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

The project aims to create a prediction model on the New York Airbnb Dataset that is capable of predicting which price group a particular house falls into. The approach we use is to build a classification model (such as decision tree) to predict a discrete version of the price (For instance, expensive vs cheap). One of the deliverable this project will include is determining what factors affect the price.

1. CURRENT PROGRESS

1.1 The Data Mining task

As mentioned in the previous phase, the data mining task selected was to build a classification model that is capable of predicting the price group a listing would fall into. This decision is based on a number of reasons. One, a model capable of predicting the price of a new listing has real-world applications. A landlord looking to put up a listing for a new house in New York would be able to use this model to determine the expected range of prices for his house and price it accordingly to make the most profit. Two, the exploration of the dataset revealed that there are a number of attributes in the data that correlate to the price. This indicates that these attributes do help determine the price and validates our hypothesis that the price can be predicted reasonably well by a data mining algorithm. Figure 1 depicts the correlation of some of the attributes with the price, which will help choose the right features for the model.

1.2 Categorizing Price

Price the New York Airbnb Dataset is a continuous attribute with the following details -
Min. : 0.0
1st Qu.: 69.0
Median : 106.0
Mean : 152.7
3rd Qu.: 175.0
Max. :10000.0

To perform a classification task on the price, this attribute has to be made categorical. We used quartiles to create four different classes ranging from cheap to expensive. The quartiles were as follows, Category 1 : 0 - 25% (0- 69)
Category 2 : 26 - 50% (70 -106)
Category 3 : 50 - 75% (107- 175)
Category 4 : 75 - 100% (176- 10000)
Figure 2 shows the distribution of the data after the categorical split.

1.3 Cleaning the dataset

1.4 Feature Engineering

1.5 Train and test split

1.6 Basic model

2. FUTURE WORK