

A Review of House Price Regression Prediction Methods

Xiaofei Cai

*Faculty of Commerce, Hong Kong Shue Yan University, Hong Kong, China
cxiaofei8681@163.com*

Abstract. Real estate is an important sector of the national economy. Fluctuations in housing prices affect residents' lives, financial security and the overall economy. Accurate prediction of housing prices can provide support for market entities' decision-making and government housing regulation. With the development of econometrics and machine learning, housing price regression prediction has gradually evolved from traditional characteristic price models to ensemble learning. This paper classifies existing prediction methods into traditional econometric methods, classical machine learning, hybrid and deep learning. It analyzes the principles, applicable scenarios and advantages and disadvantages of each method, and summarizes empirical conclusions from feature engineering - multi-source data fusion and the model applicability in different market environments. The results show that tree ensemble models such as random forest and XGBoost generally have better prediction effects than traditional linear regression. Hybrid models and multi-source data fusion further enhance the ability to predict housing prices. Currently, there are still problems in this field such as limited data quality, weak model interpretability, and difficulty in modeling temporal and spatial dynamic relationships. Future research can focus on exploring housing price prediction under causal inference, the combination of spatial econometrics and machine learning, and non-structured data mining, etc.

Keywords: Econometrics, Machine Learning, Ensemble Learning, Spatial Econometrics, Model Interpretability

1. Introduction

Real estate has both consumption and investment attributes, and changes in housing prices are an important indicator of economic prosperity. Fluctuations in housing prices can affect household wealth and consumption allocation, and are related to the security of commercial bank mortgage assets. They can also provide a basis for the formulation of government land, tax, and housing security policies. Therefore, housing price prediction is a key research direction in urban economics, real estate finance, and econometrics.

Housing price prediction essentially belongs to a regression problem. It estimates housing prices by relying on variables such as buildings, location, neighborhood, and macroeconomic indicators. Early studies were based on the characteristic price theory and used multiple linear regression to calculate the implicit prices of various attributes of housing. They were widely used in the compilation of housing price indices and property valuation [1]. However, this model has strict

assumptions about the form of the function and is difficult to capture nonlinear relationships and variable interactions. When dealing with high-dimensional data, it often encounters difficulties in variable selection and overfitting.

After the 21st century, big data and computing power have driven machine learning into the field of housing price prediction research. Machine learning does not require presetting the form of the function and can independently discover nonlinear correlations in data, offering more advantages in prediction accuracy [2]. Existing literature reviews show that housing price prediction is continuously absorbing new technologies and new data sources, with the application of deep learning increasing; traditional machine learning is suitable for small samples, while deep learning is good at handling large-scale complex data, and the potential for improving prediction through mixed models is prominent [3]. Current research no longer solely pursues prediction accuracy but increasingly emphasizes model interpretability and economic content [4]. Some empirical studies have compared multiple algorithms using the Ames dataset, confirming that XGBoost has the best comprehensive performance; other studies have integrated textual descriptions of houses and numerical features, confirming that unstructured text can effectively improve prediction results [5].

Domestic localized research has also been continuously advancing. Relevant reviews have found that domestic research variables selection has gradually focused on location and neighborhood characteristics, but there are still subjective and insufficient representative defects in variable selection [6]; the integrated prediction approach has also been developed, selecting benchmark models based on data characteristics, with better results than single models [7].

At present, there are still many shortcomings in this field: There is a large difference in the scenarios to which various models are adapted, and there is a lack of a unified comparison and selection framework; the black box problem of machine learning restricts its application at the policy level; most studies focus on cross-sectional data, and the modeling of housing price spatio-temporal dynamics and spillover effects is relatively weak [2]. Based on this, this paper systematically reviews the regression prediction methods and research progress of housing prices, with the aim of providing methodological references for subsequent research.

2. The main technical methods for predicting house prices

The development of housing price prediction methods has roughly gone through an evolution process from traditional econometric models to machine learning models, and then to hybrid and deep learning models. Each type of method has its theoretical basis and applicable scenarios, and there are different trade-offs between prediction accuracy, interpretability, and computational complexity [3].

2.1. Traditional econometric methods: the characteristic price model

The characteristic price model is a classic econometric method for predicting house prices. Its theory is derived from Lancaster's consumer theory and Rosen's characteristic market equilibrium model. This model treats housing as heterogeneous goods, and the house price is composed of the implicit prices of various attributes such as construction, neighborhood, location, etc. Through regression, the marginal prices of each attribute are estimated [1]. The model's form is ($P = f(S, N, L, C, T)$), and the function form includes linear, semi-logarithmic, double logarithmic, and Box-Cox transformations. Among them, the semi-logarithmic form is the most widely used in empirical applications, and the coefficients can be interpreted as the percentage change in price, which is convenient for comparative studies [8]. This model has a solid economic

theoretical foundation, and the economic meanings of the regression coefficients are clear, which can serve policy evaluation and benefit analysis. It has advantages in handling housing heterogeneity, compiling house price indices, and has low requirements for sample size and can be used in scenarios with limited transaction data [9]. However, its limitations are also obvious: the prediction effect highly depends on the setting of the function form, and when the real relationship is nonlinear, it will generate systematic deviations; the model is sensitive to the sample and settings, ignoring the temporal and spatial changes in the prices of housing attributes; it is difficult to depict the interaction effects of variables, and when dealing with high-dimensional characteristics, it is prone to multicollinearity and overfitting [6]. These deficiencies have driven the academic community to turn to nonlinear and non-parametric methods for research.

2.2. Classic machine learning regression methods

With the development of machine learning, classic algorithms such as random forests, support vector regression, and gradient boosting trees are widely used for housing price prediction. These algorithms do not require the pre-definition of the function form and can automatically discover the nonlinear and interaction relationships of variables, resulting in generally better prediction performance than traditional linear models [10].

Random forests belong to the tree ensemble method. By averaging multiple decision trees, they reduce variance and improve generalization ability, are robust to outliers, and can output variable importance. They perform stably in housing price prediction [11].

Support vector regression (SVR) performs well in high-dimensional and small-sample scenarios, but its prediction performance is greatly affected by the kernel function and parameter tuning. As the sample size increases, the training cost significantly rises. Some empirical studies show that the least squares support vector regression has good prediction accuracy [12].

Gradient boosting trees iteratively fit residuals to reduce errors. XGBoost is an efficient implementation of this method. Multiple comparative studies have confirmed that its performance in housing price prediction is the best [4]. With the help of TreeSHAP, it can partially compensate for the interpretability shortcomings; this model has high accuracy, is suitable for multiple types of features, but is sensitive to parameters and has weaker interpretability than linear models [13].

Overall, tree ensemble models such as random forests and XGBoost have become the mainstream due to their accuracy and stability. However, these methods share common drawbacks such as dependence on data quality, poor prediction performance for extreme samples, and insufficient model interpretability [11].

2.3. Hybrid models and deep learning methods

To further improve the accuracy of housing price predictions, the academic community has developed hybrid models and deep learning methods. Hybrid models integrate the advantages of multiple algorithms to overcome the shortcomings of a single model; deep learning relies on multi-layer neural networks to capture complex data patterns [3].

Hybrid models achieve complementary advantages through algorithm combination. Some studies have combined clustering, linear regression, K-nearest neighbors, and support vector regression. They first cluster and group housing samples, independently model each group, and then determine the sample's group to complete the prediction. Empirical evidence shows that this clustering hybrid model has superior multiple evaluation indicators compared to multiple single models [14]. This

approach enables data-driven segmentation of housing sub-markets and adapts to different sub-market differentiated price formation mechanisms.

Deep learning has developed rapidly in the field of housing price prediction. Basic artificial neural networks rely on multiple layers of neurons to fit complex relationships. In some empirical studies, the prediction effect of ANN is superior to models like XGBoost [15]. However, the prediction effect of various models for high-priced properties will decline due to the scarcity of samples. Deep learning can also handle unstructured text data. By using TF-IDF, Word2Vec, and BERT to extract housing text information, only text features can achieve good prediction results. The integration of text and numerical features can further optimize model performance and expand the data sources for housing price predictions [5].

Although hybrid models and deep learning have outstanding prediction accuracy, they also have shortcomings: Hybrid models have high construction and parameter tuning costs and are complex; deep learning requires a large amount of training data, and its performance in small sample scenarios is inferior to traditional machine learning, and it also has a significant black box problem, making it difficult to apply in scenarios requiring transparency [16].

3. Application perspective and empirical findings

3.1. Feature engineering and multi-source data fusion

The accuracy of housing price prediction not only depends on the model algorithm, but also relies on the quality of features and the richness of data. Feature engineering, as the core connection point between data and models, is a key in the study of housing price prediction. Traditional housing price prediction only relies on structured numerical features, covering basic attributes such as house structure, location, and neighborhood facilities. These are common inputs for classic models and machine learning models, but such data is difficult to fully capture the complex connotations of house value [2].

With the expansion of data channels, the integration of unstructured data and multi-source data has become the mainstream trend in research. Text data is an important new data source. Soft information such as decoration, environment, and living experience in real estate descriptions cannot be quantified by traditional numerical variables. Relevant research has confirmed that text features are strong predictors of housing prices, and the integration of text and structured numerical features can significantly improve prediction accuracy, confirming the impact of soft attributes of houses on the willingness of buyers to pay [5].

The application of spatial and geographic data further optimizes the characterization of location features. Traditional models only rely on simple distance variables to represent location, which cannot reflect spatial heterogeneity and dependence. Current research integrates multi-dimensional geographic big data such as street views, transportation, and neighborhood social economy, improving the modeling system for the laws of housing price appreciation [16]. At the same time, the integration of spatiotemporal features effectively captures the dynamic evolution patterns of housing prices. Empirical evidence shows that external features such as location and neighborhood have a much greater impact on housing prices than the physical internal features of the house, which is in line with the core location theory of real estate value [9].

Although multi-source data integration significantly improves prediction accuracy, it also brings many challenges. Various data have differences in scale, missing patterns, and noise. Feature fusion and dimensionality reduction are difficult. At the same time, data sources such as street views and

social media involve issues of information privacy, and the compliant use of data has become a practical problem that needs to be urgently addressed [16].

3.2. Applicability of models in different market environments

The performance of the housing price prediction model is not fixed and is influenced by various factors such as market maturity, external shocks, price ranges, and regional institutional differences. Selecting the appropriate model based on different scenarios is crucial for improving the accuracy of housing price predictions [3].

Market maturity directly determines the adaptability of the model. In mature markets, there is abundant and high-quality transaction data, which significantly enhances the predictive advantages of complex machine learning models such as random forests and gradient boosting machines [11]. In contrast, in emerging markets or areas with scarce transactions, traditional linear regression models tend to fail. Combining aggregated sub-market data with artificial neural networks for non-linear modeling can achieve more reasonable property valuation [17].

External shocks in the market can affect the stability and accuracy of the model. For example, during the COVID-19 pandemic, studies have shown that machine learning models outperform traditional linear models and are more suitable for the non-linear characteristics of the housing market. Among them, Boosting-type models have lower overfitting risks and more stable performance [18]. The impact of the pandemic has reshaped the importance of key factors influencing housing prices, indicating that the prediction model needs to be continuously calibrated and updated to adapt to changes in market structure.

There are significant differences in housing price prediction performance across different price ranges. Overall, ensemble tree models have better predictive performance, but all models' prediction accuracy for extreme high-end and low-end housing prices will decline. Only mid-range property predictions perform better. The main reason is that high-end properties have strong heterogeneity, few comparable transaction samples, and contain a large amount of difficult-to-quantify scarce and additional value attributes, making the prediction more challenging [11].

In addition, geographical environment and institutional differences can lead to differentiated model adaptability. The optimal parameters and variable importance rankings of the same model in different regional markets vary significantly [19]. In conclusion, there is no universal optimal model for housing price prediction. It is necessary to flexibly select and adjust the model based on specific market characteristics and data conditions.

4. Discussion: challenges and prospects

Although current research on housing price prediction has achieved significant results, it still has many limitations and has given rise to several cutting-edge research directions. Firstly, the problem of data quality and accessibility is the primary constraint. Most markets have insufficient transparency in the availability of real transaction data, and studies often use listed prices and assessed prices instead of transaction prices, resulting in systematic biases; soft features such as house renovations and internal conditions are missing, and the existing data has a coarse time granularity, making it difficult to capture short-term fluctuations in housing prices [12]. Future development of real estate technology can improve data conditions, but it is necessary to balance data application with privacy and commercial confidentiality issues.

Secondly, there is a dilemma in the trade-off between model accuracy and interpretability. Machine learning and deep learning models have better prediction accuracy, but their black-box

nature makes it difficult to explain the price driving mechanism from an economic perspective. SHAP and LIME, among other interpretable artificial intelligence methods, can only provide statistical-level explanations and cannot meet the causal explanation requirements for policy research [3].

Furthermore, existing research mostly focuses on correlation prediction and lacks causal identification capabilities, making it difficult to respond to practical issues such as policy shocks and unable to quantify the real impact of variables such as subway construction and supporting facilities on housing prices [9]. In the future, it is possible to combine the advantages of machine learning prediction with econometric causal inference methods to build a model system that combines prediction accuracy and causal explanation capabilities.

In addition, there are still shortcomings in spatio-temporal dynamic modeling. Housing prices have significant spatial correlations and temporal inertia, while traditional models often assume spatio-temporal stationarity and independence, which does not match reality. Spatial econometrics and time series analysis methods have their own focuses and have not yet formed a unified spatio-temporal fusion prediction framework [16]. New technologies such as graph neural networks also require data-driven modeling to be combined with economic theories for improvement.

Finally, the selection of automated models under market heterogeneity is an important research direction. The optimal models for different markets, periods, and price ranges vary, and the integrated prediction approach driven by data characteristics and AutoML technology can achieve intelligent model selection and combination, improving the adaptability and practicality of predictions [7].

5. Conclusion

This article systematically reviews the methodological framework and research progress of housing price regression prediction. At the methodological level, housing price prediction has evolved from the characteristic price model to classical machine learning, and then to hybrid and deep learning models. The characteristic price model has a solid theoretical foundation and strong interpretability, and is suitable for the compilation of housing price indices and policy evaluation, but it is difficult to handle nonlinear and high-dimensional data. Random forest, XGBoost, and other tree ensemble machine learning models have higher prediction accuracy than traditional models and have become mainstream; hybrid and deep learning are good at handling complex and unstructured data, but they have high requirements for data and poor interpretability.

At the application level, feature engineering and multi-source data fusion, such as text, spatial, and street view data, improve the prediction effect and deepen the understanding of the formation mechanism of housing prices. The model effect is context-dependent, and market maturity, external shocks, price range, and institutional environment will all affect the model performance. There is no universal optimal model, and it is necessary to select based on the actual scenario.

This field is still constrained by data quality, interpretability defects, and problems in spatio-temporal dynamic modeling. In the future, it can be advanced in four aspects: constructing models that balance prediction and causal explanation by combining causal inference; integrating spatial econometrics and machine learning to build a spatio-temporal prediction framework suitable for real estate characteristics; deeply mining non-structured data such as text and images; and developing an adaptive model selection mechanism. With the progress of data and algorithms, housing price prediction will continuously optimize in terms of accuracy and interpretability, providing decision support for the stable operation of the real estate market.

References

- [1] Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. *Housing Economics and Public Policy*, 67-89.
- [2] Geerts, M., vanden Broucke, S., & De Weerd, J. (2023). A survey of methods and input data types for house price prediction. *International Journal of Geo-Information*, 12(5), 200.
- [3] Naz, R., & Ijaz, H. (2024). Machine learning, deep learning, and hybrid approaches in real estate price prediction: A comprehensive systematic literature review. *Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences*, 61(2), 129-144.
- [4] Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics*, 3(1), 3.
- [5] Zhang, H., Li, Y., & Branco, P. (2023). Describe the house and I will tell you the price: House price prediction with textual description data. *Natural Language Engineering*, 29(6), 1563-1586.
- [6] Mingke, H. X. (2017). A review of the impact of residential characteristics on real estate value based on the price-characteristics model. *Business Economics*(12), 172-174.
- [7] Cui, M., Liu, X., Li, X., et al. (2020). Research on integrated forecasting of the real estate market driven by data characteristics. *Management Review*, 32(7), 33-43.
- [8] Hill, R. J. (2012). Hedonic price indexes for housing. *OECD Statistics Working Papers*, No. 2012/01.
- [9] Redfearn, C. L. (2009). How informative are average effects? Hedonic regression and amenity capitalization in complex urban housing markets. *Regional Science and Urban Economics*, 39(3), 297-306.
- [10] Jaafar, N. S., Mohamad, J., & Ismail, S. (2022). Machine learning for property price prediction and price valuation: A systematic literature review. *Planning Malaysia: Journal of the Malaysian Institute of Planners*, 20(2), 145-160.
- [11] Ho, W. K., Tang, B., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 38-59.
- [12] Pai, P. F., & Wang, W. C. (2020). Using machine learning models and actual transaction data for predicting real estate prices. *Applied Sciences*, 10(12), 4260.
- [13] Zhang, L. (2023). Housing price prediction using machine learning algorithm. *Journal of World Economy*, 5(3), 34-41.
- [14] Özögür Akyüz, S., Eygi Erdogan, B., Yıldız, Ö., et al. (2023). A novel hybrid house price prediction model. *Computational Economics*, 61(3), 1087-1112.
- [15] Rampini, L., & Re Cecconi, F. (2022). Artificial intelligence algorithms to predict Italian real estate market prices. *Journal of Property Investment & Finance*, 40(4), 350-364.
- [16] Kang, Y., Zhang, F., Peng, W., et al. (2020). Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, 99, 104994.
- [17] Yang, X. (2025). Research on house price prediction based on machine learning. *ITM Web of Conferences*, 70, 02018.
- [18] Mora-García, R. T., Céspedes-Lopez, M. F., Perez-Sanchez, V. R., et al. (2022). Housing price prediction using machine learning algorithms in COVID-19 times. *Land*, 11(8), 1277.
- [19] Tan, F., Cheng, C., Wei, Z. (2018). Modeling and elucidation of housing price. *Data Mining and Knowledge Discovery*, 32(6), 1628-1651.