Research on Renting Price Prediction Based on Machine Learning

Shixuan Cao[1,a], Weize Liao[2,3,b*], Jianrong Huang[2,c*]

{1377623677@qq.com[a], 1375480613@qq.com[b], 20074020@qq.com[c]}

United International College of Beijing Normal University and Hong Kong Baptist University (UIC),
519087, ZhuHai, China[1]
Wuzhou University, 543002, Wuzhou,China[2]
Guangxi Key Laboratory of Machine Vision and Intelligent Control, 543002 China[3]

Abstract. This study uses housing data from the Beijing region retrieved from the rental portal website, Fang Tian Xia, as the research subject. The objective is to predict rental prices of various types of housing, approaching it as a regression problem in rental price prediction. Initially, an approximately 140,000-entry dataset was constructed by collecting housing data using a web crawler. Subsequently, three machine learning models—KNN, Random Forest, and XGBoost—were separately trained using the dataset. Based on evaluation metrics, the most effective model was selected, with results indicating that the Random Forest model demonstrated optimal performance. Finally, the most effective model was used to predict rental prices for various types of housing.

Keywords: Rent; KNN; XGBoost; Random Forests; Forecasting

1 Introduction

With the progression of urbanization in China, studies on 35 major cities in the country have shown a mutual influence between population, economy, and rental markets[1]. In recent years, there has been a substantial influx of residents into urban areas, a majority of them being young people, who'd choose to rent a home due to high property prices. In 2015, affected by the national policy of "equal importance to renting and purchasing", the rental market continued to expand[2], thereby giving rise to various anomalies and hence, the urgency to regulate the rental market. Consequently, housing price and rent became the focus of scholarly research. Guo Rumeng collected data regarding shared rental housing in Beijing in 2018[3] and conducted research on rental prediction using Random Forest and XGBoost models.

Foreign rental markets have a longer history of development, and their qualitative research and theories on renting are relatively mature. Generally, many foreigners prefer renting a home. M2 Presswire[4] conducted an analysis related to the rapid growth of the rental market. Mr. Adrian Alter and Zaki Dernaoui[5] carried out detailed research on assessing the impact of the price dynamics of used houses in the United States, using time hypothesis methods and instrumental variable methods, showing that both second-hand house buyers and short-term investors are magnifying the real estate cycle. Ms. Yuko Hashimoto[6] and others, using the durable housing model, researched the factors contributing to Japan's falling house prices and

MSIEID 2023, December 08-10, Guangzhou, People's Republic of China
Copyright © 2024 EAI
DOI 10.4108/eai.8-12-2023.2344718

concluded that the decline in the number of Japanese families has a correlation with falling house prices.

After crawling housing data and cleaning the gathered information, this paper selects the optimal model using machine-learning methods. The result of the experiment reveals that the Random Forest model outperforms others in terms of performance. The optimal model is then used to predict the rent of various types of housing, resulting in fairly satisfactory forecasting results.

2 DataSet

2.1 Data Collection

The primary source of data in this study is procured through web crawling technology[7] in Python, which is used to collect rental listings data from Fang Tian Xia. The dataset includes nine features: title, city, rental method, housing type, area of the property, orientation, address, rent, and link. The data features are as illustrated in Table 1:

Table 1. Data Features

Feature Name	description	Example
title	Property Introduction	Let's rent it out! 3 rooms and 1 hall
city	City where the property is located	Beijing, Shanghai
lease	Rental method (whole lease/joint lease)	Whole lease
rooms	Housing type (several halls and rooms)	2 rooms and 1 hall
area	Housing area (square meters)	80 ㎡
towards	Housing orientation	Facing south
adress	Property Address	Chaoyang Shilihe Champs Elysees
price	Rent (yuan/month)	6100 yuan/month
link	link	

2.2 Feature Analysis

The completeness and accuracy of the data need to be scrutinized by observing ranges of various data attributes. The descriptive analysis of the data is shown in Table 2:

Table 2. Data Feature Analysis

Statistics	Room Num	Hall Num	City	Area	Towards	District	Price	Lease
count	147384	147384	147384	147384	147384	147384	147384	147384
mean	2.12	1.25	7.25	58.59	2.36	130.87	2871.94	0.50
std	0.69	0.48	4.59	46.91	2.44	62.29	3739.38	0.50
min	0.00	0.00	1.00	11.00	1.00	1.00	102.00	0.00
25%	2.00	1.00	4.00	21.00	1.00	92.00	730.00	0.00
50%	2.00	1.00	6.00	40.00	1.00	125.00	1699.00	0.50
75%	2.00	2.00	11.00	88.00	2.00	176.00	3200.00	1.00
max	9.00	5.00	16.00	299.00	10.00	239.00	29999.00	1.00

From the table above, we can observe the information range for each attribute clearly. The comparison of data might be performed by analyzing the range between the minimum (min) and the maximum (max) values. Upon observation, preliminary processing has been implemented on data among various features, exhibiting reasonably aligned data as per the requirements.

3 Model and Method

In establishing a machine learning model, an overabundance of independent variable values leads to the reduction of the model's degree of freedom[9]. Therefore, high-quality data are demanded in model analysis. Looking into the data, there are 16 variables for city (city), representing distinct cities, and 239 variables for the district, which have rather numerous values. After normalization, the data are in a normal state. Thus, we randomly select 75% of the data as training set and 25% as test set. Meanwhile, appropriate models such as KNN(k-nearest neighbors), Random Forests, and XGBoost model etc. are selected for model training.

3.1 KNN

3.1.1.Selection of Parameters

Knn algorithm is one of the popular algorithms in machine learning models. It not only serves for discrete value prediction but also trains for continuous values by calculating the average value of the nearest data points to acquire the predicted value. There are various parameters in Knn algorithm. Among them, radius represents the radius of the nearest data points, n_neighbors represents the number of nearest neighbors, metric is the distance measure, and weights represent weight with a default value being weights = 'uniform'. The present project has over 70,000 data, and n_neighbors is set as 20, with other parameters being default.

3.1.2.Model Building

In Python, import the KNeighborsRegressor package from the sklearn.neighbors library, select feature and target values (the target value being the rent), involve certain features such as city, housing type, rental method, etc. Conduct normalization and sample treatment on the data, then split the data and carry out regression prediction on the test set.

3.2 Random Forests

3.2.1.Selection of Parameters

Random forest model is among the most popular machine learning models currently. Its primary theory is to randomly extract data as training set and consider it as decision tree inputs and eventually get the output through voting. Random forest model has several crucial parameters, namely, n_estimators which refers to the number of trees that could be adjusted according to the model performance; max_features being the number of tree features, and max_features is unique. Therefore, if when the evaluation of random forest model is not ideal, adjusting features is what we need to explore; max_depth is the maximum depth of the tree, constraining max_depth is to simplify the model. During random forest construction, n_estimators, max_features, and max_depth are our major considerations.

3.2.2.Model Building

In Python, import the RandomForestRegressor package from sklearn.ensemble library, select feature and target values (the target value being the rent), input selected features. Conduct regression prediction on the test set.

3.2.3.Importance Ranking of Features

In the training process of the random forest model, a measure of variable importance of data features will be output, i.e., the feature_importances_ parameter. The variation in feature importance indicates the degree of attention the rental population gives to that feature. Generally speaking, the larger the quantity, the more significant the feature is. The importance ranking of features in this model is as shown in Table 3:

Table 3 Importance Ranking Table of Features in Random Forests.

Ranking	Feature Name	Importance Factor
1	Area	0.561138
2	District	0.311908
3	Lease	0.053110
4	City	0.025652
5	Towards	0.020605
6	Room Num	0.019928
7	Hall Num	0.007660

In table 3, area has the most significant impact on rent, followed by the district within the city. Lease type and city have a moderate effect on the rent, while the towards, number of rooms, and number of halls have less influence on the rent. This suggests that renters may not place high emphasis on these features.

3.3 XGBoost

3.3.1.Parameter selection

The XGBoost model is highly favored by many users due to its efficient learning outcomes, alongside its quick training speed that stands as one of its prominent benefits. Fundamentally, XGBoost combines tree-based models with ensemble learning techniques. Given the rental data, we select max_depth as 5, learning_rate as 0.1, n_estimators depth as 160 for model parameters, while setting silent to the default status and objective to reg:gamma.

3.3.2.Model building

The xgb package from the XGBoost model is imported in Python, whereupon feature values and target values are determined. In this case, the target value is rent, with features like city, housing type, lease method, among others. The parameters for the model selected are: max_depth of 5, learning_rate of 0.1, depth of n_estimators set to 160, silent at the default status and objective set to reg:gamma. The model then performs regression predictions on the test dataset.

3.3.3.Feature Importance Ranking

Once the model training is complete, the XGBoost model offers a feature ranking parameter, as show in figure 1, which arranges data accordingly. The ordered list of features is then displayed through plot_importance.

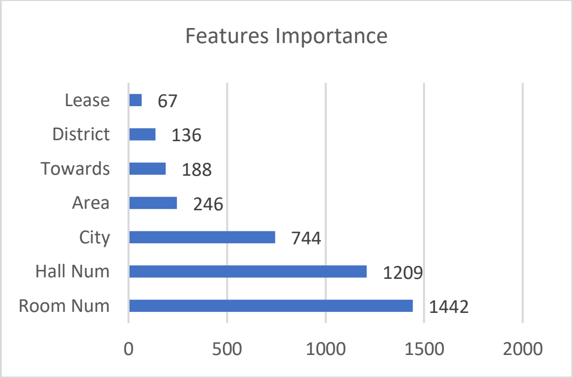

Figure 1 XGBoost Model Feature Importance Ranking

To sum up, area has the most significant impact on rent (price), followed by city and district within the city. The leasing method has a median influence on the rent, suggesting renters consider factors such as area, city, and the leasing method. Finally, Towards, numbers of rooms and halls have a lesser effect on the rent, indicating that these features may not be of high importance to renters.

4 Experimental Results and Analysis

4.1 Evaluation Metrics

To determine the goodness of a machine learning model, different metrics[8] are employed to assess the outcomes of the machine learning model. Given that the rent is a continuous variable, a regression model is used in this study. Thus, we select the Mean Absolute Error (MAE), Mean Squared Error (MSE) and R^2 (R-Squared) as evaluation metrics, assuming y_i stands for the actual value, $\hat{y}_i$ for the predicted value, and $\overline{y}$ for the average value of samples. Each evaluation metric has the following implications:

4.1.1.MAE

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

MAE stands for the absolute difference between the predicted and actual values. It enables a better analysis of the relationship between the two, where a smaller difference typically suggests that the values are closely alike, indicating a suitable model for the selected data.

4.1.2.MSE

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{2}$$

MSE refers to the ratio of the sum of squares of the difference between the actual and predicted values, over the number of observations. MSE primarily assesses the degree of variation among data: generally, the smaller the MSE value, the better the accuracy of the model.

4.1.3. R^2 (R-Squared)

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2} \tag{3}$$

R^2 , based on the fluctuation in data, demonstrates the goodness of fit of the model. R^2 ranges within [0,1]. Model fitness tends to be worse as it approaches 0, implying that the model prediction is not accurate and parameter tuning is required. As it approaches 1, data performance tends to be superior, suggesting a better model fitting. Generally, a value above 0.4 suggests that the model fitting meets requirements.

4.2 Analysis of Results

In this paper, we selected KNN , Random forests, and XGBoost regression models to forcast the rental housing data from Fang.com, while also using the evaluation metrics MAE, MSE and R-Squared of the regression model to compare the goodness of models. The comparison of model results is shown in table 4:

Table 4. Model metrics comparison

Models	MAE	MSE	R^2
KNN	648.83	2184948.13	0.8424956460983669
Random forests	560.71	1776455.18	0.8719423031959511
XGBoost	653.59	2045498.02	0.8525480580136124

To summarize, the smaller the MAE difference is, the closer the predicted value is to the real value, indicating better prediction. A smaller MSE value implies a model with good accuracy; and the larger the R^2 value, the better the model fits the data. Hence, the random forest model is superior in predicting rental prices by comparison.

4.3 Rent Prediction

Through the above model results comparison, the Random Forests model's results are better for rent prediction. This project uses the random forest model for batch prediction of the data. Feature values are Room Num, Hall Num, City, Area, Towards, district, and Lease, and the target value is the Rent (price). The predicted rent is designated as rf_price. The prediction result is shown in table5:

Table 5. Random Forest Model Rent Prediction

city	Room Num	Hall Num	Area (㎡)	Towards	District	lease	price	rf price
Beijing	1	1	80	North	Chaoyang	Whole	6100	9101.00
Beijing	3	2	165	Southwest	Haidian	Whole	18200	18742.78
Beijing	1	1	80	South	Chaoyang	Whole	4500	7604.00
Beijing	2	1	59	East and West	Fangshan	Whole	1100	1385.00
Beijing	2	1	102	East and West	Fangshan	Whole	2300	2638.50
Beijing	2	1	81	North and South	Fangshan	Whole	2100	2581.63
Beijing	2	1	78	North and South	Fangshan	Whole	2200	2562.30
Beijing	2	2	93	South	Haidian	Whole	13000	13303.60
Beijing	2	1	93	East and West	Fangshan	Whole	2200	2510.83
Beijing	2	1	64	South	Haidian	Whole	8800	8018.00

5 Conclusion

This research study exploited housing data collected from the Fang Tian Xia rental portal focusing on the Beijing region, thereby aiming to forecast rental prices of diverse housing types. Three distinct machine-learning models—KNN, Random Forest, and XGBoost—were deployed. After an evaluation based on certain metrics, the Random Forest model emerged as the superior counterpart. The reliability of the Random Forest model was not only validated through the measures of accuracy but also by its effectiveness in predicting the rental prices across different housing types with satisfactory forecasting outcomes.

The study's outcome underscores the relevance of adopting machine-learning tools in the field of real estate prediction, especially in terms of rental markets. The findings also contribute to the ongoing discussions and research in housing market dynamics, specifically in the context of urbanization and rental preferences in major Chinese cities. Future studies might explore the application of the Random Forest model in other property markets, or examine how other factors—such as tenant demographics or policy changes—might influence rental prices. Furthermore, an area of prospective focus could be to refine the model, to not only predict but also to analyze and explain what factors might trigger fluctuations in the rental markets.

References

[1] Mou, Lingling et al. Study on the Coupling and Coordination of Population-Economy-Rental Market—A Case Study of 35 Large and Medium-Sized Cities in China[J]. Tropical Geography, 2022, 42(06): 889-901.

[2] Yuan, Wenting. Analysis of the Current Situation of the Long-Term Rental Market Stimulated by China's Policy Dividend and Discussion on Establishing a Dual Trusteeship Mechanism[J]. Contemporary Economy, 2020, (01): 7-9.

[3] Guo, Rumeng. Prediction and Influential Factor Analysis of House Rent Price in Beijing[D]. Beijing University of Technology, 2019.

[4] Rental Market Sees the Fastest Growth in Nearly a Decade[J]. M2 Presswire, 2022.

[5] Mr. Adrian Alter, et al. Non-Primary Home Buyers, Shadow Banking, and the US Housing Market[J]. IMF Working Papers, 2020, 2020(174).

[6] Ms. Yuko Hashimoto, et al. Demographics and the Housing Market: Japan’s Disappearing Cities[J]. IMF Working Papers, 2020, 2020(200).

[7] Hong, Lihua et al. Research Based on Python Web Crawler Technology[J]. Value Engineering, 2022, 41(34): 154-156.

[8] Xu, Yuan. Research on Guangzhou House Rent Based on Machine Learning Model[D]. Central China Normal University, 2022: 34-36.

[9] Liu, Shuyu. Influential Factor Analysis of Urban Rent Prices Based on Machine Learning Methods[D]. Nankai University, 2021.