Beijing House Price and DOM Report
Author: Junduo Dong, Jack Shi
Date: December 2nd, 2019

Background

Buying or selling a house is one of the most important decisions in everyone's life and house pricing would significant affect their decision. As we all know, the house pricing is affected by various factors. Some of these factors are widely known, such as floor which represents the height of the house, floor space, building year, and building type, while others are not. predicting the house will not only benefit the house buyer, but also will help the house seller to have a better understanding of their house value and It could help regulators to build a more transparent and efficient real estate market. In additional, in this report we also predict DOM which means house active days on market. DOM is a crucial factor that measures market liquidity in real estate market. Indeed, at the micro level, DOM is not only a special concern of house sellers, but also a useful indicator for potential buyers to evaluate the popularity of a house. At the macro level, DOM is an important indicator of real estate market status.

Data Source

The data we use was obtained from a data file named Beijing_data.xlsx which is fetching from Lianjia.com. This dataset is house pricing of Beijing from 2008 – 2018 and it includes some house features such as trade time, DOM, bathroom, building type, construction time and so on. The detail descriptions of the data could be found in appendix.

Data Transformation

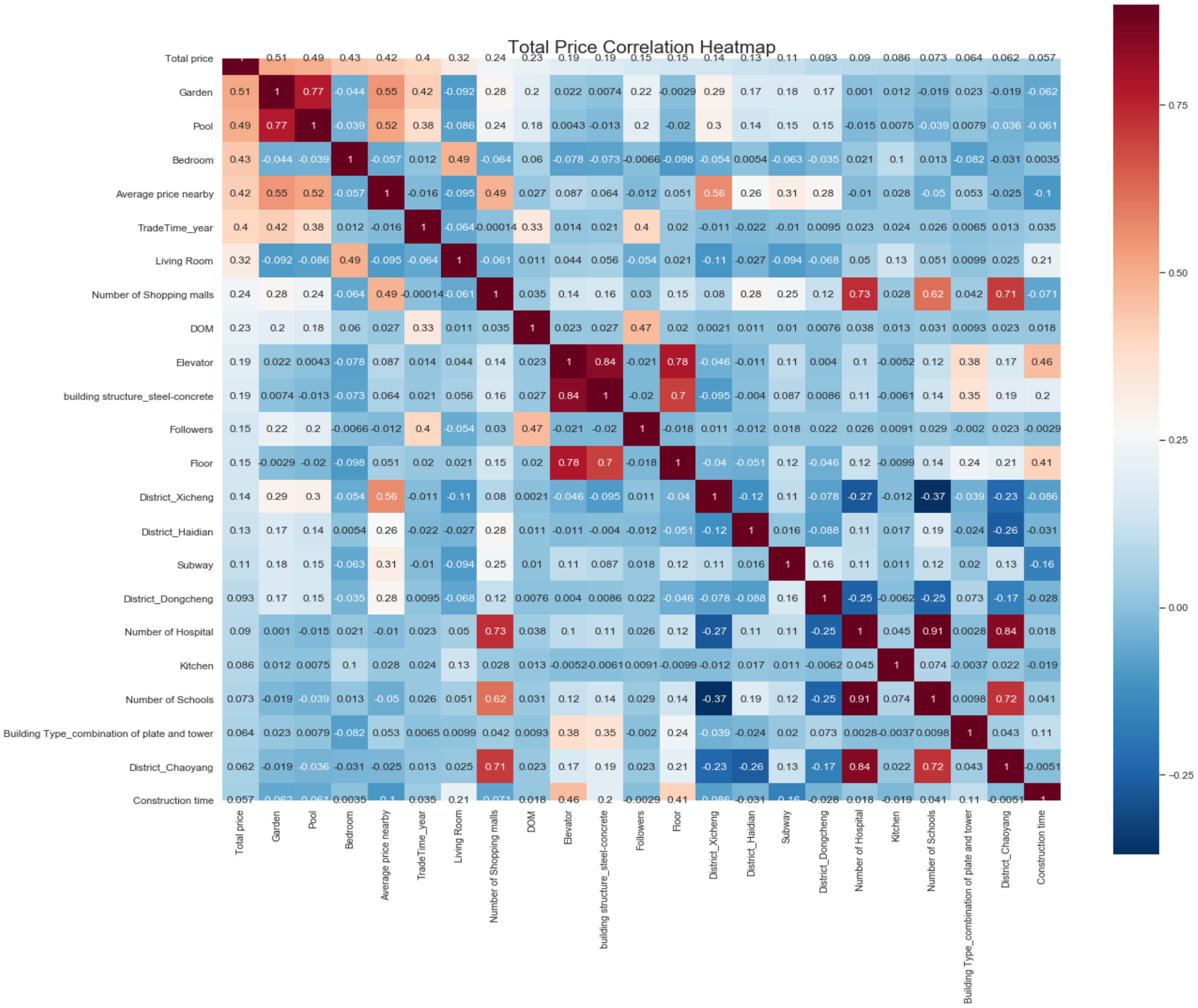
- Covert 'Living Room','Kitchen','Bathroom','Floor','Construction time','Average price nearby' to numeric
- Dummy variable for 'Floor Level','Building Type','Renovation Condition','Building Structure','District','Fire Facilities'
- Extract year, month, day from tradetime (1 variable split to 3 variables)

Exploratory data analysis

1.Analysing 'Total price' with descriptive statistics summary

```
count     318847.000000  mean        3490.319943  std          2307.815378
min            1.000000  25%         2050.000000  50%          2940.000000
75%         4255.000000  max       181300.000000
```

2.Correlation:



According to our Total price correlation heatmap, there are some variables most correlated with 'Total price'. Our thoughts on this:

- top 3 correlated: 'Garden', 'Pool' and 'Bedrooms'
- 'Garden' and 'Pool' strongly correlated with each other about 0.77, represent high-end proprities and communities
- 'Elevator' and 'building strcture_steel_concrete' strongly correlated with each other: 0.84, why?
- 'Elevator' and 'Floor' strongly correlated with each other: 0.78, the higher floor you live, the more you care about the elevator
- 'Floor' and 'building strcture_steel_concrete' strongly correlated with each other: 0.78, obviously because floor is highly correlated with elevator
- 'Number of shopping malls' and 'numebr of hosptials' stongly correlated with each other: 0.73, represent high resident population district
- 'Number of hospital' and 'Number of schools' strongly correlated with each other: 0.91, represent odler district in Beijing, where older people live with historic schools
- Interesting fact: We can tell that district_Chaoyang has the highest correlation with schools,

hospitals and shopping malls. But many famous schools and hospitals are in district_Haidian. This happens because schools in this dataset only count primary and middle schools (include international schools) whereas famous schools in Haidian are universities.

Data cleaning

1. Remove original variables (# Avoiding the Dummy variable Trap)

2. Remove all unnecessary variables:

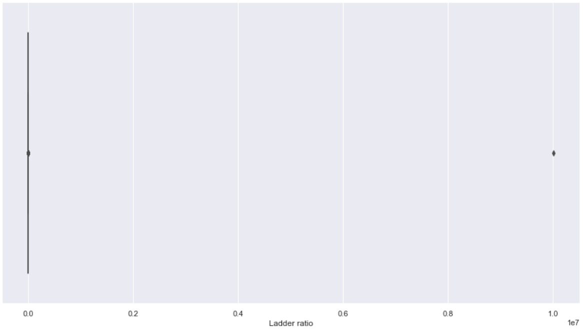
- URL: neither numeric variable nor useful for modeling
- ID: it's a unique value for each transaction, need to wipe out
- CommunityID: unique identify each Community ID
- TradeTime: Timestamp data
- Housing Ownership: the values in this column are all 70
- Price per square & Square: Price per square times square equals to total price, which we will leave our dependent variable
- Floor Level_Low & Floor Level_Medium: although they are dummy variables already, but because they are highly correlated with 'Floor', so better decision is to leave most likely the original data
- TradeTime year, Month & Day: extract year, month and day from tradetime

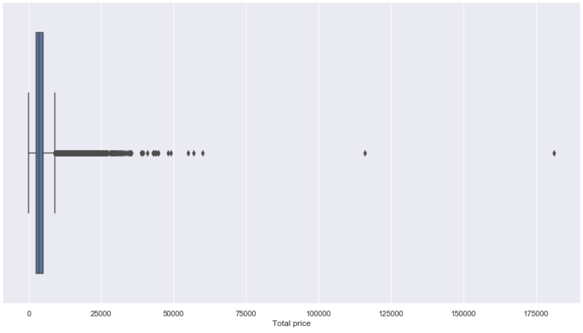
3. Remove missing value: floor, construction time, avg price nearby (only 32 rows of data in 310k)

4. For predictor 'DOM', we do consider it as an important predictor. It has 0.23 correlation coefficient with 'Total price'. But at this circumstance, we will wipe out all missing values for 'DOM', reasons are list below:

- It contains nearly 50% of missing values
- No matter how we are filling the missing values, like median or mean value. It will eventually lead us to wrong pattern
- We tried regression model without 'DOM', the accuracy is not expected

<u>Outliers</u>





After reading all columns in the file, we find out there are two columns contain outliers which are ladder ratio and total price. We tried to keep those outliers and fit them into our model below, but after we check the result that we realized that those outliers definitely effect the accuracy. If we want to keep those outliers, we can also cluster our data into different groups, but they are too small to form a group. As a result, we use 99% quantile to drop those outliers.

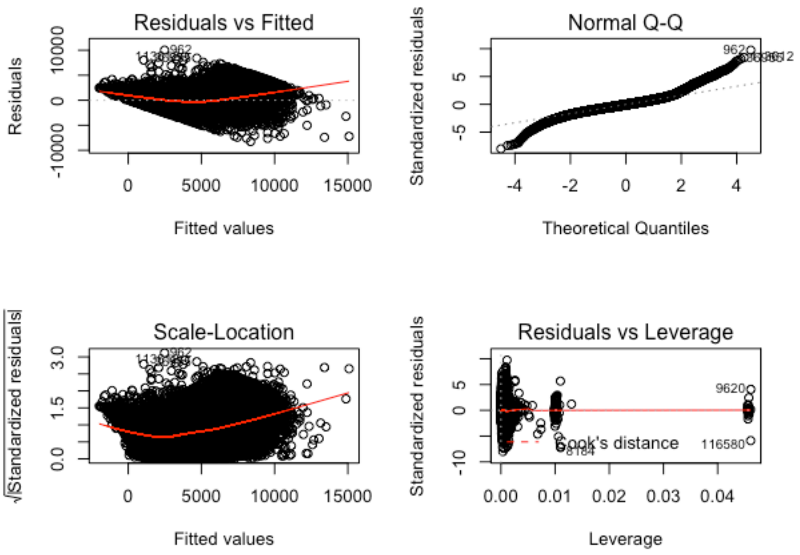
Shuffle and Split Data

For the next implementation it is required to take the Beijing housing dataset and split the data into training and testing subsets. Typically, the data is also shuffled into a random order when creating the training and testing subsets to remove any bias in the ordering of the dataset. We randomly split data into three groups which are 90%training set, 5%validate set and 5% testing set.

Models

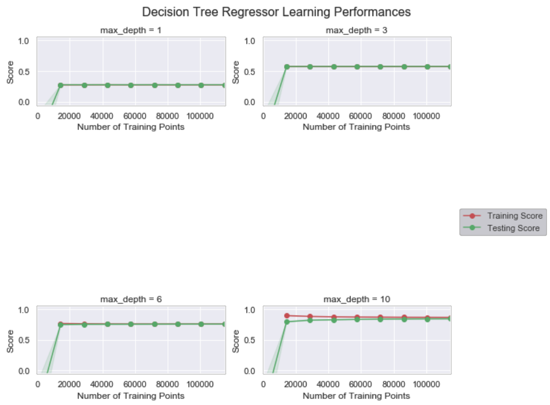
1. Multiple linear regression
   1.1 All variables included (baseline)
   We fit all data into MLR model with all variables, the R square for this model is 0.7685.
   1.2 Backward Elimination
   We use backward selection to fit our data into MLR model, three variables dropped such as building.structure_brick.and.wood, District_Tongzhou and District_Xicheng. The R square for this model is 0.7685.

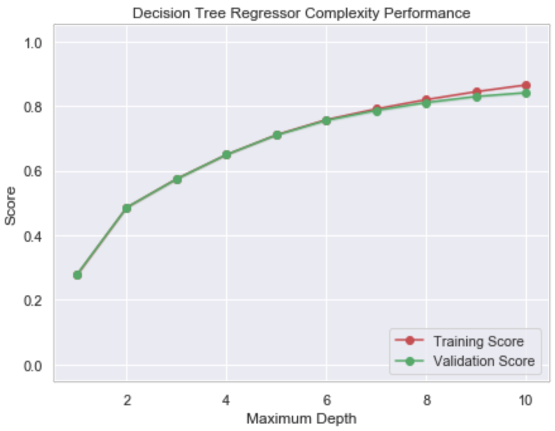
Model evaluation:



By summaries observation of both of our MLR models. We found all variables are significant (extremely small p-value for each independent t-test with dependent variable) in the model, but the R square is not as high as expected. As a result, we check if both models meet the assumptions of linear regression model. According to the graphs above, residuals vs fitted and scale-location, there appears to be no explicit pattern to the residuals, but the normal QQ plot shows that total price does not follow normal distribution, it violates the first assumption of MLR.

2. Decision Tree Regression





We choose features from backward selection model into our decision tree regression. The R square for this model is 0.87.

According to learning curves for decision tree regression, it seems the more complexity the model it, the better learning performance we got. From the 8000 points thresholds, having more training data points will not benefit the model. The training and testing score when maximum depth equals to 10 still contain unbalance between variance and bias, in this case increasing the max_depth hyperparameter to get better results. The complexity curve also determines the same performance with learning curve. We will conduct sensitivity check later to determine whether the model is either too complex or too simple to sufficiently generalize to new data.

Model evaluation:

By conducting shuffle split of training and testing dataset with test size of 20%, 10-fold cross validation and grid search. We generalize 13 as optimal max_depth, and default value of 1 for minimum sample leaf and default value of 2 for minimum sample split. We check whether our model is robust enough by sensitivity check with specific client data to predict 10 times using decision tree regressor, the result is pretty good with range of 756K through 10 trails. But we will conduct an ensemble learning model to improve performance.

3. Random Forest Regressor

We choose features right from previous decision tree model into our random forest regression. The R square for this model is 0.91. we did not conduct the learning curves and complexity curve base on time consuming, they are likely the same pattern with decision tree regressions, however, the only difference would be it needs more max_depth since size of dataset is large and random forest is complex than decision tree model. We will compare the prediction result later.

Model evaluation:
Using the same technique as decision tree regression with range of 1 to 20 for the max_depth, just in case the ensemble learning will be more complex than single decision tree algorithm. We generalize 19 as optimal max_depth, and default value of 100 for number of estimators. We didn't put sensitivity check base on time consuming, but we predicted total price for the first five properties and compare them with validate dataset. The predicted price is closer than decision trees.

Final recommendation

After checking R square of all three models, it shows that random forest regression has the best accuracy 0.91 which is a very decent result. In other hand, we think our model still have some room to improve. For example, in our dataset we use district to show the location area of each property. Instead of using district name, zip code is much better choice if we could have it in our dataset. Even two properties in same district but with different zip code they may have great price difference of each square feet. In additional, we use 3 types of feature to fit in our model such as house profile feature (like square, bedrooms, building type), residential community feature (school, shopping mall, hospital, Garden), geographical feature (District). In order to acquire better accuracy, model should contain more features such as temporal feature (DOM of recently sold properties, percentage of recently sold properties, recently crime rate). Those features could help model to improve accuracy for sure.

For DOM, we fit data into same models as house pricing, but unfortunately accuracy for all that three models are pretty low which is around 0.01. We search on google and we find DOM in Beijing is greatly affected by government policy. According to graph above, it shows that DOM sharply increased from 1.35 to 35.06 on 2015. For more information regarding DOM, please see Jupiter notebook.

References

Hengshu Zhu, Hui Xiong, Fangshuang Tang, Qi Liu, Yong Ge, Enhong Chen, Yanjie Fu (2016). Days on Market: Measuring Liquidity in Real Estate Markets. Retrieved Dec 2nd, 2019, from https://www.kdd.org/kdd2016/papers/files/adf0605-zhuA.pdf

Knoema. World Data Atlas: China topics. Retrieved Dec 2nd, 2019, from https://knoema.com/atlas/China/topics/Crime-Statistics

Knoema. World Data Atlas: China topics. Retrieved Dec 2nd, 2019, from https://knoema.com/atlas/China/topics/Education

Appendix

Variable	Description
URL	the url which fetches the data
ID	the id of transaction
Lng and Lat	coordinates, using the BD09 protocol.
CommunityID	community id
TradeTime	the time of transaction
DOM	active days on market.
Followers	the number of people follow the transaction.
Total price	the total price
Price per sqaure	the average price by square
Square	the square of house
Bedroom	the number of Bedroom
Living Room	the number of living room
Kitchen	the number of kitchens
Bathroom	the number of bathrooms
Floor	the height of the house
Floor Level	High, Medium and Low
Building Type	including tower (1), bungalow(2), combination of plate and tower(3), plate(4)
Construction time	the time of construction
renovation condition	including other (1), rough(2), Simplicity(3), hardcover(4)
building structure	including unknow (1), mixed(2), brick and wood(3), brick and concrete(4),steel(5) and steel-concrete composite(6)
Property rights for five years	if the owner has the property for less than 5 years
Ladder ratio	the proportion between number of residents on the same floor and number of elevators of ladder
Subway	have subway (1) near house or not (0)
District	13 different district area in city of Beijing
Average price nearby	average housing price nearby each specific priority
Fire Facilities	High, Medium or Low
Housing ownership	regrading to Chinese housing policy, each property has ownership of 70 years
Number of Schools	number of primary and middle schools in each district
Number of Hospital	number of hospitals in each district
Garden	have garden in community (1) or not (0)
Pool	have pool in community (1) or not (0)