

# Comparison of Real Estate Price Prediction Based on LSTM and LGBM

Songyi Shi \*

College of Horticulture, China Agricultural University, Beijing, China

\* Corresponding Author Email: shisongyi@alu.cau.edu.cn

**Abstract.** Housing price prediction is one of the widely discussed topics, and the application of machine learning and deep learning models to housing price prediction is a research hotspot. Exploring which models are suitable for predicting real housing price datasets has significant implications for guiding government and homebuyers' decision-making. In this study, the LSTM and LightGBM models were selected as research objects, and the suitability of the models was explored and compared using the second-hand housing price dataset in Beijing. Based on the analysis, in the task of pure time series prediction based on historical housing prices, the LSTM model had a better fit ( $R^2=0.91$ ), while in the task of housing price prediction based on multiple influencing factors, the LightGBM model had better comprehensive evaluation index results ( $R^2=0.53$ ). Both models can be used to predict pure time series, but the LSTM model is not suitable for predicting multi-factor input models that are not time series.

**Keywords:** Housing price prediction; LSTM; LightGBM.

## 1. Introduction

Real estate is a vital sector of the national economy, and the soaring prices of China's urban real estate market in recent years have garnered significant international attention [1]. Price prediction is crucial in real estate transactions as it improves market transparency, assists in macroeconomic regulation, and helps real estate investors make informed decisions [2]. Machine learning and artificial intelligence has the potential to enhance the precision of real estate price predictions in urban areas, which exhibit complex dynamics. Previous studies commonly used multiple features for real estate price prediction. Researchers have discovered the significance of time series in real estate price prediction, as it leads to more accurate results. Although time series problems are similar to regression problems, they have certain differences in methodology.

Traditional time series modeling includes ARMA and ARIMA models. ARMA can only perform regression on stationary univariate historical data and is not suitable for multivariate scenarios. Compared to the ARMA model, the ARIMA model requires data to be differenced before modeling after the difference is stationary [3]. These two models can handle relatively simple problems, and their ability to represent complex problems are limited. The second type refers to machine learning techniques, e.g., LightGBM and XGBoost, which use feature engineering and algorithms to transform time series problems into supervised learning tasks. They are capable of handling complex data modeling and non-linear problems, making them ideal for challenging time series prediction tasks. Machine learning models outperform traditional ARIMA models in predicting time series data affected by external factors such as store sales history, as indicated by a lower RMSE value [4]. Quang used random forest, XGBoost, and LightGBM methods, as well as mixed regression and stacked generalization regression techniques for real estate price prediction, all of which achieved ideal results [5]. There are studies indicating that the LightGBM model outperforms other models in predicting price trends and is more suitable for short to medium-length time series [6].

Finally, deep learning models are used for time series prediction. The LSTM model is designed specifically for solving time series problems. Compared to others, LSTM-based predictions purely based on time factors are more accurate, but their performance in predicting multiple feature factors is generally average [7]. Lee used LSTM to predict housing prices for 13 months in Korea, achieving a 76% prediction similarity rate [8]. Unlike other data types, real estate data also has a clear spatial

attribute. Previous research has incorporated spatial dependency analysis into housing price prediction models, such as considering community factors or adding geographical analysis [9-11]. Comparative studies on the prediction models of house prices with complex attributes are still lacking. This study aims to evaluate and compare machine learning models (LightGBM) and deep learning models (LSTM) for predicting house prices in Beijing with complex attributes. The study seeks to explore the applicability of these models under multiple factors.

## 2. Data & Method

### 2.1. Data

The dataset used in this study was obtained from <https://www.kaggle.com>, containing housing price data from 2011 to 2017, sourced from <https://bj.lianjia.com/chengjiao>. The dataset includes various features of the houses, with irrelevant features such as community and house IDs, and number of followers removed. The remaining features consist of 18 variables, including spatial information such as latitude and longitude, transaction time, price, area, building type, construction time, and others, where the unit area price is the target variable, i.e., the dependent variable. The dataset undergoes data cleaning, where feature attributes with missing values above 20% are deleted, and price data above 10% are also deleted while the rest are filled with interpolation. The processed data distribution is shown in Fig. 1, and data varies across years with the highest in 2015 and 2016 and the lowest in 2010.

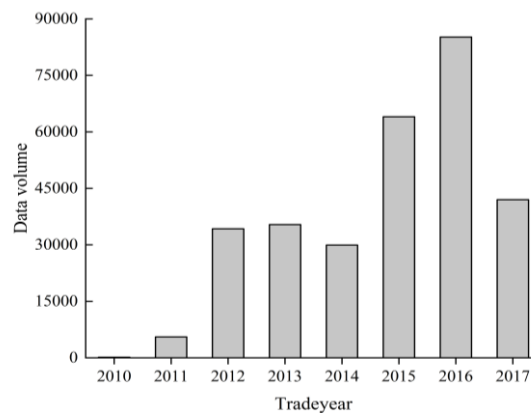


Fig 1. Data volume for each year

### 2.2. Models

#### 2.2.1 LSTM

This study utilized the TensorFlow and Keras libraries to build an LSTM prediction model. LSTM (Long Short-Term Memory) specifically designed to handle long-term dependencies in sequential data. By utilizing gate mechanisms, the model is able to control the information flow and make informed decisions on whether to retain or disregard information based on its relevance to the given task. This results in an enhanced capacity to handle lengthy sequences of data. The formulas are as shown in Eq. (1) to Eq. (6):

$$F_t = \sigma (S_f \cdot [a_{t-1}, b_t] + \beta_f) \quad (1)$$

$$I_t = \sigma (S_i \cdot [a_{t-1}, b_t] + \beta_i) \quad (2)$$

$$O_t = \sigma (S_o \cdot [a_{t-1}, b_t] + \beta_o) \quad (3)$$

$$C_t = F_t * C_{t-1} + I_t * \hat{c} \quad (4)$$

$$\hat{c} = \tanh (S_c \cdot [a_{t-1}, b_t] + \beta_c) \quad (5)$$

$$\hat{H} = O_t * \tanh (C_t) \quad (6)$$

These formulae employs  $F_t$ ,  $I_t$ , and  $O_t$  to represent the forget gate, input gate, and output gate in the same order. Additionally,  $b_t$ ,  $C_t$ , and  $a_t$  denote the new candidate state, current state, and output state, respectively.  $S$  and  $\beta$  are the model parameters, while  $\sigma$  and  $\tanh$  are the sigmoid and hyperbolic tangent functions, respectively. The symbol “\*” denotes element-wise multiplication, following conventional academic notation.

### 2.2.2 LightGBM

The model used in this study employed an interface provided by Microsoft Research. LightGBM is a GBDT framework that uses an efficient histogram-based algorithm for handling continuous features. This technique reduces the costs of sorting, making it less likely to overfit. The mathematical formulae for LightGBM is presented from (7) to (10):

$$F = (\sum_{i \in L} \lambda)^2 / ((\sum_{i \in L} \beta) + k) \quad (7)$$

$$L = (\sum_{i \in R} \lambda)^2 / ((\sum_{i \in R} \beta) + k) \quad (8)$$

$$W = (\sum_{i \in I} \lambda)^2 / ((\sum_{i \in I} \beta) + k) \quad (9)$$

$$G = 0.5 \cdot (F + L + W) \quad (10)$$

where  $G$  can be seen as a score function assessing the structural proficiency of tree structure, The variables  $\lambda$  and  $\beta$  denote the statistical values of the first and second-order gradients of the loss function, correspondingly, and for a certain tree structure, the sample sets for the left and right branches are denoted by  $L$  and  $R$ , respectively.

### 2.3. Evaluation Metrics

To assess the predictive capability of the model, this study selected the following indicators for analysis  $R^2$  (R-Squared), also known as the coefficient of determination, is utilized to assess the degree of fit of a model to data. The formula is:

$$R^2 = 1 - E/T \quad (11)$$

where  $E$  represents the sum of squared residuals and  $T$  represents the total sum of squares. RMSE is a typical metric for evaluating the performance of a regression model. The formula for RMSE is:

$$RMSE = (1/n \cdot \sum (\lambda - \beta)^2)^{1/2} \quad (12)$$

where  $\lambda$  is the predicted value,  $\beta$  is the true value,  $n$  is the number of samples. MAE calculates the average magnitude of errors in a set of predictions, regardless of their direction. It is calculated by taking the absolute difference between each predicted and true value, then averaging across the entire set. The formula is:

$$MAE = 1/n \cdot \sum |\lambda - \beta| \quad (13)$$

where  $\lambda$  is the predicted value,  $\beta$  is the corresponding true value,  $n$  is the total number of samples. The calculated MAE value represents the mean absolute deviation between the predicted values and the actual values, with a lower value indicating better performance.

## 3. Results & Discussion

### 3.1. Feature Engineering

Preprocessing of latitude and longitude data is necessary prior to conducting correlation analysis on the influencing factors. The spatial dependence of housing price data has been demonstrated; hence spatial sequence data transformation is required for the purpose of analysis [12]. The geographic information processing system ArcMap 10.8 was used to find the Moran's I index of the housing price data for each year, Moran's I is a rational number. After normalization, its values usually range from

-1 to 1, and positive Moran's I means strong positive spatial autocorrelation. The results showed that the housing price dataset exhibited high spatial autocorrelation [13].

On this basis, in order to transform the latitude and longitude information in the dataset into spatial sequence information, the spatial analysis tool in ArcMap 10.8 was used to define the geographic center points of all house locations, and then the nearest neighbor analysis tool was used to calculate the distance from each data point to the corresponding center point, which was used as a factor in constructing the spatial sequence. The analysis dataset was divided into two categories: pure time series data and multi-factor analysis data. For the former, the monthly average house prices from January 2010 to August 2017 were calculated, resulting in a total of 92 months of data. For the latter, to unleash the efficiency of model training and account for differences in data volume across different time periods, data from January 2017 to August 2017 was selected. To sort the unordered data within each month, the spatial sequence factor was added, resulting in 17,675 spatial-temporal sequence data. The data was split into a training set and a test set, with a ratio of 80:20. Spearman correlation analysis was conducted on the second category of data, and the results are given in the Fig. 2. The threshold value was set to 0.2, and the factors NEAR\_DIST, square, construction age, subway, and district were selected to represent the distance to the geographic center, total area of the house, age of the house, number of nearby subway stations, and city district, respectively. These five factors were retained as input data.

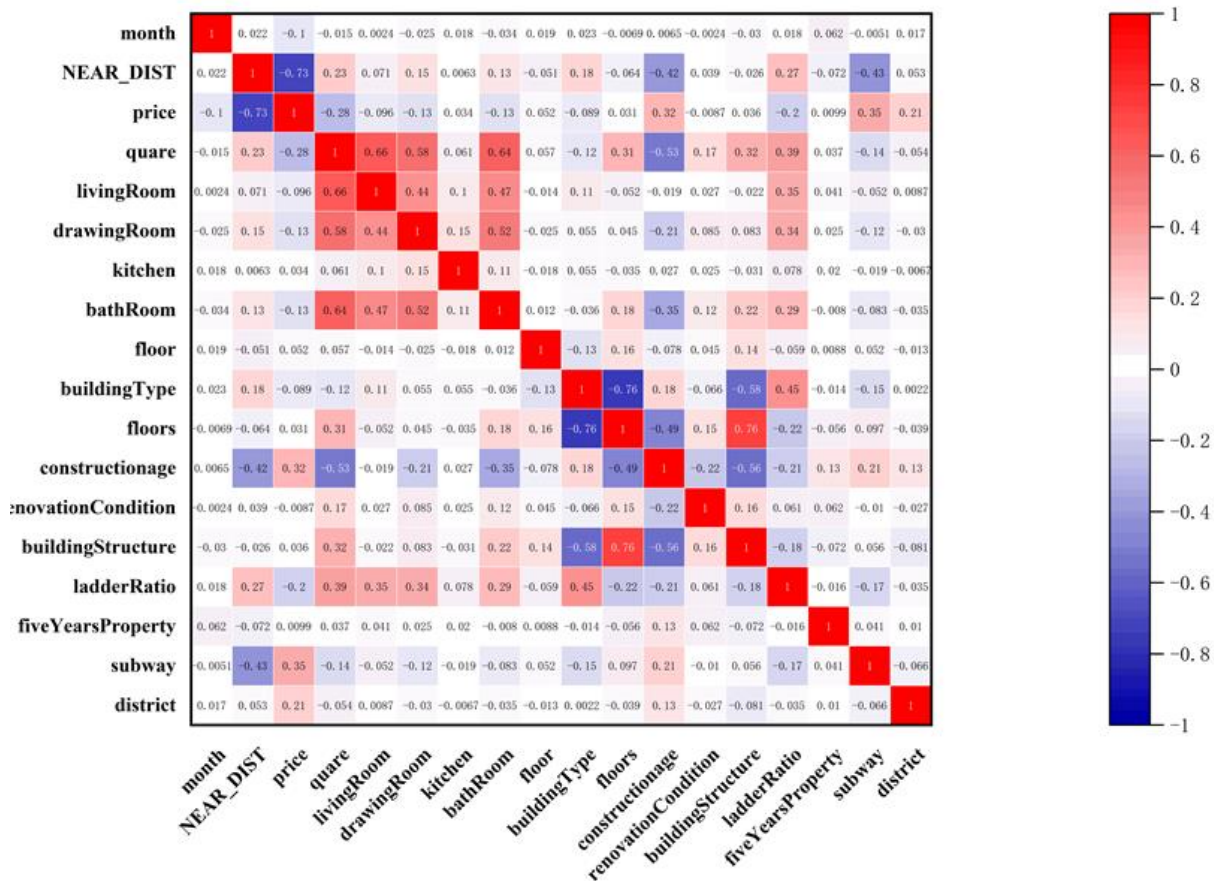


Fig 2. Correlation coefficients of housing attributes features

## 3.2. LSTM

### 3.2.1 Prediction Based on Pure Time Series

The average housing prices of 92 months were fed into the model. After multiple experiments and parameter tuning, a sliding time window of 2, a network layer of 3, a hidden layer node count of 128, 1000 iterations, and a dropout rate of 0.2 were set. The results of the test set showed an RMSE evaluation coefficient of 2569, MAE of 2240 and an  $R^2$  of 0.91 as presented in Fig. 3.

### 3.2.2 Prediction Based on Multiple Factor Inputs

After multiple rounds of parameter optimization and testing, the model performance remains poor with the val-loss value staying at a high level, despite feeding in the five influencing factors and the target variable for prediction. This indicates that the potential of the models may not have been fully explored due to the poor quality of the dataset, which exhibits low correlation between its variables.

A ConvLSTM model was constructed by adding 2D convolution and pooling layers to the base model. Suitable parameters, including a kernel number of 3 and size of 1, a pooling window size of 2, an LSTM layer number of 3, a neuron number of 128, 500 epochs, and a dropout of 0.3, were selected through experimentation. The training results of the model were tested on the test set, and a portion of the data is shown in the Fig. 4. The RMSE is 0.03, the  $R^2$  is -0.68, and the MAE is 0.02.

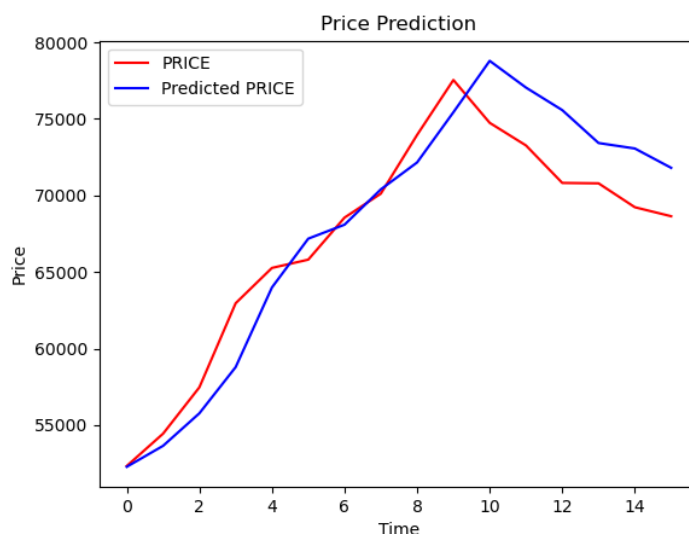


**Fig 3.** Prediction results based on time series using LSTM.

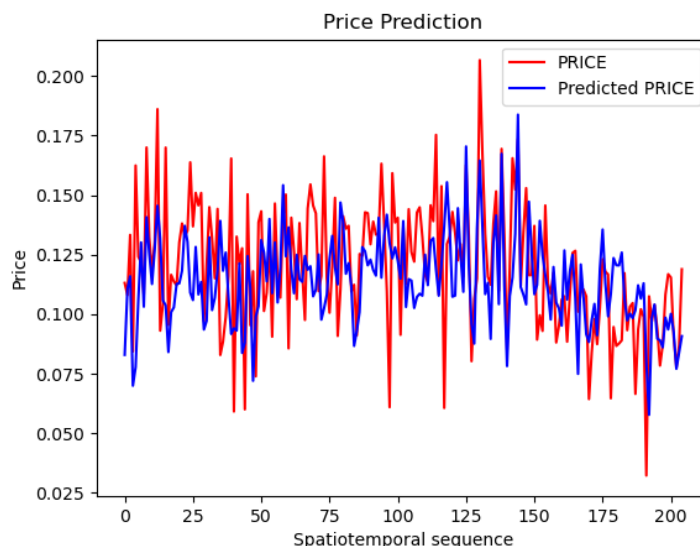


**Fig 4.** Prediction results based on multi-factor input using LSTM.

### 3.3. LGBM

#### 3.3.1 Prediction Based on Pure Time Series

For each step of prediction, the previous three steps are used as training features to predict the next price. After parameter optimization, the “boosting\_type” was set to “gbdt”, the evaluation metric was configured to “MSE”, “n\_estimators” was 50, and “max\_depth” was 3, with L1 and L2 regularization coefficients set to  $1e-5$  and  $1e-3$ , respectively. The results on the test set are demonstrated in the Fig. 5, with an RMSE of 3096,  $R^2$  of 0.81, and MAE of 2522.

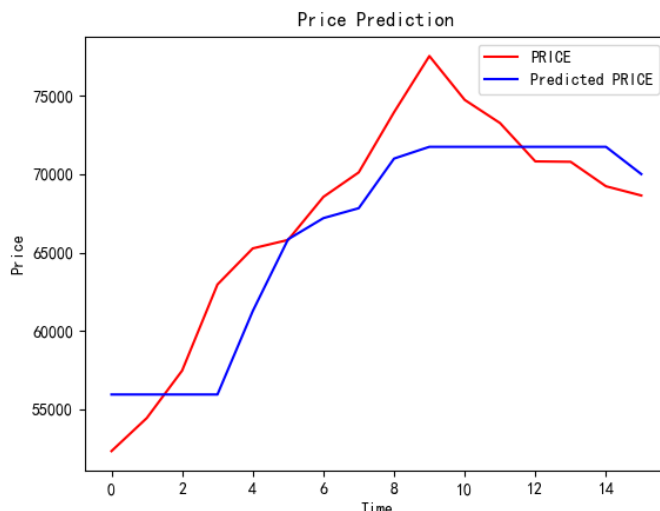


**Fig 5.** Prediction results based on time series using LGBM

### 3.3.2 Prediction Based on Multiple Factor Inputs

Due to the influence of spatiotemporal sequences on housing price trends, lagged values of time are added to reconstruct the correspondence between the original features and labels. After observing the data relationship and testing, moving two sequence units are used as the lag value. After performing a grid search, the optimal hyperparameters of the model have been determined. The L1 and L2 regularization coefficients are both 0.001, the learning rate is 0.001, the “max\_depth” is 600, the “n\_estimators” is 3000, and the “max\_leaves” is 600. The model was tested on the test set, and partial results are exhibited in the Fig. 6. The evaluation metrics are RMSE=0.61, MAE=0.48 and  $R^2=0.53$ .

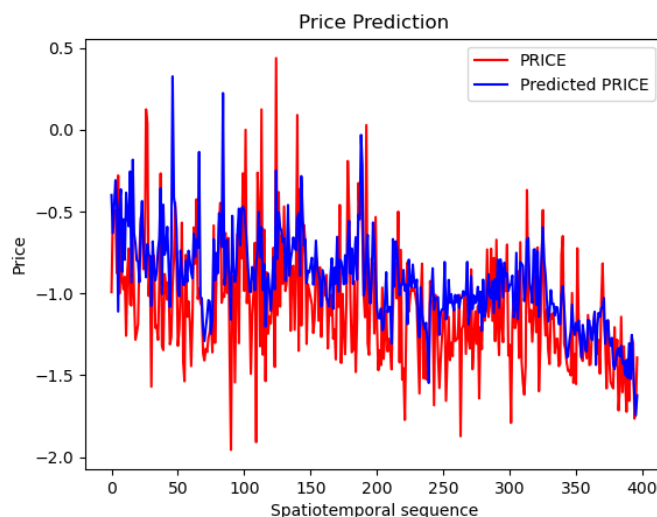


**Fig 6.** Prediction results based on multi-factor input using LGBM

### 3.4. Comparison & Suggestion

The LSTM model performed better on the prediction based on pure time series, with an  $R^2$  score of 0.91, higher than the LGBM model's 0.81, indicating a better model fit. However, both models had relatively large errors between actual and predicted values, possibly due to the size of the dataset and the lack of external time series factors for analysis, leading to relatively low accuracy. For the prediction based on multiple factor inputs, due to the large number of input factors, and the fact that they do not form an ordered sequence, the LSTM model could not take advantage of its strengths and incurred high loss rates. In contrast, the LGBM model performed well in terms of both prediction accuracy and model fit when dealing with such data. Although the added ConvLSTM model resulted in smaller differences between predicted and actual values, overfitting was still observed. To sum up,

for predicting future trends based on past house price features, the LSTM model performs better than the LGBM model. If predicting house prices based on multiple non-time-related factors, the LGBM model performs better in all aspects.

#### 4. Limitations & Future outlooks

When dealing with multi-factor input datasets, the ConvLSTM model's prediction results are relatively accurate, but after multiple parameter optimizations, it still cannot fit the model, indicating that the model has low generalizability and is not suitable for processing this type of data. Although LSTM has advantages in predicting time series data, it requires more iterations and has much lower computational efficiency than LGBM. This study has several limitations in terms of data and methodology. Firstly, the approach used to construct the spatial sequence data for housing price analysis is relatively simplistic, as it only considers the geographic distribution of sampling points without accounting for clustering effects. Secondly, although housing price data exhibits complex spatiotemporal properties, this study only aims to compare two different prediction models and does not fully explore the characteristics of the housing price data. Lastly, the utilized dataset is based on real-world data and is subject to various external factors, resulting in weak correlation among different factors and suboptimal performance of the models. It is possible that the potential of the models was not fully utilized. In future research, more appropriate spatial modeling methods should be introduced, such as considering spatial clustering, adjacency shape, and weight augmentation, to deeply explore the geographical pattern of real estate prices. Additionally, more targeted combination models or improved algorithms can be developed based on the data features to better adapt to more complex scenarios.

#### 5. Conclusion

In summary, two models were employed in this paper, i.e., LSTM and LGBM, to predict real estate prices using a dataset divided into two subsets: pure time series and multiple factors. For the pure time series subset, LSTM outperformed LGBM, whereas for the multiple-factor input subset, LGBM exhibited better overall performance. The current study acknowledges the lack of exploration on handling complex spatial-temporal data with multiple attributes and aims to enhance its evaluation through the incorporation of various composite models in the future. The research findings of this paper can contribute to the advancement of house price prediction techniques and assist decision-makers in making informed decisions based on the predictions.

#### References

- [1] Zhu L, Zhang H. Analysis of the diffusion effect of urban housing prices in China based on the spatial-temporal model. *Cities*, 2021, 109:103015.
- [2] Liu G. Research on prediction and analysis of real estate market based on the multiple linear regression model. *Scientific Programming*, 2022, 1: 1-8.
- [3] Lim W T, Wang L, Wang Y, et al. Housing price prediction using neural networks. 2016 12th International conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD). IEEE, 2016: 518-522.
- [4] Pavlyshenko B M. Linear, machine learning and probabilistic approaches for time series analysis. 2016 IEEE first international conference on data stream mining & processing (DSMP). IEEE, 2016: 377-381.
- [5] Truong Q, Nguyen M, Dang H, Mei B. Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, 2020, 174:433-442.
- [6] Sun X., Liu M., Sima Z. A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters*, 2020, 32:101084.
- [7] Yu L, Jiao C, Xin H, Wang Y, Wang K. Prediction on housing price based on deep learning. *International Journal of Computer and Information Engineering*, 2018, 12(2): 90-99.

- [8] Lee J, Kim H, Shim G. Prediction Model of Real Estate Transaction Price with the LSTM Model based on AI and Bigdata. *International Journal of Advanced Culture Technology*, 2022, 10(1): 274-283.
- [9] Ma S, Jiang G, Liu J, et al. Research and Application of Second-Hand Housing Price Prediction Model Based on LSTM. *2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*. IEEE, 2021: 320-325.
- [10] Ge C. A LSTM and graph CNN combined network for community house price forecasting. *2019 20th IEEE International Conference on Mobile Data Management (MDM)*. IEEE, 2019: 393-394.
- [11] Chiu S M, Chen Y C, Lee C. Estate price prediction system based on temporal and spatial features and lightweight deep learning model. *Applied Intelligence*, 2022, 52(1):808-834.
- [12] Wang W C, Chang Y J, Wang H C. An application of the spatial autocorrelation method on the change of real estate prices in Taitung City. *ISPRS International Journal of Geo-Information*, 2019, 8(6): 249.
- [13] Kim J, Won J, Kim H, et al. Machine-learning-based prediction of land prices in Seoul, South Korea. *Sustainability*, 2021, 13(23): 13088.