Monthly Housing Rent Forecast based on LightGBM (Light Gradient Boosting) Model

Jinze Li
College of Information Science and Engineering, Shandong Agricultural University, Tai’an, 271000, China

Abstract: In today's society, housing rent is determined by many factors such as decoration, location, type of house, convenience of transportation, market supply and demand, etc. For the relatively traditional industry of renting, the problem of serious information asymmetry has always existed. This study is based on the pain points of the rental market, based on real renting market data after desensitization. Using the historical data of monthly rent tags to establish a LightGBM (Light Gradient Boosting) model based on machine learning, the accurate forecast of housing monthly rent based on basic housing information is provided, which provides an objective measure for the city's rental market.

Keywords: Machine learning; Integrated learning; LightGBM (Light Gradient Boosting); Data mining; Rent forecasting

1. Introduction

1.1. Background

In recent years, China's economy has flourished, and more and more people are starting to rent. However, this has caused countless people to be overwhelmed. On the one hand, the landlord does not understand the real market price of renting a house, and can only endure the vacant high-rent housing; Tenants can't find a cost-effective home that meets their needs, which is a huge waste of renting resources. In this environment, almost everyone began to pay attention to the fluctuations of housing rents, trying to use some methods to make more scientific and effective predictions of housing rent. However, the reasons for the change in housing rents are extremely complicated. There are still many controversies in the perspectives of economics and other disciplines. The factors of supply and demand, decoration, location, type of housing, convenience of transportation, etc. may all be the reasons for housing rent. Therefore, how to make more reasonable and effective predictions has become a hot problem in this field. There are many random influencing factors involved, and the influencing factors are diverse. The problem becomes very complicated and difficult to predict through a simple logical model. At present, there are many methods for forecasting housing rent, such as multiple regression model, Markov prediction model, grey model, neural network, combined prediction model and classification prediction model. There are many different forecasting methods. However, it is still difficult to find an accurate forecasting method of housing rent among many forecasting methods.

1.2. Characteristics of data sets

The data in this paper is the basic rent information of the house for 4 months and the basic information of the house, and the data is desensitized. Using the housing information and monthly rent training model in the training set, the monthly rent of the houses in the test set data is predicted using the housing information in the test set. The training set is the data collected in the first 3 months, a total of 15,539 data, and the training set has a total of 18 feature columns. The test set is the data collected in the fourth month. Compared with the training set, the “id” field is added, which is the unique id of the house, and there is no “price” (monthly rent) field. The other fields are the same as the training set, a total of 56279 Data, the test set has 18 feature columns. The data is the rent price of the house for 4 months and the basic information of the house. The author has desensitized the data.

1.3. The basis for selecting the model

In the field of machine learning, integrated learning is widely used as a method to effectively improve the performance of regression models. The development of the most popular XGBoost from AdaBoost has actually become a recognized algorithm in the major machine learning competitions. This is simple because it is extremely powerful. However, the amount of data in this study is extremely large, and XGBoost takes a long time to train, and the prediction accuracy is not satisfactory. Therefore, this study uses a fast, distributed, high-performance decision tree algorithm based gradient lifting framework - LightGBM (Light Gradient Boosting) model. Lightgbm can be used for sorting, sorting, regression, and many other machine learning tasks. Because lightgbm is based on the decision tree algorithm, it uses the optimal leaf-wise strategy to split the leaf nodes. However, other lifting algorithm split trees generally use depth-wise or level-wise instead of leaf-wise. Therefore, in the LightGBM algorithm, when growing to the same leaf node, the leaf-wise algorithm reduces more loss than the level-wise algorithm. This results in higher precision than the xgboost algorithm, and no other existing lifting algorithms can. At the same time, its speed is also shocking, which is why the algorithm name Light. In addition, this paper makes a detailed explanation of data preprocessing, and according to the unique nature of the data set, the data is explored, cleaned and characterized by preprocessing. Exploratory data is the meaning of the preliminary understanding of each column of data, whether there is a missing, you can use a variety of methods to analyze the data; data cleaning is to deal with missing values and erroneous data, making the data complete science; feature engineering means cleaning The post-data is transformed into the type and structure of the data needed to build the model; finally, the test data is predicted to reach a conclusion. In the course of the experiment, cross-validation and other methods were used to avoid over-fitting, and finally a better prediction result was obtained.

2. Data Processing

2.1. Outlier processing

That is, for the unreasonable error data processing, taking the relationship between the housing area "Area" and the rent "price" as an example. There are individual outliers in the data set, which is equal to the noise, which can be deleted. If the deviation is not particularly obvious, it must be retained to ensure generalization ability. As you can see, the five points in the lower right corner are clearly significantly different from the other points. These five points are the two largest houses, but the rent is very low, although in theory it may be in the suburbs, but only five. Belonging to the minority is deleted as an outlier. The effect after processing is as follows:

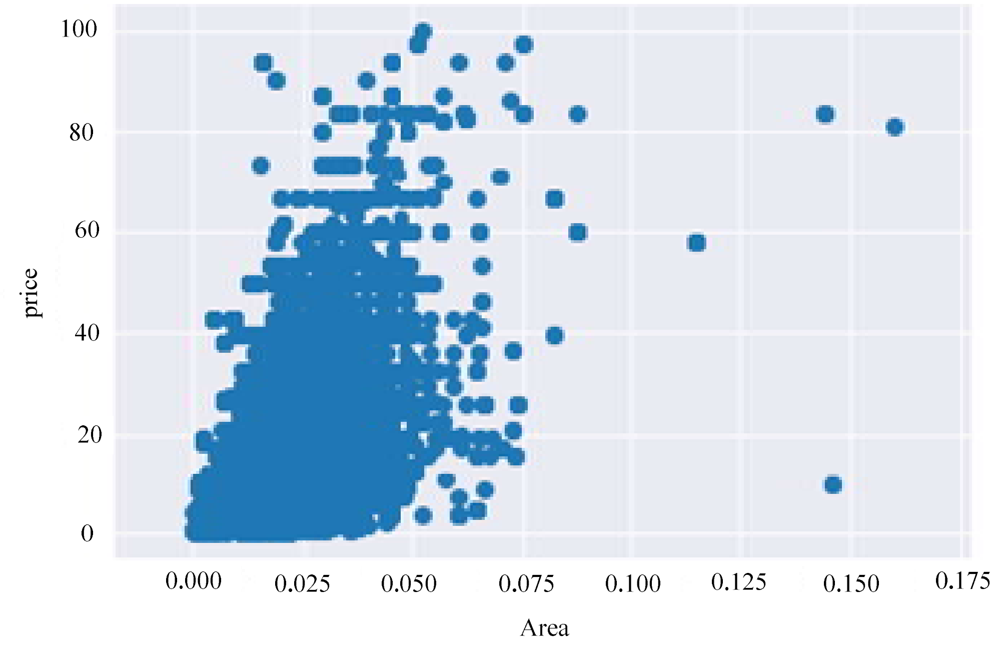


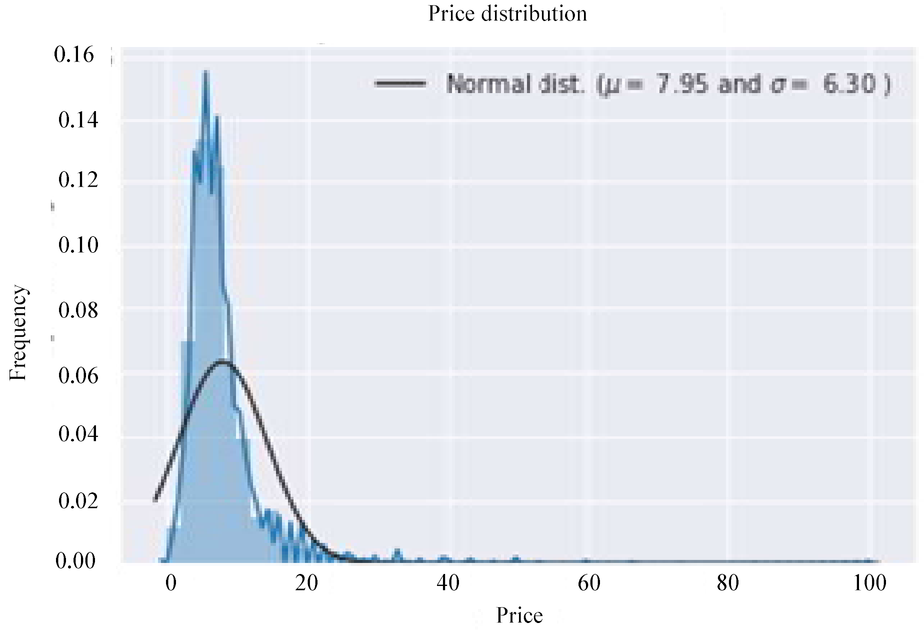
Figure 1. The effect after processing

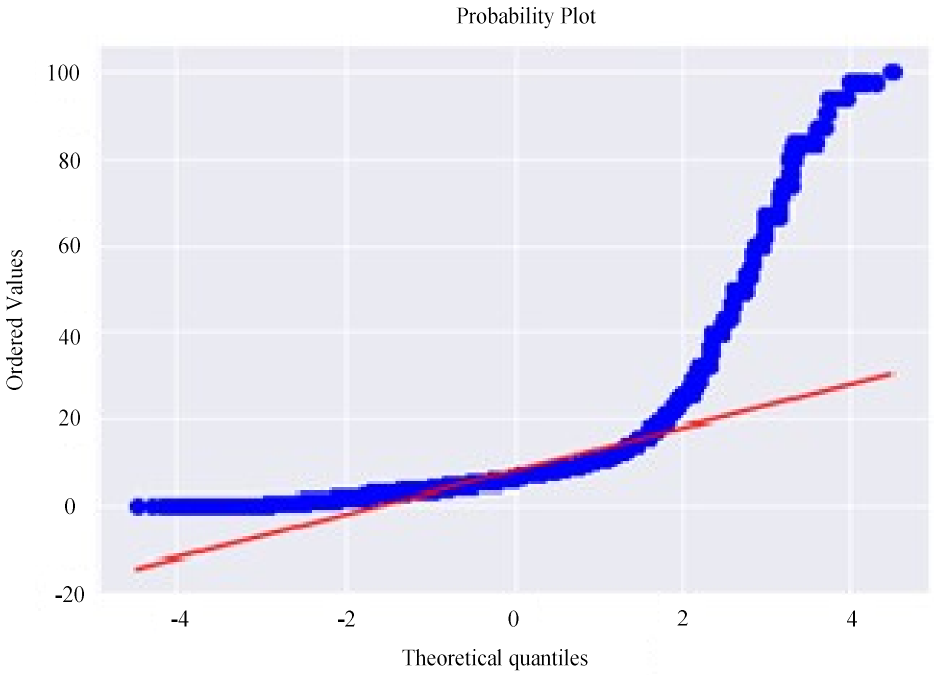
2.2. Normally distribute the target value

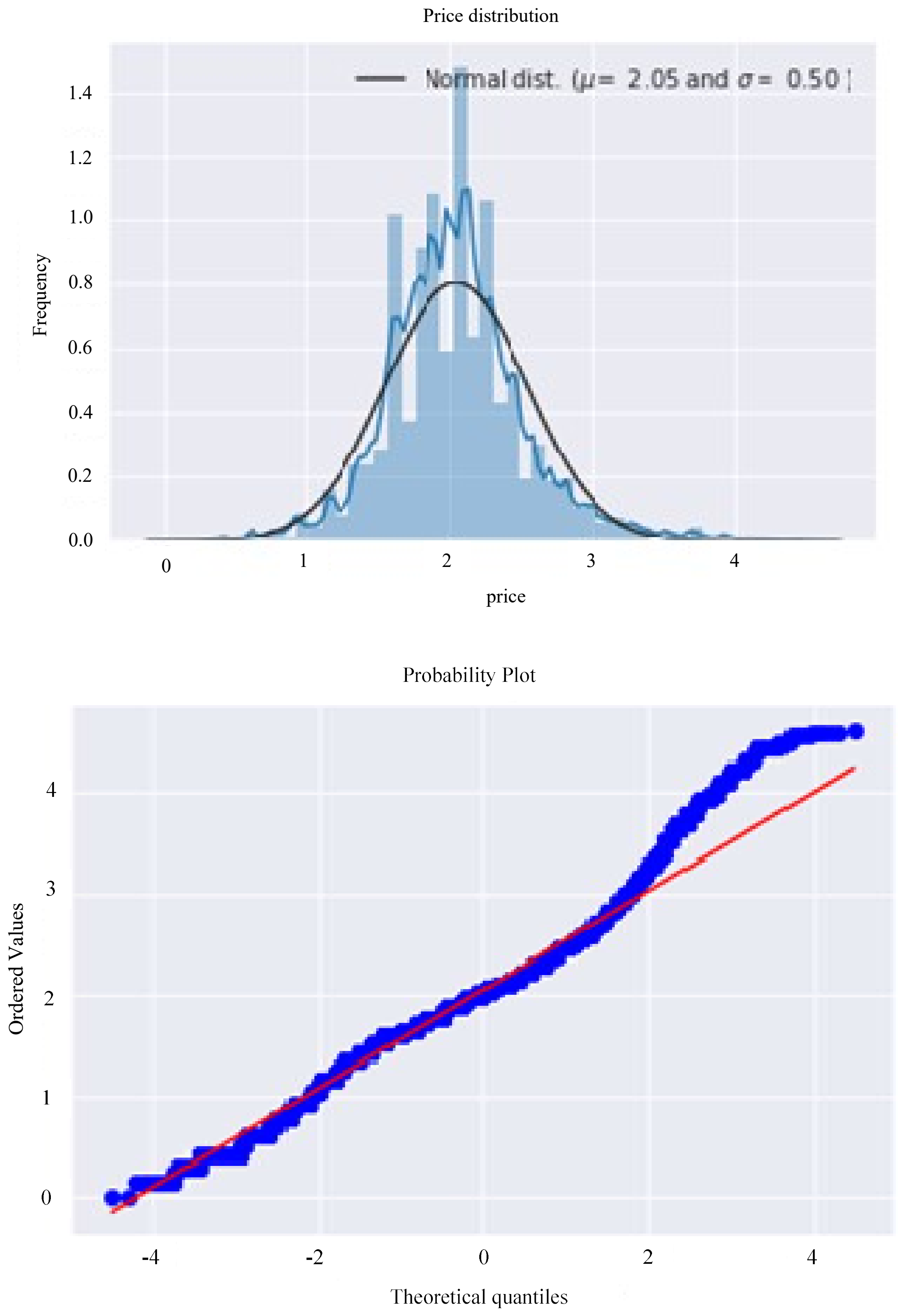
A linear model requires a normally distributed target value to be most effective. First look at the distribution of rent, as shown in Figure (a).

It can be seen from the figure a that the whole is leftward, and then the normal probability map is drawn. As shown in Figure (b), The normal probability map is used to detect whether a set of data obeys a normal distribution, and the more tends to a straight line, the more normal distribution.

It can be seen that the right skewed distribution, due to the large skewness, needs to perform log conversion on the target value to restore the normality of the target value. As shown in the Figure (c) and (d). At this time, it can be seen that the skewness is normal and the rent price is approximately normal.


Price distribution
Normal dist. (μ = 7.95 and σ = 6.30)
Frequency
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
Price


Probability Plot
Ordered Values
100
80
60
40
20
0
-20
-4
-2
0
2
4
Theoretical quantiles



Price distribution

Probability Plot

Figure 2. The distribution of rent

2.3. Missing value processing

First look at the overall missing values, by the process of visual display, we can get that:

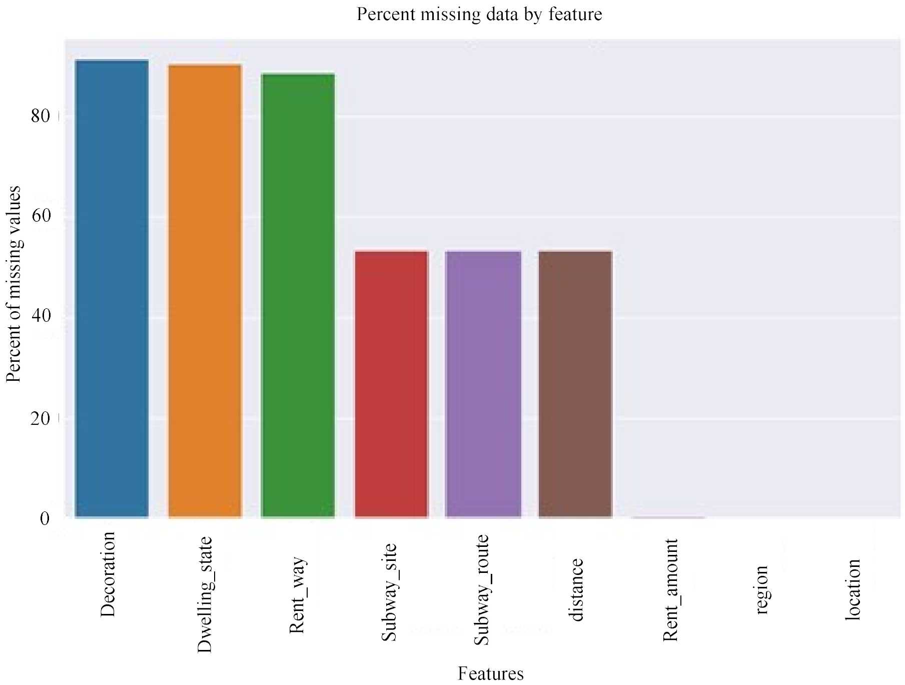


Figure 3. The overall missing values

It can be seen that the category data decoration situation "Decoration", the residence status "dwelling_state", the rental method "Rent_way" and the numerical data housing adjacent to the subway station "subway_site", the subway route "subway_route", the distance from the subway station "distance" Six features are missing.

Next, let's look at the correlation between each feature and rent. For the 18 data type features, use the heat map to analyze the correlation:

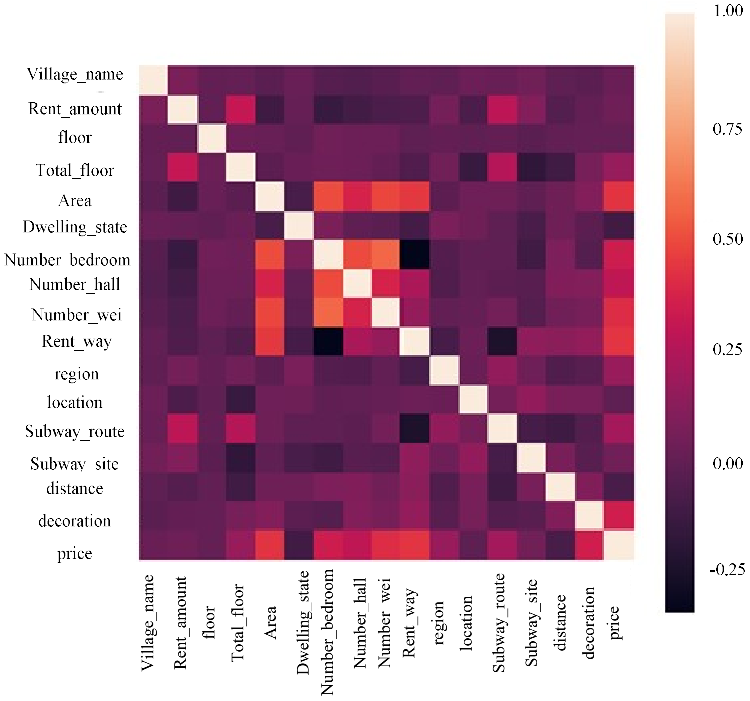


Figure 4. The correlation between each feature and rent

As shown in the figure, the lighter the color, the higher the correlation. The highlights with higher correlation are: Area and Number_bedroom, Area and Number_wei, Number_bedroom and Number_wei, Number_bedroom and Number_hall, indicating the area of the house "Area" and the number of bedrooms "Number_bedroom", The number of guards "Number_wei" is strongly positively correlated, and the number of bedrooms "Number_bedroom" is positively correlated with the number of guards "Number_wei" and the number of halls "Number_hall". When filling in missing values, you can either delete one extra feature or use one to fill the other.

When dealing with missing values, it's best to look at the properties of the features in order to fill the science, usually:

The missing is large: the category type is replaced by NA, and the data type is filled with the mean mode (depending on the characteristics).

Less missing: the category is replaced with the most attributes, and the data type is replaced by the mean mode (depending on the feature).

Category type data decoration situation "Decoration", residence status "dwelling_state", rental method "Rent_way" These three missing category data are replaced by NA. For the numeric data house neighboring subway station "subway_site", subway route "subway_route", the distance from the subway station "distance" may be similar to other houses in the surrounding area, the missing can be replaced by the median value of its neighbors. For the number of residential houses where the missing value is relatively small, "Rent_amount", the district-level administrative unit "region", and the location of the cell where the cell is located, "location" data is filled with 0.

3. Feature Engineering

The goal is to make it suitable for the model by processing the data. There are two encoding methods for the category data, LabelEncoder and OneHotEncoder. Simply put, LabelEncoder digitizes the labels of each type of feature, and uniformly converts the tag values into fixed values in the range corresponding to the range (number of tags - 1), such as 0, 1, 2...; using pd.get_dummies() OneHotEncoder is uniquely encoded, which can divide the original feature column into multiple columns by label, and each column of data is converted to 0 or 1. In general, if the number of categories of one hot encoding is not too large, it is recommended to give priority. Next, let's look at the distribution skewness of each numerical data:

Table 1. The Distribution Skewness of Each Numerical Data

Area	46.262526
Rent_amount	2.803272
Number_wei	1.778526
Number_bedroom	0.576230
location	0.288180
Village_name	0.116097
subway_site	0.011126
Total_floor	-0.145837
Number_hall	-0.165989
distance	-0.185315
subway_route	-0.537086

For non-normal distributions, we convert their log to conform to a normal distribution. Finally, the OneHotEncoder is uniquely encoded for the remaining category features. At this time, the number of features is expanded to 108, and the feature project ends.

4. Experimental Results

4.1. The advantage of housing monthly rent forecasting model based on LightGBM

LightGBM is a gradient Boosting framework that uses a decision tree-based learning algorithm. It can be said to be distributed and efficient, with the following advantages:

Faster training speed and higher efficiency: LightGBM uses a histogram-based algorithm. For example, it loads successive feature value buckets into discrete bins, which makes it faster during training.

Lower memory footprint: Using discrete bins to save and replace consecutive values results in less memory usage.

Higher accuracy (compared to any other lifting algorithm): It produces a more complex tree than the level-wise splitting method by the leaf-wise splitting method, which is the main factor for achieving higher accuracy. However, it sometimes causes over fitting, but we can prevent over fitting by setting the “max-depth” parameter.

Big data processing capability: Compared to XGBoost, it can also have the ability to process big data due to its reduction in training time.

It can be seen that the data of the rent forecasting problem is huge and needs accurate prediction. LightGBM (Light Gradient Boosting) is very suitable for dealing with this problem, and the time efficiency and prediction accuracy are very high.

4.2. Evaluation criteria for predictive models

The rent price obtained in this paper is the continuous value of the forecast, so the error and accuracy of the house rent price forecasting model are evaluated by the mean error (RMSE) and the goodness of fit(R^2). The average error can measure the deviation between the prediction result and the real result. The smaller the average error is, the closer the prediction result is to the real result, and the more the difference is. The goodness of fit can evaluate the goodness of fit of the house price prediction model. The closer the goodness of fit is to 1, the better the degree of fit between the predicted and actual results. The RMSE and R^2 two evaluation criteria can measure the pros and cons of the housing price forecasting model. The formula definitions of the two evaluation criteria are:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}\left(\widehat{y}_i-y_i\right)^2}{n}} \quad (1)$$

$$R^2=\frac{\sum_{i=1}^{n}\left(\widehat{y}_i-\overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2} \quad (2)$$

Among them : $\widehat{y}_i$ represents the predicted price of the house rent obtained from the model ; y_i is the real rent price of the house ; n is the number of samples ; $\overline{y_i}$ is y_i average value

4.3. Model establishment and predictive evaluation

In this paper, the training set (a total of 196539*108 data) is used for training. The trained model predicts the test set (56279*108 data), builds the LightGBM (Light Gradient Boosting) model in Python language, and then adjusts the parameters. In this way, the model is optimized, and the model is trained by means of 5-fold cross-validation to obtain the final house price prediction result and prediction effect.The parameters used in the final model are:

boosting_type='gbdt',
objective='regression',
metric= 'rmse',
num_leaves=1000,
learning_rate=0.05,
n_estimators=1000,
max_bin = 55, bagging_fraction = 0.8,
bagging_freq = 5, feature_fraction = 0.2319,
feature_fraction_seed=9, bagging_seed=9,
min_data_in_leaf =6, min_sum_hessian_in_leaf = 11

The average of the mean errors (RMSE) obtained by this model is 0.1429， goodness of fit(R2) is 96.13%.The entire experimental process is：

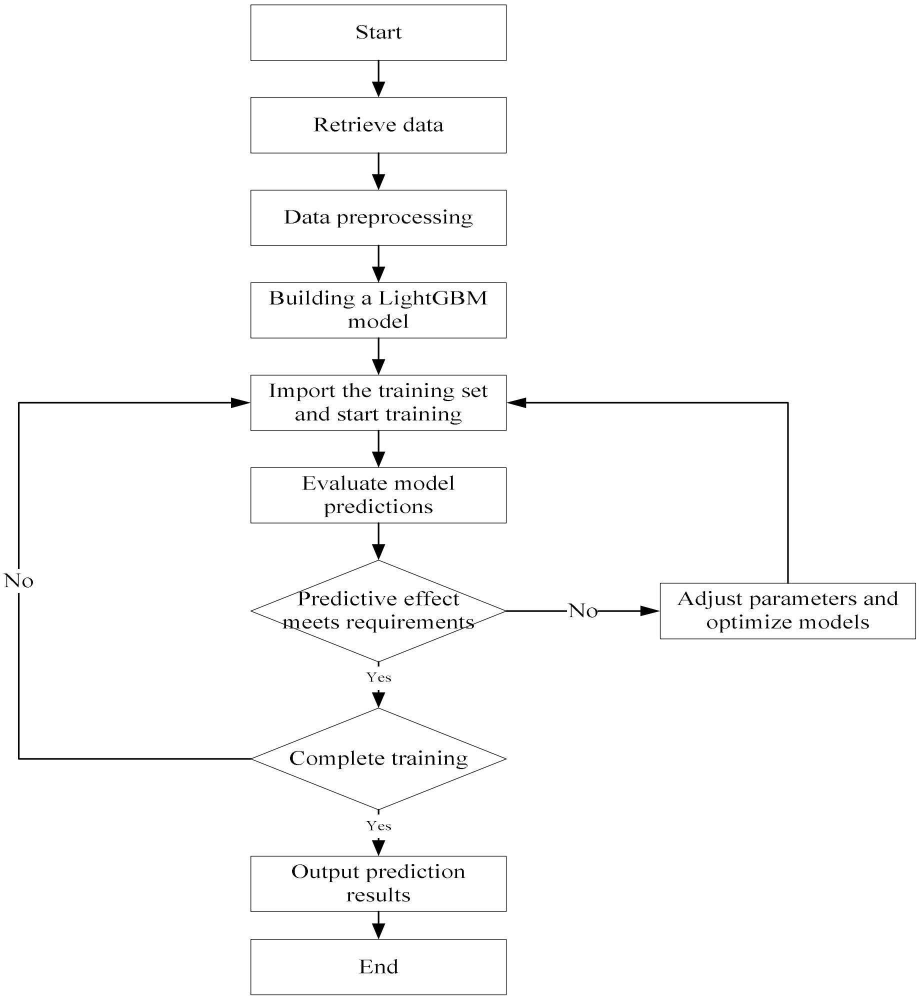

Start
Retrieve data
Data preprocessing
Building a LightGBM model
Import the training set and start training
Evaluate model predictions
Predictive effect meets requirements
No
Adjust parameters and optimize models
Yes
Complete training
No
Yes
Output prediction results
End

Figure 5. The entire experimental process

5. Conclusion

This paper establishes a model of LightGBM (Light Gradient Boosting) for rent forecasting. Python predicts the rental market data of a certain city. According to the experimental results, the error of the experimental results in the model with less training times is only 0.1429, and the prediction accuracy is high. Through the LightGBM (Light Gradient Boosting) model, the forecasting accuracy of the forecast is as high as 96%, and the results are satisfactory

References

[1] Xu Zhe. Kaggle house price actual combat summary . https://www.jianshu.com/p/76cd5ccc996f. 2018, 4, 2

[2] Lv Hao. Research on housing price forecasting model based on deep confidence network. Tianjin Science & Technology. 2018, 10.

[3] Li Jiahao, Zeng Dan. Price learning integrated learning for machine learning. Electronic Technology.

[4] Gao Wen, Li Fuxing, Niu Yongjie. Research on housing price forecast based on bp neural network. Journal of Yanan University. 2018, 3.

[5] Fan Chenchen, Cui Zechen, Zhong Xiaofeng. House prices prediction with machine learning algorithms. ICMLC. 2018, 6-10.

[6] Lu Sifei, Li Zengxiang, MongGoh RickSiow. A hybrid regression technique for house prices prediction. IEEE International Conference on Industrial Engineering & Engineering Management. 2018, 319-323.