World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3043

APPICATION OF XGBOOST ALGORITHM IN HOUSING ASSET VALUATION

BoHong Wang[1*], YiXuan Guo[2], ChaoLin Hou[1], ZhiLing Zhang[3]
[1]*Finance Management School, Shanghai University of International Business and Economics, Shanghai 201620, China.*
[2]*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, Hubei, China.*
[3]*School of International Business, Shanghai University of International Business and Economics, Shanghai 201620, China.*
Corresponding Author: BoHong Wang, Email: 18738921985@163.com

Abstract: Machine learning models supported by big data have been practiced and applied in many ways in recent years, and as a representative technology of artificial intelligence, machine learning models have been proved to be able to perform well in many predictive problems such as economics and management. This paper explores the practice in the problem of residential value assessment by using the more popular machine learning models. The Chain Home platform offers publicly available, granular data on residential property transactions, including variables such as location, area, layout, and pricing. The dataset from November 22, 2024, was selected to provide a consistent time snapshot of the housing market, facilitating reliable model training and evaluation. After that, it further compares the performance of linear regression, random forest algorithm, extreme gradient boosting tree, lightweight gradient boosting tree, classification boosting tree and other algorithms on asset pricing. The empirical results show that the machine learning algorithms can be relatively effective in assessing and pricing residential properties according to their characteristics, and the error between the predicted price and the actual price of the asset appraisal model based on the extreme boosted tree algorithm is much smaller, with an average error of about 17%. This paper attempts to introduce machine learning into the field of asset evaluation, which helps to promote the cross-fertilization research of artificial intelligence and traditional economics problems, and provides reference for promoting the application of artificial intelligence.
Keywords: Asset valuation; XGBoost; Machine learning

1 INTRODUCTION

As the real estate market continues to prosper and develop, the accurate assessment of residential prices has become a key basis for decision-making in the real estate sector, financial investment, and urban planning. Traditional residential price appraisal methods, such as the cost-based method, the market comparison method, and the income method, although capable of providing price estimates to a certain extent, are often heavily influenced by subjective factors and have limitations in dealing with large-scale data and complex market dynamics. For example, the market comparison method is highly dependent on the selection and revision of comparable examples by appraisers, and its accuracy is difficult to ensure consistency; the cost method is ambiguous in its estimation of factors such as depreciation, which makes it difficult to accurately reflect the impact of market supply and demand on prices.

In recent years, the rapid development of machine learning technology has brought new opportunities and breakthroughs in residential price assessment. Machine learning models can automatically mine potential patterns and laws from massive real estate data, which cover the physical attributes (such as area, number of rooms, building age, etc.), geographic location characteristics (such as surrounding facilities, transportation accessibility, and school districts, etc.), as well as market transaction data (such as historical transaction prices, length of listing, etc.) of residences. Through in-depth analysis and learning of these multi-dimensional data, the machine learning model can construct a more accurate and objective residential price assessment model, effectively overcoming some of the shortcomings of traditional assessment methods.

Currently, the application of machine learning to management and economics problems is mainly focused on the financial field, and for the mining of quantitative factors, Li Bin et al. systematically examined the advantages of machine learning models over traditional linear models[1], Guo Feng et al. proposed the use of machine learning models for improved policy evaluation [2] and heterogeneity causality test, etc[3], which opens up new ideas for the advantages of machine learning models in prediction. Overseas countries have recently proposed many applications that put big models on the ground, especially stock price prediction[4], financial data analysis [5] and breakthroughs in portfolio construction [6] . As the asset valuation industry needs a lot of practical accumulation in practice, most of the methods for asset valuation still focus on traditional methods such as cost method, income method and market method [7] , and there are few studies focusing on the application of data science methods such as machine learning in asset valuation [8] , and the existing new methods of machine learning lack of sufficient practice and effectiveness testing. testing. This paper hope to make new practical tests and attempts for the application of data science and asset valuation through the attempts on housing asset valuation.

This research focuses on asset assessment of residential prices using machine learning models. It will deeply explore the application of different machine learning algorithms in residential price assessment, and through a series of rigorous

research steps, such as data collection and preprocessing, model construction and training, performance evaluation and optimization, the experiment are committed to constructing a residential price assessment model with high accuracy and reliability. This will not only help to promote the technological innovation and development of real estate appraisal, but also provide more scientific and reasonable decision-making support for real estate market participants and promote the healthy and stable operation of the real estate market, which is of great significance in both theoretical research and practical application.

2 METHODOLOGIES

The task of the asset pricing module is a standard supervised learning and regression task, i.e., discovering the following functional form:

$$P_i = f(x_i; \theta) + \epsilon_i \tag{1}$$

where $f(\cdot)$ is defined as a function with parameterθ , which in this paper is the functional form of the enriched machine learning model,P_i is the total price of the house for the residence$i$, and$x_i = (x_{i,1}, x_{i,2}, ..., x_{i,N})$ is the feature vector for the residence$t$. The machine learning algorithm used will be described later in this paper.

After determining the specific functional form$f(\cdot)$, this paper will use the second-hand housing listings data presented on the Chain Home second-hand information website on November 22, 2024 to fit the model parameters. In order to ensure the effectiveness and feasibility of the calculation, stratified sampling is used to divide the data set into training set and test set with the ratio of 80% and 20%, and the random seed is set at the same time, which is of key significance in ensuring the reproducibility of the data segmentation, so as to make the results of the data segmentation have the consistency and stability under the same experimental conditions, and thus laying a solid foundation for the subsequent model training and evaluation. In addition, in view of the potential impact of the correlation between the feature variables on the performance of the model, the features that are highly correlated with the target variables, such as the "unit price" feature, are eliminated after the division is completed, aiming to optimize the input feature space of the model, reduce the interference of redundant information, and improve the effectiveness and accuracy of the model training.

2.1 Acquisition and Description of Data

For price assessment of residential houses, the data source is crucial for model accuracy. In this paper, we crawl the transaction information on Chain Home's second-hand house website on November 22, 2024 as the data source, and obtain a total of 2029×13 data.（https://sh.fang.lianjia.com/loupan/）The data field descriptions are shown in the following Table 1.

Table 1 Data Field Descriptions

field name	Meaning of the field	typical example
configuration of rooms in a residence	House Layout	3 bedrooms and 2 bathrooms
area	Building area	97.12 square meters
Qibla (Islam)	house orientation	south
renovate	Decoration status	hardcover
story	Description of the floor where it is located	Middle Floor (18 floors in total)
Floor Height	Floor Height Type	middle floor
Floor numbers	Total number of floors	18
building structure	Type of building structure	slab type building
neighborhood	Name of the neighborhood	Renegade Home
shore	Location	the bow of a ship
total price	Total price of the house ($10,000)	340
price of item	Unit price per square meter (yuan/square meter)	35009
particular year	Year of construction	2012

2.2 Feature Engineering

2.2.1 Feature separation and definition

After completing the data cleaning, the dataset was separated by features and target variables. In this case, the target variables used for prediction were explicitly specified (usually house price related columns, e.g., 'total price'), and the remaining columns were defined as the features used to predict house prices. This separation operation clearly defines the inputs (features) and outputs (target variables) of the model and lays the foundation for subsequent model training and evaluation.

2.2.2 Feature classification and recognition

Based on the nature of features, all features are classified into two categories: categorical features (e.g., 'house type', 'orientation', 'decoration', 'floor height', 'building structure', 'neighborhood', 'area') and numerical features (e.g., 'area', 'floor number', 'unit price', 'year'). This categorization process helps to adopt corresponding processing strategies for different types of features, because categorical and numerical features differ in data representation and model processing methods, and require different technical means for effective feature engineering [9].

2.2.3 Classification feature code

For categorical features with a large base (e.g., 'neighborhood', 'region'), label coding was used for processing. Label coding assigns a unique integer identifier to each category, converting the original categorical data into numerical form, thus enabling the model to process these features. This coding approach preserves the category information while transforming it into numerical inputs acceptable to the model, facilitating computation and analysis in the model.

For categorical features with a small base, the solo thermal coding technique is utilized. Solo thermal coding represents each category by creating binary vectors whose length is equal to the total number of categories, where only one element is 1, indicating the category to which the sample belongs, and the rest of the elements are 0. This coding approach effectively handles the problem of the absence of natural ordering relationships between categorical features and avoids false assumptions that may be made by the model when dealing with the categorical data, and also increases the dimensionality of the features, allowing the model to be able to capture the potential relationship between the categorical features and the target variables in more detail.

2.3 Machine Learning Prediction Algorithms

Machine learning is a collection of numerous prediction functions and various algorithms. As mentioned earlier, residential real estate pricing is a supervised learning regression task, and any of the machine learning algorithms used for the regression prediction task can be used to model asset valuation. With reference to the performance of machine learning algorithms in earlier he prediction studies, this paper intends to test the effectiveness of machine learning models in asset valuation through individual representative algorithms, focusing on the following three observations:

(1) The first observation: does the machine learning model provide a better asset valuation of residential properties?

(2) The second observation: if the prediction model $f(\cdot)$ is in the form of a nonlinear function, can the performance of the nonlinear machine learning algorithm outperform the linear model.

(3) Third observation: if the prediction model $f(\cdot)$ adopts a nonlinear functional form, which model has the best performance?

In order to verify the above three observations, this paper chooses the traditional linear regression model as the benchmark to select four machine learning modeling algorithms and traditional linear regression. To validate the first observation, this paper adopts XGBoost algorithm as the main model. The XGBoost algorithm is chosen because he has achieved better results when dealing with large sample datasets.

In order to validate the second as well as the third observation, four machine learning algorithms and linear regression algorithms are subsequently included in this paper, including Random Forest regression model (Random Forest), LightGBM (Light Gradient Boosting Machine), CatBoost (Categorical Boosting) , and OLS regression.

2.3.1 XGBoost algorithm

Core formula:

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{2}$$

where $g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the first-order derivative of the loss function, such as when squared loss, $g_i = 2(\hat{y}_i - y_i)$; $h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$ is the second-order derivative, i.e., when squared loss, $h_i = 2$; T is the number of leaf nodes of the current tree; at leaf node j, ω_j is its weight, i.e., the prediction value; γ is the minimum gain required for splitting, which is used to control the complexity of the tree; λ is the L2 regularization coefficient, which is used to inhibit the degree of overfitting phenomenon. This function optimizes the prediction error and model complexity at the same time, avoiding the overfitting of the evaluation price to the noisy data (e.g., abnormally high unit price) at the same time, and achieving the purpose of learning and prediction in a better way [10,11].

2.3.2 Random forest regression model

The basic idea of Random Forest is Bagging (Bootstrap Aggregating) and random feature subspace

Core formula:

$$\hat{y}rf = \frac{1}{B}\sum b = 1^{B} T_b(x) \tag{3}$$

where $T_b(x)$ is the predicted value of the bth tree, and the final result is the average value of all trees. Random Forest constructs an assessment model with high generalization ability through double randomness (data Bagging + feature subsampling), which can reveal the key influencing factors of house prices with its feature importance ranking driven by ΔMSE ; and quantify the reliability of the assessment results based on the prediction intervals of OOB (Out-of-Bag can compute the uncertainty error) samples; and at the same time, it possesses the ability to deal with the high-dimensional feature interactions, which can replace the cost of manually designing interaction terms. can replace

the cost of manually designing interaction terms. In the past experiments, compared with the linear model, Random Forest improves the accuracy by 12%-18% on average in the residential appraisal task (MAE reduction), and especially performs better in the non-uniform market (e.g., school districts, luxury houses) [12].

2.3.3 LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework that optimizes efficiency and accuracy by innovatively incorporating histogram-based decision trees, GOSS (Gradient-based One-Side Sampling, i.e., retaining samples with large gradients and randomly sampling samples with small gradients), and EFB (Exclusive Feature Bundling, mutually exclusive feature binding to reduce dimensionality). Exclusive Feature Bundling, mutually exclusive feature bundling to reduce dimensionality) to optimize efficiency and accuracy.

Core formula:

$$\mathrm{Obj}^{(t)} = \sum_{i=1}^{n} L\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{4}$$

where $L(y_i, \hat{y}_i)$ is the loss function; $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \text{¥} |w \text{¥}|^2$ is the regular term used to control the complexity; γ is the leaf splitting minimum gain; λ is the L2 regularity coefficient; T and w are the number of leaf nodes and the weight of the leaf, respectively.

2.3.4 CatBoost algorithm

The core innovation of CatBoost lies in the way it encodes category features (e.g., house orientation, school district rank), and in applications it does not need to encode them manually, but deals directly with high base features (e.g., neighborhood names) and preserves the intrinsic relationships of the features. Its core formula is:

$$\mathrm{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \left\| \omega \right\|^2$$

where L is the loss function (e.g., MAE, RMSE), control complexity [13].

2.3.5 OLS regression

OLS, or Ordinary Least Squares, has a core formula:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{6}$$

where β_0 is the intercept term, β_k is the coefficients of each feature, and ϵ is the error term. Compared to traditional machine learning methods, OLS regression is less complex and not as accurate as the models in 2.3.3 and 2.3.4, but it is transparent and interpretable, and can be applied at low sample sizes (>30).

Most of the machine learning models in this paper are chosen to be Random Forest Class models because the performance of 179 classification algorithms was examined and it was concluded that Random Forest Class algorithms can achieve desirable results in the vast majority of classification tasks Fernández-Delgado et al. (2014). It is worth clarifying that the algorithms selected for this paper are not the complete set of machine learning regression algorithms. Although not exhaustive of machine learning algorithms, several representative algorithms selected have achieved better predictive performance in other domains.

2.4 Model Evaluation

In the model evaluation session, a set of multi-dimensional and comprehensive evaluation system is constructed [14].

2.4.1 Construction of the indicator system

Root Mean Square Error (RMSE) selected:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{7}$$

Coefficient of determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

And the Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{9}$$

As the core evaluation index. After the model has completed the training process, the trained model is used to carry out prediction operations on the test set data, and then the quantitative values of the assessment indexes are determined with the help of the corresponding mathematical calculation paradigm based on the predicted values and the real target values of the test set. Among them, RMSE focuses on measuring the deviation of the predicted values relative to the actual values, and its magnitude intuitively reflects the accuracy of the model prediction; R^2 is mainly used to characterize the model's goodness-of-fit to the data, and its value ranges from 0 to 1, with the value closer to 1

indicating that the model is more capable of interpreting the data; and MAE focuses on the average status of the absolute errors between the predicted values and the actual values, and evaluates the performance of the model from the viewpoint of the average error. The MAE focuses on the average absolute error between the predicted and actual values, and assesses the model performance from the perspective of average error.

2.4.2 Model further assessment analysis

Further, in order to achieve a comprehensive, in-depth and integrated comparison and analysis of the accuracy of each model, some of the assessment indicators, such as the RMSE and the MAE, are normalized, and their numerical ranges are mapped to specific intervals, so as to allow comparison and comprehensive consideration under a uniform scale. On this basis, the normalized evaluation indicators are weighted and summed according to a predetermined weighting scheme to obtain a comprehensive score that reflects the overall performance of the model. In the end, through the detailed comparison and analysis of the values of the evaluation indexes and the comprehensive score, the best models with excellent performance in different evaluation dimensions are precisely identified, so as to achieve an in-depth evaluation of the accuracy of all the models participating in the experiment in the house price prediction task in an all-around, multi-level and refined manner, providing a scientific, rigorous and reliable basis and guidance for the selection, optimization and application of the models.

3 EVALUATION AND ANALYSIS OF EXPERIMENTAL PROCEDURES AND RESULTS

3.1 Residential Asset Valuation Model based on XGBoost Algorithm

This paper examines the empirical performance of the XGBoost algorithm for residential asset valuation. Table 2 demonstrates the fitting effect of the model in the training and test sets. Observing Table 2, it can be seen that although the model fitting effect is better and the goodness of fit reaches 0.86 in the test set, the prediction results fluctuate and are unstable. Figure 1 exhibits the specific distribution of the model's fitting effects, as well as the ranking of feature importance in the evaluation. It can be further seen that most of the prediction errors are small, concentrated within 250, and a few have large prediction errors. Among the feature percentages, area has a greater weight in the price assessment, and the floor feature is the least important.

Table 2 XGBoost Empirical Effects

norm	training set	test set
MSE	1831.218064	13567.41021
RMSE	42.79273378	116.4792265
MAE	29.72356542	63.32165679
R^2	0.98334536	0.861799573
Mean absolute percentage error	10.30382057	18.93482373
maximum error	346.9562988	966.1529541
Absolute error of median	20.51007843	32.37414551
Explaining variance scores	0.983345576	0.862343317

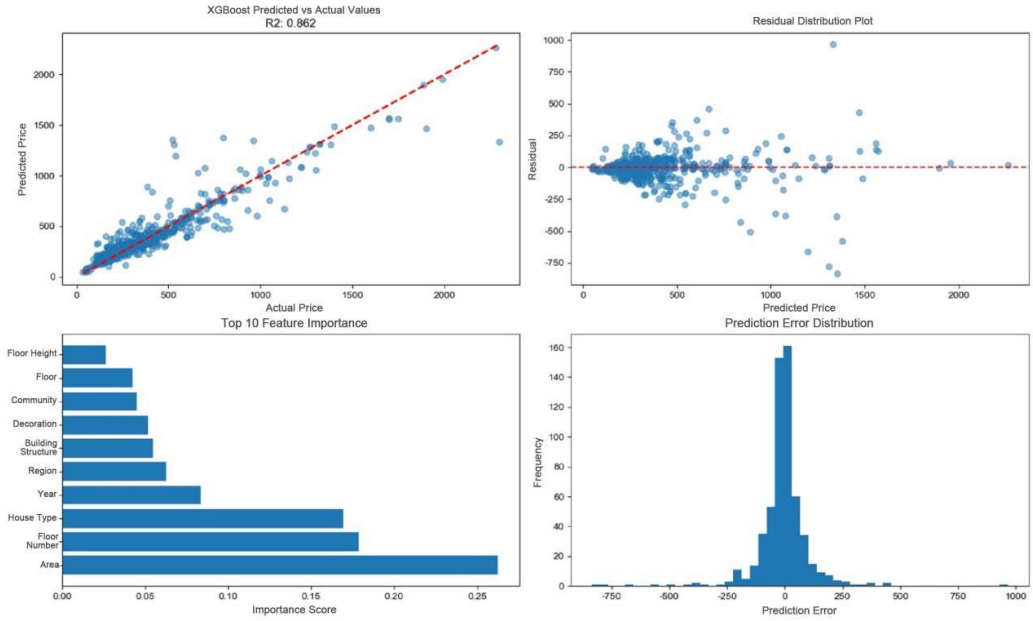

Figure 1 XGBoost Empirical Performance

3.2 Further Analysis: Sources of Error

In order to further explore the reasons for the large errors in some of the predictions, the five data sets with the largest errors are selected in this paper. Table 3 demonstrates their specific information, and through the analysis, it is found that the five listings with larger errors have relatively larger housing areas, which, as can be seen from the previous section, are more important and weighted in the model's pricing, and thus given larger predicted prices. In addition, housing price is closely related to housing area, however, due to the problem of describing the information on the website, the model has too much granularity in describing the input parameters for housing area, which leads to its weakened correlation with housing price. For example, Xiayang and Xin Jiangwan City refer to Xiayang Road in Xuhui District and Xin Jiangwan City in Yangpu District, respectively; Xin Jiangwan City has a house price of nearly 100,000/sqm due to geographic factors such as proximity to schools, while Xiayang Road in Xuhui District has only 50,000/sqm. When the model is categorized, the model will classify the listings with the same area name into one category without considering the actual location factors of Xuhui and Yangpu Districts. Especially for the listings with the same area name appearing less in the dataset, the pricing information is more likely to be distorted, and thus the pricing effect is not satisfactory enough.

Table 3 Analysis of Error Causes

signature information	Listing 1	Listing 2	Listings 3	Listings 4	Listings 5
Subdivision.	Xayanghu International Garden	City Garden	Sunrise Riverview Villa	Yanlord Yunjie Riverside Garden (Phase I)	Poly Forest Creek (Apartment)
Region.	Xia Yang (1916-1992), Chinese communist leader	Cambridge	New Jiangwan City	Xia Yang (1916-1992), Chinese communist leader	Sanlin prefecture level city in Guangxi
House type.	3 bedrooms and 2 bathrooms	3 bedrooms and 2 bathrooms	4 bedrooms and 2 bathrooms	4 bedrooms and 2 bathrooms	4 bedrooms and 2 bathrooms
Area.	146.98 square meters	147.55 square meters	234.13 square meters	157.54 square meters	141 square meters
Orientation.	south	south	south	south	south north
Decoration.	hardcover	hardcover	hardcover	simple installation	hardcover
Floors.	High floor (17 floors in total)	Lower floors (10 floors in total)	Middle Floor (7 floors)	Lower floors (17 floors in total)	18 floors.
Architecture.	slab type building	slab type building	slab type building	slab type building	slab type building
Year.	2004	2006	2014	2007	2011
Actual Price.	535.00 million	528.00 million	22,980,000	5.2 million	798.00 million
Forecast Price.	11.9556 million	1,307.09 million	13,318,500	13,546,700	13,787,500
Prediction error.	6,605,600	7,790,900	9,661,500	8,346,700	5,807,500
Relative error.	123.47%	147.55%	42.04%	160.51%	72.78%

3.3 Comparison of Evaluation Effects after Integration of Multiple Machine Learning Algorithms

In order to more intuitively show the pricing effect of machine learning models, this paper adds other mainstream machine learning models including for comparison, including Random Forest regression model (Random Forest), LightGBM (Light Gradient Boosting Machine), CatBoost (Categorical), and linear algorithm OLS regression. Boosting), and the linear algorithm OLS regression. In model evaluation, a set of multi-dimensional and comprehensive evaluation system is constructed. Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Error (MAE) are selected as the core evaluation indexes to score the effectiveness of model asset evaluation. Figure 2 takes the price of one of the listings as an example to visualize the accuracy of each algorithm in asset pricing, and it can be seen that the predictive effect of the machine learning model is indeed superior to that of ordinary linear regression algorithms, and is better able to capture the correlation between various price information.

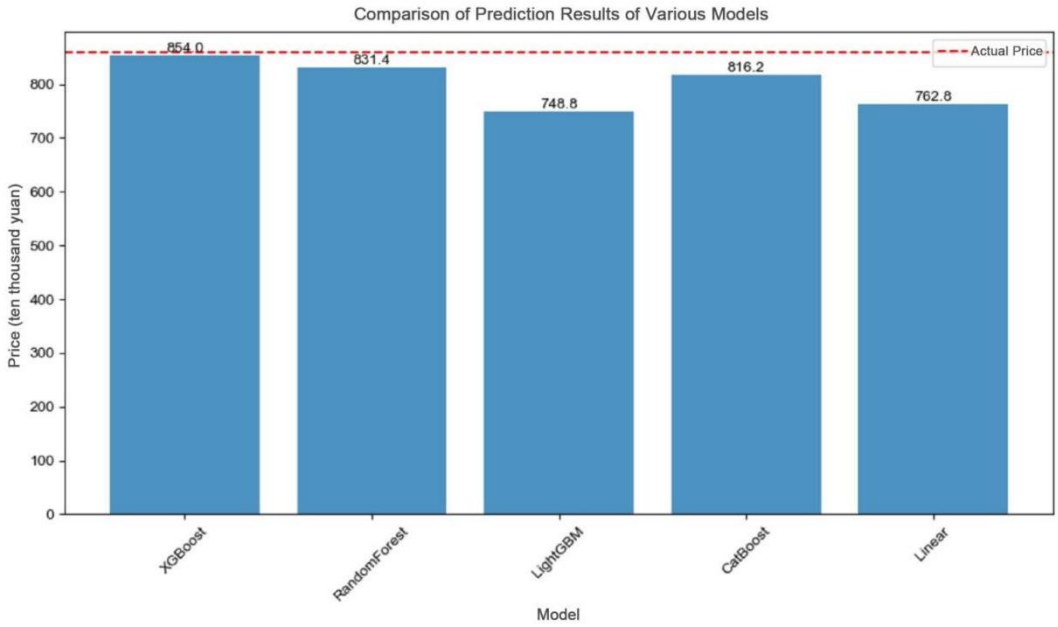

Figure 2 Comparison of Model Predictions

However, due to the limitation of the comparison of individual cases, the model evaluation system better reflects the performance of each algorithm on residential pricing. As can be seen from Table 4, the XGBoost model achieves the best scores in all three metrics, RMSE, R^2 and MAE, and naturally has the highest overall score. Therefore, compared with the other four models, the XGBoost model achieves better results in the prediction of residential asset pricing.

Table 4 Scores for Each Model

Model	RMSE	R^2	MAE	aggregate score
XGBoost	121.5662	0.8495	76.6058	0.9398
RandomForest	130.6359	0.8262	85.9488	0.8163
LightGBM	128.3093	0.8323	83.4816	0.8484
CatBoost	144.7596	0.7865	99.4019	0.6275
Linear	171.0729	0.7019	122.329	0.2808

4 CONCLUSIONS AND IMPLICATIONS

The XGBoost algorithm shows good performance in the evaluation of residential asset valuation models. Its coefficient of determination (R^2) on the test set reaches 0.86, indicating that the model is able to explain most of the variations in house prices. Meanwhile, by analyzing the listings with large errors, it is found that the listings with large housing areas are prone to high predicted prices due to the high weight of the area in the model's pricing. Meanwhile, the website information is less effective in predicting listings that have the same area name but large differences in actual location and appear less frequently in the dataset. Comparing multiple machine learning models (Random Forest Regression, LightGBM, CatBoost and Linear Regression OLS), XGBoost has the best overall performance in residential asset pricing prediction, with the highest scores in the combined assessment of its Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Error (MAE) metrics, which further proves that the machine learning model is better than ordinary linear regression in residential pricing is superior to ordinary linear regression algorithms.

This study introduces machine learning technology into the field of residential price assessment, enriching the theory and methodology of real estate assessment. The application effects of different machine learning algorithms in this field are empirically analyzed to provide empirical references for subsequent studies. The application of machine learning models can significantly improve the efficiency and accuracy of real estate assessments, lower evaluation costs, and strengthen risk assessment capabilities, thereby facilitating the healthy and stable growth of the real estate market. At the same time, there are limitations in this study: data-wise, it only relies on the second-hand house transaction information on Chain Home's website, which is a relatively single source of data and may not be able to comprehensively cover all the factors affecting house prices. In terms of modeling, although a variety of machine learning models have been compared, there are still other excellent models that have not been included in the study, and the parameter settings of the models may not be optimal. On the application side, when applying the model to real-world scenarios, it may face problems such as data updating and dynamic changes in the market. It is necessary to

establish a dynamic updating mechanism to adjust the model in time to adapt to market changes, and to strengthen the research on the interpretability of the model so that the model results are easier to understand and accept.

In future research, it can be studied in depth by focusing on multimodal data integration, covering spatio-temporal dynamic models and personalized pricing models, and considering market intervention and policy simulation. In terms of data integration, multi-source data, such as government public data, geographic information data, macroeconomic data, etc., can be used to assess house prices more comprehensively and accurately. And more advanced machine learning algorithms can be further explored with more detailed parameter tuning to improve the model performance.

This study demonstrates the potential of machine learning in residential price assessment, but there are still many aspects that need to be further explored and improved, and future research is expected to make more breakthroughs in these areas and provide stronger technical support for the development of the real estate market.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Qiu X, Ke X. The Impact of ChatGPT-like Artificial Intelligence on the Asset Appraisal Industry. China Asset Appraisal, 2024(03): 20-26.

[2] Guo F, X Tao. Machine learning and causality in the social sciences: a literature review. Economics, 2023, 23(01): 1-17. DOI:10.13821/j.cnki.ceq.2023.01.01.

[3] Li B, Shao C, Li Y. A Study of Machine Learning-Driven Quantitative Investment in Fundamentals. China Industrial Economy, 2019(08): 61-79. DOI:10.19581/j.cnki.ciejournal.2019.08.004.

[4] Tao X, Guo F. Heterogeneous policy effects assessment with machine learning methods: research progress and future directions. Management World, 2023, 39(11): 216-237. DOI:10.19744/j.cnki.11-1235/f.2023.0127.

[5] Cao S, Jiang W, Wang J, et al. From Man vs. Machine to Man + Machine: The art and AI of stock analyses. Journal of Financial Economics, 2024, 160: 103910-103910.

[6] Kim A, Muhn M, Nikolaev V. Financial statement analysis with large language models. arXiv preprint arXiv:2407.17866, 2024.

[7] Jon K, Jens L, Sendhil M, et al. Prediction Policy Problems. The American economic review, 2015, 105(5): 491-495.

[8] Wang W, Li W, Zhang N, et al. Portfolio formation with preselection using deep learning from long-term financial data. Expert Systems With Applications, 2020, 143: 113042-113042.

[9] Jie X, Yukun Z, Chunxiao X. A review of research on the application of machine learning in financial asset pricing. Computer Science, 2022.

[10] Ma J. Research on the value assessment of data assets in network transaction scenarios based on machine learning. Beijing Jiaotong University, 2024.

[11] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System.CoRR, 2016.

[12] Hu J, Szymczak S. A review on longitudinal data analysis with random forest. Briefings in bioinformatics, 2023, 24(2): bbad002.

[13] Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 2018, 31.

[14] Lu M. Research and design of asset evaluation platform based on big data. Science and Technology for Development, 2019, 15(10): 1093-1105.