Rental Price Analysis of Airbnb Amsterdam

Group id : 16
Pleader : DAI Guan

CONTENT

Project Introduction

Amsterdam Rental Price

Amsterdam is one of the most worthwhile cities to visit, and using Airbnb to rent has also become the most popular choice for tourists.

Based on Airbnb's official housing data in Amsterdam in 2018 (from kaggle), this project establishes a price prediction regression model to provide price recommendations for the latest housing and analyze the factors that most affect rental prices.

Project Introduction

Datasets Demostration

The datasets we use are: **calendar.csv, listing.csv, reviews.csv**

calendar.csv	listing.csv	review.csv
This dataset contains the availability situation in whole 365 days and the price on the available day of each housing resource.	This dataset contains the basic information about every housing resource, like: **house name, neighbourhood, coordinate, room type, etc.**	This dataset contains the information of the reviews' time which can be used to count the number of reviews in a period.

CONTENT

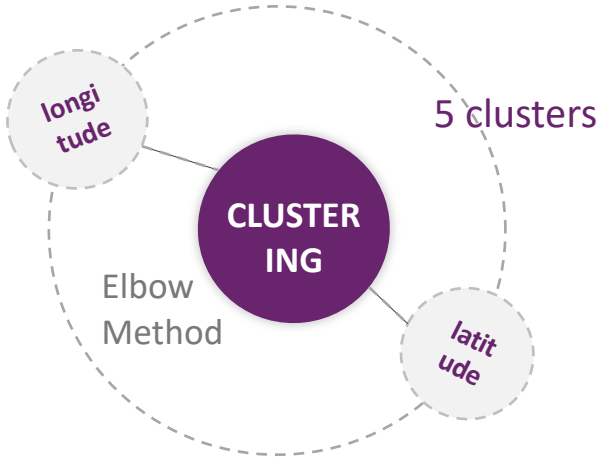
Data preprocessing -- Data Wrangling

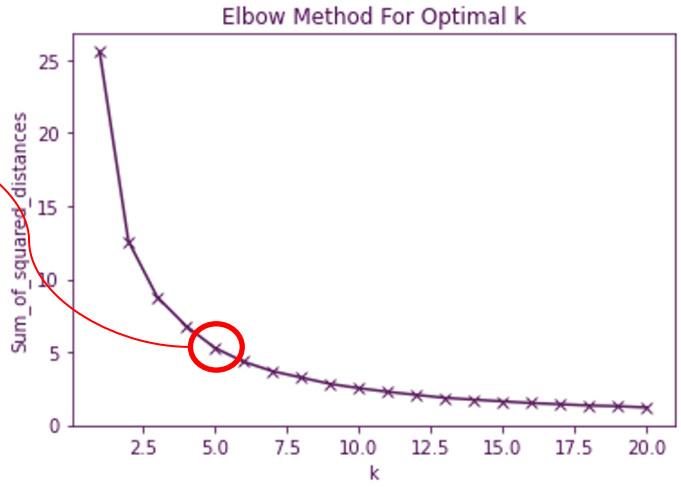
1. Drop columns used to identify (unique value)
2. Drop null data
3. Drop outlier based on “3-sigma rule”

Data Preprocessing -- Feature Engineering

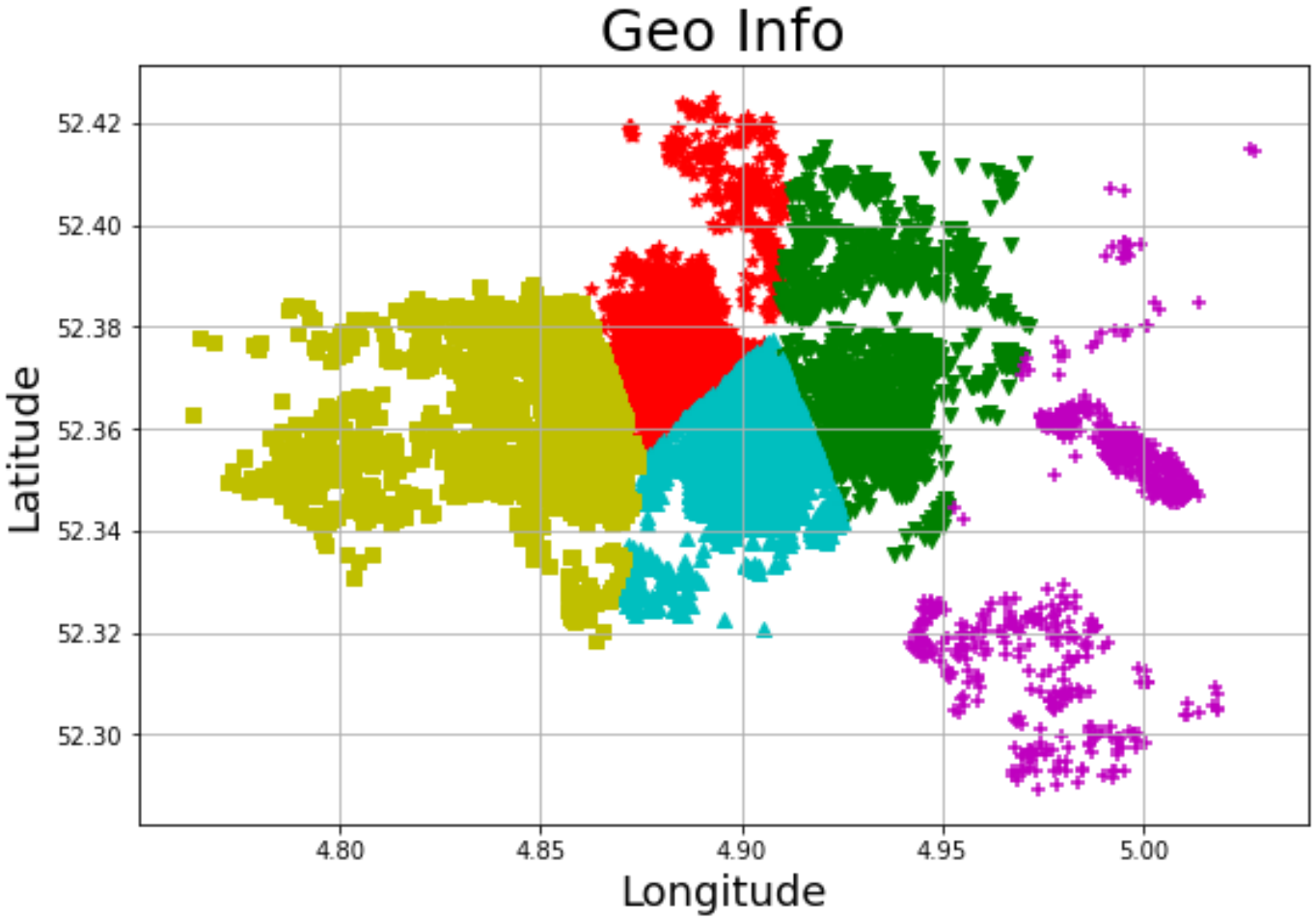
1. **Transfer categorical into dummy variables**

2. **Clustering**





Data Preprocessing -- Feature Engineering



Data Preprocessing -- Natural Language Processing

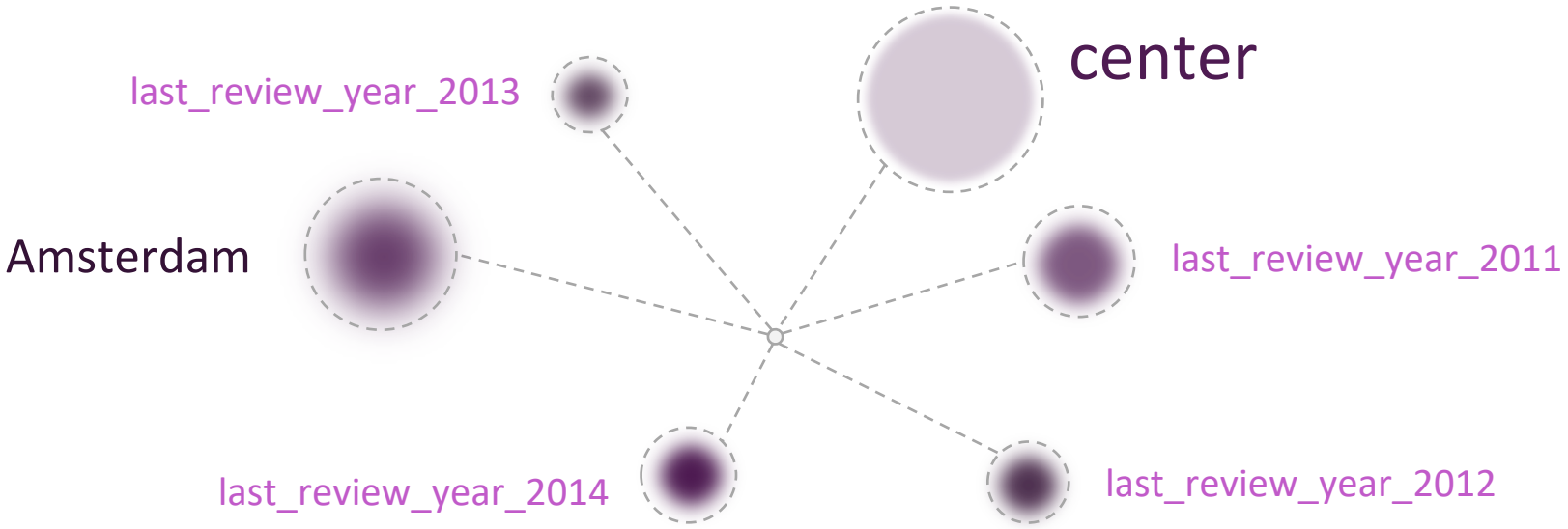
Top-10 frequent words :

'Apartment', 'Amsterdam', 'Spacious', 'City', 'Centre', 'Room', 'Near', 'House', 'Center', 'Garden'

CONTENT

Filter Feature

OLS for each feature on 'Price', to filter features that have strong relation with Y.



Regression Modelling

Model Comparison

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.

Regression Modelling: Linear Model

Linear Regression

A statistical analysis method that uses regression analysis in mathematical statistics to determine the interdependent quantitative relationship between two or more variables. It is widely used as a baseline model.

Ridge Regression

An improved least square estimation method that gives up the unbiased nature of the least square method but in return, obtains a more generalized and practical prediction result.

Lasso Regression

Constructing a penalty function to compress the coefficients of variables and lower some regression coefficients to 0. The calculation process is way less than Ridge Regression.

Elastic Regression

A combination of Ridge and Lasso.

Regression Modelling: Linear Model

Random Forest

A two-way randomization method based on decision tree. Random Forest has an excellent performance when dealing with high-dimension data.

Gradient Boosting

Flexibly handles various types of data, including continuous values and discrete values, without sacrificing the accuracy.

XGBT

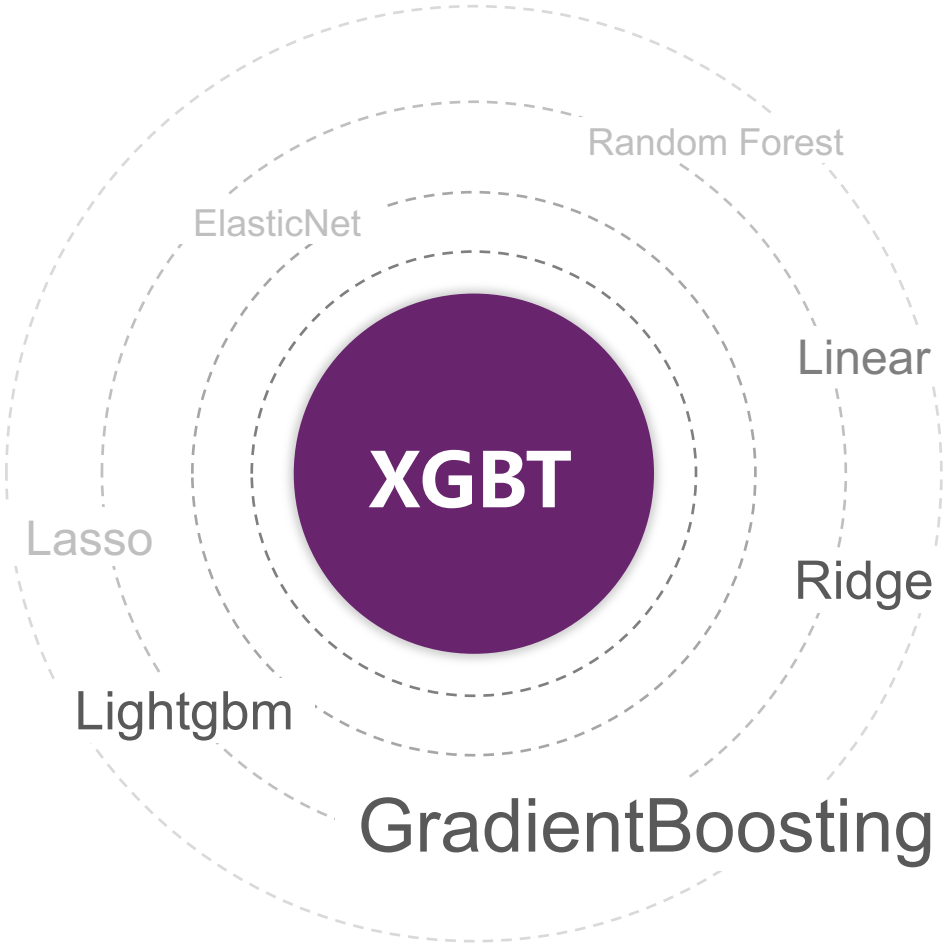
Having competitive prediction performance because of a great flexibility of fine-tuning many hyper-parameters.

Lightgbm

Faster training speed, lower memory usage and better accuracy than any other boosting algorithm.

CONTENT

Result Analysis



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:

Result Analysis

	1st Important Feature	2nd Important Feature	3rd Important Feature	4th Important Feature	5th Important Feature
ridge	neighbourhood_Centrum-West	room_type_Entire home/apt	last_review_year_2013	neighbourhood_Centrum-Oost	house
lasso	room_type_Entire home/apt	neighbourhood_Centrum-West	neighbourhood_Centrum-Oost	house	availability_365
elasticnet	room_type_Entire home/apt	neighbourhood_Centrum-West	house	neighbourhood_Centrum-Oost	neighbourhood_Zuid
linear_model	last_review_year_2011	last_review_year_2013	neighbourhood_Centrum-West	room_type_Entire home/apt	neighbourhood_Centrum-Oost
GradientBoosting	room_type_Entire home/apt	availability_365	neighbourhood_Centrum-West	last_review_days	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	last_review_days	availability_365	reviews_per_month	number_of_reviews	minimum_nights

Result Analysis

In the chart, the feature importance rankings are different depending on 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, the 4 features with highest frequency are: '**room_type**', '**neighbourhood_Centrum**', '**review_per_month**', '**house**'. The analysis is as follows:

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, probably due to the area and privacy.

4. Result Analysis

On the other hand, **reviews** also affect the price. People tend to book the property that has more reviews and reviews with higher rating. 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.

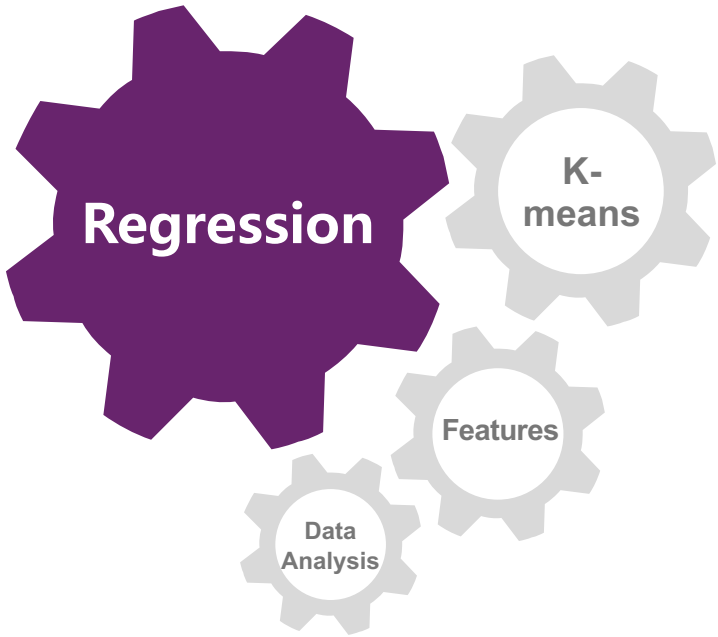
Location is another important factor to the price. The review score for location reflects the neighborhood the property locates and the convenience of the public transportation nearby. Property locating in the **centrum neighborhood** can be more expensive than the others. This can be verified by the 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.

CONTENT

Conclusion

In this project, we preprocessed a large amount of Airbnb housing data, analyzed causality, and established an Airbnb housing pricing model using 8 regression models, and established a factor contribution matrix according to linear and nonlinear rules, and finally analyzed the The four major factors that most affect the rental price of houses, and give a reasonable explanation.



Group id : 16

Group member:

1. LIANG Jiawen (sid: 1155164164)
2. WANG Xuanru (sid: 1155164156)
3. LIAO Zhengyuan (sid: 1155167386)
4. DAI Guan (sid: 1155162714)

Thanks

For Your Listening