Analysis of real estate prices in Beijing using machine learning algorithms

1. Introduction

Beijing is the capital of the People's Republic of China. By the end of 2018, the permanent population of Beijing was 21.542 million, with a GDP of 303.2 billion yuan and a per capita GDP of 140,000 yuan. In the past two decades, the real estate market in mainland China has flourished and house prices have continued to rise. Beijing has always been the vane of China's real estate market, leading the Chinese real estate market to rise and fall again and again. Among them, the housing price sharply rose in 2009 and 2015, which were transmitted from first-tier cities such as Beijing to second- and third-tier cities, which eventually led to a nationwide real estate bull market. In particular, the real estate bull market in 2015 has exerted tremendous pressure on the Chinese people's psychology. When the real estate market is the hottest, the streets and alleys are talking about buying houses after the meal, and the real estate market is almost crazy. If you want to buy a house in Beijing, you may be wondering what factors affect the price of Beijing real estate, and which factor has the most significant impact on housing prices. With these questions, we will explore the second-hand housing market in Beijing.

2. Data acquisition and cleaning

2.1 Data acquisition

The data for this topic comes from kaggle， you can download the csv file from this url (https://www.kaggle.com/ruiqurm/lianjia). It includes URL, ID, Lng, Lat, CommunityID, TradeTime, DOM(days on market), Followers, Total price, Price, Square, Living Room, number of Drawing room, Kitchen and Bathroom, Building Type, Construction time, renovation condition, building structure, Ladder ratio(which is the proportion between number of residents on the same floor and number of elevator of ladder. It describes how many ladders a resident have on average), elevator, Property rights for five years（It's related to China restricted purchase of houses policy), Subway, District, Community average price.

2.2 Data cleaning

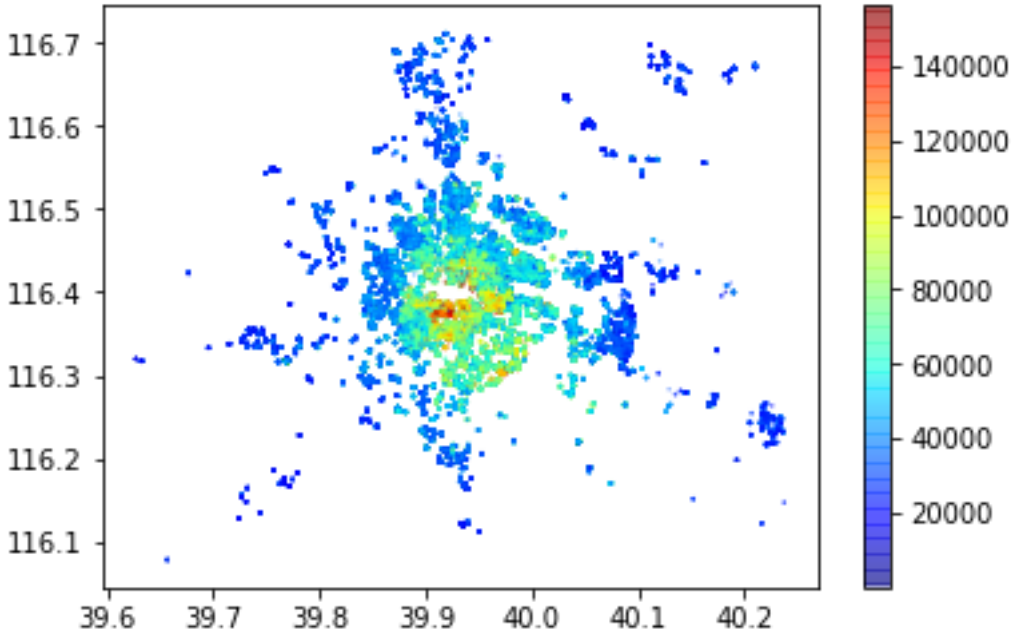
I use the pandas library in the Python language to read the csv dataset, use the dropna method of pandas to remove the missing values, select the dataset with 'Cid' as '1111030000000', and convert the data type in 'tradeTime' to datetime. Then filter the dataset with 'tradeTime'=2016 and get a dataset with 62779 rows *25 columns. I tried to use the dataset for data analysis and found that the amount of data was too large, which caused some operations to be impossible. Therefore, we randomly selected 500

samples as the research object of this study.

3. Exploratory Data Analysis

3.1 House price distribution map

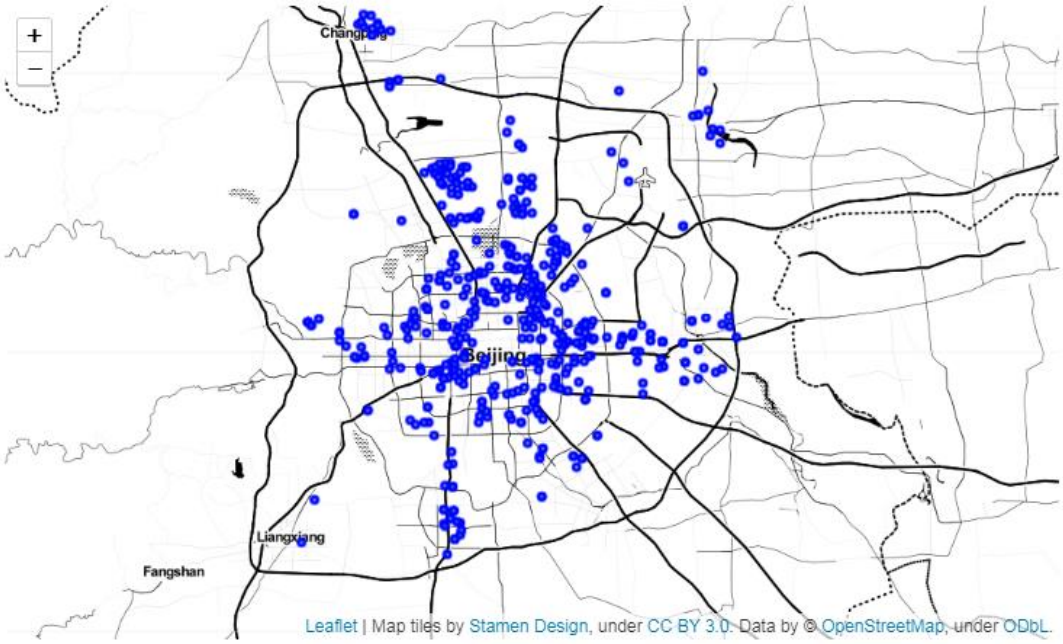
Using the longitude and latitude of the dataset as position variables, different colors represent the price of the property, and the results are shown in the figure.



The results show that Beijing's housing prices are very affected by the regional location. The closer to the core area of Beijing, the higher the house price, the most housing prices in the core area are higher than 100,000 yuan per square meter. On the contrary, the farther away from the core area of Beijing, the lower the house price, the price of some suburbs in Beijing is 20,000 yuan per square meter, and the price difference with the core area of Beijing is more than 5 times, the price difference is huge.

3.2 Sample distribution map

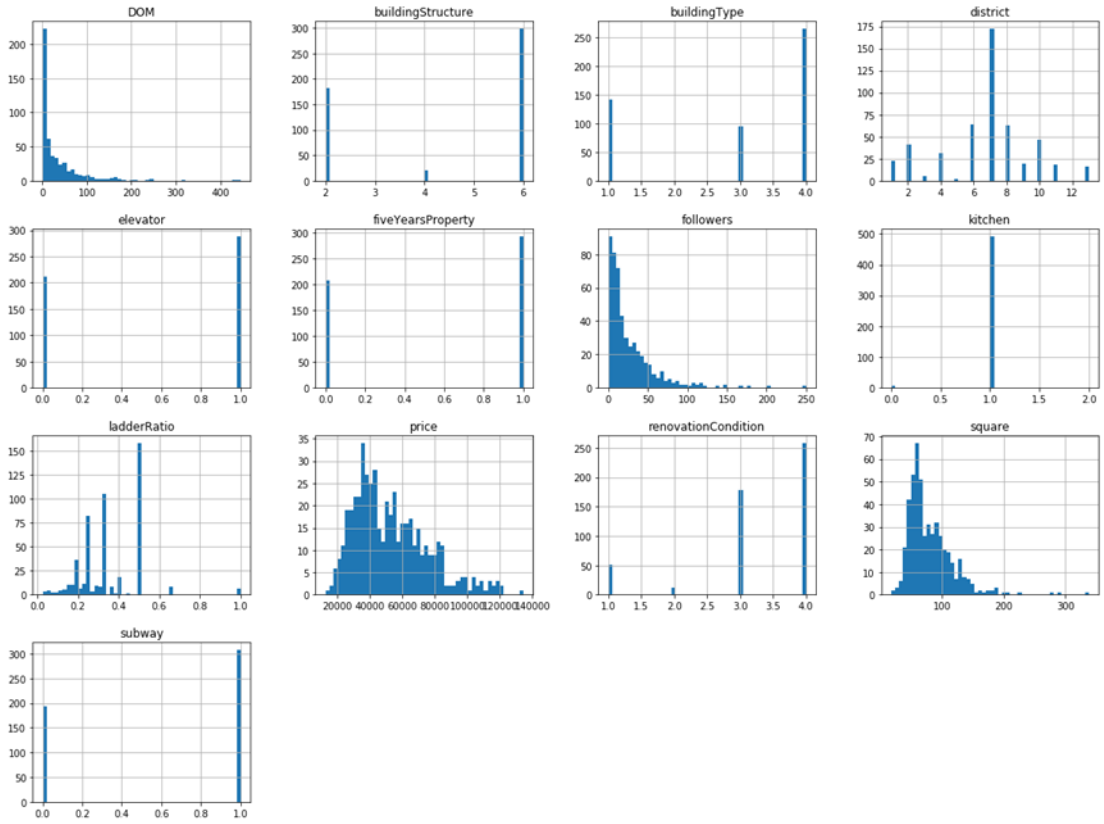
Use python's folium library to draw the distribution of 500 samples in Beijing, and test the randomness and balance of sampling. The result is shown



It can be seen from the figure that the five hundred samples are evenly distributed in various regions of Beijing. Compared with the above figure, the selected samples are well represented.

3.3 Data distribution

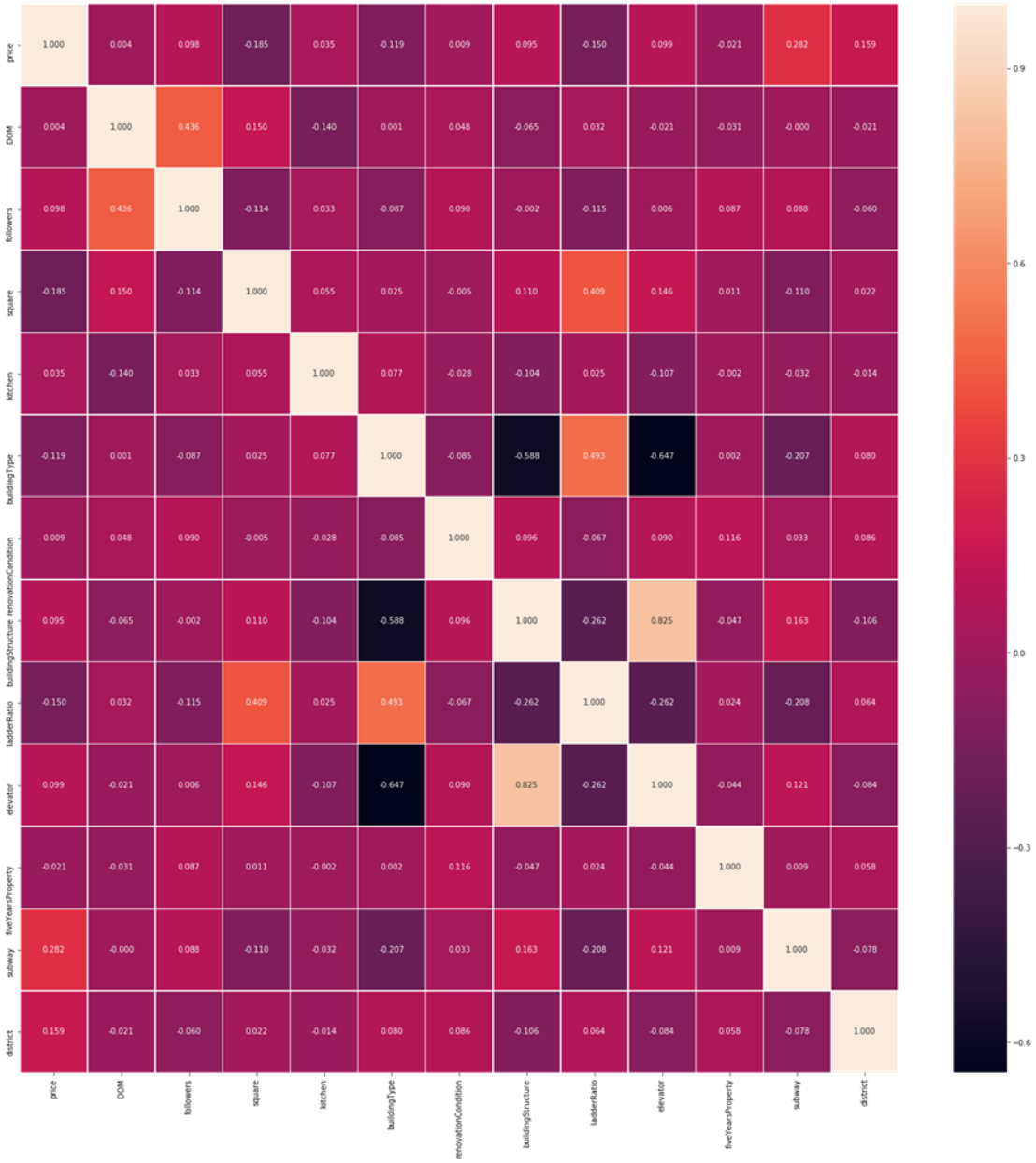
The frequency distribution figure is made for each variable of the sample, and the result is shown below.



As can be seen from the figure, the average price of the selected 500 houses is 53020 yuan / m2. Continuous variables in sample feature variables, such as DOM, followers, price, and square, are typical negative skewed distributions, indicating that most samples have small values. Mixed and steel-concrete composite are the most common building Structure. The most common building types are: plate, tower, combination of plate and tower. Most houses are concentrated in the area coded 7. There are slightly more houses with elevators than houses without elevators. More than half of the house property rights have been five years. Almost all houses have only one kitchen. Most of the houses have subways nearby. And most of this property is simply renovated or hard renovated.

3.3 Relevance between variables

We do a heat map of the correlation of the variables, and the results are shown in the figure.

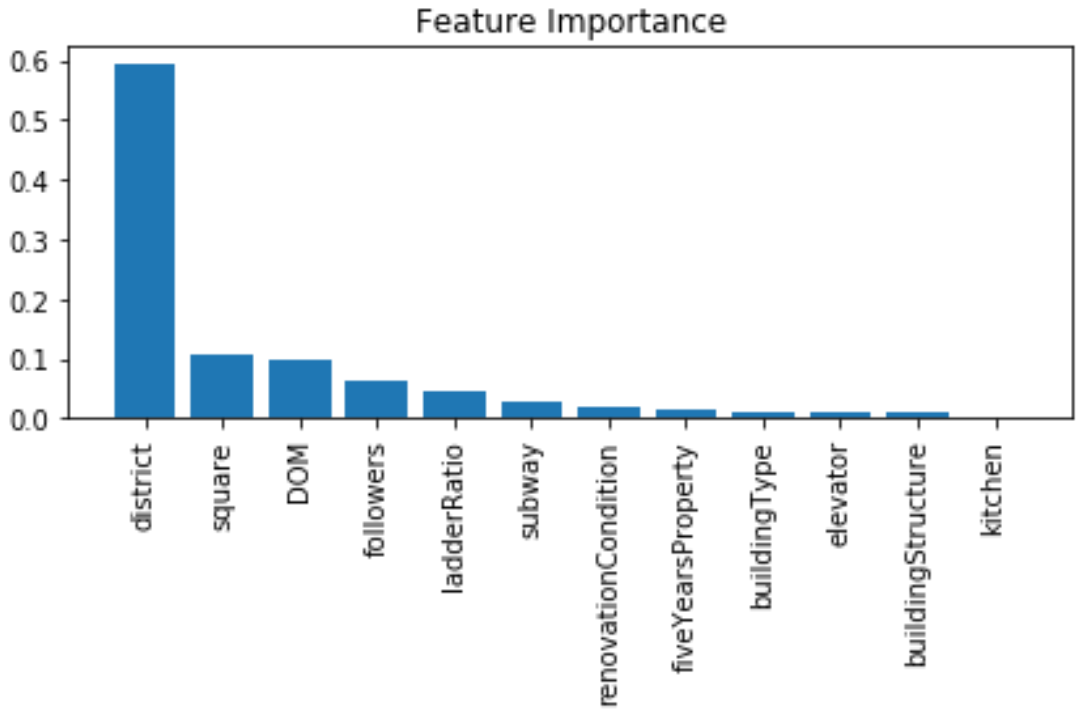


We focus on the relationship between variables and prices. As you can see, DOM, followers, kitchen, renovation Condition, building Structure, elevator, subway, district and house prices are positively correlated, while square, five Years Property, ladder Ratio and building Type and house

prices are negative related

4. Predictive model and result

I select the part of the variable whose data type is int and float as the research data set, and then use the train_test_split method in sklearn.model_selection to divide the data set into a train data set and a test data set. Then, using the RandomForestRegressor algorithm to construct the Beijing housing prediction model, and evaluate the importance of each variable. Finally, the test data set is used to calculate the MSE and R2 of the model to verify the accuracy of the model. The result is shown in the figure and table.



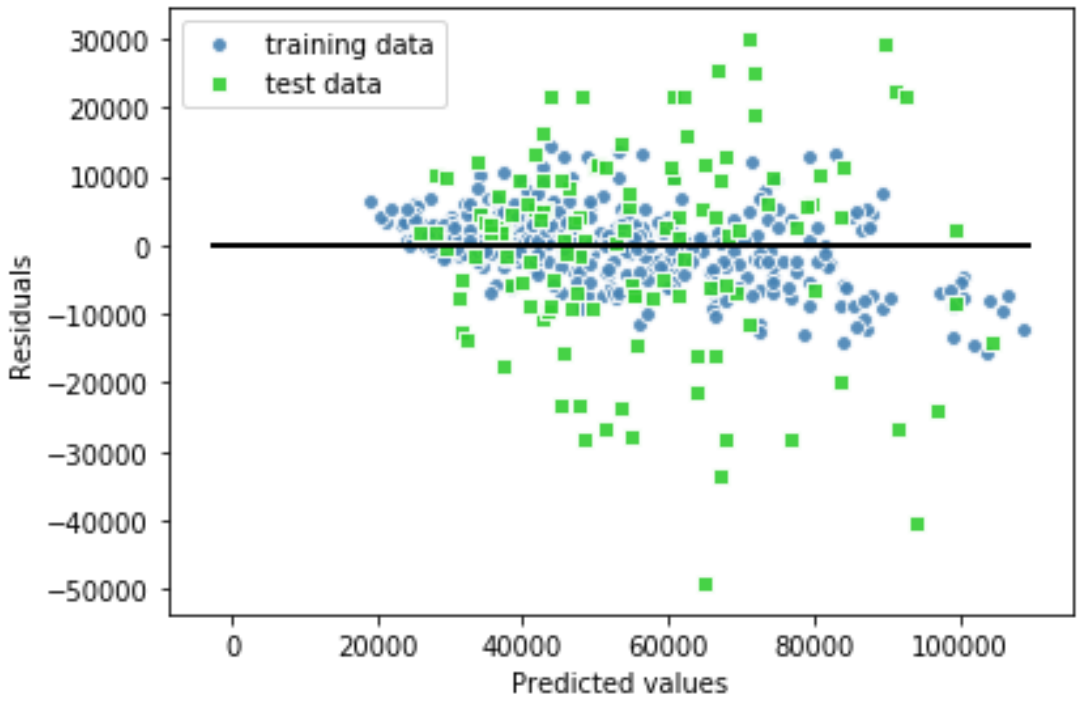
As can be seen from the above figure, the area where the property is located has the greatest impact on house prices, and its importance is close to 0.6, which is consistent with the results we saw above in the house price distribution chart. Second is the size of the property and the DOM, which are all close to 0.1. Then there is the followers and ladderRatio of the property, which shows that for a house, the more people pay attention, the higher the elevator occupancy rate and the more expensive the house price. Subway also has a big impact on house prices, and houses close to the subway are often more expensive due to convenient transportation. The degree of decoration of the house also has a greater impact on house prices. Generally speaking, the higher the degree of house decoration, the higher the house price. The rest of the property has a relatively small impact on house prices.

The MSE and R^2 of Predictive model in Train and Test data set

	MSE	R^2
Train	26446674.309	0.947
Test	205929032.236	0.629

As can be seen from the above table, the R2 of the Train data set is 0.947, indicating that the degree of fitting is high and the accuracy of the model prediction is high. The R2 of the Test dataset is 0.629, and its fitting degree is lower than that of the Train dataset. However, the constructed house price forecasting model still shows good accuracy, indicating that

the predictive ability of the model is relatively reliable.



Ideally, the residuals should be randomly distributed around the centerline. It can be seen from the residual graph that the forecast value is the horizontal axis and the residual is the vertical axis. Whether it is the Train data set or the Test data set, the residual value of the house price forecast is centered on 0, and the upper and lower sides are basically symmetrically distributed. However, the residual of the Test data set has a tendency to increase with the increase in house price forecast, indicating that the forecast model predicts a large error between the high house price and the actual house price.

5. Conclusion

This study takes Beijing 2016 'Cid' as '1111030000000' as the research object and selects 500 data as samples. The RandomForestRegressor machine learning algorithm is used to construct the house price forecasting model to study the impact of different real estate characteristics on house prices. The results show that in Beijing, the most important factor in determining housing prices is the location (location) of the property. The closer to the center of Beijing, the higher the house price, the farther away from the city center, the lower the house price. The impact of location on house prices is almost decisive. Secondly, the size of the house, the DOM and the number of people concerned also have a greater relationship with the house price, so buyers should also pay attention to the impact of these factors on house prices. Whether the house has an elevator or not, and whether it is close to the subway also has a greater impact on house prices, usually with elevators and houses near the subway. The above house price forecasting model constructed by machine learning algorithm can predict the house price better in the test data set, and its R2 is 0.629, indicating that it has a good explanatory ability for the variation of the house price. In summary, this study uses the RandomForestRegressor machine learning algorithm to successfully build a house price forecasting model, which can predict house prices through some characteristics of the property.