

The Current State and Challenges of Housing Price Prediction Models

Puxi Han^{1*}

¹ Liaoning Experimental Middle School, Shenyang, China

Email Address

pxhanchao@outlook.com (Puxi Han)

*Correspondence: pxhanchao@outlook.com

Received: 12 July 2025; **Accepted:** 28 September 2025; **Published:** 20 November 2025

Abstract:

With the development of the global economy and the acceleration of urbanization, the real estate industry has become an important pillar of the economy in many countries. Accurate prediction of housing prices is conducive to promoting reasonable investment decisions and resource allocation. For home buyers, understanding the future trend of housing prices is of vital importance. It not only affects their timing for purchasing a house, but also relates to their financial security and investment returns. Effective prediction models can help home buyers make wise decisions. Currently, many researchers have conducted a lot of research on related issues. This article reviews the current mainstream prediction models, including decision trees, random forests, and artificial neural networks, and discusses the challenges in terms of accuracy and data quality of the current models, as well as the future research directions and improvement measures. The research findings show that with the rapid development of machine learning and artificial intelligence technologies, housing price prediction models have made significant progress. The mainstream prediction models include decision trees, random forests, and artificial neural networks, etc. These models perform well in handling complex data and capturing nonlinear relationships, and can provide strong support for housing price prediction.

Keywords:

Decision Tree, Random Forest, Artificial Neural Network, Housing Price Prediction

1. Introduction

House price prediction is a core issue in real estate research. With the acceleration of urbanization and the transformation of the economic structure, its importance has become increasingly prominent. The real estate market not only affects the financial situation of individuals and families, but also plays an important role in the stability and development of the national economy. Understanding the dynamic changes of house prices can help buyers, investors, and policy makers make more rational and scientific decisions.

Traditional methods for predicting house prices mainly rely on statistical models, such as linear regression and time series analysis. These methods are usually based on historical data and are simple and easy to use, but they often show limitations when

facing market complexity and non-linear characteristics. In recent years, with the development of machine learning and deep learning technologies, researchers have begun to explore more complex and flexible prediction models. Wu et al. established a house price prediction model using SVM, and for the parameter selection problem of the SVM model, they used genetic algorithms for optimization. Through empirical analysis, it was proved that this model has a good prediction effect [1]. Zaheer et al. proposed a hybrid deep learning prediction model based on LSTM for stock data prediction [2]. These emerging methods can handle large amounts of feature data, capture the underlying patterns behind house price fluctuations, and thereby improve the accuracy of predictions.

This paper reviews the current research status of house price prediction models, analyzes the advantages and disadvantages of existing models, and discusses the challenges they face. This paper aims to provide useful references for future research and promote the development of the house price prediction field.

2. Research Progress of Housing Price Prediction Models

2.1. Decision Tree

A decision tree is a prediction model based on a tree structure. It makes decisions by dividing the data set into multiple subsets. Each node represents a feature, each branch represents the value of a feature, and each leaf node represents the final prediction result. The key to building a decision tree lies in how to select the optimal feature for division. Common selection criteria include information gain, Gini coefficient, and gain ratio. Information gain measures the degree to which the uncertainty of the system decreases after dividing the data set based on a certain feature; the Gini coefficient is used to measure the impurity of the data, and the smaller the value, the higher the purity of the data.

The commonly used decision tree classification algorithms include 1D3, C4.5, CART, and ID3. These algorithms select the attribute with the highest information gain and perform node splitting to construct the decision tree. It is believed that ID3 can only handle discrete data and that the training model is prone to overfitting. The C4.5 algorithm is an enhancement of the ID3 algorithm, introducing a pruning step to prevent the overfitting phenomenon of the ID3 algorithm. The execution process involves setting a threshold. The CART algorithm mainly performs bidirectional recursive classification on the sample data based on the criterion of minimum impurity, dividing the current sample set into two sub-sample sets. Therefore, each non-leaf node of the generated decision tree has only two branches. That is to say, the CART tree is a binary tree, while ID3 and C4.5 can be multi-branch trees [3].

The advantage of decision trees lies in their strong interpretability, ease of understanding and visualization, and suitability for handling non-linear relationships. However, decision trees are prone to being affected by noise and overfitting, especially when the data volume is small or there are many features.

In housing price prediction, decision trees can effectively identify the key factors influencing housing prices, such as location, house area, number of rooms, etc. By analyzing the impact of different features, decision trees can help users understand the logic behind the price determination. This study used housing price data from a certain city for instance analysis. The study designed a housing price prediction model based on decision tree and used grid search to obtain the optimal parameters. The

results showed that the number of houses was the most important factor affecting housing prices, followed by local population quality, geographical location, education level and crime rate [4].

The research on housing price prediction based on the decision tree algorithm has significant importance and application prospects. It can provide housing price prediction information for participants in the real estate market, guiding market investment and decision-making. At the same time, it can also provide important reference information for government departments and the public sector, guiding regional economic and urban planning development. The research on housing price prediction based on the decision tree algorithm has wide applicability and practical value in application and is an important direction in the future field of real estate research. In summary, the research on housing price prediction based on the decision tree algorithm provides scientific solutions to new problems brought about by social development. In the future, we will continue to deepen research in this field, continuously improve the prediction accuracy and adaptability of the model, and provide more scientific and reliable decision support for the real estate market and urban planning and development.

2.2. Random Forest

In recent years, random forest, as one of the machine learning methods, has received increasing attention. Random forest is a statistical learning theory that uses the bootstrap sampling method to extract multiple samples from the original dataset, builds decision trees for each bootstrap sample, and combines multiple decision trees to obtain the final prediction result [5]. A large number of theoretical and practical studies have shown that random forest has a very high prediction accuracy, is tolerant of outliers and noise, and is less likely to overfit. Random forest, as a non-linear modeling tool, is currently one of the most popular and cutting-edge research fields in data mining and bioinformatics.

Random forest is composed of decision trees and performs classification and prediction by combining multiple decision tree classifiers. Its working mechanism is roughly as follows: First, a series of training sets are selected through bootstrap sampling. On each training set, features are randomly selected and the data set is appropriately divided using indicators such as the Gini index to generate a series of unpruned decision trees. Finally, the optimal classification is determined through voting to make a prediction. Figure 1 shows the model of random forest.

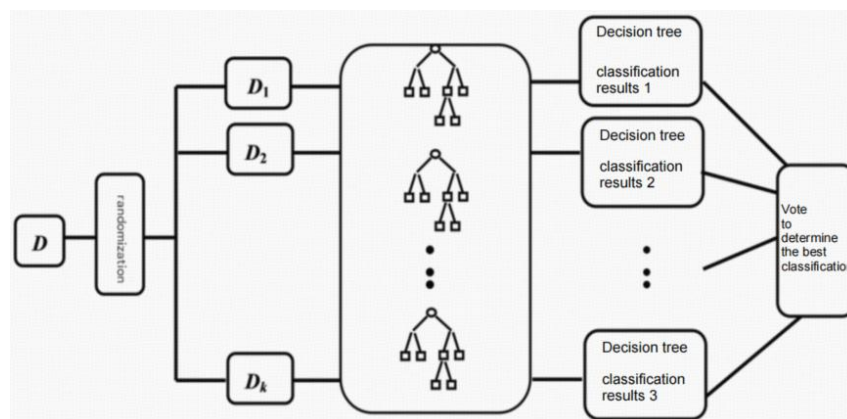


Figure 1. Random Forest model [6].

Compared to a single decision tree, random forests perform better when dealing with complex data and can effectively handle high-dimensional data and missing values. In housing price prediction, random forests can comprehensively consider the interactions of multiple features, thereby improving the accuracy of the prediction. Additionally, random forests provide importance scores that can help identify the features that have the greatest impact on housing prices, providing guidance for subsequent feature engineering. This study established random forest models for multi-category variables to conduct research on the factors influencing housing prices and make housing price predictions. The research shows that the prediction effect of random forests is the best [7].

Random forests have higher accuracy and can effectively handle large databases with many features. They perform well in data with many features. It can quickly and effectively manage a large number of input variables and balance errors. However, random forests have difficulties in processing multi-dimensional and multi-valued attributes. For data with different values, attributes with more value divisions have a greater impact on random forests.

2.3. Artificial Neural Network

Artificial Neural Network (ANN) is a computational model that mimics the structure of human brain neurons and is suitable for handling highly complex and nonlinear data relationships. ANN can automatically extract features and learn through the interconnection of multiple layers of neurons. Deep learning is an important branch of ANN, featuring a deeper network structure and stronger learning capabilities. Figure 2 shows the general model of ANN, which consists of three different layers: the input layer, the hidden layer, and the output layer.

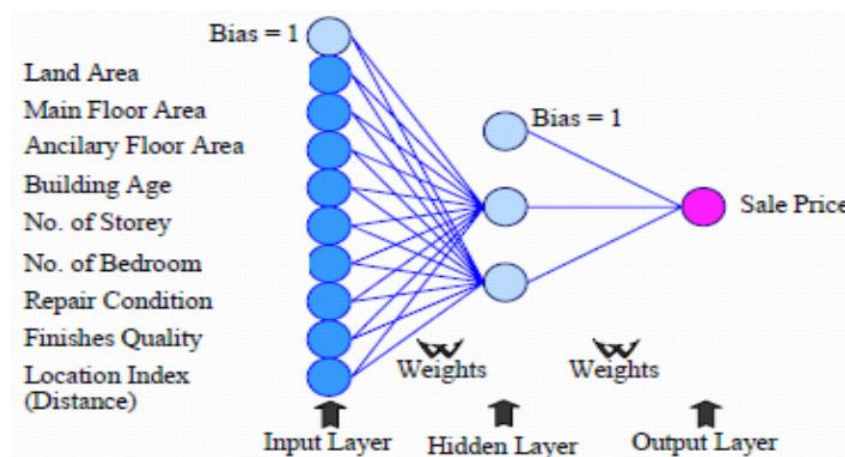


Figure 2. ANN property valuation model [7].

In housing price prediction, artificial neural networks can handle large-scale feature data and capture nonlinear relationships. By utilizing historical housing prices, economic indicators, and other related data, ANN can achieve higher prediction accuracy. However, ANN usually requires a large amount of training data, and the interpretability of the model is relatively low, which may pose a barrier for users to understand the prediction results.

In conclusion, decision trees, random forests, and artificial neural networks each have their own unique advantages and disadvantages, and are suitable for different prediction requirements and data characteristics. In practical applications, choosing

the appropriate model will directly affect the effectiveness of housing price predictions.

3. Current Challenges

3.1. Data Quality

The accuracy of housing price predictions largely depends on the quality of the input data. Missing values, outliers, and noisy data can all have a negative impact on the performance of the model. For example, missing housing feature information may cause the model to fail to fully understand the data characteristics, thereby reducing the accuracy of the prediction. Additionally, outliers may be caused by data entry errors or market events, and if not handled, will seriously distort the model's output. Therefore, ensuring the completeness and high quality of the data is the primary challenge in housing price predictions [8].

3.2. Interpretability

Many advanced prediction models, especially deep learning models, although having advantages in accuracy, have the “black box” characteristic that greatly reduces the interpretability of the results. For home buyers and policy makers, understanding how the model arrives at a certain prediction result is crucial. Lack of interpretability may lead to a decrease in users’ trust in the model, affecting its practical application. Therefore, how to improve the interpretability of the model while ensuring prediction accuracy has become an urgent problem to be solved.

3.3. Data Privacy and Security

With the increasingly strict data privacy regulations, obtaining high-quality housing price-related data has become more difficult. Many data sources may be restricted by privacy protection policies, resulting in scarce available data. This not only affects the training and validation of the model but also may limit the wide application of the model. Therefore, while following the laws and regulations, how to effectively acquire and utilize data is another important challenge in the field of housing price prediction [9].

4. Future Research Directions

First, with the continuous progress of machine learning and deep learning technologies, future research can explore the application of more innovative technologies. For example, transfer learning can help models share knowledge among different cities or regions, thereby improving the prediction ability in situations with scarce data [10]. In addition, reinforcement learning can also be applied to dynamic market environments, enabling the model to continuously optimize the prediction strategy based on real-time feedback.

Second, with the strictening of data privacy regulations, future research needs to focus on how to obtain and use housing price-related data while protecting user privacy. Emerging technologies such as differential privacy and federated learning can be explored to effectively utilize the data for model training without disclosing user information. This will provide a guarantee for the sustainable development of the housing price prediction field.

Third, while maintaining high prediction accuracy, improving the interpretability of the model will be an important direction for future research. Exploring interpretable machine learning methods, such as SHAP values and LIME (Local Explainable Model Dependence) techniques, can help users understand the decision-making process of the model. Enhancing interpretability not only increases users' trust in the model but also provides transparent data support for policy-making.

5. Conclusions

This paper summarizes the current status and challenges of housing price prediction models, emphasizing the importance of data quality, model interpretability, and data privacy security. Future research should focus on combining new technologies to improve the accuracy and practicality of predictions.

However, this article acknowledges its limitations. Specifically, it relies primarily on a literature review and lacks empirical data, which restricts the depth and applicability of the findings. The absence of empirical validation means that the theoretical insights presented may not fully align with real-world scenarios, potentially limiting the practical value of the research.

To address these limitations, future studies should incorporate empirical data through in-depth case analyses. By combining actual housing market data with theoretical models, researchers can verify the consistency between theory and practice, thereby enhancing the credibility and applicability of their findings. Empirical studies can provide concrete evidence of model performance, identify potential biases, and offer actionable insights for stakeholders in the real estate market.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

- [1] Gu, J.; Zhu, M.; Jiang, L. Housing Price Forecasting Based on Genetic Algorithm and Support Vector Machine. *Expert Systems with Applications*, 2011, 38, 3383-3386. DOI: <https://doi.org/10.1016/j.eswa.2010.08.123>.
- [2] Zaheer, S.; Anjum, N.; Hussain, S.; Algarni, A.D.; Iqbal, J.; Bourouis, S. et al. (2023). A Multi Parameter Forecasting for Stock Time Series Data Using LSTM and Deep Learning Model. *Mathematics*, 11, Article 590. <https://doi.org/10.3390/math11030590>.
- [3] Xiang, J.Y.; Wang, Z.H.; Deng, Y.Y. A review of machine learning classification based on the random forest algorithm. *Artificial Intelligence and Robotics Research*, 2024, 13, 143.
- [4] Zhang, Z. Decision trees for objective house price prediction. In 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 280-283). IEEE. 2021.

- [5] James, G. An Introduction to Statistical Learning with Applications in R. University of Southern California, 2014; pp. 303-324.
- [6] Li, X.; Guo, X.; Wang, C. Analysis of Beijing second-hand house price based on random forest. *In Data Mining*, 2017, 7(2), 9.
- [7] Sing, T.F., Yang, J.; Yu, S.M. Decision tree and boosting techniques in artificial intelligence based automated valuation models (AI-AVM). 2020.
- [8] Mathotaarachchi, K.V.; Hasan, R.; Mahmood, S. Advanced machine learning techniques for predictive modeling of property prices. *Information*, 2024, 15(6), 295.
- [9] Singh, N.; Birla, S.; Ansari, M.D.; Shukla, N.K. Intelligent Techniques for Predictive Data Analytics. John Wiley & Sons. 2024.
- [10] Sheng, Y.; Arbabi, H.; Ward, W.O.; Mayfield, M. Learning from other cities: Transfer learning based multimodal residential energy prediction for cities with limited existing data. *Energy and Buildings*, 2024, 338, 115723.



© 2025 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)