Group project for DSME6650BA

Rental Price Prediction and Analysis of Amsterdam Airbnb

Group id: 16

LIANG Jiawen (sid: 1155164164)

WANG Xuanru (sid: 1155164156)

LIAO Zhengyuan (sid: 1155167386)

DAI Guan (sid: 1155162714)

March 9, 2022

Contents

1. Problem Description

Amsterdam is the capital and most popular city of the Netherlands. It is colloquially referred to as the "Venice of the North", attributed by the large number of canals which form a UNESCO World Heritage Site. Amsterdam is the heaven of art because of its high-density distribution of museums and art galleries. It has large amount of collections of Vincent Willem van Gogh and Rembrandt Harmenszoon van Rijn. Amsterdam is also famous of its open culture to sex and cannabis.

Airbnb is an online marketplace for arranging or offering lodging, primarily homestays, or tourism experiences. More and more people choose to stay in a local house when they are travelling. The prices of the houses vary a lot depending on the location, the size, the service or the surroundings of the houses.

The data set we use for this project is downloaded from Kaggle with 6 csv files：(https://www.kaggle.com/erikbruin/airbnb-amsterdam). The '*listings*' file contains all the advertisements in Amsterdam on December 6th, 2018. Based on this dataset, our group want to estimate the price and sales of the apartment listed on Airbnb Platform. Firstly, we need to do some data wrangling work before analysis.

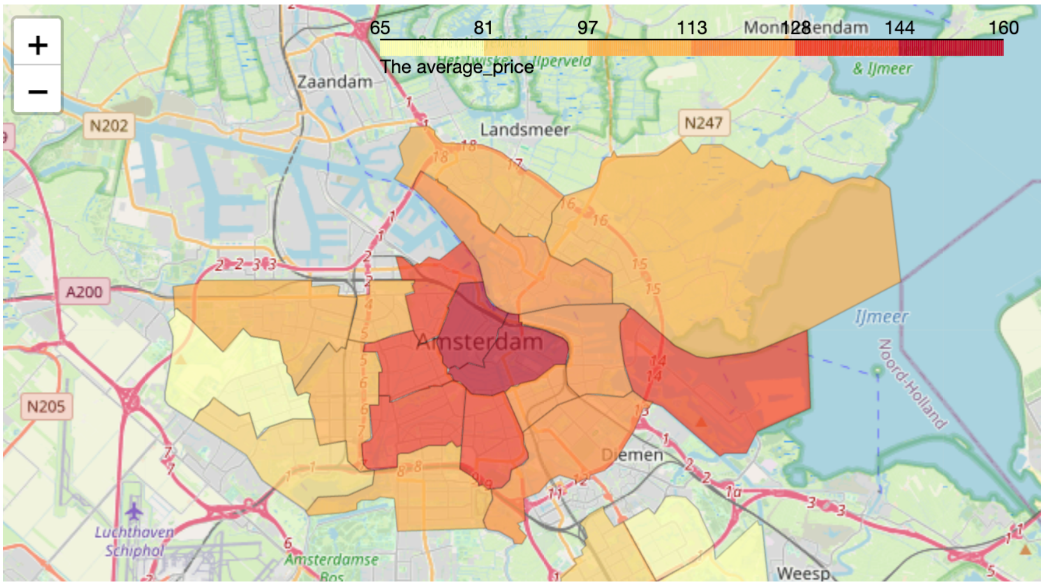


Fig 1.1 Average rental price distribution

2. Data Preprocessing

2.1 Data Wrangling:

Firstly, we drop some features such as ID and host name that contains no useful information for further analysis. Then we drop NaN values in some columns. We also noticed that there exist some outliers, here we apply the "*3-sigma rule*": Drop the data which is more than three standard deviations from the average.

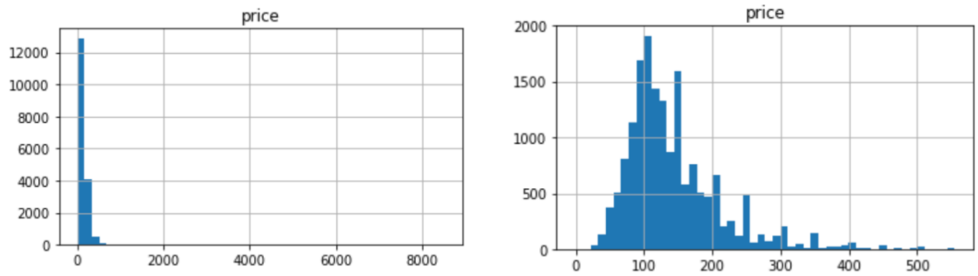


Fig 2.1 Price outliers dropping

2.2 Feature Engineering:

In this part, we find it necessary to cluster the geographical information of the apartments since we believe that the apartments in the same area should price at similar level.

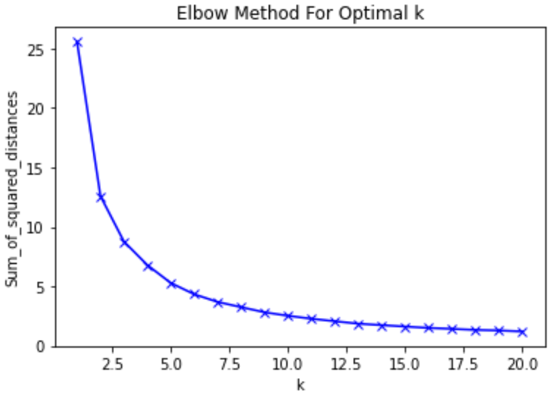


Fig 2.2 Best k choosing of cluster

Applying the Elbow Method, we decide to cluster the geographical information (longitude, latitude) into five groups with the *K-means* algorithm and use dummy variables to indicate whether the apartment is in the area.

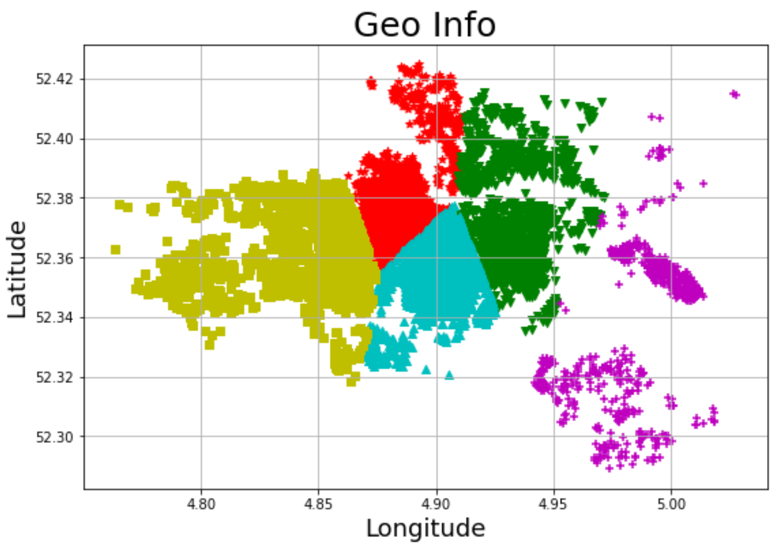


Fig 2.3 K-means results by price

Also, we transfer some categorical variables such as “*room_type*” into dummy variables so that we can do the further analysis.

2.3 Natural Language Processing

Intuitively, when the customer wants to pick an apartment, the first thing he or she would do is to check the advertisement context and absorb useful information from it. It’s reasonable to suspect that the context of the advertisement contributes a large amount to the pricing and sales. We then separate the context into words and count that how many times they appeared in our dataset. The data shows that the Top-10 frequent words used is: **‘Apartment’, ‘Amsterdam’, ‘Spacious’, ‘City’, ‘Centre’, ‘Room’, ‘Near’, ‘House’, ‘Center’, ‘Garden’**. It seems like that the householders believes that these words are the most useful to promote their apartments, however, whether these words really have positive impacts on the price and sales still needs further analysis.

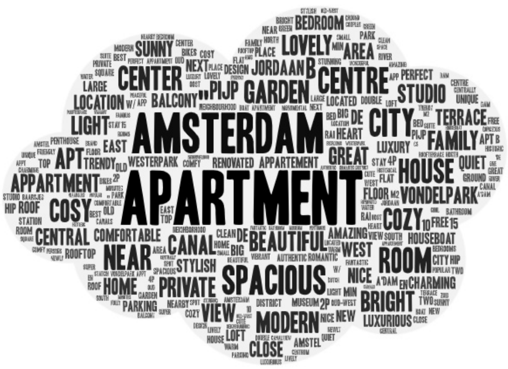


Fig 2.4 The Word Cloud of Name

3. Filter Feature

We choose and pre-process features just based on interpretation, so in this part, we would like to do OLS for each feature on ‘Price’, to filter features that have strong relation with Y.

The result of OLS is showed at the table following.

x	coef	p_value	balance or not
clust	-4.727579	0.0	False
price	1.000000	0.0	False
last_review_year_2011	115.816287	0.1	True
last_review_year_2012	25.815550	0.5	True
last_review_year_2013	21.221947	0.4	True
last_review_year_2014	-19.725799	0.1	True
last_review_year_2015	-14.379171	0.0	False
last_review_year_2016	-20.302879	0.0	False
last_review_year_2017	-11.508786	0.0	False
last_review_year_2018	16.947819	0.0	False
last_review_days	-0.022402	0.0	False
room_type_Entire home/apt	59.445732	0.0	False
room_type_Private room	-59.032636	0.0	False
room_type_Shared room	-56.554948	0.0	False
near	-9.994819	0.0	False
center	1.239406	0.6	True
city	5.944298	0.0	False
centre	5.931994	0.0	False
apartment	-4.930046	0.0	False
amsterdam	3.128865	0.8	True
room	-39.636435	0.0	False
house	40.661466	0.0	False
spacious	9.830076	0.0	False
garden	11.287386	0.0	False
minimum_nights	0.559874	0.0	False
number_of_reviews	-0.181568	0.0	False
reviews_per_month	-5.519666	0.0	False
neighbourhood_Bijlmer-Centrum	-46.753238	0.0	False
neighbourhood_Bijlmer-Oost	-53.891235	0.0	False
neighbourhood_Bos en Lommer	-31.428314	0.0	False
neighbourhood_Buitenveldert - Zuidas	-12.694571	0.0	False
neighbourhood_Centrum-Oost	32.753519	0.0	False
neighbourhood_Centrum-West	36.405758	0.0	False
neighbourhood_De Aker - Nieuw Sloten	-23.187862	0.0	False

neighbourhood_De Baarsjes - Oud-West	-4.185461	0.0	False
neighbourhood_De Pijp - Rivierenbuurt	5.837793	0.0	False
neighbourhood_Gaasperdam - Driemond	-55.169955	0.0	False
neighbourhood_Geuzenveld - Slotermeer	-33.570938	0.0	False
neighbourhood_IJburg - Zeeburgereiland	10.007314	0.0	False
neighbourhood_Noord-Oost	-28.222656	0.0	False
neighbourhood_Noord-West	-28.445396	0.0	False
neighbourhood_Oostelijk Havengebied - Indische...	-19.332551	0.0	False
neighbourhood_Osdorp	-49.864653	0.0	False
neighbourhood_Oud-Noord	-6.789925	0.0	False
neighbourhood_Oud-Oost	-8.491761	0.0	False
neighbourhood_Slotervaart	-30.223743	0.0	False
neighbourhood_Watergraafsmeer	-8.896290	0.0	False
neighbourhood_Westerpark	-6.790882	0.0	False
neighbourhood_Zuid	12.734733	0.0	False

We presume that in the OLS, the feature whose p-value is larger than 0.5, which means Y(‘price’) is not balance based on it, should be chosen as an feature. From the table above, we got the final feature list:
['clust', 'price', 'last_review_year_2015', 'last_review_year_2016', 'last_review_year_2017', 'last_review_year_2018', 'last_review_days', 'room_type_Entire home/apt', 'room_type_Private room', 'room_type_Shared room', 'near', 'city', 'centre', 'apartment', 'room', 'house', 'spacious', 'garden','minimum_nights', 'number_of_reviews', 'reviews_per_month', 'neighbourhood_Bijlmer-Centrum','neighbourhood_Bijlmer-Oost', 'neighbourhood_Bos en Lommer', 'neighbourhood_Buitenveldert - Zuidas', 'neighbourhood_Centrum-Oost', 'neighbourhood_Centrum-West', 'neighbourhood_De Aker - Nieuw Sloten', 'neighbourhood_De Baarsjes - Oud-West', 'neighbourhood_De Pijp - Rivierenbuurt', 'neighbourhood_Gaasperdam - Driemond', 'neighbourhood_Geuzenveld - Slotermeer', 'neighbourhood_IJburg - Zeeburgereiland', 'neighbourhood_Noord-Oost', 'neighbourhood_Noord-West', 'neighbourhood_Oostelijk Havengebied - Indische Buurt', 'neighbourhood_Osdorp', 'neighbourhood_Oud-Noord', 'neighbourhood_Oud-Oost', 'neighbourhood_Slotervaart', 'neighbourhood_Watergraafsmeer', 'neighbourhood_Westerpark', 'neighbourhood_Zuid']

4. Regression Modelling

Given the information of the property, we would like to solve the problem: which factors of the property could play important roles in the rental price on Airbnb in Amsterdam. So, in this section, we would fit our train set data by conducting multiple regression models and compare their performance on the price prediction. Besides, we also want to use the coefficients of the linear models and the feature importance of non-linear models to find a few factors affecting the rental price.

To evaluate the performance of models, we firstly separate the data into training set and test set using function *'train_test_split'*. The training set will be used to select and train the model, and the test set will be used to test the performance of the model. Moreover, we would like to use the mean absolute error to evaluate the model. The loss function is as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_{predict}(i) - y_{true}(i)\right|$$

There are two main reasons that we choose MAE as evaluation model:

- MAE is not sensitive to outliers since it is not of quadratic nature like MSE.
- MAE is an intuitive measure since it basically tells us about the average size of forecasting errors when ignoring negative signs.
- MAE is a good substitute for MSE when determining optimal inventory levels[1].

4.1 Model Fitting

In order to ensure the completeness of the prediction, we decide to fit the training set by using 4 linear models and 4 non-linear models as follows:

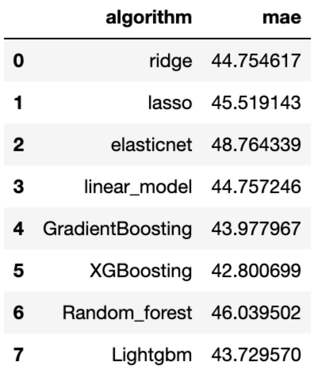
- **linear model:** Linear regression, Ridge regression, Elastic regression, Lasso regression;
- **non-linear model:** Gradient Boosting tree regression, XGBoosting tree regression, Random Forest regression, Lightgbm regression.

For linear model, the most popular model is the linear regression. However, ordinary linear regression usually suffer from insufficient prediction accuracy, which increases the complexity of the model if there are correlations among the features in the model. The emergence of Ridge regression and Lasso regression is to solve the overfitting of linear regression. And the Elastic regression is the combination of Lasso and Ridge.

For non-linear regression, Gradient Boosting tree regression, XGBoosting tree regression, Random Forest regression and Lightgbm regression are all the evolutions of decision-tree model but have different evolution direction for different goals.

[1] Brown, R.G., (1962) *Smoothing, Forecasting and Prediction*, Prentice-Hall, Englewood Cliffs, N.J.

Now we conduct all the model to fit the training set and evaluate them on the testing set by mean absolute error:



	algorithm	mae
0	ridge	44.754617
1	lasso	45.519143
2	elasticnet	48.764339
3	linear_model	44.757246
4	GradientBoosting	43.977967
5	XGBoosting	42.800699
6	Random_forest	46.039502
7	Lightgbm	43.729570

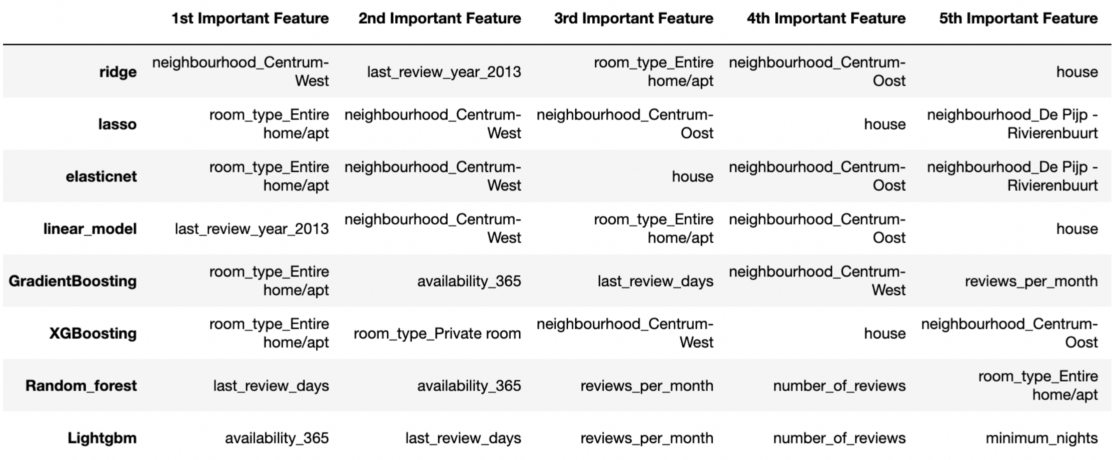
Fig 3.1 Regression model comparison by MAE

From the mae chart, we find that nearly all the models can decrease the mean absolute error to around 40. To the dataset with the mean price of 144, we think the prediction can get the rough trend of rental price and at least can find the impart factors in this problem.

4.2 Result Analysis

In the previous section, we have completed the price prediction based on 8 regression models that can get good results. In addition to that, we need to build importance matrix for these regression results, so that we can analyze to get the important factors that affect the rental price. For linear model, we can simply use the standardized coefficients of the fitting result to evaluate the importance. But for non-linear model, we need to get that through function '*feature_ importances_*'.

Therefore, we conduct the importance function operation, sort them and finally choose the first 5th important features to analyze the results as follows:



	1st Important Feature	2nd Important Feature	3rd Important Feature	4th Important Feature	5th Important Feature
ridge	neighbourhood_Centrum-West	last_review_year_2013	room_type_Entire home/apt	neighbourhood_Centrum-Oost	house
lasso	room_type_Entire home/apt	neighbourhood_Centrum-West	neighbourhood_Centrum-Oost	house	neighbourhood_De Pijp - Rivierenbuurt
elasticnet	room_type_Entire home/apt	neighbourhood_Centrum-West	house	neighbourhood_Centrum-Oost	neighbourhood_De Pijp - Rivierenbuurt
linear_model	last_review_year_2013	neighbourhood_Centrum-West	room_type_Entire home/apt	neighbourhood_Centrum-Oost	house
GradientBoosting	room_type_Entire home/apt	availability_365	last_review_days	neighbourhood_Centrum-West	reviews_per_month
XGBoosting	room_type_Entire home/apt	room_type_Private room	neighbourhood_Centrum-West	house	neighbourhood_Centrum-Oost
Random_forest	last_review_days	availability_365	reviews_per_month	number_of_reviews	room_type_Entire home/apt
Lightgbm	availability_365	last_review_days	reviews_per_month	number_of_reviews	minimum_nights

Fig 3.2 First 5th important features by different models

In the chart, we can see the different feature importance rankings from different regression models. However, some features appear for many times which means they can play important roles in the prediction model even though in different regression models. So, we choose 4 features that appears more frequently: **'room_type', 'neighbourhood_Centrum', 'review_per_month', 'house'**. The analysis is as follows:

- At first, there is no doubt that '**room_type**' would affect the rental price so much. If the room type is '**Entire home**', it will have a few rooms with larger area which leads to the higher price. If the room type is '**private room**' or '**shared room**', the price would be lower due to the area and privacy.
- On the other hand, reviews also affect the price. People tend to book the property that has **more reviews and higher reviews**. The number of reviews can describe the opening years and the popularity of the property and the review scores can reflect the condition and reliability of the property.
- Location is definitely another important factor to the price. The review scores for location can reflect the neighborhood the property locates and the how convenience the public transportation nearby. Property locating in the **centrum neighborhood** can be more expensive than the others. This can be verified by the previous visualization of the medians of the prices of different neighborhoods.
- The strange thing is that the trend shows that if Airbnb has **'house'** in its name, the price that can be rented becomes higher, maybe it has something to do with the local culture.

5. Conclusion

In this project, we analyzed the Airbnb houses in Amsterdam using the Airbnb data and the information of neighborhoods provided by Foursquare API.

In the stage of data preprocessing, we use the "3-sigma rule" to remove some housing information with abnormal prices at first. Next, in order to process the latitude and longitude information, we use the k-means algorithm to block regions with similar prices to generate regions close to various locations. Finally, we use natural language processing technology to analyze the house naming, and get the first few words with the most occurrences, so as to study the relationship between naming and price. After the initial processing of the features, we also conduct a causal analysis of these features and rental prices, looking for the features that really act as causes that affect the rental price.

In the modeling process, we used eight classic regression models (4 linear models and 4 non-linear models) to predict the price of the properties and analyzed the most influencing factors to the price of a property and compare the performance of different models. Based on the above analysis, several recommendations were proposed to people who want to rent their house on Airbnb and people who plan to live in a local place in Amsterdam. However, this analysis may not adapt to other city's Airbnb house price since different cities have their own characteristics, and thus, a different composition of price influencing factors.