3	17.173	27.071	159.295	-0.158
		4	17.173	27.071	159.295	-0.158
	ElasticNet	1	16.997	26.773	159.371	-0.133
		2	16.98	26.752	159.493	-0.131
		3	16.997	26.773	159.371	-0.133
		4	16.997	26.773	159.371	-0.133
Support Vector Regression		1	11.641	23.144	114.047	0.154
		2	8.591	19.442	70.007	0.403
		3	33.876	38.456	518.021	-1.337
		4	33.876	38.456	518.021	-1.337
Ensemble Regression	random forests Regression	1	8.545	20.819	77.764	0.315
		2	8.106	20.169	79.818	0.357
		3	8.257	20.124	79.792	0.360
		4	8.353	20.887	76.207	0.346
	XGBoost Regression	1	9.633	20.887	101.959	0.311
		2	8.994	20.373	86.963	0.344
		3	9.633	20.887	101.959	0.311
		4	9.633	20.887	101.959	0.311
	LightGBM Regression	1	9.249	19.572	71.438	0.395
		2	8.946	19.265	72.933	0.414
		3	9.249	19.572	71.438	0.395
		4	9.249	19.572	71.438	0.395
	CatBoost Regression	1	8.715	18.923	75.033	0.434
		2	7.32	17.614	57.32	0.510
		3	8.715	18.923	75.033	0.434
		4	8.715	18.923	75.033	0.434

(Table 3 Continued)

Robust Regression	**RANSAC Regression**	**1**	**16.867**	**27.994**	**145.006**	**-0.238**
		2	15.18	25.872	146.44	-0.058
		3	16.867	27.994	145.006	-0.238
		4	16.173	27.171	146.161	-0.167
	Huber Regression	1	13.72	24.483	127.344	0.053
		2	13.708	24.472	127.496	0.054
		3	13.72	24.483	127.344	0.053
		4	13.72	24.483	127.344	0.053
	Theil-Sen Regression	1	29.128	43.939	316.866	-2.051
		2	28.517	42.977	293.556	-1.918
		3	28.03	42.929	316.318	-1.912
		4	29.088	43.515	323.657	-1.992
Multi-layer Perceptron (MLP)	2 layers	1	11.43	23.782	130.47	0.090
		2	13.26	27.65	158.64	-0.210
		3	14.65	33.53	205.07	-0.780
		4	12.86	24.96	138.73	0.020
	3 layers	1	16.61	40.95	202.11	-1.650
		2	8.68	20.58	73.072	0.330
		3	16.31	45.95	238.68	-2.330
		4	12.29	23.97	130.46	0.090

The ensemble models, particularly CatBoost, LightGBM, and random forests, demonstrate superior performance across all of the metrics. CatBoost has the lowest RMSE and the highest R², thus indicating its strong predictive capability. The high MAE and MAPE values observed in linear models like the linear and Ridge regressions highlight their limitations in capturing the complexity of the data.

5.2 Regional Variations

A final summary of the model performance across the national and regional datasets is presented in Table 4.

Table 4 Final Summary Table

Type	Dataset	Model	Setting	MAE	RMSE	MAPE	R^2
Ensemble Regression	Main dataset	Catboost Regression	2	7.320	17.614	57.320	0.510
	Hanoi	XGBoost Regression		10.878	20.309	115.17	0.691
	Ho Chi Minh City	random forests Regression		6.202	18.824	33.320	0.623

A comparative analysis of the model performance across different datasets yields several noteworthy insights. The ensemble methods consistently outperform traditional linear models, which emphasizes the importance of capturing complex, non-linear relationships in the data. Among the ensemble methods, CatBoost emerges as the best-performing model for the nationwide dataset, with the lowest RMSE of 17.614 and an R^2 of 0.510 in Setting 2 (see Table 3). This superior performance can be attributed to its ability to automatically handle categorical features, thus reducing the need for extensive preprocessing and enabling this method to effectively capture the intricacies of the nationwide housing market.

In the regional datasets, XGBoost and random forests demonstrate superior performance in Hanoi and Ho Chi Minh City, respectively. XGBoost achieves the highest R^2 of 0.691 in Hanoi, thus indicating its strong predictive capability and adaptability to the specific characteristics of the Hanoi housing market. On the other hand, the random forest regression exhibits the lowest MAE of 6.202 and MAPE of 33.320% in Ho Chi Minh City, thus highlighting its robustness in capturing the complex dynamics of the housing market in this city.

The effectiveness of the ensemble methods can be attributed to their ability to combine multiple weak learners to create a strong, robust model. CatBoost, in particular, leverages gradient boosting techniques and a unique algorithm for processing categorical features, which enables this algorithm to handle the diverse and intricate nature of the nationwide dataset. Similarly, LightGBM and XGBoost, which also perform well, benefit from the efficiency of gradient boosting in handling large datasets with numerous features.

The regional variations in model performance underscore the importance of considering local market dynamics and the unique characteristics of each region when developing predictive models for house prices. The superior performance of XGBoost in Hanoi and random forests in Ho Chi Minh City suggests that these models are particularly well-suited to capture the specific nuances and patterns present in these regional markets.

Overall, the comparative analysis highlights the strengths of the ensemble methods, particularly CatBoost, XGBoost, and random forests, in predicting house prices across different geographical scales and market conditions. The results emphasize the importance of selecting appropriate models based on the specific characteristics of the dataset and the regional context to achieve optimal predictive performance.

5.3 Impact of Data Preprocessing

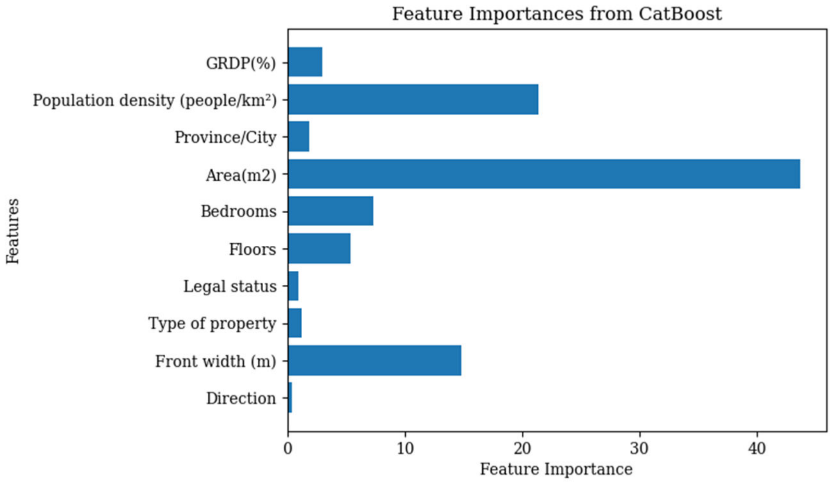
The study also investigates the impact of different data preprocessing approaches on model performance, particularly in handling the 'direction' attribute, which contains a substantial proportion of meaningless data. The comparative analysis reveals that Approach 2, which involves replacing meaningless values with random values while preserving the data distribution in the test set, consistently yields the best results across all datasets and models. The superior performance of Approach 2 can be attributed to its ability to maintain the underlying data distribution while effectively addressing the issue of meaningless data. By preserving the data distribution in the test set, Approach 2 ensures that the models are evaluated on a representative sample of the data, thereby providing a more accurate assessment of their predictive capabilities. This finding highlights the importance of carefully considering data preprocessing techniques and their potential impact on model performance.

5.4 Feature Importance Analysis

In order to gain a deeper understanding of the factors that drive house prices in the Vietnamese real estate market, we conduct a comprehensive feature importance analysis by using the best-performing models. This analysis aims to uncover the key determinants of property values and provide actionable insights for stakeholders, policymakers, and researchers.

First, we employ the CatBoost algorithm, which has superior performance on the main dataset, to identify the relative significance of each feature in predicting house prices. As illustrated in Figure 2, the CatBoost feature importance plot reveals that several attributes play a pivotal role in shaping property values. Most notably, the area of the property emerges as the dominant factor, thus highlighting the paramount influence of size on house prices. This finding aligns with fundamental market principles and confirms the common understanding that larger properties command higher values.

Figure 2 Feature Importances from CatBoost

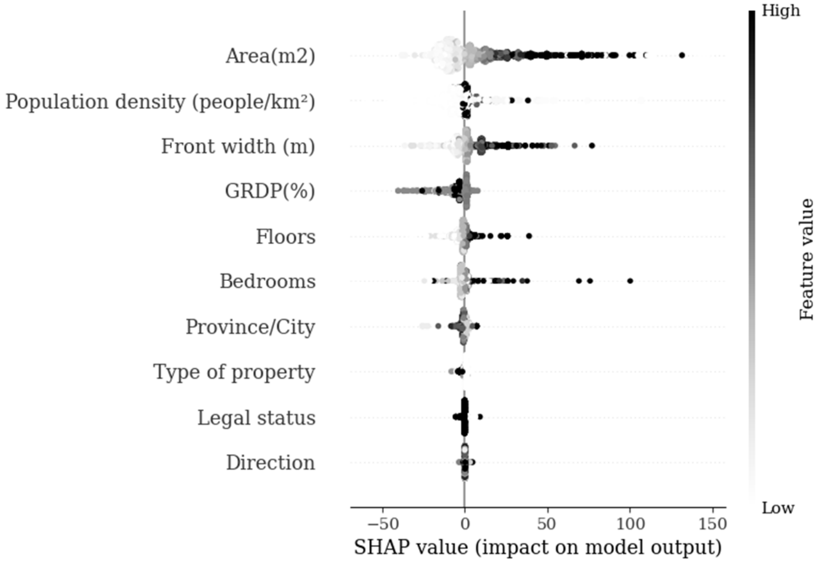


Furthermore, our analysis uncovers the substantial impact of population density on house prices. This underscores the intricate relationship between demographic characteristics and the housing market, which emphasizes the importance of considering the broader socioeconomic context when assessing property values. The prominence of population density in our feature importance analysis suggests that areas with a higher population concentration tend to experience elevated house prices, thus reflecting the increased demand and competition for housing in densely populated neighborhoods.

To further enhance the interpretability and granularity of our findings, we employ the advanced technique of SHAP. SHAP values, as depicted in Figure 3, provide a more nuanced understanding of the contribution of each feature to the predicted house prices. By quantifying the individual impact of attributes, SHAP enables us to disentangle the complex interplay of factors that influence property values.

The SHAP analysis corroborates the significance of area and population density while also highlighting the importance of additional property-specific attributes. Features such as front width and the number of bedrooms emerge as strong predictors, which emphasize the role of property dimensions and layout in determining value. These insights underscore the multifaceted nature of house price determination and the need to consider a comprehensive set of attributes when developing valuation models.

Figure 3 SHAP Analysis of Feature Importance in Predicting House Prices



Notably, the feature importance results derived from both CatBoost and SHAP align closely with domain knowledge and market expertise. This convergence of data-driven insights and established industry understanding validates the effectiveness and reliability of our ML approach. By capturing the complex dynamics of the Vietnamese real estate market, our models provide a robust foundation for data-driven decision-making and policy formulation.

The implications of our feature importance analysis extend beyond academic research. By identifying the key drivers of house prices, we empower stakeholders, including homebuyers, investors, and real estate professionals, to make more informed decisions. The insights gleaned from this analysis can guide targeted policy interventions aimed at promoting housing affordability and market stability. Moreover, the methods and findings presented here serve as a catalyst for future research by opening up new avenues for exploring the intricate relationships among property attributes, socioeconomic factors, and market dynamics.

Overall, our feature importance analysis, which leverages state-of-the-art ML techniques, sheds light on the critical determinants of house prices in Vietnam. The identification of area, population density, and property-specific attributes as key value drivers provides actionable insights for stakeholders and lays the foundation for data-driven decision-making in the real estate sector. As we continue to refine our models and incorporate additional data sources, we remain committed to advancing the understanding of the Vietnamese housing market and contributing to the development of fair, transparent, and efficient valuation practices.

5.5 Implications and Future Directions

The findings of this study have significant implications for various stakeholders in the Vietnamese real estate market. The development of accurate and reliable house price prediction models can assist homebuyers in making informed decisions, investors and developers in optimizing their strategies, and policymakers in formulating data-driven housing policies. The identification of key factors that influence house prices can guide the development of targeted initiatives to promote housing affordability and market stability.

However, the limitations of the study, such as the reliance on a single data source and the focus on a specific time period, should be acknowledged. Future research could explore the integration of multiple data sources, investigate the temporal dynamics of house prices, and delve deeper into the socioeconomic and demographic factors that influence the Vietnamese real estate market. Moreover, collaborations with domain experts from various disciplines, including urban planning, economics, and social sciences, can further enrich understanding of the complex socioeconomic and demographic factors that influence the Vietnamese housing market.

While the application of deep learning techniques is limited in this study due to the relatively small dataset, their potential for capturing complex patterns and improving predictive performance should not be overlooked. As more comprehensive datasets become available, future studies could investigate the integration of deep learning methods to further advance the field of house price prediction in Vietnam.

6. Conclusions

In conclusion, this study demonstrates the effectiveness of machine learning techniques, particularly ensemble methods, in predicting house prices in the Vietnamese real estate market. The meticulous data preprocessing approach and comparative analysis of various models and preprocessing techniques provide valuable insights for researchers and practitioners. The findings highlight the importance of addressing data complexities, selecting appropriate algorithms, and interpreting the results within the context of the specific market dynamics.

Through an extensive comparative analysis of various ML models, including state-of-the-art ensemble methods such as CatBoost, XGBoost, and random forests, we have demonstrated the superiority of these algorithms in capturing the intricate relationships between house features and prices. The application of advanced techniques like SHAP has enhanced the interpretability of our models, thus enabling stakeholders to gain a deeper understanding of the factors that influence property values.

One of the key contributions of our study lies in the meticulous handling of data complexities, exemplified by our novel approach to addressing the 'direction' attribute. By proposing and evaluating four distinct data preprocessing techniques, not only have we tackled the challenges posed by missing and inconsistent data but also set a new standard for data quality assurance in real estate price prediction research.

The feature importance analysis reveals the prominent roles of area, population density, and specific property attributes, offering a data-driven basis for more informed decision-making. These insights can support homebuyers in evaluating properties, help investors to optimize portfolios, and assist policymakers in designing strategies for sustainable urban development and housing affordability. Moreover, the proposed models form the foundation for developing automated valuation models (AVMs) in Vietnam, with potential applications in mortgage lending, property tax assessment, and digital real estate platforms.[2]

[2] **https://hathang-predicthouseprices.hf.space/**

We acknowledge the limitations of our study, such as the reliance on a single data source and focus on a specific time period. These constraints underscore the need for future research to incorporate multiple data sources, investigate the temporal dynamics of house prices, and explore the integration of deep learning techniques as more comprehensive datasets become available. Moreover, collaborations with domain experts from various disciplines, including urban planning, economics, and social sciences, can further enrich our understanding of the complex socioeconomic and demographic factors influencing the Vietnamese housing market.

References

Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., and Manjunath, B. S. (2017). 'Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery', *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 320-329.

Breiman, L. (2001). Random forests. *Machine Learning,* 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chen, K., Ding, Y., and Yu, W. (2020). 'Predicting real estate price using text mining and machine learning', *2020 International Conference on Big Data and Artificial Intelligence* (BDAI). IEEE, pp. 171-175,.

Chen, T., and Guestrin, C. (2016). 'XGBoost: A scalable tree boosting system'. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785-794. https://doi.org/10.1145/2939672.2939785

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 3146-3154.

Kok, N., Koponen, E. L., and Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. *The Journal of Portfolio Management*, 43(6), 202-211. https://doi.org/10.3905/jpm.2017.43.6.202

Ma, J., Ding, Y., Cheng, J. C., Tan, Y., Gan, V. J., and Zhang, J. (2019). Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective. *IEEE Access*, 7, 148059-148072. https://doi.org/10.1109/ACCESS.2019.2946401

Nguyen, N., and 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

Nguyen, T. A. H., Nguyen, T. V., Pham, P., and Nguyen, T. T. L. (2024). Applying machine learning in real estate prediction: The case in Vietnam. In *Proceedings of the 13th International Conference on Information Technology and Its Applications*, Vietnam, (2), 14-25. https://elib.vku.udn.vn/handle/123456789/4003

Park, B., and 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

Rahman, M. F., Murukannaiah, P., and Sharma, N. (2020). Predicting the Influence of Urban Vacant Lots on Neighborhood Property Values. *CEUR Workshop Proceedings,* 2557, 1–16. https://ceur-ws.org/Vol-2557/paper-01.pdf

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy,* 82(1), 34-55. https://www.jstor.org/stable/1830899

Savills Vietnam (2024). Viet Nam Real Estate Market Report Q4/2024. https://www.savills.com.vn/research_articles/163944/220194-0

Shahhosseini, M., Hu, G., and Pham, H. (2021). 'Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction', *Proceedings of the 54th Hawaii International Conference on System Sciences*, p. 1446. Available at: https://dr.lib.iastate.edu/server/api/core/bitstreams/d117954e-b1a6-4ed2-ad8d-f2bb8eed8e20/content

Siregar, M. U., Hardjita, P. W., Asdin, F. A., Wardani, D., Wijayanto, A., Yunitasari, Y., and Anshari, M. (2022). 'Housing price prediction using a hybrid genetic algorithm with extreme gradient boosting', *IC3INA 22: Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications,* pp. 296–300 https://doi.org/10.1145/3575882.3575939

Thi, H. V., Nguyen, P. A., Tran, T., and Le, A. N. (2025). 'Analysis and Prediction of Real Estate Prices in Hanoi Using Machine Learning', in Park, J.S., Camacho, D., Gritzalis, S. and Park J.J. (eds.) *Advances in Computer Science and Ubiquitous Computing. CSA 2024. Lecture Notes in Electrical*

Engineering,. Springer: Singapore, pp. 469–481. https://doi.org/10.1007/978-981-96-5693-6_37

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Truong, Q., Nguyen, M., Dang, H., and 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

Yacim, J. A., and Boshoff, D. G. (2018). Impact of Artificial Neural Networks Training Algorithms on Accurate Prediction of Property Values. *Journal of Real Estate Research*, 40(3), 375-418. https://doi.org/10.1080/10835547.2018.12091505

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

GitHub repository containing code and processed dataset used in this study: https://github.com/HaMThang/HousePricesPrediction.