Beijing Housing Price Prediction

Xiangyi Liu
University of Rochester
Xliu84@u.rochester.edu

Yaxin Liu
University of Rochester
yliu139@u.rochester.edu

ABSTRACT

This paper provides a detailed overview of the data mining process used to predict housing prices in Beijing, China after the year 2017 based on the data between 2012 and 2016.

1. INTRODUCTION

Since 2000, Beijing housing price has experienced three rising periods by 2016, each accompanied with short-period of price down. The first period is between 2000 and the beginning of 2009. During this time, a large population moved to Beijing and there was an increasing need in residential estates. Since 2000, the housing price in Beijing had increased from 3000 RMB to a peak of 15,000 RMB per square feet by 2007. The second period is between 2009 and 2012. The government invested in the housing market to boost the economy as a means to reduce the effect of the worldwide Great Recession. The third period is from 2012 to 2014. During the period, particularly the neighborhoods land close to schools experienced a great increment of housing prices.

It is interesting to see whether the increasing pattern would continue and to investigate how features other than the influence of the government could affect the housing price. The objective of this project is to apply prediction models to the dataset of Beijing housing prices since 2010 to create a forecast of price after 2017.

The approaches used as follows:

- *Exploratory Data Analysis*: Understand the overall structure, distribution, correlation of the dataset.
- *Data Preprocessing*: Translation of attributes, handling missing values, discretization of continuous attribute `TotalPrice`, conversion of categorical attributes to integer data.
- *Analysis and Modeling*: Linear regression, K-nearest Neighbors algorithm and Random Forest classification were used to model and predict.

2. EXPLORATORY DATA ANALYSIS

The dataset “Housing Price in Beijing” was retrieved from Kaggle and the original data was fetching from Lianjia.com, an estate agent company in China. The dataset contained 318,851 observations with 26 features, such as `renovationCondition`, `buildingStructure`, `floor`, etc.

We performed exploratory data analysis on the dataset to better understand the pattern of the data. R packages *ggplot2* and *corrplot* were used to visualize the data features.

2.1. Data Exploration

2.1.1 Data Types.

Factor (7): `Height, renovationCondition, buildingStructure, buildingTypeElevator, fiveYearsProperty, Subway, District.`
chr (2): `TradeTime, url.`
int (8): `DOM, Followers, Price, LivingRoom, DrawingRoom, Kitchen, Bathroom, CommunityAverage.`
num (8): `id, cid, Lng, Lat, TotalPrice, Square, ConstructionTime, LadderRatio.`

2.1.2 Missing Values. As shown in Figure 1, missing values were represented by the white area within each attribute. Over half of `DOM` contained missing values, and some data of `ConstructionTime` and `CommunityAverage` was missing. How to handle these missing values is illustrated in section 3.

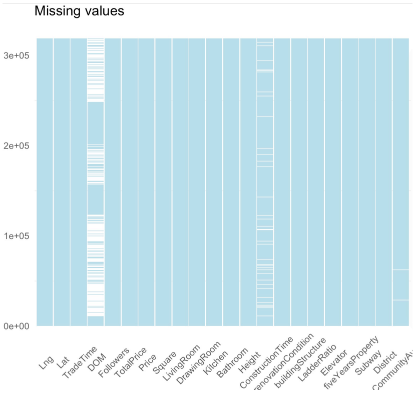


Figure 1 Distribution of Missing Values

2.1.3 *Correlation Analysis.* The Person correlation coefficient within numeric attributes was calculated and the results are presented in Figure 2.

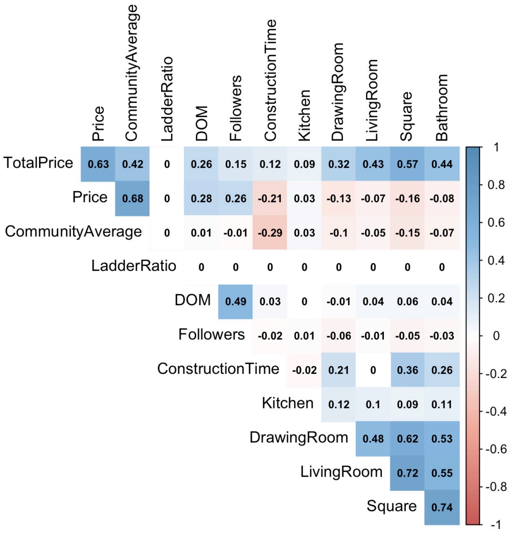
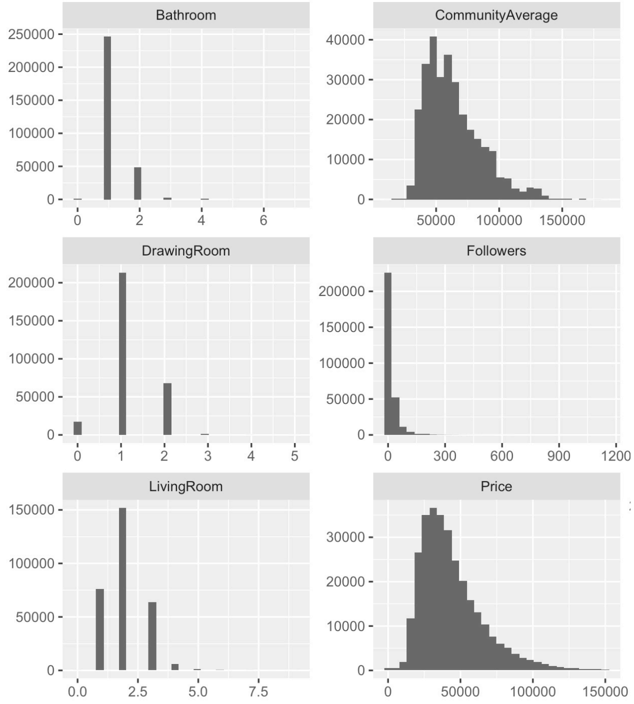


Figure 2 Correlation of Numeric Features

2.2. Distribution of Numeric Features

The distribution of each of the numeric attributes were visualized by histograms. The results are represented in Figure 3.



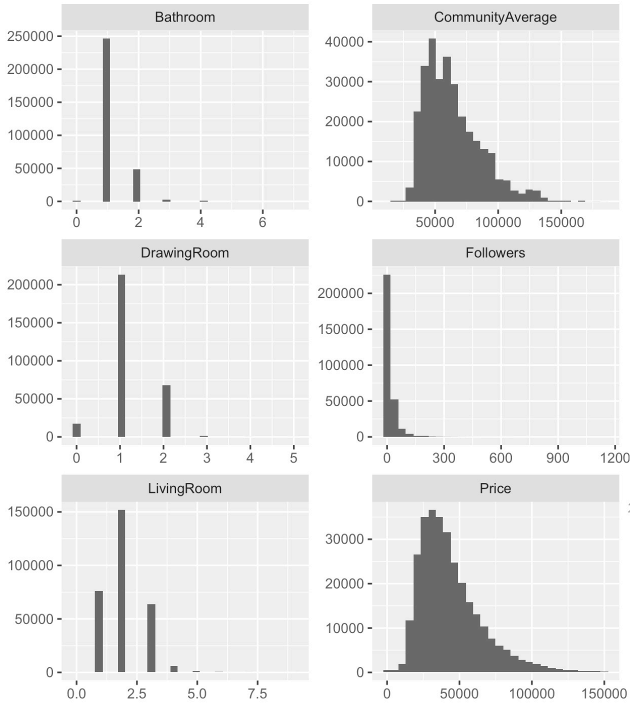


Figure 3 Distribution of Values for Numeric Attributes

2.3. Distribution of Categorical Features

The distribution of each of the categorical attributes were plotted by boxplots. The results are represented in Figure 4.

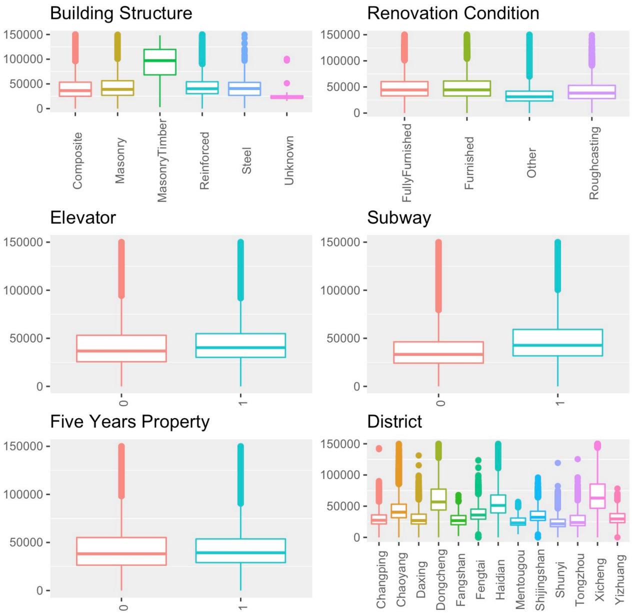


Figure 4 Distribution of Categorical Attributes

3. DATA PREPROCESSING

3.1. Data Cleaning

Filling Missing Values. The attribute DOM (Days Active On Market) contained 157,971 NAs, which constituted nearly half of the entire dataset. Directly omitting instances with NAs would cause a serious impact on the overall datasets. Thus, we substituted NAs with the median.

We noticed that there are attributes other than DOM contained NAs, which consisted of 19,746 instances. Since it only constituted a small portion of overall 320 thousand instances, we decided to omit these instances with NAs.

3.2. Data Preprocessing

3.2.1 *Translation.* Since the data is collected from Lianjia, which is an estate agent company in China, we needed to translate features from Chinese to English.

3.2.2 *Unknown Values.* We noticed that there are Unknown in several categorical attributes, such as BuildingStructure. We did not omit instances with Unknown value because we would like to consider it as one of the categories under the attributes.

3.3. Feature Engineering

3.3.1 The attribute ConstructionTime in the original dataset contained the year of construction. However, we did not think it directly represented its effect on the total price. Thus, we transformed it into years since construction.

3.3.2 *Handling Continuous Attributes.* Since TotalPrice is a continuous attribute, equal-width interval discretization was used to bin it. The number of bins was set to 10.

3.3.3 *Handling Categorical Attributes.* One-hot encoding was used to convert categorical attributes into integer data (0 or 1) in the Linear Regression model and the K-nearest Neighbor model.

4. ANALYSIS AND MODELING

The data was split into a training dataset and a test dataset. The training set contained the data whose trade time was after 2012 and before 2017. The test dataset contained the data whose trade time was after 2017.

4.1. Linear Regression

The linear regression model was chosen to be our first prediction model. After encoding the categorical attributes with one-hot encoding, all variables were numerical and could fit in the linear regression model.

The results from the test show that the model's prediction is close to the real housing price. The adjusted R-squared was 0.8903. The results of the test dataset are shown in Figure 5. However, since each instance contained 41variables after one-hot encoding using dummy function in R, the model suffers from high dimensionality.

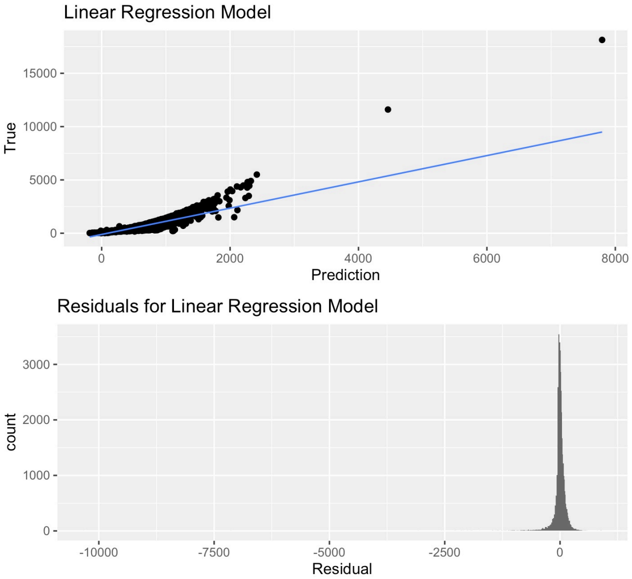


Figure 5 Results of Linear Regression Model

4.2. K-nearest Neighbor

The classification of KNN is based on the voting of the majority of 'K' nearest instances. The distance used in KNN is usually Euclidean distance. The distance is affected by the scale of attributes. In order to eliminate the effect of the magnitude of attributes on the distance, first of all we need to normalize the numerical attributes in the dataset to the same scale [0,1].

The formula used for normalization is,

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

The results of the test dataset are shown in Table 1.

Table 1 Results of K-nearest Neighbor

K	Results
3	0.8223142
5	0.8186597
10	0.8168763

4.3. Random Forest

One issue with the previous two models was the high dimensionality. Random Forest was chosen to address this problem. By equal-interval discretization of `TotalPrice`, the predication became a classification problem. Instead of converting categorical attributes using one-hot encoding, the random forest model directly took the observation's original features. The parameter *importance* in the *randomForest* function was set to be true. The parameter automatically selected the most important attributes to train the model. The model performed better than the K-nearest neighbor model, resulting in an accuracy of 0.9519664. The results of the test dataset are shown in Table 2.

Table 2 K-nearest Neighbor vs. Random Forest

Test	Accuracy
KNN (K=3)	0.8223142
Random Forest	0.9519664

5. FUTURE WORK

Overall, the Random Forest classification method has the best performance regarding to speed and accuracy. We think it is due to the Random Forest classification method automatically selected features that have the most impact on `TotalPrice`. Thus, we would like to select attributes based on their importance before we use them for training the model. We think in this way the accuracy could be improved.

The dataset contains `Longitude` and `Latitude`, so we would like to map instances on the Beijing map to create a better visualization of the distribution of the categorical attributes on the geographical level.

As the economy grows in China, Beijing as the capital and the center of the economy has expanded since 2000. Districts that seemed far away from the center of Beijing have become more connected by the construction of multiple subway lines. Also, since the economy growth takes place in more recent years in these relatively distant districts, the extent of features affect their housing price may be different to the extent of effect on district closer to the center of Beijing.

6. CONCLUSION

In this project, we employed three kinds of data mining techniques with the objective of creating the forecast of Beijing housing prices after 2017 based on 22 descriptive features from about 89 thousand instances between 2012 and 2016 as the training set. It was a challenge to train models on such a large dataset. It cost relatively significant time to train each model.

Ultimately, the Random Forest classification had the relatively high performance. We could further improve the accuracy of the other two models by a closer selection of attributes based on their extent of impact on `TotalPrice`.

7. ACKNOWLEDGEMENT

I would like to thank Professor Pawlicki and Teaching Assistants for their time and support.

REFERENCE

[1] Housing market trends 2019 - The ultimate guide. (n.d.). Retrieved December 11, 2019, from https://www.opendoor.com/w/guides/housing-market-trends-2019.

[2] Home Prices in Beijing's Best School Districts Skyrocketing Despite Housing Bubble(2017, February 23). Retrieved December 11, 2019, from https://www.huffpost.com/entry/home-prices-in-beijings-best-school-districts-skyrocketing_b_58ad98c1e4b0598627a55ebf.