Predicting the Trend of Rental Housing Prices in Shenzhen Based on Stacking Regression Models

Hongzhan Li, Hao Lin, Yuntao Jia*

Zhuhai Campus, Beijing Institute of Technology, Zhuhai, 519088 China

*Corresponding author: jyt-002@126.com

Abstract. Rent price prediction is an important research direction because it can help tenants, landlords, and real estate companies better understand market dynamics and formulate strategies. In this paper, we obtained and preprocessed the Shenzhen rent information from the Chain Home and Shell websites, and the processed data were divided into eighty-two divisions, and six single models were established, Multiple regression model, Ridge regression, LASSO regression, Multi-layer Perceptual Machine, Random Forest regression, and XGBoost regression for rent prediction comparisons, and finally, in order to improve the prediction performance of the model, finally, based on the aspect of improving the prediction performance in this paper, the Stacking regression model is used, a single model as the primary learner, and the linear regression model as the secondary learner to establish the Stacking integrated regression model as the prediction model of Shenzhen rent, the final training model has an MSE of 0.165, an R2 of 0.84, and an MRE of 1.179, which are better than the previous models in all the three evaluation indexes, which indicate that the The prediction performance of the model has been significantly improved, which brings great reference significance to home buyers, landlords and real estate companies.

Keywords: Rent projections, unique thermal code, Data preprocessing, Stacking regression model

1 Introduction

1.1 Background to the problem

As a first-tier city in China, Shenzhen's fluctuating rent prices have a direct impact on a large amount of tenants and landlords. Tenants need to budget their rental expenses, while landlords need to determine reasonable rents to ensure revenue. However, predicting rent prices is complicated by the fact that rent prices are affected by many factors, such as location, size of the house, and degree of renovation. In addition, rental prices in Shenzhen are also affected by the macroeconomic environment, policy adjustments and other factors, which are often unpredictable. Therefore, developing a model that can accurately predict rent prices in Shenzhen is of great practical importance to both tenants and landlords.

C. Chen et al. (eds.), *Proceedings of the 3rd International Conference on Digital Economy and Computer Application (DECA 2023)*, Atlantis Highlights in Computer Sciences 17,
https://doi.org/10.2991/978-94-6463-304-7_73

Then, on the problems related to predicting house prices, many scholars have utilized a single model for research. Zhu Haiyu ([1]) and others limited hotspot areas and calculated regional house prices by utilizing fine-grained house price data and POI data, and then used the XGBoost algorithm to train the fitted hotspot area house prices for prediction, and the results showed that the accuracy of the prediction was in line with expectations. Xu Dandan ([2]) utilized new commodities in Xi'an city as data and established a multiple regression model and a BP neural network model for house price prediction respectively, and finally compared the prediction effect of these two models in order to find a suitable house price prediction method. Deng Smooth ([3]) et al. used gamma regression model and inverse Gaussian model to predict the house price dataset of Xindian District, New Taipei City, Taiwan Province, and finally compared the two models by using MSE to compare the two models are good or bad, and finally concluded that the gamma regression model is more suitable for prediction. Fu Qiwe ([4]) et al. established a time series ARIMA model for house price prediction for the house price problem in Hengyang city, in order to make the data as a smooth and non-white noise series, the data were smoothed with pure random test, and finally the optimal ARIMA (1,2,0) was derived. Then in the next twelve months of Hengyang city's house prices were predicted in the short term. Ren Ziming ([5])] used the data of Xicheng District and Haidian District of Beijing respectively to carry out regression analysis first, and then used linear function and trigonometric function to analyze the growth trend and periodicity of house price, and got good prediction effect, and finally also used gray prediction GM (1,1) model for prediction, and got the model with more accurate precision. Li Yuqi ([6]) for the problem of house price, in the well-known kaggle competition website to collect house price data, through the establishment of "regression decision tree" random forest model, and choose ID3 generation algorithm, according to the information gain to calculate the importance of the features, and finally found the important factors affecting the house price. Meanwhile, in the experimental process, the method of cross-validation was used to prevent overfitting, and good prediction results were obtained in the end.

It is known through previous research that single models have achieved good prediction results in terms of predicting house prices. In order to get a better house price prediction model, this paper studies an integrated algorithm that synthesizes the advantages and disadvantages of each single model in the prediction of house price, and then gets a more accurate prediction model.

This study initially collected and cleaned rental data from the Lianjia and Beike websites for Shenzhen. Next, various individual regression models were used for an initial prediction of rental prices, and the performance of each model was compared. Finally, to further improve the accuracy of the predictions, this study adopted a Stacking regression model for ensemble learning, achieving superior prediction results.

1.2 Research significance

This study aims to use machine learning to predict rent prices in Shenzhen in order to better understand the various factors affecting prices and to provide tenants and landlords with a basis for decision-making. This is helpful for policy makers to regulate the

real estate market more accurately. Meanwhile, real estate developers can plan and price their properties more effectively accordingly. For tenants and landlords, accurate forecasts help them make more informed financial decisions. Overall, this study has practical implications for all participants in the Shenzhen real estate market.

2 Data sources

2.1 Data acquisition

The data studied in this paper comes from the Chain.com and Shell Rent website, which are the largest real estate online trading platforms in China and provide a large amount of information on rental listings. In this paper, we use a program written in Python to obtain the information of rental listings in Shenzhen. The information obtained includes the number of the listing, the city, the district, the street, the name of the district, the area, the leasing method, the orientation, the monthly rent, the billing method, the room structure (room, hall, bathroom), the time of occupancy, the lease period, the way to see the room, the floor, the total floor, whether there is an elevator, whether there is a parking space, the way to use water, the way to use electricity, the way to use gas, the way to use heating, and the way to decorate, etc. This information constitutes our original data, and we use Python to obtain the information of rental listings in the Shenzhen area. This information constitutes our original dataset and can provide a rich data resource for our study.

3 Data preprocessing

3.1 Handling of duplicate values

The next step is to preprocess the raw data, the first thing to deal with is the duplicate values of the data, the acquired data has a total of 5427, by using python to check that there are 471 duplicate values, after removing the duplicates, there are 4956 pieces of data left.

3.2 Handling of missing values

Due to the acquisition of data, there are a lot of "no data" and "no" data, this time the two data are replaced with "nan", and then in each column of data were calculated in the number of missing values and the proportion of missing([7]), so that you can get the following table 1:

Table 1. Missing value statistics

features	Number of missing values	Percentage missing
billing	196	3.954802
family or clan	2	0.040355

Lease Term	2471	49.858757
view a house	1	0.020178
…	…	…
parking	3136	63.276836
water consumption	442	8.918483
electricity consumption	428	8.635997
natural gas	90	1.815981
heating	4777	96.388216
decoration method	746	15.052462

According to Table 1, the proportion of missing values for both features "parking" and "heating" is more than 50%, of which the proportion of missing values for "parking" is about 63%, and the proportion of missing values for "heating" is about 96%. The proportion of missing values for "heating" is about 96%, and all the features of "heating" must be eliminated because the proportion of missing values is too large, and even if the original data is kept, too many categories of nan will have an impact on the results. "The reason why I don't consider deleting "parking space" is that I have consulted the customer service of Chain Home and Shell, and found that there are indeed some houses with no data or no parking space in the column of "parking space". I replaced all the missing values in the column of "parking space" with "no parking space", and then, through the observation of the data, we can see that there is only one category of "decoration" except for "nan". "and the proportion of "nan" is also small, so the feature "decoration method" is also eliminated, and finally all the remaining features are filled with mean values.

3.3 Handling of outliers

In this paper, we adopt the method of drawing box-and-line diagrams to identify outliers, and visualize and detect the characteristic variables "area" and "monthly rent", as shown in the figure1 and figure2 below:

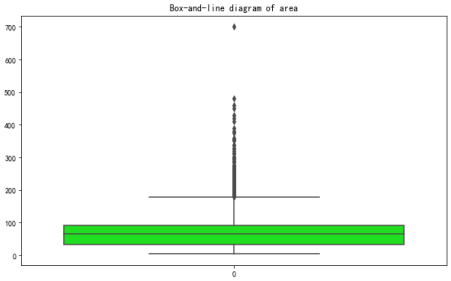

Fig. 1. Box-line diagram of area

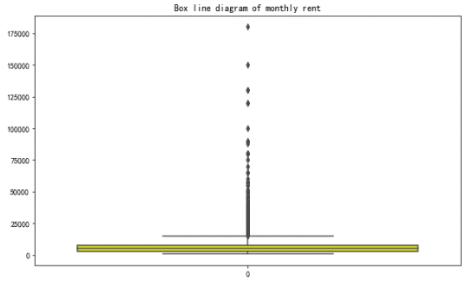

Fig. 2. Box line diagram of monthly rent

As can be seen from the above figure, there are many outliers in the area and monthly rent columns, which may be due to the fact that some areas have particularly

high rent prices and large floor areas, i.e., there are many rich neighborhoods in Shenzhen. The outliers are eliminated because they can easily affect the accuracy of the model.

3.4 Quantification of characteristic variables

Before training the machine learning model, the feature variables are quantized and there are two ways to quantize the general categorical variables converted into numerical variables, one is solo thermal coding while the other is labeled quantization as shown in the table 2 below:

Table 2. Raw data variables

variable name	Variable type
district	Hierarchical variables
Leasing method	Non-hierarchical relational variables
orientation	Non-hierarchical relational variables
billing	Non-hierarchical relational variables
family or clan	Hierarchical variables
arcade	Hierarchical variables
bathroom	Hierarchical variables
occupancy	Hierarchical variables
Lease Term	Hierarchical variables
view a house	Non-hierarchical relational variables
floor	Non-hierarchical relational variables
total floor	Hierarchical variables
escalator	Hierarchical variables
parking	Hierarchical variables
water consumption	Non-hierarchical relational variables
electricity consumption	Non-hierarchical relational variables
natural gas	Hierarchical variables

4 Screening of Characteristic Variables

4.1 Screening of uniquely hot coded variables

As this paper studies the problem of rent information in Shenzhen, then it is necessary to study which characteristic variables significantly affect the rent in Shenzhen, but also in order to better improve the accuracy of the model, here for the study of the unique heat coded over the non-hierarchical type of variables of which variables significantly affect the monthly rent in Shenzhen, utilizing the Wilcoxon rank-sum test, because for the variables that only have the labels of 0 and 1 can be turned into the monthly rent into the two main overall.

Next, using Python software and using the Wilcoxon rank sum test, the results obtained are shown in the table 3 below:

Table 3. Statistical results of Wilcoxon rank sum test

Indicators of correlation of qualitative variables	Statistic	p-value
Leasing Method_Shared Rent	45.506800	0.0
Lease Type_Whole Lease	-45.506800	0.0
Billing_Quarterly	35.788692	1.655937e-280
Billing_Monthly	-35.759041	4.787239e-280
Viewing_Generally available after hours	-8.008591	1.160304e-15
Viewing_only on weekends	-3.529601	0.000416
Viewing_Anytime		
Viewing_Advance Appoint-	34.126637	2.970645e-255
ment Required	-26.339599	6.752911e-153
Floor_Middle Floor		
Floor_Low Floor	0.177792	0.858886
Floor_High Floor	-1.869541	0.061548
Water_Commercial_Water	1.629894	0.103123
Water_Civil Water	0.556669	0.577753
Electricity_Commercial Elec-	-0.556669	0.577753
tricity	0.518531	0.604088
Electricity_Civil Power	-0.518530	0.604088
East	-6.318337	2.643918e-10
South	-10.698115	1.038694e-26
West	2.774524	0.005528
North	1.607026	0.108049
Northwestern	-2.066578	0.038773
Northwest	-2.977151	0.002909
Southwestern	-2.017806	0.043611
outheast	-8.174529	2.970215e-16

With the results of the above table, we can see that the features "North", "Floor_Middle", "Floor_Low", "Floor_High", "Water_Commercial_Water", "Water_Civil_Water" and "Electricity_Commercial_Electricity"have a p-value greater than 0.05, indicating that there is no significant correlation between these characteristics and monthly rent, so these variables can be excluded directly, while the others are significantly affecting the monthly rent, so they are retained.

4.2 Screening of hierarchical categorical and quantitative variables

For the remaining variables, most of them are hierarchical categorical variables, and at this point in time, in order to explore the relationship between the variables and the monthly rent and there are many hierarchical categorical variables([8]), Spearman's correlation coefficient is used here as the method is more suitable for exploring the hierarchical categorical variables and does not

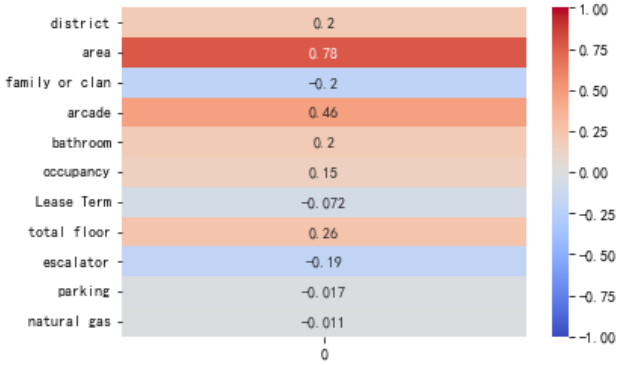

Fig. 3. Heat map of Spearman's correlation coefficient

depend on the distribution of the data, it is also stable.The following are the results obtained using Python software and using Spearman's correlation coefficient, as shown in the figure3 below([9]):

Through the results of the above figure, it can be seen that the p-value of the characteristics "parking space" and "gas" are greater than 0.05, which means that there is no significant influence between these two characteristics and monthly rent, so these two variables can be directly excluded. The other features are significantly correlated with the monthly rent, but the correlation is not necessarily linear, in which the correlation of the feature "area" is very significant, and the correlation is also very large 0.777624.

4.3 Visualization of feature variables

These variables were visualized using correlation heat maps and the results of the visualization are as follows figure4:

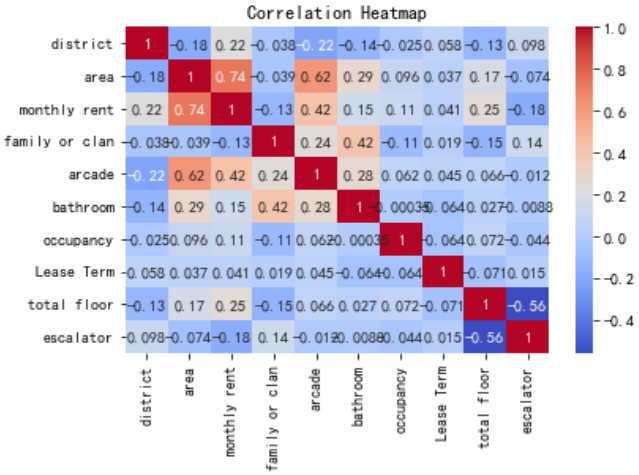

Fig. 4. Heat map of correlation

This heat map mainly reflects the correlation between any two variables, the darker the color represents the larger the correlation coefficient between these two variables. From the above figure, we can see that the feature of area has the largest correlation with monthly rent, followed by hall, rental period, while the weaker correlation is the occupancy, elevator and so on.

The next step is to visualize a histogram of the data in the columns Monthly Rent and Square Footage, as shown in the following figures5 and figures6:

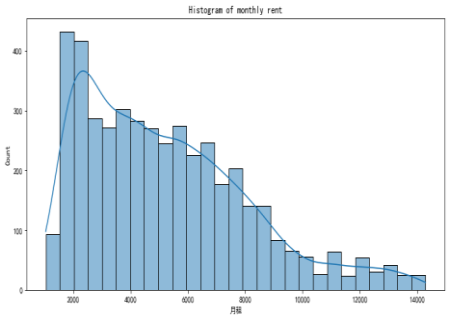

Fig. 5. Histogram of monthly rent

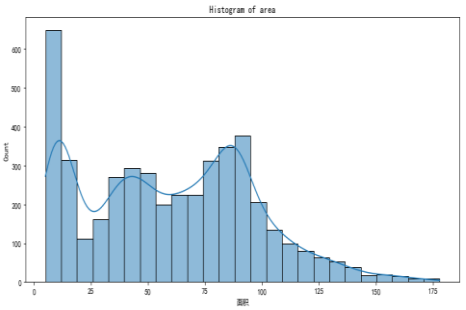

Fig. 6. Histogram of area

Because square footage significantly affects monthly rent, it is now important to start with a simple treatment of these two data.We performed descriptive statistics for both the area and monthly rent columns, as shown in the table 4 below:

Table 4. Indicator statistics

features	average	maximum	Minmum	standard deviation	Coefficient of variation	skewness factor	kurtosis coefficient
area	59.837	178.0	5.01	37.193	0.622	0.291	-0.605
monthly	5215.031	14300.0	1030.0	2887.309	0.554	0.855	0.235

The results of the above table show that the skewness coefficients of both area and monthly rent are greater than 0, which indicates that the distribution of the data are right skewed, while the data of monthly rent are more severely skewed, in terms of the kurtosis coefficient, the top of the area is more flat and the monthly rent is more pointed, in terms of the coefficient of variation. The column of data for monthly rent is more unstable. The distribution of the data will also affect the accuracy of the model, in order to improve the data distribution, here using the area column and the monthly rent column to take the logarithm to make the data distribution closer to the normal distribution, after taking the logarithm of the QQ chart is as follows figures7 and figures8:

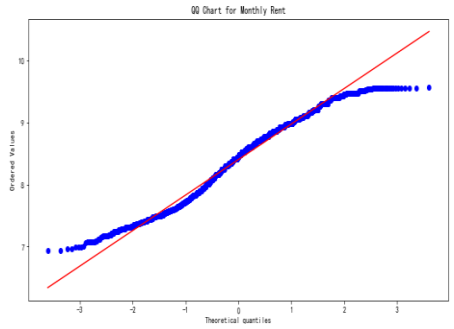

Fig. 7. QQ Chart for Monthly Rent

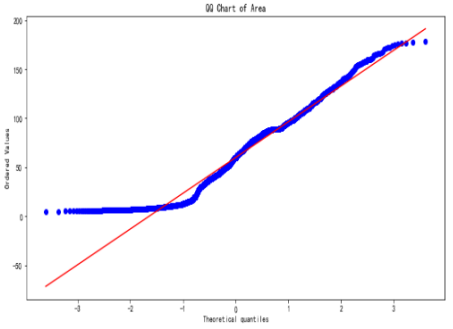

Fig. 8. QQ diagram of area

As can be seen in Figure 7, the distribution of the monthly rent data has improved, most of the data are on the fitted straight line, only some deviate a little bit, and at this time, the skewness coefficient of the monthly rent column is -0.156, and the skewness coefficient is -0.862, the skewness coefficient has become significantly smaller, which indicates that the data are closer to the normal distribution. And the data of area is not taken logarithmic, because the area takes itself the skewness coefficient and kurtosis coefficient are very good. As can be seen from Fig. 8, the data in the latter part of the large part of the data are on a straight line, while some of the data in the former part are a little off.

After the logarithmic processing, the metrics of monthly rent and square footage need to be normalized by z-scores, which are used to remove the differences in magnitude and scale from the data so that different features have the same scale. This is important for many machine learning algorithms. The formula for standardization is as follows:

$$x_{\text{standardized}} = \frac{x-\mu}{\sigma} \tag{1}$$

5 Model selection

Through the above data preprocessing, as well as the screening of variables, you can use the processed data to train the prediction model, where we intend to establish multiple regression model, ridge regression, LASSO regression, multi-layer perceptual machine, random forest regression, XGBoost regression model to predict the rent in Shenzhen, and then, through the comparison to get the optimal prediction model. We can divide our dataset into eighty-two divisions, eighty percent as our training set, twenty percent as the test set of this paper, respectively, to train these six models, and the results obtained are shown in the table 5 below:

Table 5. Model Performance

Model	MSE	R^2	MRE
Linear	0.214071	0.792880	1.420614

Ridge Regression	0.213984	0.792964	1.419311
LASSO Regression	0.954615	0.076383	1.308651
MLP Regressor	0.242660	0.765220	1.338832
Random Forest	0.169190	0.836304	1.191798
XGBoost	0.180009	0.825836	1.189564

With the table above, it can be seen that for this particular task, the Random Forest Regressor and the XGBoost Regressor perform the best because they have the smallest MSE, the closest R^2 to 1, 0.836 and 0.825, respectively, and the smallest MRE.Also, a plot of the fit between the predicted and true values of the model can be obtained as follows figures 9:

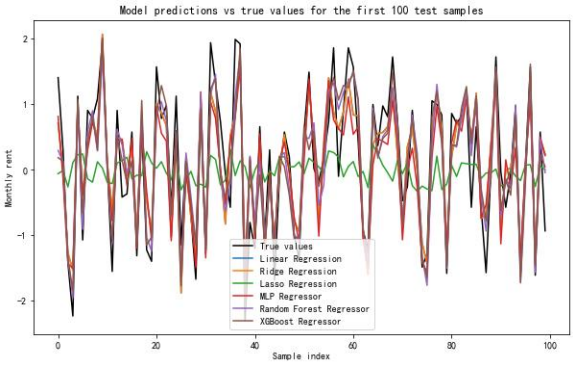

Fig. 9. Plot of the training effect of the six models

As we can know from the above figure, it is obvious that the green line is the prediction effect of the LASSO regression model, which is very poor, while the other models are almost the same, and at the same time, in order to make up for this error, this paper utilizes the Stacking regression model to predict the rent.

5.1 Stacking regression model

5.1.1 Stacking regression model fundamentals

Stacking is an integrated learning approach that uses a meta-mode l to make the final prediction by taking the prediction results of multiple base models as inputs. In regression problems, Stacking models can be used to improve the overall prediction performance by training multiple regression models to combine the prediction results([10]). The model integration approach is roughly shown in Figure 10 below:

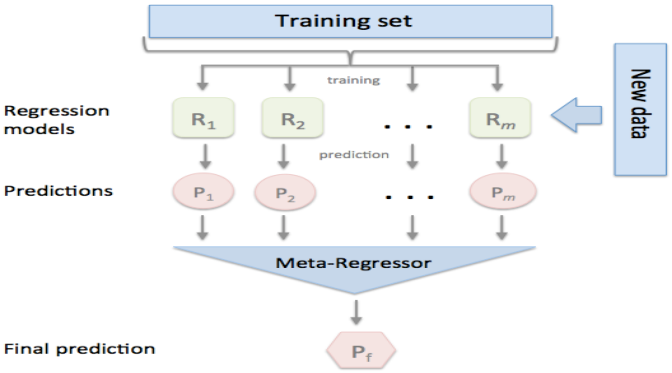

Fig. 10. Stacking integration principle

The following is the training process and explanation of the Stacking regression model:

Suppose we have a training dataset containing n samples, each consisting of d features. We use x_i to denote the feature vector of the ith sample and y_i to denote the corresponding target output.

Base model training: first, we choose multiple different regression models as base models, such as linear regression, decision tree regression, support vector regression, etc. Then, each base model is trained using the training dataset to obtain multiple base models.

Base model prediction: for each training sample, a prediction is made using each base model, and the predicted output of each base model is obtained. For the ith sample and the jth base model, denoted as $y_i^{(j)}$.

Creating the meta-feature matrix: the predicted outputs $y_i^{(j)}$ of each base model are combined into a new feature matrix called the meta-feature matrix. The size of the meta-feature matrix is $n \times m$, where m is the number of base models. Each row represents a sample and each column represents the predicted output of one base model.

Meta-model training: using the meta-feature matrix as input and the target output y as labels, a meta-model is trained, which can be any regression model, such as linear regression, ridge regression, etc. The goal of the metamodel is to learn how to combine the predictions of the base model to obtain the final predicted output.

Prediction output: for a new test sample, the prediction is first performed using each base model to obtain the prediction output of the base model. Then, the predicted outputs of the base models are combined into a meta-feature matrix. Finally, the meta-feature matrix is predicted using the trained meta-models to get the final predicted output.

5.1.2 Stacking's training results

Here, the six models above can be integrated using the Stacking regression model and the model training results obtained are as follows table 6:

Table 6. Stacking Model Performance

Metric	Value
MSE	0.165153
R^2	0.840209
MRE	1.179681

Through the above results, we can know that the prediction effect of Stacking model obviously improves the prediction performance of the model. the MSE, R^2 and MRE are all better than the training effect of the previous six single models, so the Stacking model is very suitable for the prediction of the rent information in Shenzhen, which provides a great reference and help to the buyers.

6 Conclusions

In summary, this paper proposes a prediction method based on Stacking regression model, firstly, six single models are established: multiple linear regression model, Rideg regression model, LASSO regression model, multilayer perceptual machine model, random forest regression model, XGBoost regression model. Then the three evaluation indexes of MSE, R^2, and MRE were used to evaluate the models, and it was found that the prediction ability of LASSO model was very poor, while the best performance was the random forest regression model, with the MSE of 0.16919, the R^2 of 0.836304, and the MRE of 1.191798. After that, considering that each model has its own focus as well as advantages and disadvantages, this paper finally established a model of XGBoost regression. Therefore, in the end of this paper, we established the Stacking integration model, which is good in general, and integrated the six single models, and the experimental results show that the prediction effect of the Stacking model obviously improves the prediction performance of the model, and the MSE, R^2, and MRE are better than the training effect of the previous six single models and the weighted combination model, which indicates that the Stacking integration algorithm is better for the prediction of Shenzhen than the six single models. The Stacking integration algorithm has the smallest bias, the most stable model, and the strongest interpretability of the regression prediction for predicting the rent and rent in Shenzhen, which provides a relatively large reference value for home buyers and landlords.

References

1. ZHU Haiyu,WANG Zhijie,YE Canchan. Prediction of house price in urban hotspot areas based on XGBoost algorithm--Taking Nanjing Jiangbei New District as an example[J]. Construction Economy,2022,43(S2):433-437.
2. Xu Dandan. A comparative study of house price prediction in Xi'an based on multiple linear regression model and BP neural network[J]. Real Estate World,2022(08):11-13.

3. Deng Smooth,Xie Zhizhou. Application of gamma regression model and inverse Gaussian regression model in house price prediction[J]. Green Technology,2022, 24(13): 215-218+224.
4. FU Qiwei,YI Yanchun,ZHANG Shijia et al. Forecasting analysis of house price in Hengyang city based on ARIMA model[J]. Science and Technology Innovation and Application, ,2021, 11(31):47-50.
5. Ren Ziming. Research on the prediction of house price in Beijing urban area based on gray prediction and regression model[J]. Modern Business Industry,2019,40(10).
6. Li YQ. House price prediction model based on random forest[J]. Communication World,2018(09):306-308.
7. Görne Lorenz,Reuss Hans Christian,Krätschmer Andreas,Sauerwald Ralf. Smart data preprocessing method for remote vehicle diagnostics to increase data compression efficiency[J]. Automotive and Engine Technology,2022,7(3-4).
8. Byungwook Lee. Improvement of the Semantic Information Retrieval using Ontology and Spearman Correlation Coefficients[J]. Journal of Digital Convergence,2013,11(11).
9. LI Yuan,LIU Yutian,FENG Liwei. Nonlinear dynamic process feature extraction and fault detection based on Spearman correlation analysis[J]. Journal of Shandong University of Science and Technology (Natural Science Edition),2023,42(02):98-107.
10. Zhang Xinyuan. Rent Prediction Based on Stacking Regression Model and Baidu Map API[D]. Lanzhou University, 2022.