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

Data mining and its scope has grown tremendously over the last decade and is now an integral part of our lives. This project aims to perform a data mining task to explore the field as well as understand identify the scope of prediction in the selected dataset. The phases are planned accordingly to improve the measure with each step.

1. INTRODUCTION

1.1 The New York City Airbnb Dataset

The New York City Airbnb dataset is a dataset available on Kaggle which consists of data of hosts, locations and prices for Airbnb in New York in 2019. The dataset consists of attributes such as name, location in latitude and longitude, neighborhood, room type and price. The dataset consists of around 5000 rows.

1.2 Choosing the dataset

There are a multitude of reasons why this dataset was appealing. They are as follows,

- The dataset is based on real data which makes the data mining task more practical - both in terms of use case scenarios as well as complexity of the data.
- With around 5000 rows, the dataset size is optimal for a data mining task. It is not too large to create needless challenges but not too small to prevent generalization.
- The dataset is open source and is updated annually.

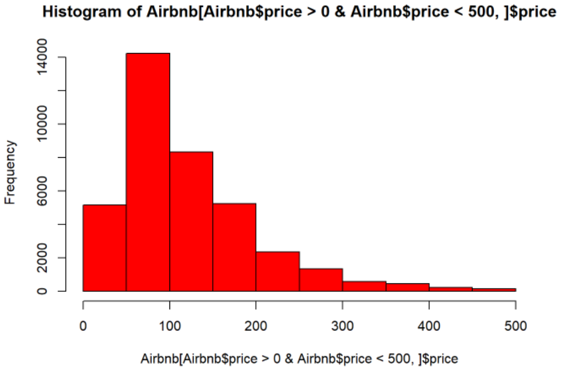
1.3 The Data Mining Task

1.3.1 Prediction

One important usage of data mining is to perform predictions, especially for commercial datasets. For our dataset, we choose to use a data mining algorithm to predict the price range for a new Airbnb host in New York City. That would allow new hosts to estimate the price for their listings for maximum interaction.

1.3.2 Evaluation

With this idea of predicting the price category in mind, we explored the dataset to confirm with our assumption that this task is actually possible through the data present. As seen in the figure, the price column is distributed reasonably well. It is possible to map these into bins for further predictions. Furthermore, through exploration of the dataset, we found correlation between price and multiple other attributes in the dataset. With this information, we decided to go forth with predicting the price in the dataset.



1.3.3 Integration

We all know that apart from the data itself, there are many other outside factors that can have impact on the data. Outside of the information in the dataset, we are considering integrating the data with some other factors like the map information of the New York City, safety level of the neighborhood and so on. This task will involve finding datasets that can be with our current dataset. Through this integration, we hope to be able to present our findings from the prediction in a more user-friendly and understandable manner.

1.4 Proposed method

Several different data mining algorithms exists for classification including decision trees, SVM and K-Means. We plan to perform classification and evaluation on the price by splitting it into bins. Evaluation will be measured by accuracy. The project will be coded in R.

2. FUTURE PLAN

Week 9: Preparation and dataset selection

Week 10-11: Data features analysis, feature engineering, shortlist classification algorithms

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

Week 13-14: Visualizations, inferences based on our final model.