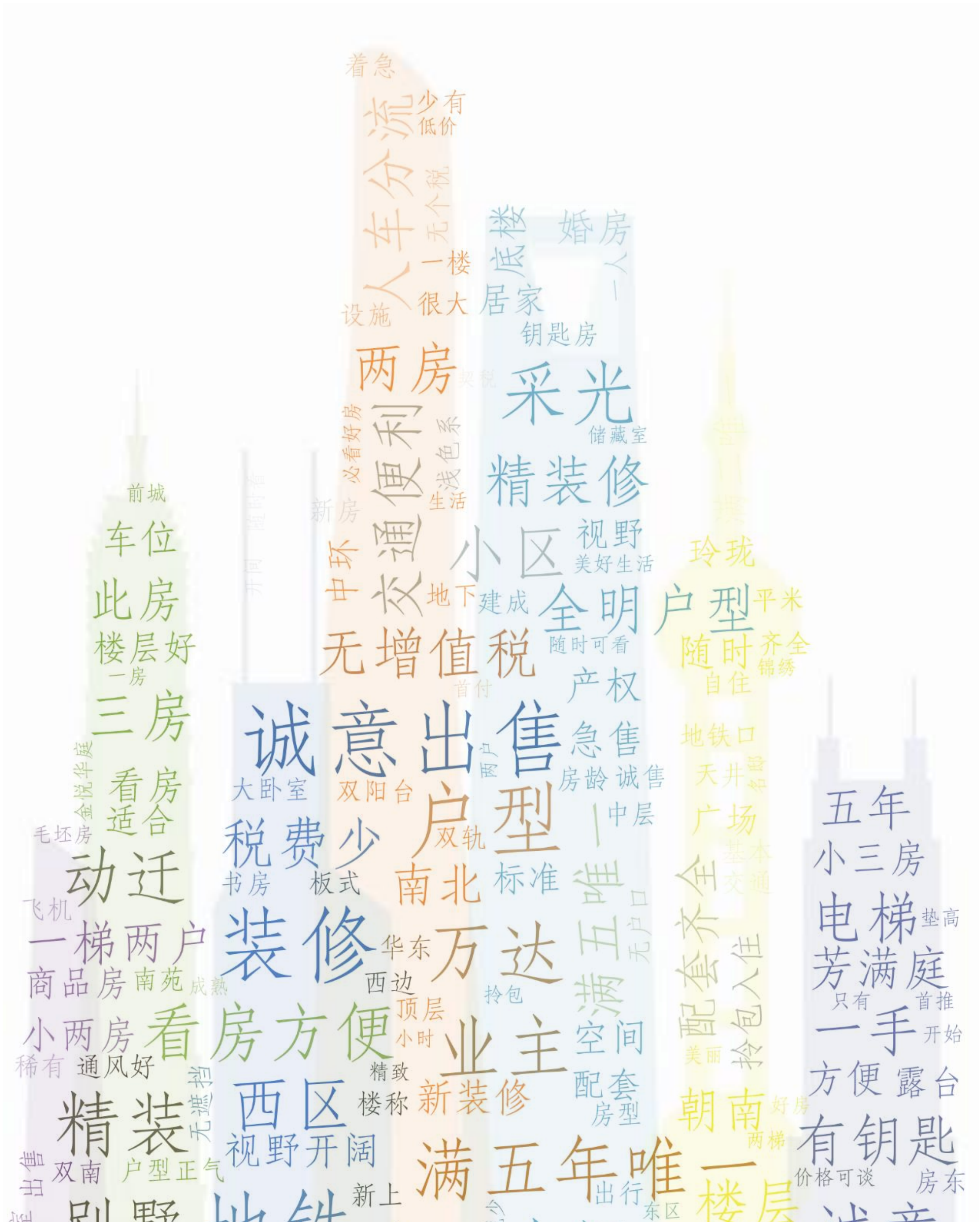
Shanghai House Price Modeling, Evaluation and Visualization

Zhen Wu

19 November 2019



Introduction

1.1 Background

As the financial centre of China, Shanghai has attracted a large influx of migrants due to its geographical location, industrial structure, educational environment and public resources., which makes it one of the most populated cities in the world. Subsequently, this results in Shanghai to be the 6th highest city [1] in terms of house price to income ratio. Selling or buying a house is no doubt the most important decision for the majority of people living in Shanghai. After the explosive growth of the real estate market in the last few decades, it is gaining more stability at the moment. Therefore, it is a great time to gather information and perform analysis. Gaining more insight into the underlying principle of how the houses are priced in Shanghai and being able to predict the value of house using mathematical models would certainly help the seller and buyers to make the right decision.

1.2 Problem

Data that contribute to determining the house price might include the community information, location information, area, orientation, floor and built year. This project aims to collect all the related data from the local real estate agency website to build a statistical model predicting the value of the houses in Shanghai and analyses the importance of each kind of information.

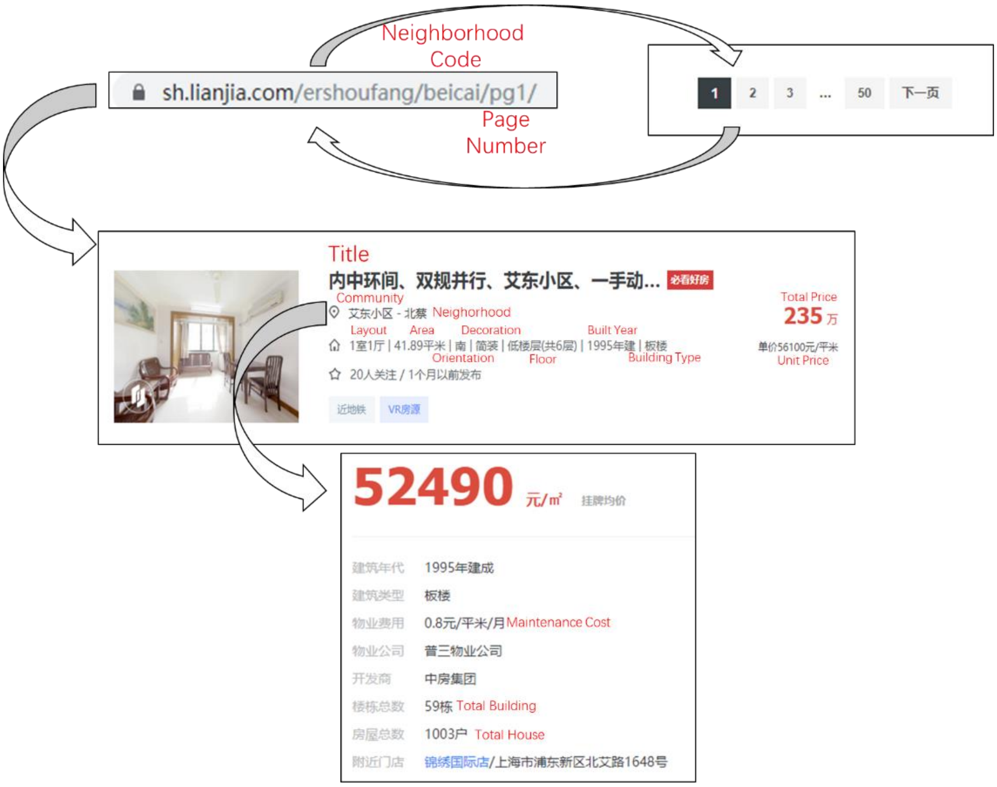
1.3 Interest

I am looking to move to Shanghai soon, so gathering as much information as possible and having a general look prior to setting off is critical for me. Other people who are looking to sell or buy the house in Shanghai might be interested in the insight provided by this report and the predictive model produced by this project.

2. Data acquisition and cleaning

2.1 Data sources

Lianjia.com is one of the biggest real estate agencies in China and its website lists all the houses for sale as well as displaying accurate information on each of the listing. By using the BeautifulSoup library, I have managed to scrap all the listing information from the website, the workflow is shown in figure 1.



sh.lianjia.com/ershoufang/beicai/pg1/

1 2 3 ... 50 下一页

内中环间、双规并行、艾东小区、一手动... 必看好房

艾东小区 - 北蔡

1室1厅 | 41.89平米 | 南 | 简装 | 低楼层(共6层) | 1995年建 | 板楼

20人关注 / 1个月以前发布

近地铁 VR房源

235 万

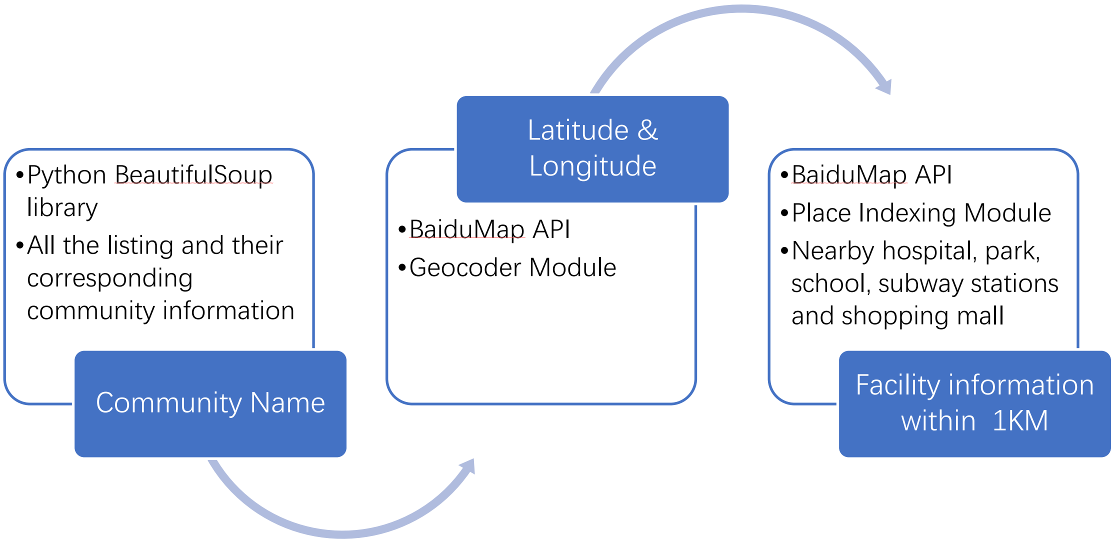
单价56100元/平米

52490 元/㎡ 挂牌均价

建筑年代	1995年建成
建筑类型	板楼
物业费用	0.8元/平米/月
物业公司	普三物业公司
开发商	中房集团
楼栋总数	59栋
房屋总数	1003户
附近门店	锦绣国际店/上海市浦东新区北艾路1648号

Figure 1

As shown in figure 1, the scraping algorithm first reads all the district name and total page number of each neighbourhood. It then crawls the listing information (name of the listing, community, neighbourhood, layout, area, orientation, floor, built year, following information, total price and unit price). Then the scrawler goes to community page to scrawl the maintenance cost, total building and total house.



Community Name
- Python BeautifulSoup library
- All the listing and their corresponding community information

Latitude & Longitude
- BaiduMap API
- Geocoder Module

Facility information within 1KM
- BaiduMap API
- Place Indexing Module
- Nearby hospital, park, school, subway stations and shopping mall

Figure 2

Next, instead of using foursquare API that IBM professional data science course recommended, a Chinese equivalent API (BaiduMap API) is used to gather the location information of each community. As shown in figure 2, the latitude and longitude of each community are obtained using API's geocoder function, this piece of information is then fed into the API's place index module to gather the hospital, shopping mall, school, park and subway station information within 1km radius of each community.

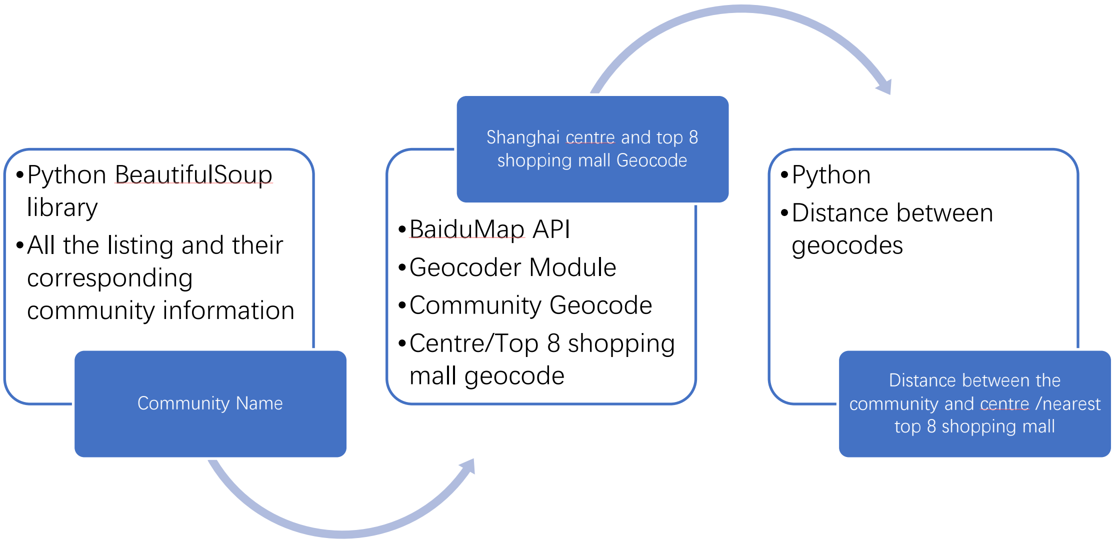

•Python BeautifulSoup library
•All the listing and their corresponding community information
Community Name
Shanghai centre and top 8 shopping mall Geocode
•BaiduMap API
•Geocoder Module
•Community Geocode
•Centre/Top 8 shopping mall geocode
•Python
•Distance between geocodes
Distance between the community and centre /nearest top 8 shopping mall

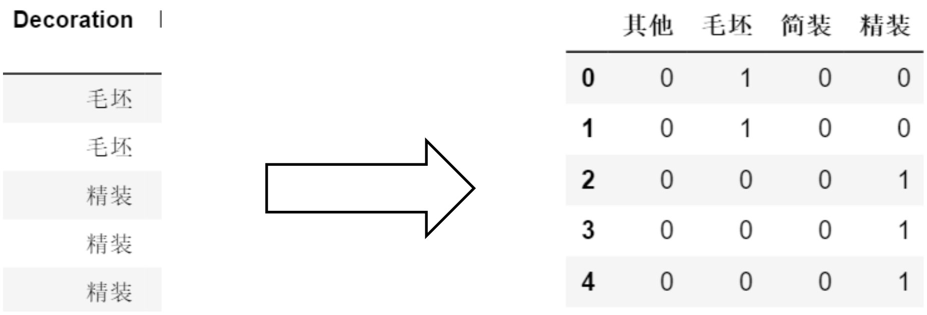
Figure 3

Finally, as shown in figure 3, the reverse geocoding function of BaiduMap API is used to get the latitude and longitude of the centre of Shanghai and the 8 biggest shopping areas. The distance between each community and center as well as the nearest shopping area is then calculated using Python

2.2 Data Cleaning

All the information scraped or returned from API are consolidated into a table. The csv table was encoded using the GBK standard as it contains Chinese characters. Before further modelling and analysis, the following processes are applied to the data collected:

1. Drop listings with missing values: a few communities are lacking maintenance cost information and some houses are missing information such as floor, orientation and built year. These informations are vital when determining the house price and there is no good way to fill in these informations.
2. Format the data: the listing data scraped is a mix of characters and numbers in string format. For example, the built year scraped are strings like "1997 年建", in this case, the last 2 Chinese characters are dropped, and the numbers are converted from string to float.
3. Create dummies for all descriptive features, take decoration as an example showing in figure 4



Decoration
毛坯
毛坯
精装
精装
精装

	其他	毛坯	简装	精装
0	0	1	0	0
1	0	1	0	0
2	0	0	0	1
3	0	0	0	1
4	0	0	0	1

Figure 4

2.3. Feature Selection

After data cleaning, there were 62,400 samples and 230 features in the data. There was some redundancy in the features after having a close look at the data. For example, the total price of each house collected, which can be calculated from the unit price and area in the data collected. Therefore, the total price of the house is redundant. In addition, the neighborhood information can be treated as a more specific version of district information with less entropy, so the district information can be dropped. Finally, the information such as the floor is categorized as high, mid, low and whole building. As these are all the categories we have, we can safely drop one of these features. Finally, there is information that is irrelevant to the house value, such as the title of the listing, name of the community, etc. The irrelevant information is dropped as well.

3. Exploratory Data analysis

3.1 Shanghai House Price Heatmap

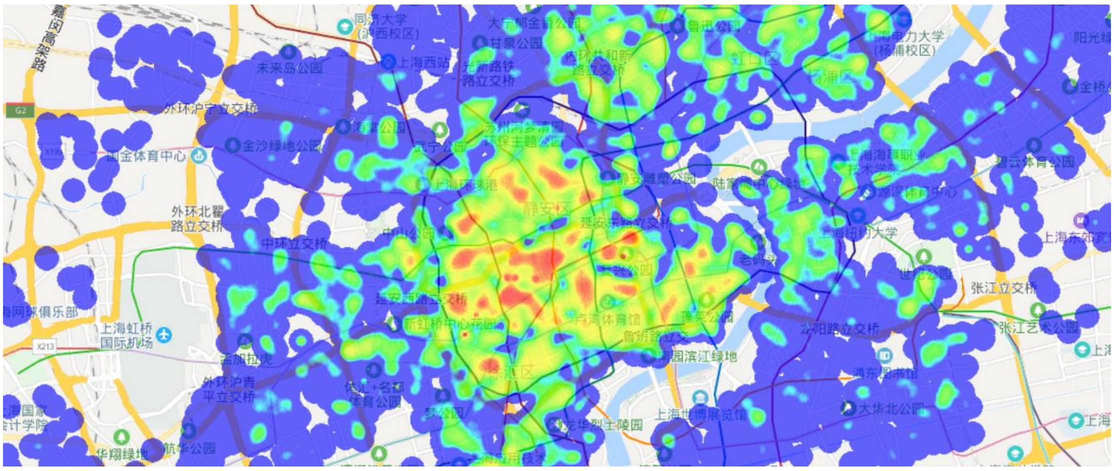


Figure 5

The heatmap of the unit price is shown in figure 5. The west side of the river has a much higher average unit price than the east. The high unit price areas in the west are concentrated around Lujiazui, which is the financial center of Shanghai. On the west side, it is evident that the house price decreases as moving further away from the center.

3.2 Violin Plot of Unit Price in each District

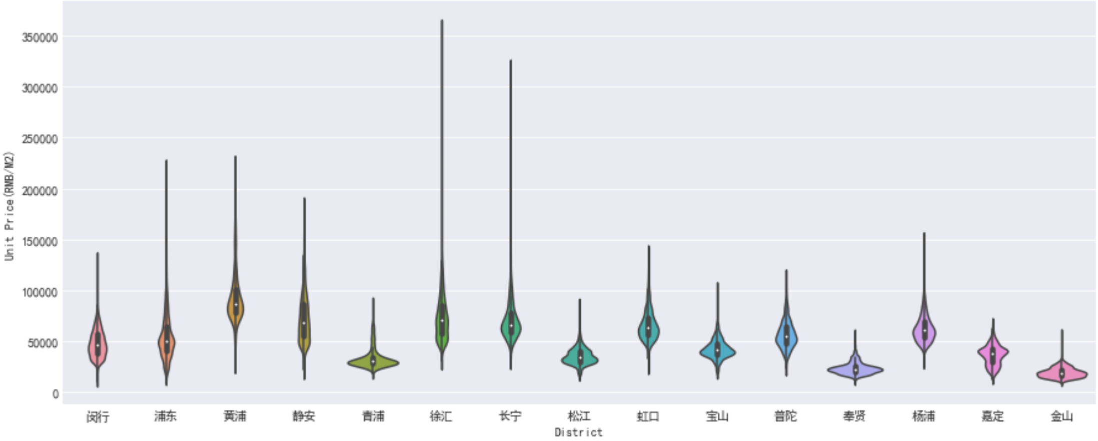


Figure 6

It is shown that urban districts such as Xuhui and Changning have more spreaded and overall higher unit house price, sub-urban areas such as Qingpu, Fengxian and Jingshan, on the other hand, house prices are lower and within a narrower range.

3.2 Unit Price Distribution

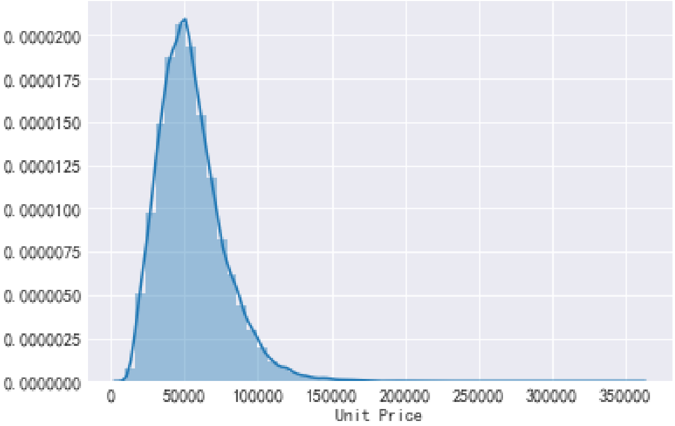


Figure 7

As shown in figure 7, the unit price of most houses are ranged from 20000RMB to 50000RMB.

3.4 The Correlation between the Unit Price and the distance to the Centre

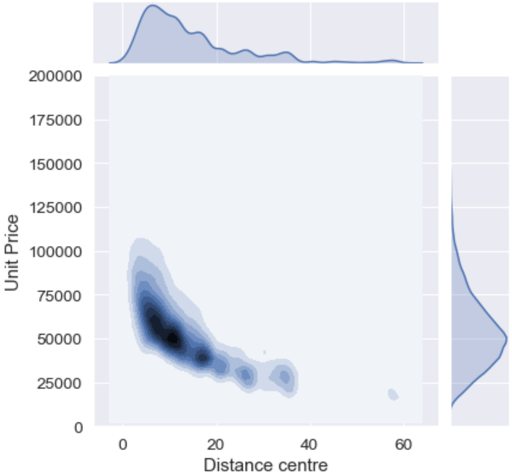


Figure 8

As the house further away from the centre, the lower unit price it has. The correlation between the distance to the centre and unit price is 0.63.

3.5 The correlation between the Unit Price and the Area

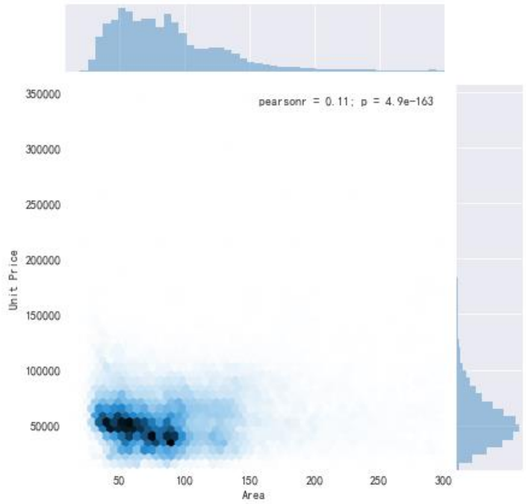


Figure 9

In this Hexbin plot, we can tell that the unit price has almost no correlation with the unit price.

3.6 The Correlation between the Unit Price and the Built Year

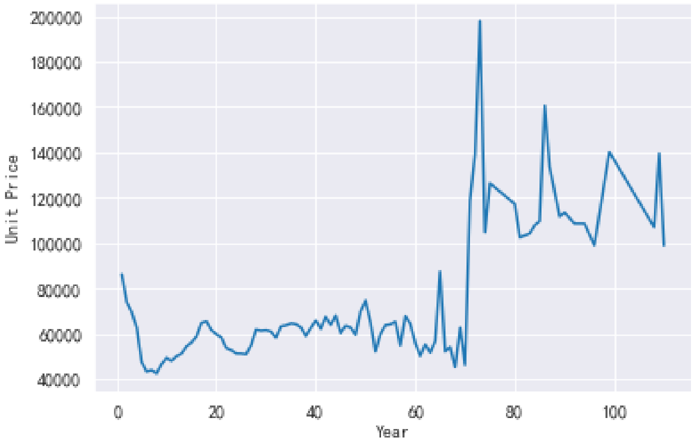


Figure 10

As shown in figure 10, there is a dramatic jump at around 70 years. The houses are more than 70 years old were before the People's Republic of China, so these houses may have special historical values, which results in such high price

3.7 Heatmap of Correlations between each Features.

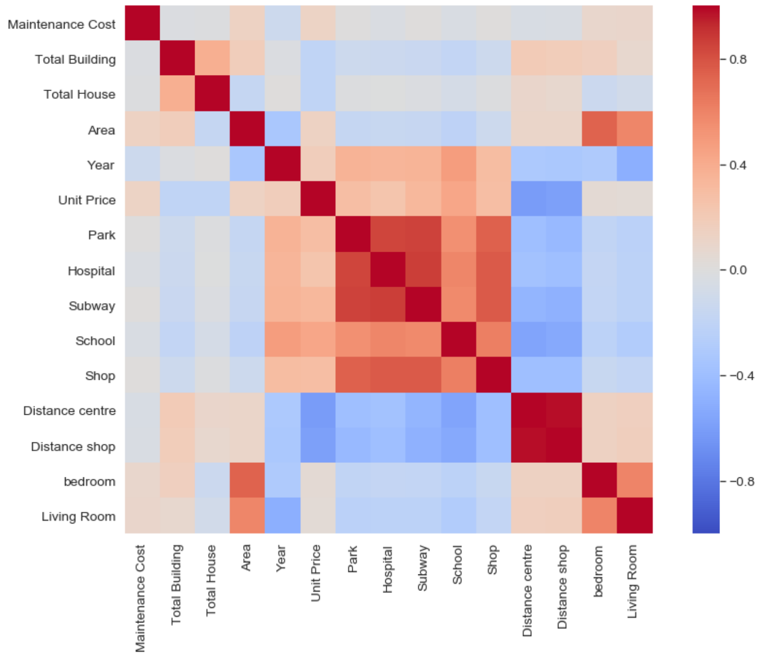
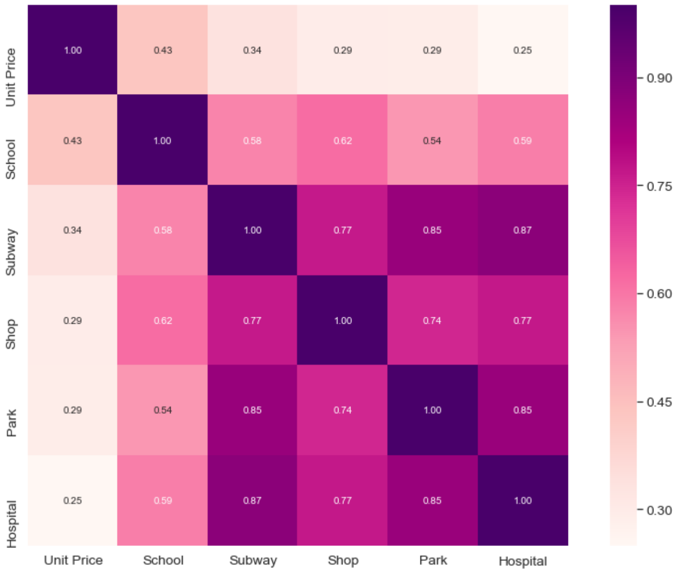


Figure 11

From the figure above, several interesting observations can be obtained. Firstly, further away from the city center, the fewer facilities (park hospital, subway, school and shop) a house has around. Secondly, the distance to the center and the distance to the nearest commercial area looks like are strongly related, this will be investigated further later. Thirdly, the house built year seems to be inversely correlated to the number of bedrooms, living room and area, this suggests that the total area of the house is getting bigger as time goes by, this is probably resulting from the cheaper land price during the expansion of Shanghai.

3.8 Top 5 Unit Price correlated Features



	Unit Price	School	Subway	Shop	Park	Hospital
Unit Price	1.00	0.43	0.34	0.29	0.29	0.25
School	0.43	1.00	0.58	0.62	0.54	0.59
Subway	0.34	0.58	1.00	0.77	0.85	0.87
Shop	0.29	0.62	0.77	1.00	0.74	0.77
Park	0.29	0.54	0.85	0.74	1.00	0.85
Hospital	0.25	0.59	0.87	0.77	0.85	1.00

Figure 12

As shown in figure 12, the factors contribute to the house price the most are School, subway, shop park and hospital.

3.9 The Correlation between the Distance to the Center and the Distance to the Commercial Area

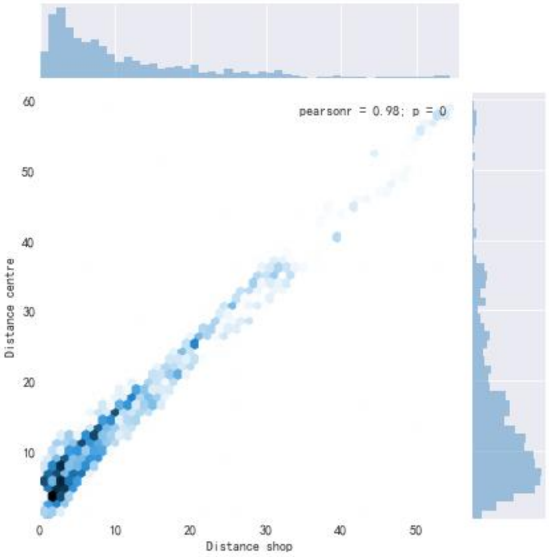


Figure 13

We can see the correlation between these two is almost 1. Which means these two variables are heavily correlated and one of these should not be used in the predictive model building.

4. Predictive Modeling and Discussion

4.1 Regression models

12 regression models (RandomForest, LinearRegression, KNNRegressor, Ridge, Lasso, MLPRegressor, DecisionTree, ExtraTree, XGBoost, AdaBoost, GradientBoost, Bagging) supplied by Scikit-learn package were used.

4.2 Applying standard algorithms and Their Performance

All 12 algorithms mentioned were applied using default settings to the house with an area less than 300m^2 and the evaluation score of each regression model is listed in table 1. The data was first divided into a training set and a testing set at a 9-to-1 ratio. All the score listed in the table are resulted from the testing set.

Table 1

Model	Score
RandomForest	0.936019752777
MLPRegressor	0.707103025454
Lasso	0.770117813651
Ridge	0.770395504881
LinearRegression	0.770414901791
KNNRegressor	0.780702909711
DecisionTree	0.903531033266
ExtraTree	0.871744614878
XGBoost	0.76924551180
AdaBoost	0.242632409602
GradientBoost	0.77066957433
Bagging	0.932413937567

As we can see from the table, random forest regression is giving the best evaluation score, around 0.936, so it is selected as the model for further tuning.

4.3 parameter tuning of regression models.

n_estimators parameter determines the number of child model in the random forest regression model. The parameter is tuned from 10 to 190 with stepsize of 10. The result is plotted in figure. As there is no more significant improvement in the score after the parameter passing 150, n_estimator = 180 is arbitrarily chosen. Finally, all the data (including the training set and testing set) are fed into the randomforestregressor for the best performance.

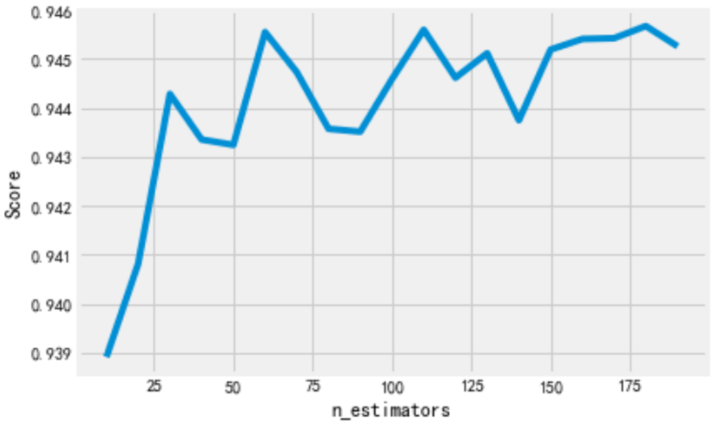


Figure 14

4.4 Random Forest Regression Visualization

Machine learning algorithms are often interpreted as black boxes. However, a glance at what inside the box would help us to understand the data better.

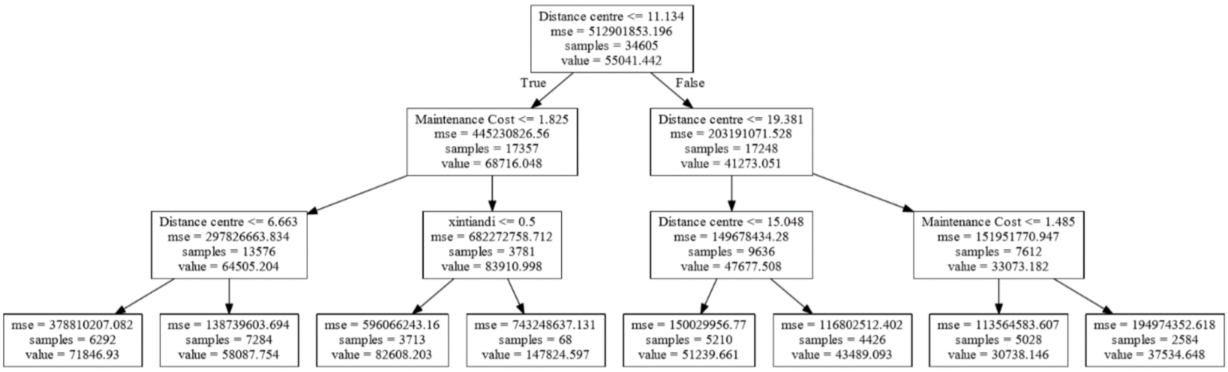


Figure 15

Figure 15 shows the first 4 layers of one of the branches of the predictive model built in this project. In the first 4 layers, we can see maintenance cost, neighborhood and distance to the center are important differentiators.

4.5 Feature Importance

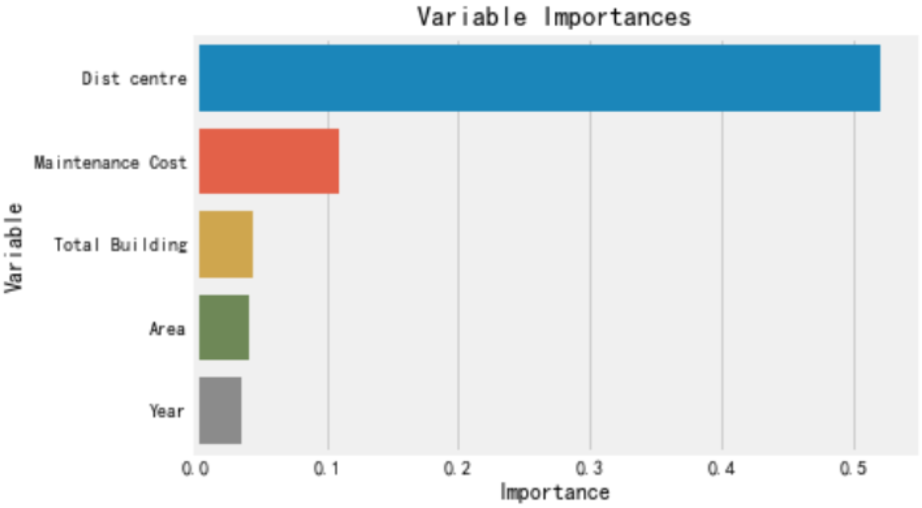


Figure 16

As shown in figure 16, the most important variable is the distance to the center. The second most important feature is the maintenance cost. Higher maintenance cost normally means greater community, this would certainly add value to the house.

5. Conclusions

The west side of Shanghai has a much higher average unit price than the east. Urban districts have more spread and overall higher unit house price. On the contrary, House prices in sub-urban areas are lower and fall within a narrower range. The further away from the city center, the less facilities (park hospital, subway, school and shop) a house has around, which all together contribute to a lower house price. Furthermore, interestingly, the houses are more than 70 years old have a way higher price than the rest.

Finally, among 12 regression models used, random forest regression model was found to have the best performance and it is found out that distance to the center and maintenance cost are the two most important differentiator in this model.

6. Future directions

- Improving the data collected: by click into each listing, more information about the house will be displayed. This step was not been done due to the time restriction.
- Improving the modelling: only one parameter is tuned in this project. GridSearch can be used in the future for multiple parameter tuning to get better performance.
- Application of the model: developing an application to automatically search for the most cost-effective house by calculating the percentage difference between the predicted price and actual price using the model.

7. Reference:

[1]https://www.numbeo.com/property-investment/rankings.jsp