Machine Learning-based Models for House Price Prediction in Provincial Administrative Regions of China

Xiafei Ding[1,*,†] Weiya Wang[2,†] Yiqian Zhang[3,†] Xiaoyuqian Zhong[4,†]

[1] *Shandong University, Shandong, China*
[2] *Renmin University of China, Beijing, China*
[3] *Hubei University, Hubei, China*
[4] *American University, USA*
[*] *Corresponding author. Email: 201900820057@mail.sdu.edu.cn*
[†] *These authors contributed equally.*

ABSTRACT

House price is an intractable problem with numerous influence factors. In this paper, boosting and traditional algorithms are compared to screen out the optimal model for house price prediction in provincial administrative regions of China. Based on provincial house price data in China from 2000 to 2019, the data is preprocessed by doing statistical analysis, dealing the missing values as well as choosing characteristic features for analysis. Then the data is imported into different models, comparing the prediction effects to pick out the best and then optimizing the hyper-parameters. Using mean squared error (RMSE), root mean absolute percentage error (MAPE), R-square (R^2), and explained variance score (EV) as evaluation indicators to appraise the models, the result presents that CatBoost is better than any other models, whose MAPE is 12.5%, R^2 is 87.81% and EV is 90.5%. Then sub-sample test is used to examine the robustness, whose result shows that CatBoost is always effective. The empirical findings mainly show that CatBoost is effective in predicting house prices with complex variables and the feature importance graph generated by CatBoost presents that demand and macro environment factors can explain the major fluctuation of house price and that in macro environment factors, macro-economic and education indicators are obviously important than other macro indicators.

Keywords: *Machine learning, House price prediction, Regression analysis.*

1. INTRODUCTION

Sun Yat-sen once put forward the people's livelihood ideology of "land for the tiller and house for the householder", which has been an ideal of Chinese. Owning a house is very important to most families in China. On the one hand, house has become the largest component of household wealth; On the other hand, "living and working in peace and contentment" has long been respected [1]. The real estate industry is also important to society. Booms and busts in the housing market have historically been bad for financial stability and real economy. Many biggest events in the banking crisis were linked to the boom and bust cycle of property prices. Research by the International Monetary Fund shows that more than two-thirds of the nearly 50 systemic banking crises in recent decades were preceded by boom-bust patterns in house prices [2].

Therefore, accurate prediction of house prices is of great significance to society. The efficient allocation of house prices through forecast models is essential to financial markets to help avoid risks and make investments in real estate viable [3]. For local governments, they can adjust the market effectively and achieve stable growth of fiscal revenue. For developers, investment direction can be determined to maximize their profits [4].

At present, domestic discussion on influence factors of house price mainly focuses on supply factor, demand factor and other macro factors. Zhang and Fan found that income is the main factor driving house price, and the

Copyright © 2022 The Authors. Published by Atlantis Press International B.V.
This is an open access article distributed under the CC BY-NC 4.0 license -http://creativecommons.org/licenses/by-nc/4.0/.

higher of house price, the larger of the effect of income while the smaller of the effect of construction cost and per capita GDP [5]. 35 large and medium cities' empirical results also show that due to more benefits associated, such as better living facilities, more concentrated social resources in medical care and education, etc., more and more people are attracted to the first-tier cities, leading to the increase of house price [6]. On the supply side, once real estate development enterprises purchase land at a higher price than normal, they will try to transfer the extra cost to buyers by extending the development period to find the house price peak or forcing up the house price by virtue of the monopoly position, so as to protect their own return rate [7]. Through multiple regression analysis, Yang and Li found that 7 indicators, such as house completion area, per capita disposable income, land price, and consumer price index, are valid to explain the changes in house prices in 35 cities from 2003 to 2017 [8].

Traditional house price forecasting models focus on factors such as property size, location and amenities. And traditional house price forecasting models mainly use regression techniques, such as multiple linear regression (least square), Lasso and Ridge regression models, and support vector regression [3]. The disadvantage of Lasso regression is that if two or more highly collinear variables are present at the same time, it randomly selects one of them. There are some variables with high collinearity in the data we used which is shown in Fig 1. The least Square's disadvantage is that it does not work with censored data and is very sensitive to the choice of starting values. The disadvantage of support vector regression is that it does not perform well when faced with large data sets because of high training time required [9]. In this study, the CatBoost model consistently showed the best predictive performance, superior to other popular traditional machine learning models and other Boost models. It has a fast inference, its propulsion scheme helps reduce overfitting and improve model quality, and it also supports complex classification features [10].

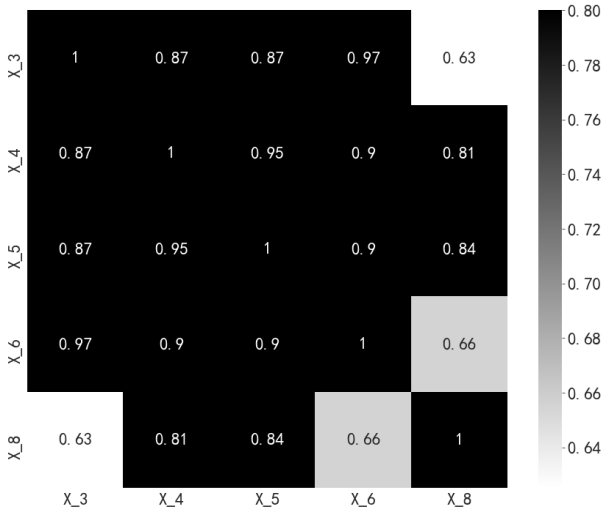

Figure 1 Pearson correlation coefficient heat map

In this paper, the variables are selected and screened by referring to relevant research, and house price data of 31 provincial administrative regions in China from 2006 to 2019 are analyzed statistically. Using boosting models and traditional models, the explanatory variables of previous year are used to predict house prices in the next year, and test sets and training sets are divided in a 10:3 ratio. The performance of model prediction is evaluated by R^2, RMSE, MAPE, and EV, and the stability of the results is ensured by subsample analysis. Contribution of features is examined by CatBoost, which showed best performance. Finally, house price in 2019 are predicted, and results before and after adjustment are compared. The research conclusions are that CatBoost can efficiently process more explanatory variables and predict more accurately, and that demand and macro environment factors can explain the major fluctuation of house price. So, in subsequent research and house price control, the factors that really affect effective demand are needed more attention.

This paper is divided into following sections: Section 2 describes data and variables. Section 3 explains methodology. Section 4 concludes empirical results. Section 5 sums up the work.

2. DATA AND VARIABLES

2.1. Data description

The data shown in Table I are scraped from National Bureau of Statistics, which include data from 31 provinces of China and range from 2006 to 2019.

Table 1. Variables

No.	Variables	Category
X_1	completed house cost of real estate development enterprises (¥/m²)	supply cost

X_2	house completion area (10,000 m²)	supply volume
X_3	actual funds in place of real estate development enterprises this year (100 million ¥)icp	supply ability
X_4	paid-in capital of real estate development enterprises (100 million ¥)	
X_5	owner's equity of real estate development enterprises (100 million ¥)	
X_6	main business income of real estate development enterprises (100 million ¥)	
X_7	per capita disposable income (¥/person) (urban residential disposable income (2004 to 2012) residential disposable income (2013 to 2017))	income level
X_8	residential consumption level (¥)	consumption level
X_9	permanent resident population at year-end (10,000)	population
X_10	per capita GDP (¥/person)	macroeconomic
X_11	registered urban unemployment rate (%)	
X_12	number of regular institutions of higher education	education
X_13	number of regular primary schools	
X_14	number of regular senior high schools	
X_15	number of medical and health institutions	medical service
X_16	number of health technicians per 10,000 person	
X_17	completed investment in industrial pollution control (10,000 ¥)	environment
X_18	district number	region
Y	average selling price of residential commercial houses (¥/m²) (2007 to 2019)	target variable
Y_1	house price to income ratio(Y/X7, 2013-2019)	

2.2. Descriptive statistics

Some representative features are selected to describe the statistics which are shown in Table II.

Table 2. Statistic description

statistic	X_7	X_2	X_17	Y
Count	496	496	496	495
Mean	19112.81	10072.18	40264.82	5539.76
Std	9580.92	13492.42	27025.94	4753.92
Min	7217.87	84	4317	1158
25%	12865.46	2413.32	19842.75	2822.5
50%	17491.44	5759.86	35187	4330
75%	22569.25	12618.2	51813.5	6391.43
max	69442	77823.4	164220	38433

By observing the standard deviation, maximum and minimum values of the data, it can be seen that the dispersion degree of this group of data is very large and unstable. The preliminary judgment is the result of the wide variation between different provinces.

2.2.1. Per capita disposable income

The per capita disposable income grew steadily and accelerated from 2004 to 2012, the variance was small and the increase was slow. During this period, the data set was urban resident disposable income, but since 2013, it has been changed to resident disposable income, increasing the non-urban data, so there is a breakpoint which is shown in Fig.2 From 2013 to 2019, it still showed an increasing trend. The standard deviation increased significantly compared with that from 2004 to 2012, indicating that the sample difference increased.

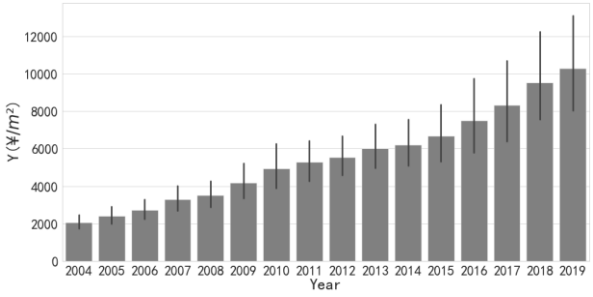

Figure 2 Average selling price of residential commercial houses.

Average selling price of residential commercial houses.

Notes: The black lines on the chart are the error lines, representing the standard deviation of this data set.

2.2.2. Target variable

As shown in Fig.3, the average selling price of residential commercial houses of 31 Provinces in China showed a continuous upward trend from 2004 to 2019. The variance also increases with time, indicating that the average house price difference among provinces is increasing year by year.

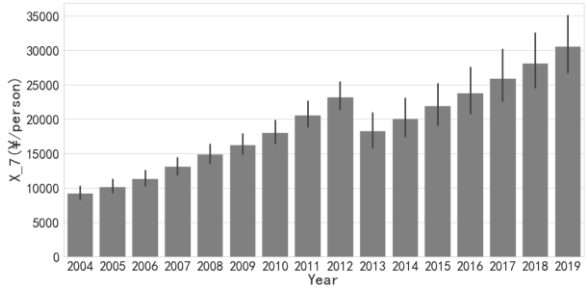

Figure 3 Average selling price of residential commercial houses.

2.2.3. Regional differences

Ranking the house price to income ratio from 2013 to 2019, which is after the breakpoint. It was found that Beijing had the highest data each year, while Inner Mongolia had the lowest. However, Tibet, as an economically underdeveloped region, ranked fourth.

As shown in Table III, Beijing, as the capital, has a large population density and insufficient land supply, resulting in the transaction of house price at a premium. Inner Mongolia is vast and sparsely populated, with sufficient land supply. In addition, the urbanization level is low, the degree of economic aggregation is low, and the influx of capital leads to the low house price [11].

However, Tibet is still in the development stage, and its economic development is extremely weak, resulting in low per capita disposable income. However, the lagging development of the real estate market, the serious lack of market mechanism, the chaotic structure of the main body of the market, and insufficient market competition lead to a serious house price premium [12].

Table 3. Reginal differences

Region	X_7	Y	Y_1
Beijing	53379.14	28158.86	0.52
Inner Mongolia	24404.57	4482.57	0.18
Tibet	14086.71	5365.14	0.38

Notes: Data used are averages from 2013 to 2019.

As shown in Fig.4, with the development of Tibet in recent years, the price-to-income ratio tends to be stable and starts to decline in 2019.

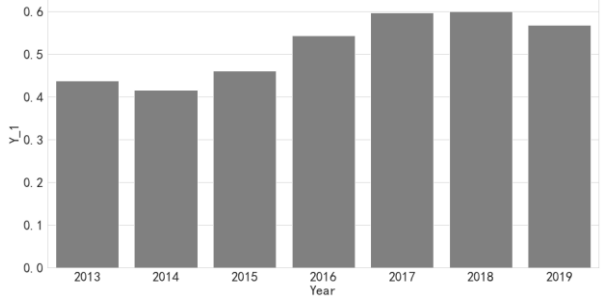

Figure 4 Tibet's house price to income ratio.

2.3. Data Preprocessing

The variables of previous year are used to predict house price in current year. And 3 variables have missing values. For X_17, missing values of 2018 in each province are filled with the average values from 2004 to 2017 of each province for no significant trend over time. Because of obvious increasing trend of X_16 over years, we split the differences of each province between 2004 and 2008 and fill the years from 2005 to 2007, so that values from 2004 to 2008 form an arithmetic sequence. The data of X_8 in 2018 and 2019 are missing, so they are moved two years later in the regression, for example: data in 2016 are used to predict house price in 2019. Finally, few missing values are dropped.

3. METHODOLOGY

There are three existing state-of-the-art implementations of gradient boosted decision trees — eXtreme Gradient Boosting （XGBoost）, LightGBM and CatBoost. In this paper, we used CatBoost, which outperforms the other two gradient boosted decision trees. Gradient boosting is a method to build strong predictors by iteratively combining weaker models (base predictors) in a greedy fashion. It is essentially a process of constructing an ensemble predictor by performing gradient descent in a functional space [13].In addition, we also compared the results with some traditional models like RandomForest, BayesianRidge, LASSO, ElasticNet and GAM.

3.1. Model Description

XGBoost was issued by T. Chen and C. Guestrin in 2016, which is a scalable machine learning system for end-to-end tree boosting and supports parallel computing [14].Issued by Q. Meng in 2017, LightGBM is based on histogram, adopting the generation strategy of splitting by leaf to split the nodes of the sub-model of decision tree [15].

Different from the other two, CatBoost uses binary decision trees as base predictors. The two main algorithmic features of CatBoost are orderly promotion with permutation boosting as well as a novel algorithm to deal with classification variables [16].

3.1.1. GBDT

Gradient Boosting Decision Tree (GBDT) is the algorithm obtaining a strong learner through serial iteration of a group of classifiers to carry out higher precision classification [17]. It uses forward stagewise algorithm and weak classifier uses Classification and Regression Trees (CART) [18].

A gradient boosting procedure builds iteratively a sequence of approximations $F^t: \mathbb{R}^m \to \mathbb{R}, t = 0,1, \ldots$ in a greedy fashion [19]. Assume that the strong classifier obtained in the previous iteration is $F^{t-1}(x)$ and the loss function is $L\left(y, F^{t-1}(x)\right)$, the purpose of this iteration is to find a weak classifier h^t of CART and minimize the loss function of this iteration. Equation (1) represents the objective function h^t of this iteration.

$$h^t = arg\min_{h \in H} \mathbb{E}(-g^t(x,y) - h(x))^2 \tag{1}$$

$GBDT$ uses the negative gradient of the loss function to fit the approximate value of the loss of each round. In equation (2), $g^t(x,y)$ represents the above gradient.

$$g^t(x,y) = \frac{\partial L(y,s)}{\partial s} \Big|_{s=F^{t-1}(x)} \tag{2}$$

Usually, h_t is approximately fitted by equation (3).

$$h^t = arg\min_{h \in H} \mathbb{E}(L(y, F^{t-1}(x) + h(x))) \tag{3}$$

Finally, the strong classifier of this iteration is obtained, as shown in equation (4).

$$F^t(x) = F^{t-1}(x) + h^t \tag{4}$$

3.1.2. CatBoost

In each iteration of GDBT, the loss function uses the same data set to get the gradient of the current model and then trains the base predictor. However, this will generate gradient estimation bias, which lead to overfitting. CatBoost replaces the gradient estimation method in the traditional algorithm by ordering boosting, thereby reducing the bias of gradient estimation and improving the generalization ability of the model. The procedure is as follows.

Firstly, CatBoost generate a random permutation σ of the training examples. Secondly, it maintains n different supporting models $M_1, \ldots, M_n$ that the model M_i is learned using only the first i examples in the permutation. Finally, at each step of the iteration, model M_{j-1} is used to obtain the residual for j-th sample.

Ordered boosting algorithm process is shown in Fig.5.

Algorithm 1: Ordered boosting

input : $\{(\mathbf{x}_k, y_k)\}_{k=1}^n, I;$
$\sigma \leftarrow$ random permutation of $[1, n]$;
$M_i \leftarrow 0$ for $i = 1..n;$
for $t \leftarrow 1$ **to** I **do**
  for $i \leftarrow 1$ **to** n **do**
    $r_i \leftarrow y_i - M_{\sigma(i)-1}(\mathbf{x}_i);$
  for $i \leftarrow 1$ **to** n **do**
    $\Delta M \leftarrow$
     $LearnModel((\mathbf{x}_j, r_j) :$
     $\sigma(j) \leq i);$
    $M_i \leftarrow M_i + \Delta M$;
return M_n

Figure 5 Ordered boosting algorithm process. Notes: Cite [16].

3.1.3. Feature importance

The feature importance ShapValues is a vector v with contributions of each feature to the prediction for every input object and the expected value of the model prediction for the object [20].

v_i is the contribution of the i-th feature. It is calculated as follows for each feature i.

$$v_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \tag{5}$$

Notes: M is the number of input features, N is the set of all input features, S is the set of non-zero feature indices (the features that are being observed and not unknown), $f_x(S) = \mathbb{E}[f(x) \mid x_s]$ is the model's prediction for the input x, where $\mathbb{E}[f(x) \mid x_s]$ is the expected value of the function conditioned on a subset S of the input features.

$v_{feature\_count}$ is the expected value of the model prediction.

For a given object the $\sum_{i=0}^{feature\_count} v_i$ is equal to the prediction on this object.

3.2. Evaluation Indicators

It is important to evaluate the performance, such as model accuracy and training time, of the prediction model. This paper chooses four indicators shown as follows. *Note*: y_i symbolizes the original values of 'average selling price of residential commercial housing', while y_i^{pred} is its predicted values and $\bar{y}_i$ is its average value.

3.2.1. R-square (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \quad (6)$$

R^2 illustrates how the variance of y_i explained by the y_i^{pred}. The closer R^2 gets to 1, the better the model is.

3.2.2. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{pred} - y_i\right)^2}{t}} \quad (7)$$

RMSE represents the expected value of error squared, which can reflect the degree of dispersion of a data set. The smaller it is, the better the model is.

3.2.3. Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i^{pred} - y_i}{y_i}\right| \times 100\% \quad (8)$$

The closer MAPE gets to 0, the better the model is.

3.2.4. Explained Variance Score (EV)

$$EV = 1 - \frac{Var\{y_i - y_i^{pred}\}}{Var\{y_i\}} \quad (9)$$

The closer MAPE gets to 1, the better the model is. It's the same thing as R^2 when the mean of the residuals is 0.

4. EMPIRICAL RESULTS

4.1. Model Selection

The data were processed from 2006 to 2018, and were grouped in a 10:3 ratio. The eight basic models use features of the previous year to predict average selling price of residential commercial houses in the next year. The prediction effect of the models ranked from small to large according to MAPE is shown in Table IV below.

Table 4. Perdition Effect

Models	RMSE	MAPE	EV	R^2	time
CatBoost	2410.03	12.5%	90.5%	87.81%	3.26
LinearGAM	2042.92	16.35%	91.73%	91.25%	0.23
RandomForest	2972.55	17.88%	82.15%	81.47%	0.27
XGBoost	2571.47	19.07%	87.16%	86.13%	0.12
ElasticNet	2921.44	19.97%	82.13%	82.1%	0.01
Lasso	2825.12	20.63%	83.27%	83.26%	0.01
LightGBM	3958.36	21.7%	69.17%	67.13%	0.06
BayesianRidge	3025.65	23.89%	81.7%	80.8%	0.03

Notes: Train data set is 2006 to 2015, test data set is 2016 to 2018.

As the Table 4 presents, CatBoost has the best prediction effect, while LinearGAM is the model with the highest fitting degree.

4.2. Subsample analysis

It can be seen from II B that the per capita disposable income showed a breakpoint in 2013. Therefore, multiple time segments were considered for CatBoost basic model robustness test. The first three paragraphs were segmented around 2013 using Sub-sample regression method [21]. The last two periods are grouped in 2006-2012 and 2013-2018 with a shortened time window [22].

Table 5. CatBoost basic model Segmented inspection

Periods	RMSE	MAPE	EV	R^2
Period1	1563.17	11.47%	93.28%	93%
Period2	1826.4	10.8%	93.05%	91.89%
Period3	1857.63	11.65%	90.04%	89%
Period4	1172.91	12.3%	88.38%	88.25%
Period5	2411.7	11.89%	91.52%	88.6%

Notes: Train data used for Period1 are 2006-2010 & 2013-2016, while test data are 2011-2012 & 2017-2018; Train data used for period2 are 2006-2012 & 2014-2016, while test data are 2013 & 2017-201; Train data used for period3 are 2006-2009 & 2013-2015, while test data are 2010-2012 & 2016-2018. Train data used for Period4 are 2006-2011, while test data are 2012; Train data used for period5 are 2013-2016, while test data are 2017-2018.

It can be seen from Table V that under the two stability test methods, the changes of each index are not obvious, so the results are relatively stable. The breakpoint of per capita disposable income in 2013 has no significant effect on the model.

4.3. Feature Importance

There are many independent variables used in the model, hence the contribution degree of each feature is a problem that must be considered. CatBoost is used to rank the contribution of each feature and the result has been shown in Fig.6.

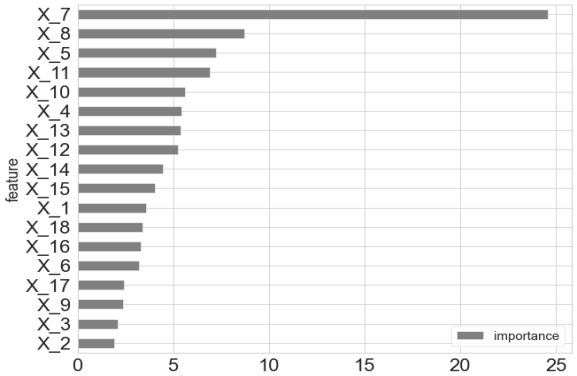

Figure 6 Contribution of each feature. Notes: Train data set is 2006 to 2017, test data set is 2018.

As Fig. 4 presents, the most important factor is

economic fundamentals, which consistent with the results of previous studies [24]. Per capita disposable income is the feature with the highest contribution, accounting for approximately 25%. Resident consumption level has the second highest contribution. The growth of the above two increases the purchasing power of residents, stimulates the growth of market demand, and the house price rises accordingly [25].

4.4. Prediction results

The train data used in the models below are from 2006 to 2017, and data of 2018 are used to predict the effect of 2019 house prices.

4.4.1. Hyper-parameter optimization

The CatBoost's built-in auto-adjust function Randomized Search is used, which gave the best set of Hyper-parameter after learning shown on Fig.7. It can be seen that the MAPE value of the best parameter combination given in auto-callback is as low as 7.8% on the test set. The parameter set tested was recommended by the CatBoost website [20].

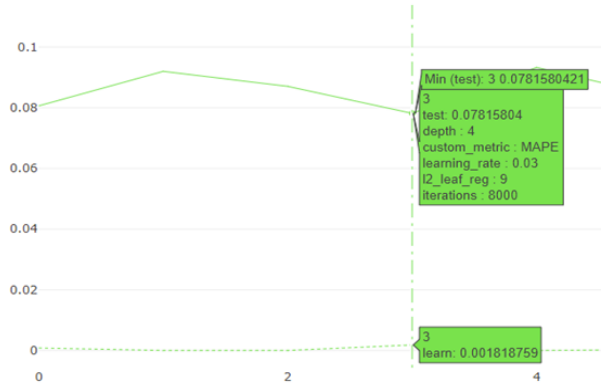

Figure 7 MAPE adjustment.

Notes: The Hyper-parameters sets we tried are as below: iterations: [6000, 7000, 8000]; learning rate: [0.03, 0.1]; depth: [4, 6, 10]; leaf reg: [1, 3, 5, 7, 9]

4.4.2. Prediction power comparison

The RMSE and R^2 values of the adjusted model are better than those of the base model, but the MAPE is much higher(Table VI). Scatter plots (Fig.8) of predicted and true values respectively are drawn. The difference is small after adjustment. The adjusted model has a higher R^2, EV, smaller RMSE, while the basic model has a smaller MAPE. The prediction ability of the CatBoost model is great.

Table 6. Comparison of parameter tuning models

Models	RMSE	MAPE	EV	R^2
Basic	1476.18	8.28%	96.53%	95.94%
Hyper	1334.9	9.29%	96.95%	96.68%

Notes: Basic stands for basic CatBoost model while Hyper stands for the CatBoost model after auto-tuning.

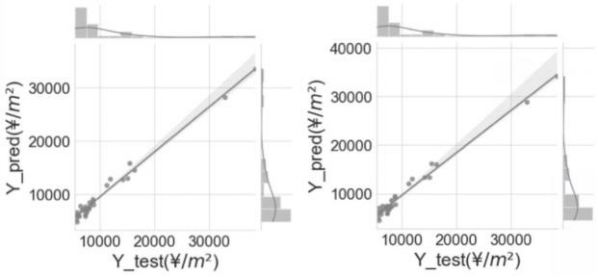

Figure 8 Predicted and true values respectively.

Notes: The abscissa represents the real value of house prices, and the ordinate represents the predicted value of house prices. The basic model is on the left, and the adjusted model is on the right.

5. CONCLUSION

This paper uses machine learning to predict house prices in a given year in an area more accurately. To select out the optimal prediction model, 17 explanatory variables in demand, supply, and macro environment aspects are considered and data from 2006 to 2019 from 31 provincial-level administrative regions in China Statistical Yearbook are used to train. The final results are evaluated by several indicators, including MAPE, RMSE and R^2. Empirical results show that CatBoost has the best prediction effect, with average MAPE less than 10%, while LinearGAM is the model with the highest fitting degree, with R^2 reaching 92.87%. To ensure the stability, 5 subsample tests are conducted and MAPE fluctuates around 12%, which show that CatBoost is always effective.

In general, this research has two major conclusions. Firstly, CatBoost model is effective in predicting house price. Compared to previous research, typically only 5-7 explanatory variables are measured, but in this study, CatBoost efficiently processed nearly 20 explanatory variables. Secondly, the feature importance graph generated by CatBoost model shows that demand and macro environment factors can explain the major fluctuation of house price and that in macro environment factors, macro-economic and education indicators are obviously important than other macro indicators. Because in China, high-speed urbanization and enrollment expansion of colleges make the large and medium-sized cities constantly faced with a large number of housing demand brought by migrants. Though high house-price-to-income ratios are formed in some cites because too many people chase few houses, the result of population migration to big cities will not change, as long as the economic benefits brought by

population aggregation are higher. And high-income families are increasingly willing to invest in the education of their children. These results indicate that demand plays a decisive role in real estate price compared with supply, so the regulation policy need to pay more attention to the factors that really affect effective demand, such as disposable income and education. Residents can also make wiser and more appropriate development decisions through the house-price-to-income ratio.

Though we do several robustness tests, there are some concerns about overfitting. Because the number of provincial districts is too low, though the data is from 2006 to 2019, the total amount of data to train and test is still not enough. For machine learning method, using data at municipal level to do further research can make model of better generalization effect and do more accurate prediction.

AUTHORS' CONTRIBUTIONS

W. Wang conducted the research; X. Ding constructed the model; Y. Zhang analyzed the data; X. Zhong wrote the paper. All authors had approved the final version.

ACKNOWLEDGMENTS

The authors declare no conflict of interest.

REFERENCES

[1] Q. Yu, W. Sun, and S. Zheng, "Housing ownership rate and economic development level: Based on the empirical analysis of 31 provinces and regions in China," (translate in Chinese) China Real Estate, vol. 2014, no. 14.

[2] Housing markets. (2014, June 11). financial stability and the economy. IMF. [Online] Available: https://www.imf.org/en/News/Articles/2015/09/28/04/53/sp060514

[3] J. Manasa, R. Gupta, and N. S. Narahari, "Machine learning based predicting house prices using regression techniques," In 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), IEEE, pp. 624-630, March, 2020,.

[4] J. Schuetz. (2019). How can government make housing more affordable? [Online] Available: https://www.brookings.edu/policy2020/votervital/how-can-government-make-housing-more-affordable/.

[5] S. Zhang, and X. Fan, "The study of the influence of income and interest rate on house price based on panel quantile regression model," (translate in Chinese) Journal of Applied of Statistics and Management, vol. 34, no. 6, pp. 1057-1065, 2015.

[6] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.

[7] C. Li, and C. Zhang, "The formation of high house price to income ratio and its impact on Urban population agglomeration in China: Theory and Empirical Study," (translate in Chinese) Journal of South China Normal University (Social Science Edition), vol. 2015, no. 6, pp.116-123+191.

[8] J. Wu, H. Li, and B. Hu, "The study of the impact of land cost on the price of new commercial House," (translate in Chinese) Price: Theory & Practice, vol. 2015, no. 9, pp. 52-54, 2015.

[9] SVM | Support Vector Machine Algorithm in Machine Learning. (2021). Analytics Vidhya. [Online] Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.

[10] CatBoost, The Spirit of a True Racer. Datascience.foundation. (2021). Retrieved 21 November 2021, from https://datascience.foundation/datatalk/catboost-the-spirit-of-a-true-racer.

[11] H. Li, and F. He, "An empirical analysis of the impact of land finance and urbanization factors on housing prices——based on panel data of housing prices in China from 1999 to 2014," (translate in Chinese) PRICE: THEORY & PRACTICE, pp. 89-92, 2016.

[12] T. Wang, "Research on the regional differences in the level of real estate price bubbles——macro data from 31 inter-provincial units in China," (translate in Chinese) Review of investment Studies, vol. 37, no. 3, pp. 24-49, 2018.

[13] R. Chen, R. Caraka, and A. Piliang, et al., "An end to end of Scalable Tree Boosting System," Sylwan, vol. 164, no. 5, pp.140-151, 2020.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu., "LightGBM: a highly efficient gradient boosting decision tree," In Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA, pp. 3149–3157, 2017.

[16] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. "CatBoost: unbiased boosting with categorical features," In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18), Curran Associates Inc., Red Hook, NY, USA, pp. 6639–6649, 2018.

[17] H. Friedman, "Greedy function approximation: A Gradient Boosting Machine," (translate in Chinese) Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[18] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, "Classification and regression trees," Biometrics, vol. 40, no. 3, pp. 358, 1984

[19] F. Miao, Y. Li, C. Gao, M. Wang, and D. Li, "Diabetes prediction method based on CatBoost algorithm," (translate in Chinese) Computer Systems & Applications, vol. 28, no. 9, pp. 215-218, 2019.

[20] A. Gulin, "CatBoost," Yandex, [Online] Available: https://CatBoost.ai/en/docs/concepts/contacts]

[21] C. Lu, Z. Ma, Y. Han, and X. Guo, "House price data analysis based on support vector regression," (translate in Chinese) Journal of North China University of Science and Technology (Natural Science Edition), vol. 43, no. 4, pp.76-82,2021.

[22] Y. Hong, and H. Huang, "Theoretical Research on the Influencing Factors of House Prices," (translate in Chinese) China Economic & Trade Herald, pp. 72, 2010.

[23] W. Li, and K. Zhang, "The impact of air pollution on enterprise productivity: Evidence from Chinese industrial enterprises," (translate in Chinese) JOURNAL OF MANAGEMENT WORLD, vol. 35, no. 10, pp. 95-112+119, 2019.

[24] F. Ding. "Research on influencing factors and forecast of real estate price," (translate in Chinese) Ph.D. dissertation, Dept. Elect. Eng., Anhui University of Finance & Economics, 2014.

[25] X. Wang, and L. Bu, "International export trade and enterprise innovation——A quasi-natural experimental research based on the opening of "China-Europe Railway Express," (translate in Chinese) CHINA INDUSTRIAL ECNOMICS, pp. 80-98, 2019.