CSCI-620 Data Mining with the Airbnb Dataset

[Exploring and Mining the dataset of NY Airbnbs]

Aishwarya Rao
ar2711@rit.edu

Apurav Khare
ak2816@rit.edu

Martin Qian
jq3513@rit.edu

Prateek Kalasannavar
pk6685@rit.edu

ABSTRACT

Nowadays, data mining is more and more involved in every-day life. This project aims to do data mining on a data set, so that we have to understand a dataset, then give meaning to every value in the data set, finally with algorithms to find out valuable imformations. This project also hope to build a visualization of the data set as final presentation.

1. INTRODUCTION

2. THE NEW YORK CITY AIRBNB DATASET

This is the dataset that is used in our project. It is a dataset of about 50k real-world records of Airbnbs in New York City along with 10k different hosts. It is a public and open-source dataset, which can be found here:Kaggle.

It contains all needed information to find out more about hosts, geographical availability, price, and rating in order to make predictions and draw conclusions.

For an example of data, is shown below.

2.1 Why we choose this dataset

First of all, this dataset is real-life, every attribute has it's own meaning which make it more practical. Secondly, it is not a huge dataset, which means it can be handled more quickly yet still challenging. Last but not least, it is an open-source dataset having a quite good organization, which is a great example to explore.

2.2 What we plan to use this dataset

2.2.1 Prediction

One important usage of data mining is to do predictions, especially for commercial datasets. For our dataset, the prediction means to give possible prices for a new Airbnb host in New York City.

2.2.2 Evaluation

There is another attribute called rating which is also a good measure for data mining. We can do data mining based on the factors that will influent the rating by doing correlation analysis.

2.2.3 Integration

We all know that apart from the data itself, there are many other outside factors that can have impact on the data. What we want to do here is to itegrate the data with some other factors like the map information of the New York City, safety level of the neighborhood and so on. In the end, we will try to do visualization of these data to give a more readable and direct result to everyone.

2.3 How

Several different data mining algorithms that we might use are like: Decision Tree and Forest, SVM and K-Means. We plan to do classification and evaluation by R and python. It depends on futher implementation.

3. FUTHER PLANNING

Week 9: Preparation and dataset selection

Week 10-11: Data features anlysis and algorithm try out

Week 12-13: Run algorithm and make improvements on precision and accuracy.

Week 13-14: Final test and do visualization, integration and so on. Making result more meaningful and reasonable.