International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 4005-4017

https://doi.org/10.70517/ijhsa464518a

Machine Learning-Based Prediction of the Impact of the Opening of New Rail Transit Lines on Price Fluctuations in Housing Markets

Binghe Zhang[1] and Shilin Zhang[1,*]
[1] Qingdao Metro Line 8 Co., Ltd., Qingdao, Shandong, 266100, China
Corresponding authors: (e-mail: zhangbinghe_Metro@163.com).

Abstract Python calls Baidu API interface to crawl the newly opened data of rail transit lines corresponding to each housing address, and for the raw data containing many interfering information, it is processed with data cleaning and quantization, followed by screening the spatial feature variables based on Pearson correlation coefficient. In order to improve the prediction accuracy and generalization ability of the algorithm combination, the three algorithms XgBoost, LightGBM and CatBoost in the Boosting series of algorithms are selected as the primary learners of the prediction model, and the prediction model based on machine learning is constructed. The research data and prediction models are assembled to examine the mechanism of the role of the opening of new rail transit lines on price fluctuations in the housing market. The correlation values of the four types of spatial characteristic variables of rail transit are 0.06, -0.112, -0.33, and -0.164, respectively, which concludes that the opening of new lines of subway transportation has the greatest role in influencing the housing market price, and it provides a reference for the prediction of the housing market price; in addition, the integrated error rate of the prediction model is 0.3697%, which indicates that the model has an excellent performance in the prediction of the housing market price, and it helps to urban planning, design and construction economy to a more desirable direction.

Index Terms rail transit, housing market price, machine learning, urban planning and design, prediction modeling

I. Introduction

With the acceleration of urbanization and the rapid growth of population, urban transportation has become a global challenge. As an efficient and environmentally friendly mode of transportation, urban rail transit plays an important role in improving urban traffic conditions, reducing traffic congestion, and enhancing the travel experience of urban residents. In addition to the improvement of traffic conditions, urban rail transit also has a profound impact on the real estate market [1]-[4].

The construction and operation of urban rail transit can make the transportation convenience greatly improved, which also means that the convenience of transportation will become an important consideration when people buy houses. Therefore, when the city planning department announces that a new rail transit line will pass through a certain area, the real estate price of the area will usually rise significantly [5]-[8]. Home buyers are willing to choose the high price area for the convenience of transportation, which makes the heat of the real estate market in that area increase. In addition, the construction of urban rail transit usually requires land and space, which makes the land along the transit lines become precious [9]-[12]. In order to utilize the potential of these lands along the lines, developers usually carry out large-scale real estate project development near the rail transit lines. The construction of rail transit has a significant impact on the price of housing in the surrounding area. The positive impact is mainly reflected in the enhancement of the transportation convenience of the surrounding area, which attracts more home buyers to choose to buy homes in the vicinity [13]-[16]. The development along the rail transit will also lead to the rise of the surrounding house prices. With the development of rail transit, the surrounding house prices may also have the negative impact of excessive growth leading to house price bubbles [17], [18].

Literature [19] presents a quantitative review of studies on the impact of rail transit on land and housing values. The results show that there are large variations in the estimates of the magnitude of the impacts of rail transit access from one study to another. And these variations are related to the type of railroad projects, data, etc. Insights are provided for assessing the impact of rail transit improvements on real estate. Literature [20] used a series of econometric methods to examine the impact of BRT accessibility and proximity on house prices in a particular location. The study obtains conclusions such as “the co-existence of BRT accessibility premium and proximity penalty” and “the spatial heterogeneity of the impact of BRT on house prices”, and discusses the policy implications.

Literature [21] explores the impact of high-speed rail (HSR) on urban house prices. Based on the study of HSR operation and house price data in Chinese cities, it is pointed out that HSR not only raises urban house prices, but also alleviates regional imbalances and shows a certain degree of synergistic effect. Literature [22] explores the relationship between public transportation and house prices using the GBRT methodology with Shanghai residences as the object of study. The positive impact of public transportation accessibility on most house prices and the negative impact on some house prices are emphasized. Literature [23] explored the impact of rail transit on urban residential prices based on a semi-logarithmic eigenprice model, combining facility point-of-interest data and residential unit transaction data. The results show that the spatial effect of rail transit on residential prices varies in different locations, and relevant recommendations are made. Literature [24] explores the heterogeneous effects of the subway system on housing using a hedonic price model as a complementary approach to quantile regression. The results show that the distance of a home from a metro station influences the price of a home, while transportation accessibility has a decreasing effect on low-, medium-, and high-priced homes. Literature [25] takes Xuzhou properties as an example and analyzes the impact mechanism of urban rail transportation on housing prices by establishing a multiple regression model. The results show that an increase in the distance between transportation and properties decreases house prices, while a positive transportation coefficient indicates that transit stations increase house prices. Literature [26] takes Wuhan as the research object and points out that the more mixed land areas, the higher the transportation proximity premium of house prices. It also derives the policy implications of TOD planning and land value acquisition near transportation stations. Literature [27] aims to investigate the impact of bus accessibility on urban house prices. A spatial econometric model was developed based on a database of some apartment units in Xiamen. The results found that bus accessibility is positively correlated with neighborhood house prices and has a significant impact on surrounding house prices. Literature [28] analyzed the change of land use around metro stations before and after the construction of metro lines based on the information provided by Google Earth. It shows that suburbs are greater than urban areas in the revitalization effect of new subway stations on land. It shows that the location choice and spatial distribution of metro stations should not only take into account the interests of the land sector, but also the interests of residents. Literature [29] used the SNA method to study the impact of high-speed rail on urban housing prices. Based on multiple city samples, it was found that there is a positive impact of HSR network accessibility on house prices. And this impact varies by region and housing type. The spatio-temporal economic law of spatial differences and regional economic convergence is revealed, which has important policy implications. Literature [30] constructed a hedonic pricing model based on the price data of residential neighborhoods near rail transit lines, and empirically drew conclusions including “urban rail transit has a positive impact on the surrounding house prices”.

Combining three basic machine learning algorithms to predict the relationship between the opening of new rail transit lines on the role of housing market prices, so that the prediction accuracy and robustness are guaranteed. In this paper, the main urban area of a city is taken as the research object, and the Python crawler program is used to grab the data of rail transit characteristics of four urban areas from the Chain Home platform. In the process of obtaining the raw data, some incomplete data and unpublished information lead to problems such as overfitting of the model, high prediction error and low prediction accuracy. Accordingly, data cleaning is carried out from three aspects, namely, duplicate value processing, missing value processing and outlier processing, after which three algorithms, XgBoost, LightGBM and CatBoost, are combined to construct the housing market price prediction model. After completing the model parameter setting, the impact of the opening of new rail transit lines on the price fluctuation of the housing market is explored.

II. Data pre-processing

This chapter mainly discusses the related work of data preprocessing, including: data acquisition, data cleaning, data quantification, data visualization and feature screening five processes, data preprocessing is shown in Figure 1. Firstly, the way of data acquisition and the process of crawling data are introduced, followed by data cleaning, quantification and visualization analysis of the original data, and then feature screening based on the Pearson correlation coefficient, and finally, the original dataset obtained from crawling is transformed into a high-quality dataset that can be directly substituted into the subsequent model for modeling and analysis.

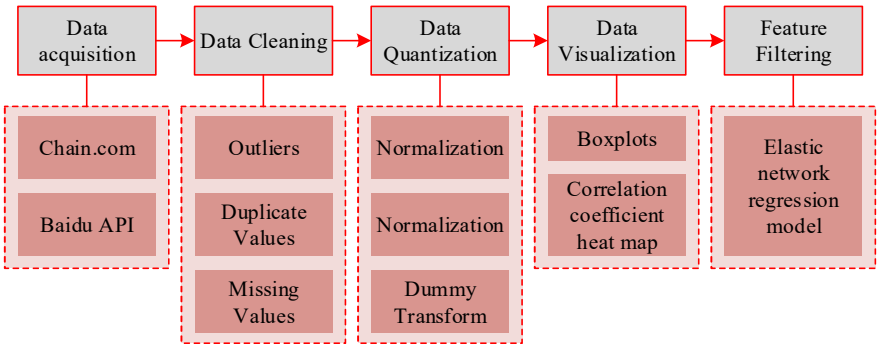


Figure 1: Data preprocessing flow chart

II. A.Data acquisition

First, Python is used to crawl the data of railroad transportation characteristics and housing market prices of four urban areas (Area A, B, C, and D) in a city in Chain.com, which was collected in November 2020. In Chain.com, each webpage only displays at most 80 pages of housing information, and if the webpage of housing rentals is crawled directly, the data that can be extracted is limited. However, in the webpage of house price, each region will then correspond to a new webpage displaying at most 80 pages of listing information, so the amount of extracted data will be greatly increased if crawling by region. Therefore, in this paper, we choose to implement the crawling work by region. The specific process of housing information in a city on the crawler chain home network is as follows: first, select the first page of the website where the housing market price information of the required crawler area is located, as the initial URL link, and then obtain detailed housing rental information by parsing the link, and then add pg*/ after the initial URL link, so as to obtain the housing market price information on the first page, and so on, and finally obtain 14387 housing price data in 4 administrative districts of a city.

The second is to use Python to call the Baidu API interface to crawl the new opening data of the rail transportation lines corresponding to the address of each listing. Combined with the actual situation, there will be differences between the housing prices of different transportation trips, so the new opening of transportation lines will have a certain impact on housing prices, but the address of the listings is text data and contains many different categories, which need to be analyzed by substituting them into the prediction model in order to understand the mechanism of the role between the two.

II. B.Data cleansing

In the process of collecting data on rail transportation characteristics and housing market price data, the data obtained are prone to problems such as duplications, missing and anomalies due to the fact that some of the data are incomplete and the information is not published in the process of obtaining the original data. If these problematic data are used directly for modeling, it will lead to overfitting of the model and problems such as poor generalization ability, high prediction error and low prediction accuracy. Therefore, it is necessary to clean the rail transportation characteristic data and housing market price data to find out the problematic data and deal with them to improve the overall prediction effect of the model. Data cleaning refers to the cleaning of problematic data according to certain criteria, which mainly deals with missing values, outliers and duplicated values in the data, and is an important step in data preprocessing. Therefore, this next subsection will carry out data cleaning from three aspects: duplicate value processing, missing value processing and outlier processing.

II. B. 1) Repeat value processing

On Chain Home Online, each listing information will correspond to a specific listing link, so this paper filters and deletes duplicate data with exactly the same information according to the listing link via Python. Finally, 2972 pieces of duplicate data are filtered out and 11415 pieces of data remain after deletion.

II. B. 2) Missing value processing

Missing values refer to the fact that the data for an attribute or attributes are vacant in the rail transit characteristics data and housing market price data. In the process of acquiring rail transportation characteristics data and housing market price data, there are often missing values. There are a number of reasons for missing values, including, on the one hand, the failure of the system to save the rail transportation characteristics and housing market price data, and on the other hand, subjective human errors or historical limitations that lead to missing values. If the missing

values are not handled, the effectiveness and accuracy of the model will be affected, and the results will not meet the expectations and cannot be interpreted. Therefore, it is necessary to deal with missing values. The processing methods of missing values mainly include eliminating missing values and utilizing interpolation to supplement missing values. Combined with the content of the study, this study adopts the method of eliminating missing values.

Elimination of missing values. If there is a large amount of missing data for a certain indicator in the rail transportation characteristics data and housing market price data, and at the same time, the indicator provides little valid information and contributes little to the process of model building and forecasting, the indicator can be deleted directly. However, if there are only a small number of missing values for an indicator in the rail transit characteristics data and housing market price data, there are two ways to deal with it: first, if the sample size of the rail transit characteristics data and housing market price data is large enough, and if the deletion of this small number of missing values has a negligible impact on the model as a whole, the method of deleting this small number of missing values can be taken directly. Secondly, if the data set provides a small sample size, the direct deletion of this small number of missing values may damage the integrity of the data, and at the same time, due to the insufficient amount of data, resulting in the model can not obtain comprehensive information, then you can use interpolation to fill the missing values.

II. B. 3) Handling of outliers

Outliers are usually data where some values in a feature of the rail transit characteristic data and the housing market price data deviate significantly from the rest of the values; it is also known as an outlier. Outliers generally exhibit non-conforming characteristics in a data set. If these outliers are ignored and not properly handled during the model building process, the model will deviate from the reasonable prediction direction due to fitting these outliers during the training process, and eventually problems such as large prediction error, prediction results that do not meet the expectations, and low prediction accuracy may occur. Therefore, before constructing the model, it is necessary to carry out the work of outlier analysis, including the detection and treatment of outliers. Among them, box-and-line diagrams, the principle of 3σ, and simple statistical methods are often used to detect whether there are outliers in the data set, and to deal with outliers by directly deleting outliers and calculating the average value to correct the outliers.

II. C.Quantification of data

The quantification of rail transportation characteristic data and housing market price data includes three aspects of qualitative data dummy transformation, data normalization processing and data normalization processing. In the original dataset obtained in this paper, it contains not only numerical data, but also a large number of category data. However, most of the models built based on machine learning algorithms can only predict and analyze numerical data. Therefore, in order to facilitate the subsequent use of machine learning models for predictive analysis of the data in this paper, it is necessary to transform the category-type data in this paper into data types acceptable to machine learning models by means of qualitative data dummy transformations. In addition, many models are built on the premise of assuming that the data obeys a normal distribution, which is mainly due to the fact that if the data does not obey a normal distribution, it will lead to a large error in the model prediction, so it is very necessary to normalize the data. Moreover, since the scale of different variables is not the same, if the variables are directly brought into the model for training, it may have an impact on the model prediction effect. Therefore, it is necessary to normalize the data of rail transportation characteristics and housing market price data to avoid the problem that the model may have low prediction performance due to the influence of the scale.

III. Machine learning-based predictive models

III. A. Primary Learner Selection

In order to improve the prediction accuracy and generalization ability of the combination of algorithms for rail transportation and housing market prices, the primary learners are selected by considering the correlation of the algorithms' arithmetic laws and the performance of regression, and three algorithms in the Boosting series of algorithms, namely, XgBoost, LightGBM, and CatBoost, are finally selected as the primary learners of the prediction model.

III. A. 1) XgBoost algorithm

The XgBoost algorithm is based on the GBDT algorithm framework, which has a great improvement in rail transportation and housing market price computing time and accuracy, and its principle is also to input rail transportation data and housing market price data in the first decision tree, randomly get the first sample predicted value and calculate the residuals between the true value and the predicted value, and the second tree's task is to fit the parametric residuals of the first decision tree, get the residuals of the second decision tree, the next tree to fit

the residuals of the former, and so on, until the overall desired objective set to stop the iteration [31], [32]. The overall desired objective function is:

$$L(\Phi) = \sum_i l(\hat{y}, y) + \sum_k \Omega(f_k) \tag{1}$$

Among them:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{2}$$

Equation (1) where l represents the loss function, which represents the degree of difference between the true value y and the predicted value $\hat{y}_i$ of the i rd sample, usually second order derivable. Equation (1) represents the regular term that limits the complexity of the model, where λ represents the hyperparameter, T is the number of leaf nodes, the larger the hyperparameter the smaller the number of leaf nodes. w is the weight of the leaf nodes, the larger the weight the stronger the restriction, in avoiding or fitting at the same time will make the model more conservative. Generally want to directly optimize the objective function is more difficult, so decomposed into multiple steps, add a decision tree in each step, complete the small goal of fitting the previous tree, stepwise reduction of the loss function, so as to gradually optimize the predicted value.

Compared with the GBDT algorithm XgBoost innovates by adding regular terms, which reduces the risk of overfitting by limiting the number of leaf nodes, and at the same time reduces the complexity of the model. The learning process also uses Taylor's formula expansion to join the first-order derivative and second-order derivative, while the GBDT algorithm uses the first-order derivative expansion, this change can be customized when tuning the parameters of the loss function, which also improves the accuracy of the prediction.XgBoost can be calculated by sampling, rather than all the features are operated to reduce the risk of overfitting. When there are more missing values in the sample XgBoost can divide the direction of the missing values according to the gain, which improves the efficiency of the model operation.

III. A. 2) LightGBM Algorithm

XgBoost algorithm and LightGBM algorithm leaf node splitting method is shown in Fig. 2, LightGBM is also an innovative algorithm relying on the GBDT framework for improvement, and its significant features are fast computing speed, small operating memory, but the accuracy rate has not decreased [33], [34]. LightGBM innovation is mainly focused on the rail transportation characteristics data and housing market price data optimization. For the optimization of the number of samples, LightGBM introduces one-sided gradient sampling algorithms (GOSs), which compressed the training dataset by discarding samples with lower weights during the model training, which greatly reduces the computing power without affecting the accuracy of the computation, and at the same time enhances the complexity of the base model to improve the generalization ability. For the part of feature number optimization, LightGBM adds the mutually exclusive feature bundle (EFB) algorithm, which turns the feature set into an ensemble problem for graph coloring, and bundles two mutually exclusive or nearly exclusive features together into one feature, the bundle is discretized using a histogram algorithm for continuous data.

There is also a difference between XgBoost algorithm and LightGBM algorithm in the way of splitting the leaf nodes, XgBoost algorithm splits each leaf node of the same layer without any difference, while LightGBM algorithm splits the continuous type rail transit feature and housing market price feature into multiple discrete features using histogram algorithm, and uses histograms to count the values of the discretization and find the leaf with the largest splitting gain from the same layer of leaves Find the leaf with the largest splitting gain in the same layer of leaves, and only let this leaf node split, which undoubtedly greatly improves the efficiency of rail transportation and housing market price prediction, the leaf node splitting schematic of the two algorithms is shown in the following figure:

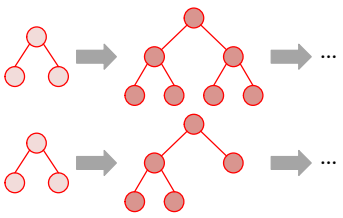


Figure 2: leaf node splitting mode of XgBoost algorithm and LightGBM algorithm

III. A. 3) CatBoost Algorithm

CatBoost algorithm is also an improved algorithm of GBDT algorithm, which is notable for its ability to deal with categorical features, fewer parameters, and at the same time innovatively solves the problem of machine learning's bias in predicting the price of rail transit special and housing market, which further enhances the algorithm's ability to generalize and accuracy [35].

First of all, unlike the previous two is CatBoost algorithm is based on symmetric decision trees, ordinary decision trees with different leaves in the same layer are composed of different judgment conditions, while the symmetric decision trees with different leaves in the same layer of the same judgment conditions, i.e., the leaves are symmetric to each other, because of the large number of decision trees so it does not therefore appear overfitting problems, but it simplifies the fitting pattern, XgBoost algorithm and the LightGBM algorithms need to go through the whole tree from beginning to end when predicting each sample, while the CatBoost algorithm only needs to carry out array operations, which speeds up the prediction speed.

Second, the CatBoost algorithm makes breakthrough improvements in the processing of rail transportation feature data and housing market price features. Machine learning is often used to deal with category-type features with unique heat coding and label coding, unique heat coding is only applicable to the case of a small number of feature categories, and if a large number of feature categories use unique heat coding, it will cause dimensionality disaster due to a large number of redundant features. Labeling encoding converts numerical features with large differences and is only applicable to some cases. Usually we need to preprocess the category-based features in advance, which increases the workload of the study.CatBoost algorithm adds a priori distribution terms based on the GBDT algorithm that takes the mean of the category labels as the node splits, which reduces the influence of the noisy data, and its calculation formula is as follows:

$$x_{\sigma_y,k}=\frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k}=x_{\sigma_j,k}\right]\cdot Y_{x_{\sigma y}}+\alpha\cdot P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k}=x_{\sigma_j,k}\right]+\alpha} \tag{3}$$

where $\alpha>0$, represents the weighting coefficient, and P is the a priori term.

CatBoost algorithm in the processing of rail transportation features and housing market price features at the same time also automatically combines features, the category features of any combination of the new category is also considered as a new feature, in the category-type features into numerical features will be lost part of the information, while the category combination of features to make up for this defect at the same time but also the features of the more effective refining.

III. B. Combinatorial model construction based on Boosting algorithm

After confirming the selection of primary learners, the next step is to combine the models by integrating learning.Bagging and Boosting integration requires that individual learners are homogeneous, while the learning algorithms of the three kinds of Boosting are not the same, which belongs to heterogeneous learners, and Stacking integration is more suitable. In this paper, we plan to train XgBoost, LightGBM and CatBoost models as the primary learners in the first stage, and usually choose a simple basic model as the meta-model in the second stage, and this paper solves the regression problem, so we use the multiple linear regression model as the meta-model, and the overall idea is shown in Figure 3.

The specific steps of combinatorial prediction model construction are as follows:

(1) Divide the cleaned sample data into training set and test set.

(2) The XgBoost, LightGBM and CatBoost algorithms are used as primary learners to train prediction on the training set samples through cross-validation, respectively, and obtain the respective prediction results.

(3) The training results are used as the training set for the meta-model, and the prediction results of the impact of rail transportation on housing market price are output after training.

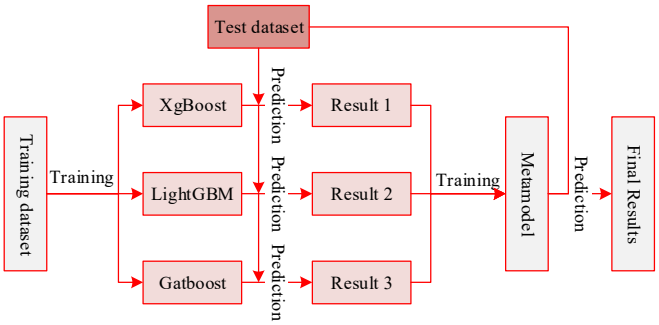


Figure 3: overall idea

IV. Forecasting the impact of rail transportation on housing prices

IV. A. Feature Measurement

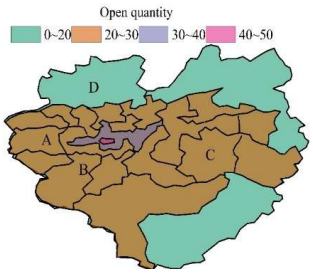
It is especially important to select and define the rail transportation types in the prediction model scientifically and accurately. On the basis of crawling data, four types of rail transportation, namely, subway, bus, light rail and tram, are selected as rail transportation characteristic variables. These four types of rail transportation characteristics are closely related to the fluctuation of housing prices and the construction planning of the city, which is of great significance for the study. The time threshold of life circle travel is generally 30 minutes, and the corresponding walking distance threshold is about 2000 meters, so the maximum walking distance can be set to 2000 meters, and the buffer zone for pedestrian travel is set by Arcgis tool, and the feature size is measured by counting the number of rail transit lines opened in four urban areas of a city.

IV. A. 1) Study area

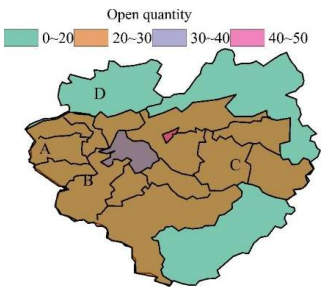
In this paper, the main urban area of a city is selected as the research object, including parts of four districts, namely, District A, District B, District C and District D, which covers most of the urban area in the center of a city, and is the core area of social and economic development of a city. The main urban area of a city is currently built with one main and one secondary city centers, of which Quan Cheng Square and its surrounding areas are the main center and the old city center, integrating commercial and business, culture and entertainment, tourism and leisure, and public service facilities. The secondary center is the central business district, including the CBD, convention and exhibition center, etc., which undertakes the central functions of financial services, sports and leisure, and commercial and trade services.

IV. A. 2) Spatial characteristics of transportation tracks

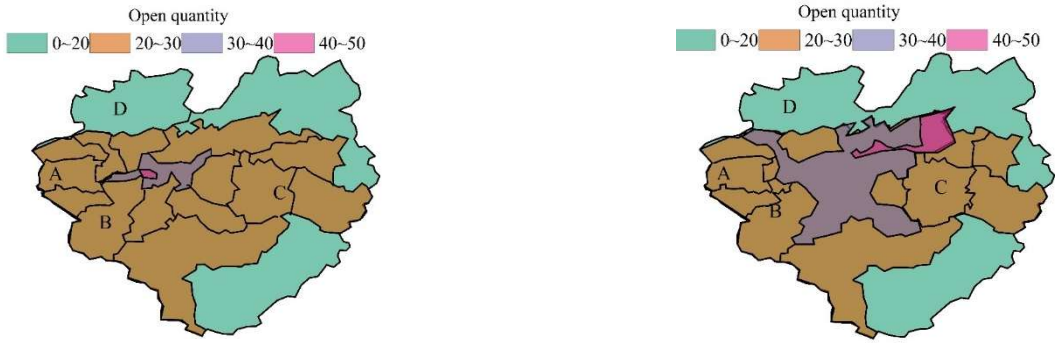
The rail transportation involved in this study includes four types of public services: subway, bus, light rail, and tram. The urban structure of a city is similar to a monocentric city, and it is measured by ArcGIS10.3 software that about 13% of the housing is located within 3,000 meters from the city center, and about 45% of the housing is located within 5,000 meters from the city center, and the spatial characteristics of the rail line traffic in a certain city are shown in Fig. 4, in which (a) to (d) denote the subway, bus, light rail, and tram, respectively, and the rail traffic line spatial characteristics can be divided into four levels. Taking subway as an example, the number of subway tracks opened in urban area D is relatively small, while the number of subway tracks opened in urban areas A, B, and C is mostly 20-30, a small portion of 30-40, and a very small portion of 40-50, and urban area D is a suburb with a large number of industrial buildings.



(a)Subway



(b)Public transportation



(c)Light rail (d)Tram

Figure 4: Spatial characteristics of rail line traffic in a city

IV. A. 3) Spatial characteristics of housing prices

In recent years, guided by the “Overall Territorial Spatial Plan of a Certain City (2020-2035)”, and on the basis of the industrial plan of a certain city, and in line with the deployment of the new pattern of urban development of “Strong in the East, Rising in the West, South America, Rising in the North, and Superior in the Middle”, the management of the total amount of land supply and spatial coordination have been strengthened, resulting in a staggered and integrated development. According to the statistics of a city's Real Estate Registration Center, by the end of 2020, the total number of supporting housing units in the built-up area of a city will be approximately 3.5 million, essentially meeting the housing needs of a wide range of types of households.

(1) Global analysis

The results of the global analysis of the spatial characteristics of housing prices are shown in Figure 5, where the average unit price of each housing neighborhood is calculated in order to reveal the spatial pattern of housing prices. The results found that the spatial variation of housing prices in a city is large, with an average housing price of 1.49 million yuan and an average unit price of 16,442 yuan/m², with the total price mainly distributed in the range of 600,000-2,000,000 yuan (accounting for about 43.82% of the total number of housing units) and the unit price mainly distributed in the range of 13,000-28,000 yuan/m² (accounting for about 49.5% of the total number of housing units). The number and price of housing in different areas of a city vary, with high-priced housing mainly located in the southeast of the city, and housing prices in the periphery of the city are relatively low, with an increasing trend in housing prices from the west to the east, which reflects the city's “strong east” urban development strategy. The global spatial correlation of housing prices in a city is further analyzed. Spatial autocorrelation refers to the correlation between the values of variables in similar regions. In this paper, we use the spatial autocorrelation analysis command in Geoda software to calculate Moran`s I index to test the global spatial autocorrelation of housing prices in a city, and further analyze the global spatial correlation and agglomeration characteristics of housing prices in a city. The results yielded a Moran`s I index of 0.708, which is significant at the 0.01 significance level, indicating that housing prices in a city are not randomly distributed in space but have positive spatial correlation, with closer neighborhoods having similar prices and a stronger degree of agglomeration of different grades of housing.

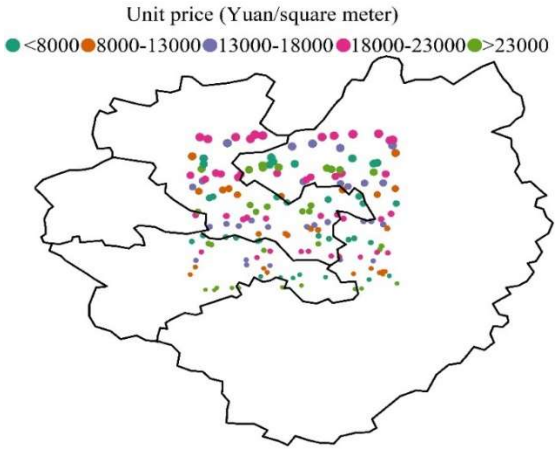


Figure 5: Results of global analysis of spatial characteristics of housing prices

(2) Local analysis

In order to detect the local correlation law of housing prices in a city, this paper uses the univariate local autocorrelation analysis tool in Geoda software to identify the local clustering characteristics of housing prices, and combines with the spatial local clustering map to further explore the spatial clustering of housing prices in local areas, the results of the spatial characteristics of the housing prices in localized analyses are shown in Fig. 6, where the high-high refers to the clustering of high housing prices, i.e., high housing prices are also surrounded by high housing prices, and low-low refers to the agglomeration of low housing prices, i.e., low housing prices are surrounded by low housing prices, and high-low and low-high respectively indicate that the surrounding housing prices are significantly different. The high and low agglomeration areas of housing prices in the main city of a certain city are obvious, while the low value agglomeration areas are mainly located in the northwestern part of the main city and other urban fringe areas. It can be seen that the urban development of a city shows a trend of eastward expansion, while the clustering of housing prices in the old city area of a city is more of a “non-significant” feature. The city's urban master plan focuses on urban preservation, which largely limits urban renewal in the inner city.

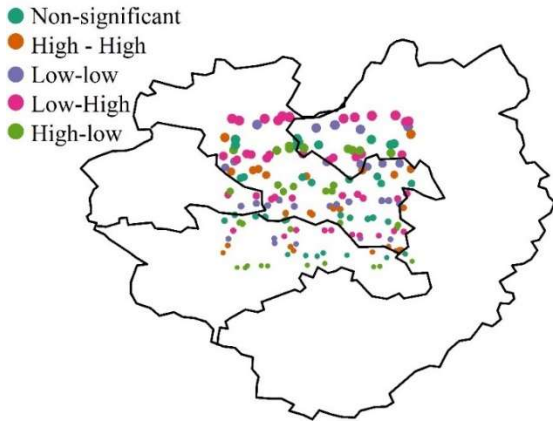


Figure 6: Housing price spatial agglomeration map

IV. B. Analysis of forecasting model applications

IV. B. 1) Characterization screening analysis

With the help of Pearson's correlation coefficient on the above four types of rail transportation spatial characteristics of variables and housing price characteristics correlation analysis, the calculated Pearson's correlation coefficient as an important indicator of feature screening, the results of the feature screening analysis is shown in Figure 7, in which X1 ~ X4 in the figure respectively represents the subway, buses, light rail, tram, and Y is the housing market price. The data analysis in the figure shows that the Pearson correlation coefficients of the four types of spatial characteristics of rail transportation variables and housing price characteristics are 0.06, -0.112, -0.33, -0.164, which all show significant correlation, and in the four types of spatial characteristics of rail transportation variables, the opening of the subway's new line has the greatest impact on the housing price characteristics on the price

fluctuations of the housing market in the urban areas, and the subway around the housing and buses The impact on housing market price is also larger, reflecting the importance of housing transportation convenience for housing market price, accurately grasping the factors that have a greater impact on housing market price, but also for housing providers and government departments to provide support for housing price estimation, to develop a more reasonable market price to meet the expectations of consumers. The metro spatial characteristic variables are used as inputs to the housing market price prediction model based on the Boosting algorithm, and the output of the model is the predicted value of housing prices. The following section will analyze the application effectiveness of the model in this paper in detail.

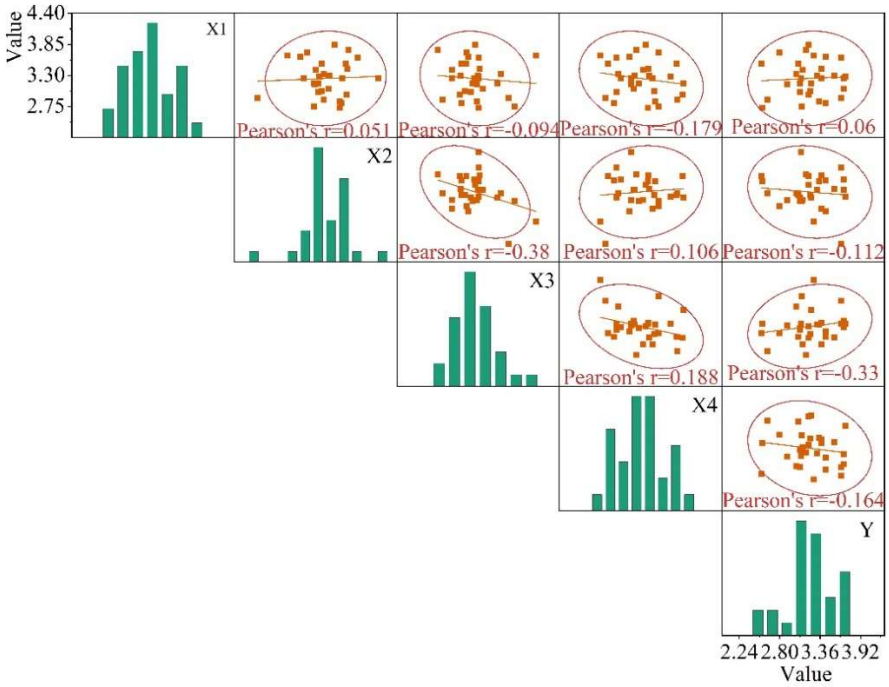


Figure 7: Feature screening analysis results

IV. B. 2) Model parameters

The most important step in the model application analysis process is to tune the parameters, and the model parameters are shown below:

(1) General parameters: the main parameters.

(2) Booster parameters: can adjust the prediction effect of the model, the most important part.

(3) Learning target parameters: for us to get a good result first, by adjusting the Booster parameters we can get a more excellent prediction value. In python Scikit-learn provides a function that can help us better tune the parameters, through sklearn.model_selection.GridSearchCV this function can get the best parameters. First we will use the default parameters set to initial values, for example:

(a) learning_rate: 0.5.

(b) n_estimators:: 1000.

(c) max_depth: 10.

(d) min_child_weight: 5.

(e) subsample: 1.6.

(f) colsample_bytree: 1.6.

(g) gamma: 0.

(h) reg_alpha: 0.

(i) reg lambda: 1.

By initializing our parameters with these common parameters, parameter searching through Scikit-learn's sklearn.model_selection.GridSearchCV inside python is an excellent means of optimizing parameters.

IV. B. 3) Application effectiveness analysis

The housing market price of a city in November to December 2020 is selected as the real value of this study, while the output of the housing market price prediction model based on Boosting algorithm is the predicted value, and the

model application effectiveness is detected through the error analysis of the real value and the predicted value, and the prediction analysis of the housing market price of a city in November to December 2020 is shown in Table 1, and the red numbers in parentheses in the table are Negative values (this is the case that the model predicted value is greater than the actual value, resulting in a negative error). Based on the data in the table, it can be seen that the price fluctuation range of a city's housing market from November to December 2020 is 8,000 to 9,000 yuan/square meter, and it can also be seen that the error value of the two is in the range of 0 to 64, and the final calculation of the two comprehensive error rate is 0.3697%, although the actual house price prediction results and the real house price can't be completely coincident with the actual house price, but the comprehensive error rate is completely in the acceptable range. Although the actual house price prediction results and the real house price cannot be completely matched, the combined error rate is completely within the acceptable range, and the combined prediction method using Boosting algorithm can have high prediction accuracy and robustness in the housing market price prediction, which provides theoretical guidance for the city planning, design and construction economy.

Table 1: Forecast of housing market price in a city from November to December 2020

Date	Actual value(Yuan per square meter)	Predictive value(Yuan per square meter)	Error	Date	Actual value(Yuan per square meter)	Predictive value(Yuan per square meter)	Error
11/1	8128	8107	21	12/2	8912	8931	(19)
11/2	8192	8205	(13)	12/3	8164	8141	23
11/3	8355	8355	0	12/4	8327	8330	(3)
11/4	8875	8891	(16)	12/5	8780	8803	(23)
11/5	8571	8589	(18)	12/6	8076	8057	19
11/6	8920	8949	(29)	12/7	8562	8549	13
11/7	8124	8101	23	12/8	8696	8721	(25)
11/8	8556	8560	(4)	12/9	8577	8584	(7)
11/9	8159	8143	16	12/10	8715	8694	21
11/10	8716	8732	(16)	12/11	8154	8169	(15)
11/11	8504	8520	(16)	12/12	8649	8651	(2)
11/12	8847	8856	(9)	12/13	8546	8541	5
11/13	8497	8527	(30)	12/14	8337	8343	(6)
11/14	8433	8424	9	12/15	8509	8505	4
11/15	8888	8898	(10)	12/16	8619	8606	13
11/16	8119	8132	(13)	12/17	8836	8852	(16)
11/17	8976	8964	12	12/18	8893	8900	(7)
11/18	8551	8539	12	12/19	8817	8795	22
11/19	8729	8736	(7)	12/20	8123	8106	17
11/20	8882	8818	64	12/21	8550	8556	(6)
11/21	8426	8409	17	12/22	8226	8252	(26)
11/22	8547	8551	(4)	12/23	8443	8459	(16)
11/23	8101	8054	47	12/24	8113	8123	(10)
11/24	8329	8309	20	12/25	8198	8203	(5)
11/25	8860	8838	22	12/26	8674	8672	2
11/26	8927	8950	(23)	12/27	8699	8704	(5)
11/27	8819	8790	29	12/28	8758	8765	(7)
11/28	8519	8492	27	12/29	8607	8628	(21)
11/29	8054	8046	8	12/30	8633	8639	(6)
11/30	8767	8760	7	12/31	8141	8099	42
12/1	8551	8523	28				

V. Conclusion

In this paper, we use Python to crawl the rail transit feature data of four urban areas (Area A, Area B, Area C, and Area D) in a city in Chain.com, and then pre-process and analyze this data. The feature variables with the highest degree of correlation were screened out through the features as the input variables of the prediction model, and the housing market price prediction model was constructed by integrating the XgBoost algorithm, LightGBM algorithm,

and CatBoost algorithm, and the model was applied to predict the impact of the new lines of rail transit on the housing price. Among the four types of rail transportation spatial characteristic variables, the subway spatial characteristic variable has the highest correlation, with a specific value of 0.06, which reflects that the opening of new subway lines around housing has a greater degree of influence on housing prices. In addition, the combined error rate between the predicted value and the actual value is 0.3697%, which is within the permissible range of the error rate, indicating that the Boosting combination algorithm has a particularly significant application in the prediction of price fluctuations in the housing market, which provides guidance to the government and housing companies in the regulation of housing prices, and improves the urban planning, design, and construction economy.

Funding

National Natural Science Foundation of China (Project No. 52178436).

References

[1] Zhao, J., Liu, J., Yang, L., Ai, B., & Ni, S. (2021). Future 5G-oriented system for urban rail transit: Opportunities and challenges. China Communications, 18(2), 1-12.

[2] Awad, F. A., Graham, D. J., AitBihiOuali, L., & Singh, R. (2023). Performance of urban rail transit: a review of measures and interdependencies. Transport Reviews, 43(4), 698-725.

[3] Yan, H., Gao, C., Elzarka, H., Mostafa, K., & Tang, W. (2019). Risk assessment for construction of urban rail transit projects. Safety science, 118, 583-594.

[4] AlQuhtani, S., & Anjomani, A. (2019). Do rail transit stations affect housing value changes? The Dallas Fort-Worth metropolitan area case and implications. Journal of Transport Geography, 79, 102463.

[5] Pan, Q., Pan, H., Zhang, M., & Zhong, B. (2014). Effects of rail transit on residential property values: Comparison study on the rail transit lines in Houston, Texas, and Shanghai, China. Transportation Research Record, 2453(1), 118-127.

[6] Kang, C. D. (2019). Spatial access to metro transit villages and housing prices in Seoul, Korea. Journal of Urban Planning and Development, 145(3), 05019010.

[7] Rennert, L. (2022). A meta-analysis of the impact of rail stations on property values: Applying a transit planning lens. Transportation Research Part A: Policy and Practice, 163, 165-180.

[8] Qiu, F., & Tong, Q. (2021). A spatial difference-in-differences approach to evaluate the impact of light rail transit on property values. Economic Modelling, 99, 105496.

[9] Mulley, C., & Tsai, C. H. (2017). Impact of bus rapid transit on housing price and accessibility changes in Sydney: A repeat sales approach. International Journal of Sustainable Transportation, 11(1), 3-10.

[10] Lee, J. K. (2022). New rail transit projects and land values: The difference in the impact of rail transit investment on different land types, values and locations. Land Use Policy, 112, 105807.

[11] Cengiz, E. C., & Çelik, H. M. (2019). Investigation of the impact of railways on housing values; the case of Istanbul, Turkey. International Journal of Transport Development and Integration, 3(4), 295-305.

[12] Forouhar, A. (2016). Estimating the impact of metro rail stations on residential property values: evidence from Tehran. Public Transport, 8(3), 427-451.

[13] Pan, Q. (2019). The impacts of light rail on residential property values in a non-zoning city. Journal of Transport and Land Use, 12(1), 241-264.

[14] Trojanek, R., & Gluszak, M. (2018). Spatial and time effect of subway on property prices. Journal of Housing and the Built Environment, 33, 359-384.

[15] Gadziński, J., & Radzimski, A. (2016). The first rapid tram line in Poland: How has it affected travel behaviours, housing choices and satisfaction, and apartment prices?. Journal of Transport Geography, 54, 451-463.

[16] Rojas, A. (2024). Train stations' impact on housing prices: Direct and indirect effects. Transportation Research Part A: Policy and Practice, 181, 103979.

[17] Zhou, Y., Tian, Y., Jim, C. Y., Liu, X., Luan, J., & Yan, M. (2022). Effects of public transport accessibility and property attributes on housing prices in Polycentric Beijing. Sustainability, 14(22), 14743.

[18] Vichiensan, V., Wasuntarasook, V., Hayashi, Y., Kii, M., & Prakayaphun, T. (2021). Urban rail transit in Bangkok: Chronological development review and impact on residential property value. Sustainability, 14(1), 284.

[19] Wu, W., Zheng, S., Wang, B., & Du, M. (2020). Impacts of rail transit access on land and housing values in China: a quantitative synthesis. Transport Reviews, 40(5), 629-645.

[20] Yang, L., Chu, X., Gou, Z., Yang, H., Lu, Y., & Huang, W. (2020). Accessibility and proximity effects of bus rapid transit on housing prices: Heterogeneity across price quantiles and space. Journal of Transport Geography, 88, 102850.

[21] Wang, Y., Liu, X., & Wang, F. (2018). Economic impact of the high-speed railway on housing prices in China. Sustainability, 10(12), 4799.

[22] Jin, T., Cheng, L., Liu, Z., Cao, J., Huang, H., & Witlox, F. (2022). Nonlinear public transit accessibility effects on housing prices: Heterogeneity across price segments. Transport Policy, 117, 48-59.

[23] Shi, D., & Fu, M. (2022). How Does Rail Transit Affect the Spatial Differentiation of Urban Residential Prices? A Case Study of Beijing Subway. Land, 11(10), 1729.

[24] Ren, P., Li, Z., Cai, W., Ran, L., & Gan, L. (2021). Heterogeneity Analysis of Urban Rail Transit on Housing with Different Price Levels: A Case Study of Chengdu, China. Land, 10(12), 1330.

[25] Zhu, Z., Zhu, Y., Liu, R., Zhang, L., & Yuan, J. (2022). Examining the Effect of Urban Rail Transit on Property Prices from the Perspective of Sustainable Development: Evidence from Xuzhou, China. Buildings, 12(10), 1760.

[26] Li, J., & Huang, H. (2020). Effects of transit-oriented development (TOD) on housing prices: A case study in Wuhan, China. Research in Transportation Economics, 80, 100813.

[27] Yang, L., Chau, K. W., Szeto, W. Y., Cui, X., & Wang, X. (2020). Accessibility to transit, by transit, and property prices: Spatially varying relationships. Transportation Research Part D: Transport and Environment, 85, 102387.
[28] Tan, R., He, Q., Zhou, K., & Xie, P. (2019). The effect of new metro stations on local land use and housing prices: The case of Wuhan, China. Journal of Transport Geography, 79, 102488.
[29] Liu, X., Jiang, C., Wang, F., & Yao, S. (2021). The impact of high-speed railway on urban housing prices in China: A network accessibility perspective. Transportation Research Part A: Policy and Practice, 152, 84-99.
[30] Yang, L., Chen, Y., Xu, N., Zhao, R., Chau, K. W., & Hong, S. (2020). Place-varying impacts of urban rail transit on property prices in Shenzhen, China: Insights for value capture. Sustainable Cities and Society, 58, 102140.
[31] Jianxing Liao,Yachen Xie,Pengfei Zhao,Kaiwen Xia,Bin Xu,Hong Wang... & Hejuan Liu. (2024). Probabilistic assessment of the thermal performance of low-enthalpy geothermal system under impact of spatially correlated heterogeneity by using XGBoost algorithms. Energy133947-133947.
[32] Min Huang,Hang Zhao & Yazhou Chen. (2024). Research on SAR image quality evaluation method based on improved harris hawk optimization algorithm and XGBoost. Scientific reports(1),28364.
[33] Huihui Lian,Ying Ji,Menghan Niu,Jiefan Gu,Jingchao Xie & Jiaping Liu. (2025). A hybrid load prediction method of office buildings based on physical simulation database and LightGBM algorithm. Applied Energy(PC),124620-124620.
[34] Long Li. (2024). LightGBM integration with modified data balancing and whale optimization algorithm for rock mass classification. Scientific Reports(1),23028-23028.
[35] Fang Xing,Hui Li & Tianyu Li. (2024). Deformation Modeling and Prediction of Concrete Dam Using Observed Air Temperature and Enhanced CatBoost Algorithm. Water(23),3341-3341.