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.

1.2 Targeted Knowledge

This data mining task is based on a number of factors. 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.

1.3 Exploration and visualization

Correlation plots were plotted on attributes against the price to see how they impact it. As a starting point, the attributes that have business meaning in the context of the data are used to plot the correlation plots. These attributes are: "neighbourhood", "neighbourhood_group", "room_type", "minimum_nights", and "availability_365".
Following are the notable observations:

1. The price of an Airbnb listing varies by the neighbourhood group that it is in. As observed from the correlation plot, the prices of the Airbnb listings in Brooklyn and Manhattan are relatively higher than those in Broknx, Queens, and Staten Island, with the listings in Staten Island being the least priced.

2. The room type of the Airbnb listing affects its price. There are three room types in the dataset: Entire home/apartment, Private room, and Shared room. The correlation plot clearly shows that Entire home/apartments are priced higher than private rooms, and Shared rooms are the least expensive of them all.

3. The neighbourhood impacts the price of the Airbnb listing as well. From the correlation plot, it is observed that listings in neighbourhoods like "Upper West Side", "Upper East Side", and "Easts Harlem" have a higher price compared to those in other areas.

4. Some attributes which initially seemed to have business meaning did not have a strong correlation to the price. These attributes are "availability_365", and "minimum_nights".

Figure 1 depicts the correlation of some of the attributes with the price, which will help choose the right features for the model.

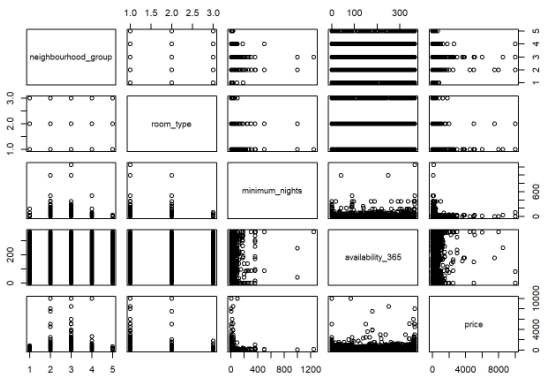

Figure 1: Correlation of price between other attributes

1.4 Categorizing Price

Price the New York Airbnb Dataset is a continuous attribute with the following details -

Min. : 10.0
1st Qu.: 69.0
Median : 101.0
Mean : 141.8
3rd Qu.: 170.0
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 : 25% (10 - 69)
Category 2 : 50% (70 - 101)
Category 3 : 75% (102 - 170)
Category 4 : 100% (171 - 10000)
Figure 2 shows the distribution of the data after the categorical split. As seen below, the distribution is well balanced with all classes having approximately the same number of instances, eliminating any class skewed dataset problems.

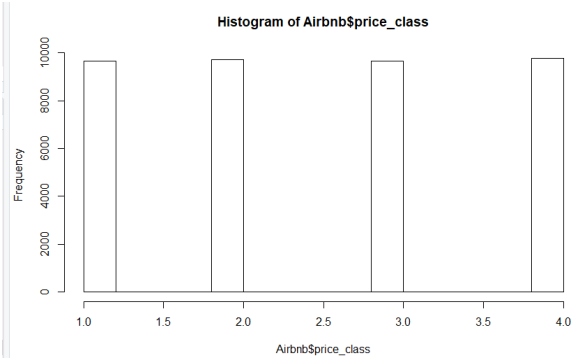

Figure 2: Distribution of price after categorizing

1.5 Cleaning the dataset

Since the dataset consists of both missing and inconsistent values, the first stage in the data mining process is to clean or eliminate these records. The total number of records in the dataset is 48,895 records. 10,052 records had missing values in last_review and reviews_per_month. However, since we foresee some significance in using these attributes to determine price, these missing records have to be eliminated to prevent incorrect data. The size of the dataset after eliminating the missing values is 38,843 records. Further, there are 10 records in the dataset that have the price value as 0. While these data records may be accurate, they form outliers that can interfere with the model's accuracy. The final dataset consists of 38,833 records.

1.6 Feature Engineering

1.6.1 Dimensionality Reduction

In an attempt to reduce the dimensions of the dataset, attributes that are unlikely to contribute to the prediction of the price group without intensive preprocessing are eliminated. For instance, while the host name and the description of the listing could indirectly contribute to the price (An unknown pattern that could reveal that customers are more likely to pick a specific listing based on description - old ancient house versus new house in the suburb), these features are not easily found through simple classification techniques.

1.6.2 Factoring attributes

Factoring attributes allows R to encode the vector (in this case, the attribute column) into a categorical attribute. For the attributes room type, neighborhood, neighborhood group are all categorical values and hence these are factored to make them understandable to an R algorithm. The attribute availability 365 is a real valued number. Since this value would not easily be used to make a decision tree split, the attribute is split into 3 bins based on the 33rd and 66th percentile and then factored.

1.6.3 Leveling attributes

Leveling an attribute in R allows categorical attributes to be ordered in a meaningful manner. For instance, the attribute room type is ordered for 3 levels.

- Level 1 - Shared room
- Level 2 - Private room
- Level 3 - Entire home/apartment

1.7 Train and test split

The final stage of the preprocessing is splitting the dataset into train and test in a 80-20 ratio. The dataset is shuffled to ensure that there is representation of the data distribution in both the parts. As seen in figures 3 and 4, the distribution for all categories of price group remains the same in both train and test. All models that will be built will use train and test accuracy as a metric to determine performance. These accuracy measures will also be used to identify potential problems such as overfitting and underfitting if they occur and in turn work towards a model that eliminates such issues.

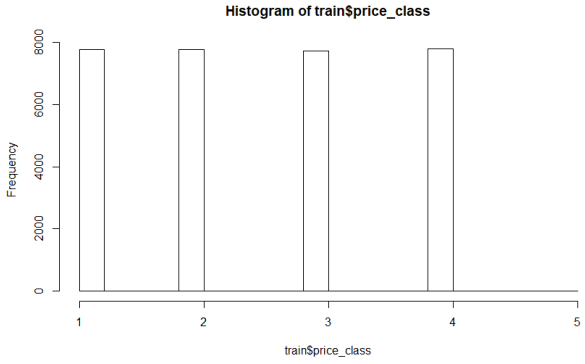

Figure 3: Distribution of price in Train

2. NEXT STEPS

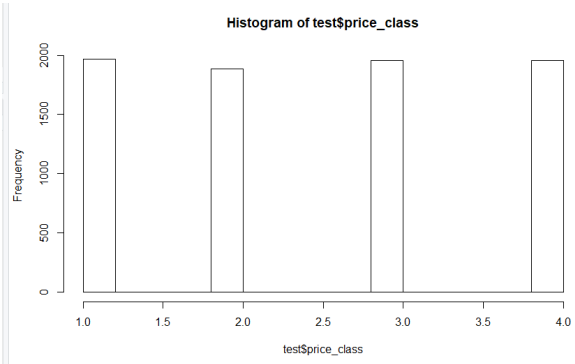

Figure 4: Distribution of price in Test