Forecast Analysis of Urban Housing Prices in China Based on Multiple Models

Zelin Qu[a]

School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, China

Keywords: Housing Prices Prediction, Machine Learning.

Abstract: Social attention has long been focused on the cost of housing in China. Rising urbanization and quick economic growth have made housing costs crucial to social stability and the standard of living for locals. This study compared the performance of four machine learning models - eXtreme Gradient Boosting (XGBoost), Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), and Long Short-Term Memory network (LSTM) - in detail in order to increase the accuracy of housing price prediction. The models were then combined with nominal GDP data to forecast housing price. The experimental findings demonstrate that the XGBoost model, which is used to forecast future home prices, performs well across a range of evaluation indices. In addition, this study predicts that housing prices in Chinese cities will show a slight upward trend in 2024-2025. This study fills the gap of the existing research in comparing and integrating multiple models and provides a reference for the government to make more accurate real estate policies and investors' decisions.

1 INTRODUCTION

China's housing price problem has always been a focus of social attention. The swift advancement of the economy and the quickening pace of urbanization have made housing costs a crucial element influencing social stability and the standard of living of inhabitants. China's housing market has been booming recently; the majority of residents bring a heavy burden with them, and housing expenses are still high. Therefore, in order to maintain social order and foster the healthy expansion of the real estate sector, it is imperative to investigate the trends and fluctuations in China's home prices. The changes in housing prices in different cities exhibit significant heterogeneity due to the different levels of economic development and policy orientation in different regions of China. As a result, it is challenging for a traditional single forecasting model to accurately predict the housing price trend in other regions. Existing housing price forecasting models usually use a single machine learning model, such as multiple linear regression, but these models often face the problem of insufficient forecasting accuracy. In addition, most studies do not profoundly explore the comparison and integration between different models, so it isn't easy to fully use the advantages of various models to deal with regional differences. In order to address this issue, this study will compare the effectiveness of four machine learning models - Support Vector Regression (SVR), Multi-Layer Perceptron (MLP), eXtreme Gradient Boosting (XGBoost), and Long Short-Term Memory network (LSTM)—in predicting housing prices. The goal is to identify the best model through experimental comparison. The final study shows that the XGBoost model performs best on various evaluation indicators and uses this model to forecast future housing prices. By comparing several models, this work not only closes the gap in the body of research on housing price forecasting but also makes an effort to expand the use of machine learning in economic forecasting. The research results will provide a reference for the government to make more accurate real estate policies and also offer a more reliable basis for real estate investors to make decisions. The study background and pertinent literature are reviewed in the second section of this paper. The data preprocessing procedure is presented in the third section. The fourth section introduces comparison and model selection approaches. The fifth section examines the findings of the prediction process and

[a] https://orcid.org/0009-0008-0410-4460

Qu, Z.
Forecast Analysis of Urban Housing Prices in China Based on Multiple Models.
DOI: 10.5220/0013214400004568
In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)*, pages 257-262
ISBN: 978-989-758-726-9

the experimental setup. The last section contains the overview and recommendations for future work.

2 RESEARCH AND LITERATURE REVIEW

2.1 Research Premise

At the Central Economic Work Conference held in Beijing from December 14 to 16, 2016, the "house is for living in, not for speculation" policy was proposed, clarifying the residential attributes of China's real estate and aimed to curb speculative property purchases. At that time, China's real estate market faced an out-of-control situation, and many speculators unquestioningly speculated on real estate. The proposal of this policy laid the foundation for the formulation of subsequent real estate control policies. Measures like purchase restrictions, lending restrictions, and price caps are just a few of the laws and guidelines that the Chinese and local governments have successfully implemented since 2016 to control and regulate housing prices nationwide. Many cities have steadily adopted the policies restricting purchases and loans over time. The deleveraging and monetary resettlement aspects of the shed reform have also been executed successfully. In 2020, the People's Bank of China and the Ministry of Housing and Urban-Rural Development jointly launched the "three red lines" policy. These regulations, which were put into place gradually, have successfully stopped the major cities' real estate markets from overheating. Apart from the execution of policies, China's economy has experienced a transition from rapid expansion to moderate expansion since 2016. As a result, there is now less demand for real estate on the market. However, a mismatch in the supply and demand structure has allowed several first and second tier cities to maintain the increase in real estate prices. In contrast, third - and fourth-tier cities have declined due to oversupply. Many studies often choose a single model for the existing research on Chinese housing price forecasting, ignoring the possibility of multi-model comparison and integration, resulting in insufficient forecasting accuracy. In addition, most existing studies focus on a city or a province and do not fully consider the impact of regional differences in housing price changes in different regions across the country. Therefore, this study focuses on the changes in housing prices in other areas in China and comprehensively considers the differences between various regions to analyze and forecast. To improve the accuracy and reliability of home price estimates, the research combines GDP and housing price data from multiple Chinese regions with a variety of machine learning models. Four models - XGBoost, SVR, MLP, and LSTM - are compared in order to identify which performs best in predicting the trend of property prices in various regions over the following few years. The findings of this study have the potential to be useful in regional planning, real estate investment, and policy formation. They also contribute to the improvement of the national housing price projection.

2.2 Literature Review

In the past, scholars have created many outstanding results in studying machine learning housing price prediction. The suitability of the random forest model for predicting housing prices was confirmed by Adetunji et al. (2020). According to Chen et al. (2021), Bayes, support vector machines, and backpropagation neural networks are better options for predicting home prices. According to Goel et al. (2023), the LSSVM model outperforms SVM, CNN, and other models in predicting home prices. Henriksson and Werlinder (2021) discovered that random forests, which take a long time to train and infer, fared worse on small and big data sets than XGBoost. After analysing and contrasting the SVM, random forest, and GBM algorithms, Ho et al. (2020) came to the conclusion that while SVM can yield remarkably accurate predictions, random forest and GBM perform better. The random forest model can fully capture the complexity and nonlinearity of the actual housing market, as demonstrated by the study by Hong et al. (2020), which shows that the average percentage deviation between the predicted and actual market prices is only 5.5%, and that the probability that the expected price is within 5% of the actual market price is 72%. In order to create a home price evaluation and prediction model based on factors influencing prices, Manasa et al. (2020) took inspiration from the multiple linear regression models, the Lasso regression model, the Ridge regression model, the support vector machine model, and the XGBoost model. By contrasting the model mistakes, they were able to choose the best model. Ming et al. (2020) investigated rent prices in Chengdu, China, using three machine learning models, and discovered that XGBoost was the most accurate in forecasting them. According to Sheng and Yu (2020), the comprehensive learning algorithm predicts more accurately than a linear regression model, and the

PSO-XGBoost model has the best effect overall. It also has the highest prediction accuracy. In their 2020 study, Truong et al. examined a variety of housing price prediction models, encompassing two regression techniques, stacking generalization regression and mixed regression, as well as three machine learning approaches: Random Forest, XGBoost, and LightGBM. They discovered that the mixed regression approach performs well, the random forest error is low but prone to overfitting, and stacking generalized regression is the best option for the greatest accuracy requirements—but at an excessively high time complexity. In order to forecast property prices in 100 cities over a ten-year period, Xu and Zhang (2021) investigated a basic neural network with just four delays and three hidden neurons. The neural network worked remarkably well.

3 DATA ACQUISITION AND PREPROCESSING

This paper adopts and processes the housing price data of various prefecture-level cities in China from 2016 to 2023 and introduces the corresponding annual nominal GDP data as an auxiliary variable. In most scholars' studies, housing price forecasting always uses existing housing prices to predict future prices. At the same time, the author introduces nominal GDP as an auxiliary variable for housing price forecasting. Nominal GDP reflects the overall development level of the local economy. Theoretically, the higher the nominal GDP, the better the economic development, and the higher the housing price. Therefore, introducing nominal GDP data may help improve model prediction accuracy, especially in cities with stable economic development or large fluctuations, where there may be a stronger influence of GDP fluctuations on housing prices. The original housing price data set came from anjuke.com, a Chinese property rental and sales service, while the original GDP data came from data.stats.gov.cn, a website run by the National Bureau of Statistics. The data sets used in this article and some actual data are shown in Table 1 and Table 2.

Table 1: Data labels and descriptions.

Label	Description
Province	The province where the data is located
City	Prefecture-level city corresponding to the data
Year	Year of data statistics
Nominal GDP	Nominal GDP data of the city for the corresponding year
Average Annual Housing price	The city's average annual home price information for the relevant year

Table 2: Excerpts from data instances.

Province	City	Year	Nominal GDP ($\times10^6$ CNY)	Average Annual Housing price (CNY/m^2)
Zhejiang	Hangzhou	2016	11710	17,301
Zhejiang	Hangzhou	2017	13161	23,693
Zhejiang	Hangzhou	2018	14307	25,007
Zhejiang	Hangzhou	2019	15419	23,515
Zhejiang	Hangzhou	2020	16207	24,590
Zhejiang	Hangzhou	2021	18247	21,393
Zhejiang	Hangzhou	2022	18753	22,305
Zhejiang	Hangzhou	2023	20059	23,185

The author used the Anjuke website to gather 37,874 monthly average home prices for 350 perfection-level Chinese cities between January 2010 and July 2024, and the national data website to gather GDP information for 296 perfection-level cities. After preliminary processing, the author selected the data from 2016 to 2023. The author also summarized monthly housing prices into annual housing prices and then screened and merged the two data sets to obtain the "housing price-nominal GDP" data set of prefecture-level cities, including data from 244 cities. The author processed 29 missing values in data cleaning, accounting for about 1.2% of the total data. The author chose the linear interpolation method to fill in these missing values. Still, for the missing data of some cities throughout the year, the author decided to delete these data. The authors use Z-score to detect outliers and filter out rows with absolute Z-score values greater than 3.

4 MODEL INTRODUCTION

Four models were used in this paper's experiment. The following is the introduction of these models.

4.1 eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is an optimization implementation based on the Gradient Boosting Decision Tree (GBDT). Gradient lifting is an iterative technique for ensemble learning that combines several weak learners to gradually optimize model performance. The main notion is that the errors of the preceding model are corrected by the latter model in each training round. Boosting is a sequential training procedure where each model is trained one after the other, trying to adjust for the residual of the model that came before it. Overfitting is avoided by XGBoost by regularizing the model's complexity. XGBoost allows custom loss functions. XGBoost automatically handles missing values. XGBoost implements parallel computing through a column-based block structure. XGBoost weights the contribution of each tree. XGBoost selects the best-split point through pre-sorting and approximation algorithms. XGBoost offers significant advantages over traditional GBDT.

4.2 Support Vector Regression

Support Vector Regression (SVR) is a regression model based on the Support Vector Machine (SVM) algorithm. SVR fits around as many data points as possible by constructing a nonlinear function but ignores some errors within a specific range. SVR introduces a ε- insensitive loss function where ε is an artificially set threshold. Prediction errors within this threshold range are ignored, and only prediction errors over ε will contribute to the loss function. This gives SVR a certain tolerance, i.e., minor errors do not affect model optimization. SVR aims to find a smooth function within the range of error ε. This function consists of minimizing model complexity and minimizing errors beyond ε. In SVR, only points with errors more significant than ε become support vectors. SVR is good at dealing with high-dimensional data and linear indivisible problems and is often used in nonlinear regression scenarios.

4.3 Multilayer Perceptron

Multilayer Perceptron (MLP) is a feedforward neural network model consisting of multiple layers of neurons capable of processing both linear and non-linear data. It gradually extracts higher-order data features through hidden layers to build complex mapping relationships between inputs and outputs. An MLP's fundamental structure consists of an input layer, one or more hidden layers, and an output layer. Weighted connections link the nodes in each layer to the corresponding layer's nodes. MLP works through Forward Propagation, loss function calculation, Backpropagation, and iterative updating. MLP has the characteristics of nonlinear modeling, multi-layer, and fully connected.

4.4 Long Short-term Memory

Long Short-term Memory Network (LSTM) is a Recurrent Neural Network (RNN) variant designed explicitly for processing and predicting time series data. Long- and short-term time dependencies are both captured by LSTM. The Memory Cell, which is the central component of the LSTM, is equipped with three gates to regulate the information flow in addition to a Cell State. These systems decide what should be stored in memory, what should be deleted, and how fresh input data should modify the state of the cell. There are three different kinds of gates: input, output, and forget gates. To determine which memories to retain and which to trash, the LSTM initially employs the oblivion gate. The input gate decides how the data being entered should be stored in memory. Based on the output of the input gate and the forgetting gate, the cell state is updated. The output gate ascertains the hidden state output at that precise moment by combining the cell state with the current time step LSTM output.

5 EXPERIMENTAL PRACTICE, PREDICTION AND ANALYSIS

This research builds a regression model to forecast the average yearly home prices of Chinese cities in 2024 and 2025 utilizing the four models discussed in the preceding section. Three sets of pre-processed data—a training set, a verification set, and a test set—represent 60%, 20%, and 20% of the total data in the third section. Data standardization was done by the author to satisfy the model's needs. The StandardScaler was used to normalize the feature and target variables, and the subsequent model predictions were reversed and normalized. On XGBoost and SVR, GridSearchCV tunes hyperparameters for grid search. The input layer, five hidden layers, and output layer are the several levels of fully connected neural networks that the experiment builds into a model on the MLP. Leaky ReLU is the activation function used in each layer, while a linear activation function is used in the output layer. The model is trained using the Adam optimizer, and overfitting is avoided by using the EarlyStopping and ReduceLROnPlateau callback methods. The housing price prediction problem is treated as a time series prediction problem on LSTM in this experiment. Multiple Bidirectional LSTM layers, Conv1D convolution, and Dropout layers make up the model. The learning rate is adjusted dynamically by using the CyclicalLearningRate callback function. Mean square error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) were used to assess each model's prediction outcomes in the validation and test sets. The results are shown in Table 3.

Table 3: Evaluation results of each model.

	XGBoost	SVR	MLP	LSTM
R^2	0.9543	0.7269	0.7133	0.7212
MSE	1.926×10^{6}	1.394×10^{7}	1.464×10^{7}	1.430×10^{7}
MAE	898	2061	2149	2418

It is evident from the evaluation findings that XGBoost performs noticeably better than the other three models. XGBoost effectively handles high-dimensional data and can capture complex nonlinear relationships, while SVR and MLP struggle to handle highly nonlinear parts of the data. XGBoost uses L1 and L2 regularization to prevent model overfitting, whereas MLP and LSTM rely on manually adjusting regularization parameters to avoid overfitting. XGBoost's feature importance analysis automatically selects the most valuable features for optimization, a critical factor in its superior performance. XGBoost uses second-order Taylor expansion to accelerate the gradient calculation and improve convergence speed. This allows it to find the optimal solution faster while training. In contrast, an MLP or LSTM takes longer to adjust the weights and perform gradient descent, making it sensitive to the selection of hyperparameters. If there are noise or outliers in the price data, XGBoost can handle these noise and outliers better due to its robustness. These advantages give XGBoost certain predictive advantages. The four models used by the author predict are shown in Table 4.

Table 4: Prediction of future urban housing prices by each model (excerpt).

Year	City	XGBoost	SVR	MLP	LSTM
2024	Sanming	10271	8428	9127	10736
2025	Sanming	10273	8798	9018	11350
2024	Sanmenxia	5948	7171	8697	9701
2025	Sanmenxia	6408	7606	8481	9776
2024	Shanghai	56344	49496	74049	43372
2025	Shanghai	56344	42038	79407	39481
2024	Shangrao	8476	8827	9240	7266
2025	Shangrao	8521	9269	9204	7034

In addition, the author uses the XGBoost forecast of China's urban housing price forecast (2024-2025) to compare with the overall housing price over the years (2016-2023) and shown in Figure 1.

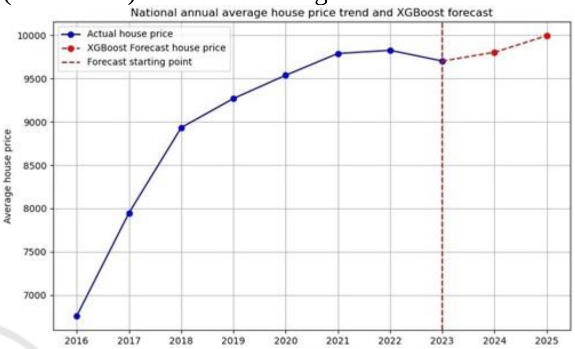

Figure 1: Average annual housing price trends combined with XGBoost forecasts (Photo/Picture credit : Original).

The figure shows that China's housing prices continued to increase from 2016 to 2021 and slowed down or declined from 2022 to 2023. The forecast suggests that housing prices will rise slightly in 2024-2025. The changes in the historical data are consistent with the actual economic situation. China's slowing economy and aging population have gradually weakened the upward trend in property prices. The real estate market regulations implemented by the state have also contributed to a decrease in home prices. Due to the substantial decline in home demand brought about by the COVID-19 pandemic, China's housing prices are expected to decline in 2022–2023. The experiment's gradual increase in property prices is partially explained by the Chinese government's July 2024 easing of the housing purchase policy and the anticipated rebound of the Chinese economy. Therefore, the predicted results of this experiment are reasonable from the economic and policy perspectives.

The contribution of this experiment is to demonstrate the excellent performance of XGBoost in processing large-scale complex data, providing a comparative analysis of multiple models, and providing a decision-making basis for policymakers and investors. In addition, this experiment has numerous flaws, including inadequate generalization due to the exclusion of numerous other significant variables (such as population migration) and an incomplete understanding of the effect of emergencies on future home prices. More macroeconomic variables, deep learning model optimization, and the use of external risk prediction models, like Monte Carlo simulations, to increase robustness are examples of future research directions that might be consulted.

6 CONCLUSIONS

This study compares the performance of four machine learning models (XGBoost, SVR, MLP, and LSTM) to predict property prices in Chinese cities. The XGBoost model is ultimately determined to be the best model. The experimental results show that XGBoost performs better than other models in terms of prediction accuracy. Concurrently, adding nominal GDP as an auxiliary variable improves the accuracy of home price forecasts, especially in areas where significant shifts in economic development have occurred. Projecting home prices for 2024 and 2025, this study validates the applicability of machine learning technologies in home price forecasting. It provides a scientific resource for decision-making to legislators and real estate investors. In order to improve forecast accuracy over the long run, future research might look into how other economic variables affect property values or use more sophisticated deep learning models.

REFERENCES

Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., Oluwadara, G., 2022. House price prediction using Random Forest machine learning technique. Procedia Computer Science, 199:806-813.

Chen, Y., Xue, R., Zhang, Y., 2021. House price prediction based on machine learning and deep learning methods. 2021 International Conference on Electronic Information Engineering and Computer Science, 699-702.

Goel, Y. K., Swaminathen, A. N., Yadav, R., Kanthamma, B., Kant, R., Chuahan, A., 2023. An innovative method for housing price prediction using Least Square – SVM. 2023 4th International Conference on Electronic and Sustainable Communication Systems, 928-933.

Henriksson, E., Werlinder, K., 2021. Housing price prediction over countrywide data: A comparison of XGBoost and Random Forest regressor models. KTH Royal Institute of Technology, Stockholm, Sweden.

Ho, W. K. O., Tang, B. S., Wong, S. W., 2020. Predicting property prices with machine learning algorithms. Journal of Property Research, 38(1):48–70.

Hong, J., Choi, H., Kim, W., 2020. A house price valuation based on the Random Forest approach: The mass appraisal of residential property in South Korea. International Journal of Strategic Property Management, 24(3):140-152.

Manasa, J., Gupta, R., Narahari, N. S., 2020. Machine learning based predicting house prices using regression techniques. International Conference on Innovative Mechanisms for Industry Applications, 2020(2):624-630.

Ming, Y., Zhang, J., Qi, J., Liao, T., Wang, M., Zhang, L., 2020. Prediction and analysis of Chengdu housing rent based on XGBoost algorithm. Proceedings of the 3rd International Conference on Big Data Technologies, 1-5.

Sheng, C., Yu, H., 2022. An optimized prediction algorithm based on XGBoost. International Conference on Networking and Network Applications, 2022:1-6.

Truong, Q., Nguyen, M., Dang, H., Mei, B., 2020. Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174:433-442.

Xu, X., Zhang, Y., 2021. House price forecasting with neural networks. Intelligent Systems with Applications, 12.