Analysis and Comparison of House Price Prediction Based on XGboost and LightGBM

Shengquan Chen[1,a,*,†], Huihui Jin[2,†], Ling Li[3,†]

[1] *Business School, PSB Academy, Singapore, 039594, Singapore*
[2] *International College, Jiangxi University of Finance and Economics, Nanchang, 330013, China*
[3] *School of Accounting and Finance, Hong Kong Polytechnic University, Hong Kong, 99907, China*
a. 2201901573@stu.jxufe.edu.cn
**corresponding author*
[†]These authors contributed equally.

Abstract: Real estate price prediction is one of the key research topics contemporarily. Based on the rapid development of Big Data, machine learning has gradually become the mainstream tool for housing price prediction. The XGboost and LightGBM models, as new advanced models in recent years, have received widespread attention in the application in housing price prediction. Therefore, this study identifies the house price prediction based on XGboost model and LightGBM model and compares them with other models in order to obtain an analysis of the advantages and disadvantages of these two models in housing price prediction. According to the analysis, both models have ad-vantages such as high accuracy, high efficiency, and fast training speed. However, although XGboost has the smallest error prediction, it requires more computational time, thereby increasing computational costs. In addition, LightGBM has disadvantages such as high overfitting risk in small sample sizes and increased sensitivity in noisy datasets. Therefore, besides the model studied in this article, feature selection methods such as Filter and Wrapper can also be introduced in subsequent studies to further improve the prediction accuracy.

Keywords: house price prediction, LightGBM, XGboost

1. Introduction

Real estate finance is one of the most popular research topics in finance. With the arrival of information age, traditional data processing methods are unable to adapt to such a large amount of data. Compared with traditional data processing tools, Big Data technology has the characteristics of massification, versatility, value, timeliness and authenticity, and can efficiently and accurately process massive data [1, 2]. Effective use of machine learning can enhance the functionality of Big Data technology [3]. Therefore, machine learning has been widely applied in various fields of society. In the era of Big Data, machine learning needs to have the ability of generalization and rapid learning, easy understanding and cost perception, data utilization and knowledge transfer [4]. As one of the most important contents of Big Data, machine learning has been widely used and deeply developed in predicting housing prices.

Recently, machine learning has developed rapidly. Many basic models in machine learning, such as ElasticNet, GBDT, linear regression, Feedforward neural network, have been tested in the field

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/46/20230317

of housing price prediction. However, the ElasticNet model has poor accuracy when using linear models to fit nonlinear data; GBDT is difficult to train data in parallel due to the dependency relationship between weak learners; the linear regression model has limited ability to handle complex data; Feedforward neural network is expensive and prone to errors. In terms of housing price prediction, the experiment shows that the average RMSE of the average test error of 100 times of training is 0.0178. Compared with the multi-layer Feedforward neural network and multiple linear regression model, this data is a lower value [5]. Zhang and Du argued that the average test error calculated by XGBoost model is lower than that of linear regression and Feedforward neural network based on housing price prediction, which also shows that the accuracy of the integrated model is superior to the traditional machine learning model [6]. Using integrated multiple model fusion can learn dataset features from multiple different dimensions and has strong transfer and generalization abilities [7]. In data processing, there will be some problems and different assumptions at the statistical, computational and presentation levels, which will lead to more complex situations. Simple equal voting using Ensemble learning can avoid these problems [8]. Therefore, this article chooses two integrated models (XGBoost and LightGBM) as the main re-search tools for housing price prediction.

The XGBoost model is an integrated machine learning algorithm based on decision trees (GBDT) [9]. It aggregates a single decision tree to form a highly accurate decision tree. In many experimental comparisons, the error of XGBoost model in predicting housing prices is far lower than that of linear regression and Feed-forward neural network. The LightGBM model comes from Microsoft and is an integrated new algorithm [10]. Firstly, the model uses a single edge gradient sampling algorithm technology to reduce the amount of data processing. Secondly, the mutually exclusive feature binding algorithm helps the model improve processing efficiency in multi-dimensional data. Meanwhile, the LightGBM model can effectively improve its training speed by using a histogram algorithm to find the optimal branch point during training. The two models both have high efficiency, high accuracy, and low cost. Compared to other models, the two models have the ability to process richer data and make more accurate judgments.

This study uses a systematic literature review method, analyzing the advantages and disadvantages of the XGBoost and LightGBM in predicting housing prices. This research is mainly divided into six parts. The first part is a brief introduction. The second part will introduce the independent and dependent variables used in the research model. The third and the fourth part will analyze the application scenarios of the XGboost model and LightGBM model in predicting housing prices. The fifth part will summarize and make future prospects.

2. Independent & Dependent Variables

This article mainly uses region, house type, floor, orientation, and transportation (surrounding facilities) as the independent variables for the study. In the scenario of the XGboost model, the decision cost is used as a measure for predicting housing prices in the Chengdu region, and the application experience of multiple linear regression, decision tree, and XGboost model is compared; Compare the Mean squared error of four basic models of extra tree, Random forest, GBDT, and XGboost for the housing price forecast in California; By comparing Kuala Lumpur's advanced XGboost, LightGBM and traditional multiple regression analysis, ridge regression uses average absolute error (MAE), root mean square error (RMSE) and adjusted as the measurement indicators, the advantages and disadvantages of XGboost are obtained. In the LightGBM model, taking the real estate situation in Japan as an example for analysis, the purchase contract price is used as the target variable, and the location of the house, surrounding transportation facilities, etc. are used as explanatory variables to analyze their impact.

DOI: 10.54254/2754-1169/46/20230317

3. Scenarios for the Application of XGboost Model

3.1. Introduction to the Model

XGboost (Extreme Gradient Boosting) is one of the boost algorithms and can be seen as a variant of the GBDT algorithm. XGboost is incremental, where each step adds a tree to the previous one, and the new tree learns a new function to fit the residuals of the last prediction, which is used to fix the deficiencies of the previous tree [10]. The XGboost algorithm consists of an objective function that seeks to minimize the loss function and a regularization parameter that prevents overfitting. In addition to the regularization parameter, XGboost has two features that prevent overfitting. In addition, XGboost optimizes the loss function by using the second order gradient of the loss function to speed up the optimization [11].

3.2. Advantages of Models in Forecasting House Prices

XGboost is known for its flexibility, performance and speed, which is fast compared to other models. This model is best suited for tabular and structured datasets, and it works best for classification and regression models. XGboost has parallel processing functions, making it at least 10 times faster than other models for trees. It avoids over-fitting through regularization and can handles missing values internally. Zhen et al. used multiple linear regression, decision tree and XGboost model, and screened out 10 factors affecting house prices to fit the second-hand house prices in Chengdu. Using decision coefficient as the evaluation metric, the final scores of the three models were found to be 0.848, 0.896 and 0.9251 respectively, with the XGboost model being the most effective prediction among the three models. From this, it is concluded that XGboost has good classification and regression results, and the performance of the training algorithm is also very high because its basic principle is based on gradients to promote trees, which can be better adapted to unbalanced data sets. Moreover, the model is not easy to over-fitted, has better generalization ability, and also performs well for many non-linear regression problems. Thus, the model has a very wide range of use. Not all factors affecting second-hand property prices are linear, hence XGboost has a great advantage over the other two models in dealing with such unbalanced data [10].

Yang et al. chose the extra tree, random forest, GBDT, and XGboost models, and they used California housing price data to train and test these four models [12]. Among them, random forest is able to handle high-dimensional data without feature selection, which can obtain unbiased estimates of the internally generated error, and has good generalization ability. However, random forests can overfit on some classification or regression problems that appear noisy. The extra tree is similar to a random forest in that some features are randomly selected to build a tree. It directly uses the training set data to build a random tree, modifying the way bagging is done. Therefore, when the data set is noisy or large, it can perform better than a normal random forest. GBDT is more generalizable and its core is made up of regression trees, so it is mostly used for regression prediction. However, GBDT is not suitable for high-dimensional features. In comparison, XGboost is a more efficient way to train the model and can obtain better prediction results. This paper uses MSE as a measure. As shown in Table.1, XGboost has the lowest MSE. Therefore, XGboost is able to provide the most accurate prediction results [12].

DOI: 10.54254/2754-1169/46/20230317

Table 1: Mean squared error of the four base models.

Predict Model	Mean Square Error
Extra Tree	44216.900081
Random Forest	44625.093537
Gradient Boosting	43764.930335
XGboosting	43279.231065

Table 2: Comparison of the results of the four models.

Measures / Predict Model	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Adjusted R^2
XGboost	0.148	0.197	0.911
LightGBM	0.161	0.210	0.898
Multiple regression Analysis	0.181	0.238	0.869
Ridge Regression	0.195	0.252	0.853

In addition, Shuzlina et al. compared two state-of-the-art advanced machine learning algorithms, XGboost and LightGBM, with two traditional algorithms, multiple regression analysis and ridge regression, using the secondary dataset of "Kuala Lumpur Real Estate Listings" [13]. Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and adjusted R^2 were chosen as the measures, the final results are shown in Table. 2. Ultimately, it was found that XGboost had better fitting and predictive power, producing more consistent and reasonable results and achieving better predictive power than other models used for housing market data, with the lowest MAE and RMSE and the adjusted R^2 closest to 1, which indicated that the XGboost model was the most accurate model.

Rolli used the gradient boosting model XGboost to analyse real estate prices in three counties of California (Los Angeles, Ventura and Orange) and the advantages of the XGboost model can be summarized as follow: (1) XGboost uses a second order gradient of the loss function, which speeds up the optimisation; (2) XGboost transforms the loss function into a complex objective function with regularisation terms, and this new extension of the transformed loss function can add a new decision tree to the model; (3) XGboost also runs much faster due to its use of caching and memory optimization [11].

3.3. Limitations of XGboost in Predicting House Prices

Although XGboost obtains the smallest prediction error, the model requires more computational time due to the complexity of the parameters. Although the GBDT model has a larger prediction error than XGboost, if requires much less computational time than XGboost [13]. In addition, XGboost uses a horizontal or deep growth strategy, which makes it computationally expensive [14]. Besides, XGBoost cannot capture a linear relationship and extend it beyond the training set compare to the linear regression. That is, if there is a linear relationship between X and Y, the linear model is able to learn that linear relationship and predict the value of Y. However, the XGboost model does not assume a linear relationship between the individual features in the dataset, but instead looks for non-linear relationships between the features by constructing decision trees and ultimately outputs the expectation of the conditional distribution. Furthermore, XGBoost cannot capture the interactions between the independent variables compare with neural networks. If there is an interaction between the independent variables, for example, the ratio of X1 and X2 is a fixed value, XGboost is unable to identify this interaction, but instead mechanically splits each feature using the

DOI: 10.54254/2754-1169/46/20230317

data splitting rule. Therefore, whenever the prediction set has a different distribution of features than the original training set, there is a risk of inaccurate prediction.

4. Scenarios for the Application of LightGBM Model

4.1. Model Introduction

LightGBM is a gradient boosting framework that uses tree based learning algorithms. This is one of the most popular types of algorithms among Kaggle2 gold medal winners. The main difference is that while other algorithms use the "LevelWise" strategy to develop trees, LightGBM uses the "LeafWise" strategy (as shown in Fig. 1), resulting in shorter training time and higher accuracy of the model. In addition, since it arranges the training data hierarchically based on the feature histogram, there is theoretically no need to calculate the best points of the tree branches, which reduces the computational cost even for large datasets. This is a very suit-able method for real estate investment if a large training dataset is available.

The primary distinction is that, whereas other algorithms create decision trees using the "LevelWise" method, LightGBM employs the "LeafWise" technique (as illustrated in Fig. 1), which reduces training time and increases model accuracy. Additionally, since the training data is organized hierarchically based on the feature histogram, it is potentially unnecessary to determine the best locations for the tree branches, which lowers the computational cost even for sizable datasets. This is a very suitable method for real estate investment if a large training dataset is available [15, 16].

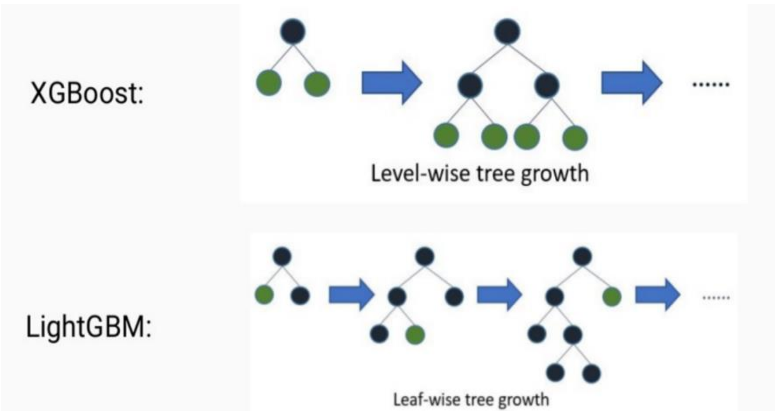

Figure 1: Growth of LightGBM tree.

4.2. Model Dataset & Parameters

The model uses a publicly available dataset of Japanese real estate transactions that contains a large amount of data on real estate contract prices (e.g., Table 3). By learning and forecasting and improving the variables, the model enables real estate market participants and decision makers to make more accurate decisions. This model has a number of parameters. Some of the qualities are as follows. The boosting type ='gbdt' is a gradient boosting decision tree. The goal 'binary' denotes that this technique will be used for binary classification. "n_estimators" this parameter determines the amount of change in the estimates 'feaure_fraction' the size of the change in the estimations is determined by this parameter. 'bagging_fraction' this parameter determines the size of the change in the estimations. learning rate = '0.07': this parameter controls the magnitude of the change in the estimations. The entire number of leaves in the tree is represented by num_leaves = 128.

Table 3: Japanese real estate contract price data set.

Attribute name	Data example	Type	Role
Contract price	40 million yen	Numerical	Objective variable
Location	1-Chome Minato, Minato, Tokyo	Categorical	Explanatory variable
Distance from station	10 minutes	Numerical	Explanatory variable
Date of construction	Jan, 2000	Date	Explanatory variable
Date of information posted	March, 2018	Date	Explanatory variable
Parking lot	Yes	Categorical	Explanatory variable

4.3. Model Advantages and Disadvantages

LightGBM has several advantages over other house price prediction methods. First, LightGBM is easier to handle non-numerical property data than neural net-work methods because it eliminates the complexity of pre-processing and trans-formation. Second, LightGBM offers better performance, faster training, and higher accuracy when dealing with large datasets compared to XGboost methods. In addition, LightGBM does not require image data of objects, so predictions can be made with a minimum amount of information.

However, it should be noted that LightGBM has some limitations, such as higher risk of overfitting with small sample sizes, increased sensitivity and susceptibility in the case of noisy datasets. As well as for unbalanced category datasets, the default objective function of LightGBM is based on quadratic loss, which may lead to poor prediction results for some categories. If the categories are unbalanced, they must be adjusted using an appropriate objective function (e.g., log loss or exponential loss). Therefore, in practical applications, the user must choose the appropriate algorithm according to the specific problem and the characteristics of the data set.

5. Conclusion

In summary, real estate is the pillar industry of a country, and has a close relationship with residents' life. The rapid development of Internet and network information technology has brought many opportunities to the society, and people have put forward higher requirements for the efficiency and accuracy of using information technology to predict housing prices. The emergence of machine learning meets this need. Among them, XGBoost and LightGBM are two very popular machine learning algorithms. They are two powerful machine learning algorithms known for their flexibility, efficiency, and speed, performing extremely well on both tabular and structured data sets. Both models can achieve good results in both classification and regression problems. But at the same time, they also have their own shortcomings. For example, XGboost has the defects of complex parameters, high calculation cost, and inability to deal with unstructured data. LightGBM is easy to overfit and sensitive to noisy data. Although XGBoost and LightGBM are local problems based on heuristic algorithms rather than global optimal solutions, we can further improve the prediction accuracy by doing more research and introducing other feature selection methods, e.g., Filter and Wrapper. With the development of The Times, with the promotion of big data technology, there are higher expectations for the quantity, efficiency and accuracy of data processing, and the future housing price prediction is more refined. In this case, residents have more simplified data to choose their own houses.

References

[1] Yin, Q., Shen, X., Xia, Y.: The application of machine learning in data mining in the context of Big data Digital Technology and Applications, 5, 21-23 (2022).

[2] Zhao, J., Bai, Z., Zhao, J.: Machine Learning data analysis and processing methods for Big data technology Shanxi Electronic Technology, 3, 9-11+17 (2022).
[3] Tan, C., Zhou, X., Zhu Y.: Research on Data Mining Methods Based on Weka and Collaborative Machine Learning Technology Journal of Changchun University 12, 5-9 (2020).
[4] Pan, Z.: Discuss the application and development of machine learning in the era of Big data Electronic Components and Information Technology, 4, 66-69 (2022).
[5] Li, T., Xu, C., Cao, L., Wang Y.: Sales volume prediction based on BP neural network China New Communications, 1, 137 (2020).
[6] Zhang, J., Du, J.: A housing price prediction model based on XGBoost and multiple machine learning methods Modern Information Technology 10, 15-18 (2020).
[7] Shen, J., Zhao, X.: Research on Data Resource Value Evaluation Method Based on Dynamic Stacked GBDT Algorithm Research on Science and Technology Management, 1, 53-61 (2023).
[8] Chen, T., He, T., Benesty, M., et al.: Xgboost: extreme gradient boosting R package version 0.4-2, 1(4), 1-4 (2021).
[9] Ke, G., Meng, Q., Finley, T., et al.: Lightgbm: A highly effective gradient boosting decision tree Advances in neural information processing systems, 30 (2017).
[10] Peng, Z., Huang, Q., Han, Y.: Model research on forecast of second-hand house price in Chengdu based on XGboost algorithm. In 2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT), 168-172, (2019, October).
[11] Rolli, C. S.: Zillow Home Value Prediction (Zestimate) By Using XGBoost (2020).
[12] Cao, B., Yang, B.: Research on ensemble learning-based housing price prediction model. Big Geospatial Data and Data Science, 1(1), 1-8, (2018).
[13] Rampini, L., Re Cecconi, F.: Artificial intelligence algorithms to predict Italian real estate market prices. Journal of Property Investment & Finance, 40(6), 588-611 (2022).
[14] Sibindi, R., Mwangi, R. W., Waititu, A. G.: A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. Engineering Reports, e12599 (2022).
[15] Li, T., Akiyama, T., Wei, L.: Constructing a highly accurate price prediction model in real estate investment using LightGBM. In 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR), 273-276 (2021).
[16] John, L., Shinde, R., Shaikh, S., Ashar, D.: Predicting House Prices using Machine Learning and LightGBM. SSRN, 8, 1 (2022).