Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.20931

Review of Housing Price Forecasting Methods Based on Machine Learning and Deep Learning

Chenyu Li[1,a,*]

[1]*Macau University of Science and Technology, Macau, China*
a. 1179351251@qq.com
**corresponding author*

Abstract: Housing price forecasting holds profound significance in the development, investment, and policy formulation within the real estate market. Precise forecasts not only assist real estate developers in evaluating the potential value of projects but also furnish investors with a scientific basis to optimize their investment decisions. At the same time, home buyers can also gauge the price trend and appreciation potential of properties based on the forecast results. Moreover, governments and financial institutions rely on reliable housing price forecast results to formulate reasonable market supervision policies and control loan risks. This article discusses the status of house price forecasting in society in recent years and the latest research progress in applying machine learning and deep learning. Notably, it zeroes in on the role that the fusion of numerical features, textual descriptions, and image features plays in predicting housing prices. Through a methodical systematic of the existing research, this paper analyzes the characteristics, advantages and disadvantages of different models and their application scenarios. The results show that the fusion of multi-source data can significantly improve the accuracy of model prediction, but it also faces challenges in finding high-quality data and data processing and computing resources. Future studies should further optimize the feature extraction method to enhance the interpretability and adaptability of the model.

Keywords: house price prediction, machine learning, deep learning, data fusion, feature extraction

1. Introduction

Traditional housing price forecasting methods, epitomized by the hedonic pricing model, customarily assume a linear relationship among feature variables. Despite the fact that this model affords high interpretability, it encounters difficulties when dealing with the non-linear elements that impact housing prices. Especially when faced with a substantial quantity of intricate features, its shortcomings become conspicuously apparent [1]. In tandem with the rapid escalation of data scale and computing prowess, machine learning and deep learning technologies have introduced novel vistas to the realm of housing price forecasting. In recent years, researchers have embarked on exploring the means by which these techniques can be harnessed to process multifarious data forms, namely numerical, textual, and image features, all with the objective of augmenting prediction accuracy [2]. Significantly, multi-source data fusion has progressively crystallized as a pivotal trend in housing price forecasting research. By amalgamating fundamental numerical data like square

DOI: 10.54254/2755-2721/2025.20931

footage and the number of bedrooms, textual portrayals such as location and surrounding amenities, and image traits, for example appearance, and decoration quality, investigations have evinced that this methodology substantially augments the predictive efficacy of the model. The present study is dedicated to methodically encapsulating the research advancements within the housing price forecasting domain. It painstakingly dissects the characteristics, merits, and demerits of diverse forecasting methodologies, in addition to their pertinent application scenarios. Moreover, it identifies the limitations and hurdles extant in current research. Through juxtaposing the performance of traditional statistical methods, machine learning methods, and deep learning methods, this study probes into the potential of multi-source data fusion and proffers conceivable future research trajectories, thereby proffering invaluable references for researchers and practitioners engaged in related fields.

2. Evolution of housing price forecasting methods and exploration of multi-source data fusion

2.1. Forecast the development history of housing prices

House price forecasting methodologies have undergone a gradual evolution from traditional statistical methods to modern machine learning and deep learning techniques. Initially, housing price forecasts predominantly relied on statistics-based models, such as the Hedonic Pricing Model and linear regression. These traditional methods make price estimates by analyzing a home's structural characteristics, geographic location, and neighborhood attributes. The Herdunik model is a classical method that realizes the prediction by establishing a linear relationship between the various characteristic attributes of a house and the price [1]. These methods have good interpretability, and the model parameters are easy to understand, but there are significant shortcomings in dealing with complex nonlinear relations and high-dimensional characteristics, and it is difficult to effectively deal with various influencing factors and market fluctuations in the real estate market.

With the advancement of data scale and computing capabilities, machine learning methods have gradually been introduced into the field of housing price forecasting. Compared with traditional methods, machine learning methods such as Random Forest and limit gradient Boost (XGBoost) and other ensemble learning models have stronger nonlinear processing ability and enhanced generalization performance [3]. These methods can effectively capture complex interactions in data and significantly improve the accuracy of predictions. However, machine learning models require a lot of computational resources and rich data sets in the training process, especially in the case of a large number of features, and the computational cost and complexity increase significantly [4]. In addition, although these models have high predictive accuracy, they lack interpretation compared with traditional methods.

Deep learning models, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory Network (LSTM), have found application in housing price prediction tasks. The foremost advantage of deep learning lies in its ability to process multi-source data, especially in feature extraction of image, text, and time series data. For instance, CNN is able to extract visual features such as decoration quality and structure of houses by analyzing exterior images of houses, while LSTM is able to process time series data of house prices and capture dynamic trends of price changes. The introduction of deep learning methods has further improved the accuracy of house price prediction, but its high dependence on large-scale data and computing resources, as well as the "black box" of the model, makes the transparency of interpretation and application become a direction of optimization.

In summary, the traditional method has good interpretability and easy to understand model parameters, but it is difficult to deal with multiple influencing factors and fluctuations of the real

DOI: 10.54254/2755-2721/2025.20931

estate market when dealing with complex relationships and high-dimensional characteristics. Machine learning models necessitate substantial computing resources and rich data sets during training, with costs escalating when features abound. Although their prediction accuracy is high, their interpretability pales in comparison to that of traditional methods. Deep learning methods improve the prediction accuracy of housing prices but rely on large-scale data and computing resources, and there is a "black box", explanation and application transparency need to be optimized.

2.2. Multi-source data fusion and the latest methods

In recent years, researchers have begun to explore ways to improve the accuracy of housing price forecasting through multi-source data fusion. This method improves the forecasting performance by integrating statistical features like basic house information, textual features such as house description, and spatial features like nearby points of interest (POIs) [5]. By fusing multi-source data, the accuracy of the model can be effectively improved, but the complexity of data acquisition and processing also increases significantly. The model combining self-attention mechanism and heterogeneous data has further enriched the methods of housing price forecasting. The proposed model combines heterogeneous data, including satellite maps, public facility data, and so on, and a joint self-attention mechanism to improve prediction accuracy. By combining multi-source data and using the self-attention mechanism to capture the complex interrelationships among features, the method is more efficient and accurate in aggregating and utilizing information. The advantage of this method is that it can automatically focus on the features that have the greatest impact on housing prices through the attention mechanism, thus enhancing the adaptability of the model to different features and overcoming the limitations of traditional methods in feature weights [6].

3. Multi-dimensional comparative analysis of different housing price forecasting methods

3.1. Applicable scenarios of different housing price forecasting methods

The traditional method is suitable for housing price forecasting tasks with small data amount and relatively simple feature relationship, such as the housing price assessment of some specific communities. Machine learning methods are suitable for larger data sets, especially when there are complex interaction relationships between features, and machine learning methods are better able to capture these relationships. Deep learning methods are more suitable for the fusion of multi-source data, such as the simultaneous use of numerical, image and text features to predict housing prices [6]. These methods perform well when dealing with high-dimensional data and complex non-linear relationships, but they require sufficient data and computational resources to support the training of the model.

3.2. Comparison of three different prediction methods from traditional methods, machine learning and deep learning

Traditional statistical methods, such as the Hedonic pricing model and linear regression, have been widely employed in housing price forecasting. However, they typically presume a linear relationship among features. Consequently, when confronted with nonlinear and high-dimensional data, their performance leaves much to be desired [1]. In contrast, machine learning methods like Random Forest and XGBoost possess stronger nonlinear processing capabilities and are able to significantly improve forecasting accuracy [3]. Deep learning methods, including CNN and LSTM, can further improve the accuracy of predictions by processing image, text, and time series data [2]. Nevertheless, the performance of deep learning models is highly dependent on the quality and quantity of data, and is

DOI: 10.54254/2755-2721/2025.20931

prone to overfitting problems during training, especially when the data is insufficient, the performance is inferior to that of machine learning methods.

Traditional methods generally have relatively modest data requirements and often rely on structured numerical features. When dealing with complex, high-dimensional data, machine learning and deep learning methods necessitate more extensive training data [7]. Although Random Forest and XGBoost can effectively process high-dimensional data, they require a large amount of data to ensure the generalization performance of the model. Deep learning methods are especially reliant on large-scale data. For example, CNN requires a large amount of annotated image data in image feature extraction in order to extract effective features. Text embedding methods such as Word2Vec and BERT also require enough text data to train meaningful embedding vectors.

Regarding computational resources and complexity, the computational overhead of traditional methods is relatively small, and it is suitable for application scenarios with limited resources. Although machine learning methods such as Random Forest and XGBoost offer higher prediction accuracy, their model complexity also rises significantly. Especially when the number of features is large, the computational burden becomes considerable [3]. Deep learning methods rely more on powerful computing resources, such as CNN and LSTM, which require higher computing power and longer training time in the training process, which may be a bottleneck for ordinary researchers or small and medium-sized enterprises. It is worth noting that both convolutional operations and backpropagation processes in deep learning models require efficient parallel computing resources; otherwise, the training time may be extremely long [3].

In terms of model interpretability, traditional statistical methods such as linear regression and the Herdnick pricing model have high transparency, and their parameters have clear economic significance. For example, each parameter coefficient in the Herdnik model can be interpreted as the marginal effect of a certain feature on housing prices. However, the interpretability of machine learning methods and deep learning methods is relatively poor, especially for deep learning models. Although CNN has excellent performance in image feature extraction, its internal convolution kernel is difficult to intuitively interpret the predicted results [8]. In the processing of text features, the complex structure of BERT and other models also makes it difficult to understand the specific contribution of each feature to the final prediction.

Traditional methods are suitable for housing price forecasting tasks with small data volumes and relatively simple feature relationships, such as housing price assessment in some specific communities. Machine learning methods are suitable for larger data sets, especially when there are complex interaction relationships between features, and machine learning methods are better able to capture these relationships. Deep learning methods are more suitable for the fusion of multi-source data, such as the simultaneous use of numerical, image, and text features to predict housing prices [9]. These methods perform well when dealing with high-dimensional data and complex non-linear relationships, but they require sufficient data and computational resources to support the training of the model.

3.3. Summary

In general, different methods have their advantages and disadvantages in house price forecasting. The traditional method has good interpretability and low computational resource requirements, but it has shortcomings in precision and processing complex data. Machine learning methods have significant advantages in capturing nonlinear relationships and improving prediction accuracy but have higher computing resource requirements; deep learning is best at handling multi-source data and high-dimensional complex features, but its dependence on data and computing resources also limits the breadth of its application.

DOI: 10.54254/2755-2721/2025.20931

4. Limitations and challenges of existing research

Multi-source data fusion holds the promise of enhancing the accuracy of model predictions. Nevertheless, it is accompanied by several thorny issues. Firstly, the acquisition of such diverse data incurs high costs, and the quality of the obtained data is often uneven. This heterogeneity renders the tasks of data cleaning and preprocessing extremely challenging. Consider text description data sourced from disparate origins; they typically display variations in writing style and informational content. Consequently, unified preprocessing procedures, such as text normalization and the elimination of stop words, are essential. Failure to implement these steps would lead to inconsistent quality in text embedding, as documented in [10]. Secondly, deep learning models have large requirements on computing resources and are difficult to apply in resource-limited environments. For example, when training convolutional neural networks (CNNs), the computational complexity of convolutional layers is $O(n^2 \cdot k^2 \cdot d)$. As the input size is large, the computational complexity increases sharply, the training time is extended, and the computing power is high. The amount of computation increases exponentially with the increase of input [11]. Furthermore, the "black box" nature inherent to deep learning models makes it difficult in result interpretation. For example, although the gating unit of the LSTM model can control the flow of information, the complexity makes it difficult to explain the specific contribution of each input. Although the self-attention mechanism can provide the importance of clues of features through weights, the meaning of weights is not intuitive in multiple data types. This limits its use in sensitive application scenarios, such as financial scenarios. Finally, existing deep learning and machine learning models are susceptible to data bias and lack generalization ability, and their performance will be significantly degraded when the data distribution of training and actual application scenarios is inconsistent. For example, the housing price forecasting model may have poor prediction accuracy in other cities after training based on the data of a specific city. This requires more cross-regional data and more effective domain adaptation technology to improve the robustness and adaptability of the model [12].

5. Conclusion

Centering around machine learning and deep learning technologies, this paper delves into the issue of housing price prediction, with particular emphasis on the crucial role that the fusion of multi-source data, encompassing numerical, textual, and image features, plays in enhancing prediction accuracy. It systematically summarizes the research progress within the domain of housing price forecasting and analyzes the advantages and disadvantages of different methods. The main conclusions are as follows: The traditional method has good interpretability and computational efficiency, but it is insufficient in dealing with complex nonlinear relations and high dimensional features. Machine learning methods, such as random Forest and XGBoost, are able to capture interactions between complex features but require large amounts of data and computational resources. Deep learning approaches (like CNN and LSTM) further improve prediction accuracy through multi-source data fusion, but the model's dependence on data size and computational resources is strong, while the "black box" limits its interpretation and application transparency.

However, this paper has some shortcomings in the following aspects. Firstly, the discussion on the evaluation of the model is limited, focusing more on the prediction accuracy and lacking a comprehensive analysis of practical application scenarios. Secondly, the exploration of the challenges related to feature selection and computational efficiency is relatively simple and needs to be further explored. Future research needs to improve on these aspects. In light of the aforementioned shortcomings, the author puts forward the following future research directions. To begin with, the interpretability of the model should be enhanced, for example, through visualization techniques (such as LIME and SHAP) and the introduction of domain knowledge to improve the transparency and user

DOI: 10.54254/2755-2721/2025.20931

trust of the model. Secondly, the feature fusion approach needs to be optimized to reduce computational complexity and resource dependence by developing lightweight fusion architectures. In addition, improving the generalization ability of the model through domain adaptation and small sample learning techniques can reduce the dependence on large-scale labeled data. To sum up, machine learning and deep learning methods show great potential in housing price forecasting. Future research should focus on optimizing feature fusion methods and enhancing model interpretability in order to better adapt to practical application scenarios and promote the further development of real estate market forecasting models.

References

[1] *Liu, J. G., & Wu, W. P. (2008, Jul 25-27). FNN Model of House Price Prediction Based on Hedonic Price Theory. [Iccse 2008: Proceedings of the third international conference on computer science & education: Advanced computer technology, new education]. 3rd International Conference on Computer Science and Education, Kaifeng, PEOPLES R CHINA.*

[2] *Shen, H., Li, L., Zhu, H. H., & Li, F. (2022). A Pricing Model for Urban Rental Housing Based on Convolutional Neural Networks and Spatial Density: A Case Study of Wuhan, China. Isprs International Journal of Geo-Information, 11(1), Article 53. [https://doi.org/10.3390/ijgi11010053] (https://doi.org/10.3390/ijgi11010053)*

[3] *Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2021, Jul 09-11). House Price Prediction using Random Forest Machine Learning Technique.Procedia Computer Science [8th international conference on information technology and quantitative management (itqm 2020 & 2021): Developing global digital economy after covid-19]. 8th International Conference on Information Technology and Quantitative Management (ITQM) - Developing Global Digital Economy after COVID-19, Chengdu, PEOPLES R CHINA.*

[4] *Zhou, X. L., Tong, W. T., & Li, D. Y. (2019). Modeling Housing Rent in the Atlanta Metropolitan Area Using Textual Information and Deep Learning. Isprs International Journal of Geo-Information, 8(8), Article 349. [https://doi.org/10.3390/ijgi8080349] (https://doi.org/10.3390/ijgi8080349)*

[5] *Jiang, L., Li, Y. H., Luo, N., Wang, J. A., & Ning, Q. (2022, Nov 28-Dec 01). A Multi-Source Information Learning Framework for Airbnb Price Prediction. International Conference on Data Mining Workshops [2022 ieee international conference on data mining workshops, icdmw]. 22nd IEEE International Conference on Data Mining (ICDM), Orlando, FL.*

[6] *Wang, P. Y., Chen, C. T., Su, J. W., Wang, T. Y., & Huang, S. H. (2021). Deep Learning Model for House Price Prediction Using Heterogeneous Data Analysis Along With Joint Self-Attention Mechanism. Ieee Access, 9, 55244-55259. [https://doi.org/10.1109/access.2021.3071306] (https://doi.org/10.1109/access.2021.3071306)*

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

[8] *Zhang, H. X., Li, Y. S., & Branco, P. (2024). Describe the house and I will tell you the price: House price prediction with textual description data. Natural Language Engineering, 30(4), 661-695, The Article Pii s1351324923000360. https://doi.org/10.1017/s1351324923000360*

[9] *Wang, F., Zou, Y., Zhang, H. Y., Shi, H. D., & Ieee. (2019, Oct 19-20). House Price Prediction Approach based on Deep Learning and ARIMA Model. [Proceedings of 2019 ieee 7th international conference on computer science and network technology (iccsnt 2019)]. 7th IEEE International Conference on Computer Science and Network Technology (ICCSNT), Dalian, PEOPLES R CHINA.*

[10] *Abdul-Rahman, S., Mutalib, S., Zulkifley, N. H., & Ibrahim, I. (2021). Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur. International Journal of Advanced Computer Science and Applications, 12(12), 736-745. <Go to ISI>://WOS:000738710600091*

[11] *Zhan, C. J., Wu, Z. Q., Liu, Y. L., Xie, Z. F., Chen, W. L., & Ieee. (2020, Jul 21-23). Housing prices prediction with deep learning: an application for the real estate market in Taiwan.IEEE International Conference on Industrial Informatics INDIN [2020 ieee 18th international conference on industrial informatics (indin), vol 1]. 18th IEEE International Conference on Industrial Informatics (INDIN), Electr Network.*

[12] *Fourkiotis, K. P., & Tsadiras, A. (2023, Jun 14-17). Comparing Machine Learning Techniques for House Price Prediction.IFIP Advances in Information and Communication Technology [Artificial intelligence applications and innovations, aiai 2023, pt ii]. 19th International Conference on Artificial Intelligence Applications and Innovations (AIAI), Leon, SPAIN.*