Available Online: 27 November 2025 DOI: 10.54254/3029-0880/2025.30128

The application of machine learning in house price prediction

Zihao Chen

Huanggang Middle School, Huanggang, China

13477672699@189.cn

Abstract. In recent years, with the significant impact of housing price changes on economic and social stability, scientifically predicting housing price trends can prompt local governments to formulate policies related to housing prices, and investors can make investments based on the corresponding housing prices. At the same time, homebuyers can also make housing purchase plans according to housing prices. Traditional methods for predicting housing prices are primarily based on experience or simple statistical models, which cannot reasonably consider complex factors. This paper investigates the efficacy of machine learning methodologies in the prediction of housing market valuations. Firstly, it discusses the characteristics and value of supervised, unsupervised, and decision tree algorithms in the machine learning field when applied to housing price prediction. Secondly, it explores the reasons affecting housing prices, such as economic characteristics like national income, regional features, and the influence of supply and demand. Then, it applies a public housing price dataset to establish a decision tree for predicting housing prices. Based on basic features, it expands the feature library. It seeks suitable feature combinations to analyze housing prices better and draw conclusions on the value of machine learning methods in housing price prediction. This article summarizes successful experiences from existing application cases and studies the problems and deficiencies in the application process.

Keywords: machine learning, housing price prediction, linear regression, financial prediction, data modeling

1. Introduction

With the increasingly prominent position of the real estate industry in the modern economic system, housing price fluctuations concern residents' asset allocation and directly affect the financial market's stability and the formulation of macroeconomic policies [1]. As a special asset with both commodity and financial attributes, the formation mechanism of housing prices involves a dynamic game of multiple factors such as supply and demand, policy regulation, and regional characteristics [2]. Its instability makes it difficult for classic economic models to explain and predict it effectively, and these models often encounter problems such as weak explanatory power, poor dynamic adaptability, and prediction bias when explaining and predicting the trend of housing prices. However, machine learning's robust nonlinear modeling capabilities offer effective solutions to these challenge [3].

Therefore, housing price prediction aims to find the inherent patterns and relationships between modeling variables in historical housing price data.Traditional statistical methodologies necessitate the manual specification of model architectures contingent upon the observed phenomena, frequently proving inadequate for capturing the interaction effects inherent in high-dimensional datasets [4]. The assumptions of artificial methods on feature representation and patterns are restrictive, meaning that if the original data cannot be processed for feature representation using artificial methods, the problem or application becomes unrecognizable by machines, such as hierarchical feature processing, discriminant function methods, discrete and association analysis methods, and artificial neural networks. However, machine learning is a principle of representing features through automatic learning of patterns (i.e., generating suitable models or decision-making processes for unknown input signals or data based on the model), such as Support Vector Regression (SVR), which can map nonlinear relationships through kernel functions [5]. XGBoost and other ensemble learning methods have shown superior generalization performance in predicting second-hand housing prices by optimizing loss functions and regularization terms [4]. Hybrid models combining time series analysis with machine learning algorithms have further improved prediction accuracy [6].

This study focuses on constructing housing price prediction models within the framework of supervised learning, with a particular emphasis on exploring the influence mechanism of feature engineering on model performance. A multi-source data perspective covering spatial location, building attributes, and market supply and demand is established by systematically sorting out the key variable system affecting housing prices. The algorithm's parameters are set through cross-validation and grid search optimization algorithms. This approach provides new tools for analyzing the formation mechanism of housing prices and enables

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

the established model to provide a basis for formulating government macro-control policies [2]. It is worth noting that although machine learning has been well demonstrated in valuation, it actually faces problems such as data issues, feature collinearity, and interpretability, which all require further research and improvement by researchers and the organic combination with expert knowledge [3, 7].

2. Basic theories of machine learning

2.1. Definition and classification of machine learning (supervised learning, unsupervised learning)

Machine learning is an algorithm and technology that enables computers to improve their performance by analyzing data and discovering patterns. Its essence is a learning method that allows computers to extract patterns from experience without the need for human-written programs. By processing extensive data, its ability to judge and predict things will somewhat improve [3]. From the implementation perspective, machine learning mainly relies on statistical learning theory and optimization algorithms, and uses mathematical models to describe the potential patterns in the data.

According to the type of data involved by the learning object and the type of the learning object's goal, machine learning is mainly divided into three types. Supervised learning is the most common type of machine learning. The learning object of this type requires corresponding input-output relationships. The system learns from known sample objects, establishes the relationship between input and output, and achieves the mapping relationship between the corresponding input and output [4]. Typical supervised learning tasks include regression problems and classification problems. The former predicts continuous values, and the latter predicts discrete categories.

Unsupervised learning is used to solve data sets without specific output data. It aims to mine data through the internal structure or form features of the data. This learning method has no preset output at the beginning and achieves it through data learning [8]. Due to the absence of manually labeled data, unsupervised learning has unique advantages in processing massive raw data. Still, the learning process and results are often difficult to explain and evaluate.

2.2. Principles and characteristics of common algorithms

The commonly used algorithms in machine learning can be classified into different categories based on their principles and application scenarios. Various algorithms have their own characteristics when processing data and performing prediction tasks. Supervised learning algorithms are the most well-known due to their application in prediction tasks. Among them, linear regression is the simplest model, which predicts by establishing a linear relationship between input features and the target. It has a fast operation speed and is relatively simple and easy to interpret, but it is not suitable for fitting nonlinear relationships. Decision tree algorithms recursively divide the data into tree-like structures to handle nonlinear relationships. They are insensitive to abnormal data but prone to overfitting. Support Vector Regression (SVR) uses kernel functions to map data to high-dimensional spaces, making it easier to establish fitting functions in non-linear spaces. Moreover, SVR has strong generalization ability, but the type of kernel function and parameter selection, as well as the computational resource requirements, are relatively high [9].

Many unsupervised machine learning algorithms, such as clustering analysis algorithms, are used for data exploration and feature extraction. The K-means algorithm is commonly used. In the K-means algorithm, the data set is adjusted multiple times to obtain the desired cluster data, which is used to explore the potential structure of the data [10]. Principal Component Analysis (PCA) reduces the dimension of the data through linear transformation and can effectively remove redundant information, but it loses some interpretability [8]. Reinforcement learning algorithms learn the optimal strategy by interacting with the environment. They are suitable for sequential decision-making problems, but the training process is complex and requires a lot of trial and error [11].

Different algorithms have different features and advantages for various application scenarios. For example, in terms of operation speed, traditional algorithms such as linear regression and decision tree algorithms have a fast calculation speed. In contrast, ensemble algorithms and deep learning methods have high accuracy. Based on specific data applications and problem sizes, select the corresponding machine learning algorithms, such as choosing the decision tree model when the data volume is small and structured data is involved, and selecting the neural network model when dealing with big data, structured and unstructured data. Grad gradient-boosting algorithms often achieve better results when predicting tasks of structured data. Understanding the principles and characteristics of these algorithms helps make reasonable choices in practical applications [12].

3. The application of machine learning in house price prediction

3.1. Analysis of application fields

3.1.1. House price forecast

Predicting house prices is another critical application direction of machine learning in the financial field. Using machine learning methods, patterns can be mined from historical data on house prices to construct prediction models, providing decision support for personnel in the related real estate market. Regarding technical implementation, house price prediction is a regression task, where given feature variables are used to predict the continuous value of house prices.

Firstly, accurate house price predictions benefit multiple stakeholders. Buyers can assess price reasonableness, while financial institutions can mitigate mortgage loan risks. Governments gain support for real estate market policies, and tenants can avoid rent increases [1].

Secondly, from a practical perspective, machine learning-based housing market prediction has some technical flaws. The housing market is subject to macro-control by the government, and some external policy shocks may cause errors in the predictions of machine learning models. Due to different cities, locations, and construction years, parameters in the machine learning model may need to be reset during house price formation in some areas. In summary, people can use machine learning technology for new predictions but must make reasonable judgments based on the housing market situation [12].

Thirdly, feasibility. House price-related data often exhibits relatively obvious structural characteristics, such as house area, location, and construction year, which can be easily transformed into quantitative models. Deep machine learning can establish stronger models and expression capabilities for the nonlinear relationship between house prices and influencing factors. At the same time, machine learning models, informed by real estate transaction big data, iteratively adapt to price fluctuations, enhancing predictive accuracy through data feedback [13].

3.1.2. Construction of predictive model

In constructing the housing price prediction model, the data source is the foundation of the model's reliability. Usually, the data collected for housing price prediction includes house characteristic attribute data (such as area, age of the house, floor, decoration condition, etc.), location characteristic attribute data (such as nearby facilities, transportation conditions, school district resources, etc.) and macroeconomic attribute data (such as interest rates, inflation, GDP growth rate, etc.). The sources include relevant data from the National Bureau of Statistics on second-hand house transactions, Lianjia's second-hand house data, etc.

The decision tree model is a traditional model. This model divides the data in a tree-like form to obtain a series of rules for housing price prediction. The criterion for establishing the decision tree is to select the optimal feature splitting node based on information gain until the termination (reaching the maximum depth or the minimum sample number). The decision tree can also predict features that do not satisfy a linear relationship and is not sensitive to outliers. However, the decision tree may risk overfitting in practical applications, and methods such as pruning can be used to control its occurrence [13].

3.2. Application advantages and challenges

The application of machine learning in housing price prediction has its advantages, but it also faces some challenges. From an advantages perspective, machine learning can handle many complex nonlinear data and has stronger modeling capabilities than traditional statistical methods. Various factors influence housing prices, such as location, house area, and surrounding facilities. Machine learning algorithms can automatically learn the complex relationships between these features, improving prediction accuracy. Machine learning models have strong generalization ability and can adapt to changes in housing price data in different cities and times.

Another significant advantage is the flexibility of feature engineering. Machine learning allows for various raw data transformations and combinations, extracting more valuable features. By converting geographic coordinates into distances from the city center, the impact of location on housing prices can be better captured. At the same time, machine learning supports online learning, which can continuously update the model as new data is added, maintaining the timeliness of predictions [13].

The application of machine learning in housing price prediction also faces several challenges. Firstly, some factors change significantly, and these sudden changes can disrupt the regularity of historical data. Implementing sudden purchase restrictions policies can break the regularity of historical data; the interpretability of machine learning models is low, and they cannot provide detailed explanations of influencing factors like traditional regression models [1].

Another challenge is the problem of overfitting the model. When the amount of training data is insufficient or there are too many features, the model may perform well on the training set but have a decreased prediction effect on the test set [8]. Solving this problem requires the use of techniques such as regularization and cross-validation. The formation mechanisms of housing

prices in different regions may vary, and a model trained in one city may not directly apply to other cities. This requires researchers to adjust the model structure and parameters according to specific circumstances [10].

3.3. Existing application case studies

Many studies have adopted machine learning methods for analysis in housing price prediction and achieved specific prediction results. By summarizing these studies, it can be found that different models have differences in prediction accuracy and applicability. Although the linear regression model is simple and clear, it is often used as a preliminary estimation and cannot reflect nonlinear relationships. Compared with the linear regression model, the Support Vector Machine Regression (SVR) can adapt to more complex structures and is a generalization of the linear model.The method demonstrates robustness in high-dimensional data and superior performance in small sample learning scenarios [9].

Time series models also have significant value in housing price prediction. The AutoRegressive Integrated Moving Average (ARIMA) model is suitable for analyzing data with time trends, but its adaptability to unexpected events is poor [14]. The SARIMAX model, combined with seasonal adjustment, can better capture cyclical fluctuations and improve the stability of predictions. Hybrid models such as Empirical Mode Decomposition-Particle Swarm Optimization (EMD-PSO)-ARIMA further optimize the prediction results by decomposing the fluctuation characteristics of the original data [14].

The application of deep learning models in housing price prediction is also increasing. Neural networks can automatically extract complex features from data, but their training process requires large data support and has high computational costs [15]. Long Short-Term Memory (LSTM) networks effectively extract long-range dependencies, making them suitable for time-series data like housing prices. However, the limited interpretability of deep neural networks restricts their application in certain contexts.

Current housing price research primarily focuses on model selection, data, and feature analysis, considering regional variations in geography, economics, and macro-controls. Inconsistent housing price patterns necessitate careful city selection for model construction. Machine learning has been applied to housing price prediction, but data noise and model generalization issues require further investigation and improvement.

4. Conclusion

The research on machine learning for predicting house prices provides a paradigm for us to solve the problem of house prices efficiently and accurately. At the same time, the advantages of machine learning algorithms in dealing with more complex data relationships and nonlinear features have been identified based on this research. Moreover, the regression algorithm in supervised learning algorithms is more adept at discovering the relationship between house prices and influencing feature data. Additionally, the integration learning technology (eXtreme Gradient Boosting (XGBoost)) can be used to optimize the algorithm model further to improve the stability and accuracy of the model.

Limitations of the model include the impact of dynamic economic, policy, and location factors on house price predictions, the difficulty in obtaining comprehensive historical data, and interpretability issues. Future research should integrate Geographic Information System (GIS) and geographic economics to incorporate spatial characteristics, and utilize deep learning for improved house quality assessment. Despite the broad prospects of machine learning in house price prediction, further investigation is needed to address data quality, model optimization, and application scenarios.

References

[1] Chen, X., Chen, Y., & Wang, Z. (2024). Research on the influencing factors of housing price differentiation between cities. *Economic Theory and Economic Management*, 44(2), 49.

[2] Zhang, J., & Guo, H. (2024). A review and outlook on batch appraisal of real estate driven by data. *Journal of Xihua University (Philosophy and Social Sciences Edition), 43*(3), 13-27.

[3] Zhou, L. J., & Zhao, M. Y. (2022). Analysis of housing price prediction based on several machine learning models. *National Circulation Economy*, (6), 111–116.

[4] Wang, Y. (2024). Research on housing price prediction based on machine learning. *Finance, 14*, 1552.

[5] Mo, P., & Lin, G. (2024). Research on second-hand housing transaction price prediction based on the XGBoost model. *Advances in Applied Mathematics, 13*, 4417.

[6] Zhu, Y., Zheng, L., & Zhu, W. (2024). Construction of urban housing price prediction model from a spatiotemporal perspective: A case study of Beijing. *Statistics and Application, 13*, 276.

[7] Wang, Y., & Chen, M. (2025). Research on housing rent prediction in Chengdu based on machine learning. *Computer Science and Application, 15*, 138.

[8] Li, J., Li, Z., & Chen, J. (2022). Forecasting tourism demand in Beijing based on online search data. *Hans Journal of Data Mining, 12*, 133.

[9] Qin, Q., Chen, Y., & Xiang, Z. (2022). An improved SVR-SARIMAX hybrid sales prediction model for excavators. *Construction Machinery & Equipment, 53*(9), 10-15.
[10] Li, X., & Ding, Z. (2024). Hybrid strategy improved Harris hawk optimization algorithm. *Journal of Yunnan University (Natural Science Edition), 47*(1), 60-69.
[11] Zhang, X., & Hu, S. (2024). Comparison of machine learning-based monetary policy models and traditional rule-based models. *World Economic Research, 13*, 222.
[12] Zheng, Y. (2024). Real estate price analysis and prediction based on GM(1, 1) model: A case study of Shanghai. *Modern Management, 14*, 1359.
[13] Wang, Y. (2024). Research on housing price prediction based on machine learning. *Finance, 14*, 1552.
[14] Shang, J., Li, W., & Xi, L. (2024). Agricultural product price prediction based on EMD-PSO-ARIMA model. *Hubei Agricultural Sciences, 63*(8), 121.
[15] Wang, X., Chen, X., & Lin, X. (2024). Design and implementation of a housing price prediction system based on deep learning. *Computer and Digital Engineering, 52*(9), 2572-2567.