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

1. TRIALS ON THE DATASET

1.1 Decision Tree

Decision Tree is a algorithm about

Because the data has many attributes of discrete, in order to do decision tree on our dataset, At first we did some simple modifications to the data. These modifications will be applied to the other 2 algorithms, too. What we did first was to remove all the extreme data, such as the data with prices that are 0 meaning these hosts are totally free or that are too expensive that are higher than 1000 dollars. These data is generated in abnormal conditions and will greatly impact the result of decision tree model. Secondly, we kept all the attributes of numberics and discarded other attributes. This step might looks like controversial yet as we are doing just test, we think it is okay. Moreover, our aim is to make predictions yet decision tree can only do classification, so we converted the aim attribute price from numberics to boolean value. In the end, we divided dataset to train and test by 9:1.

The final result is about 72% and concrete report can be found here

1.2 k nearest neighbours

k nearest neighbours(kNN) is about

the data processed by kNN needs to be normalized because it used distance to show varieties. What's more, k of kNN is a hyper-parameter, and a popular choice of k is between 5-20.

after doing the same process, the final result of kNN is 64% with k=20. Report

1.3 Naive Bayes

Naive Bayes is about Naive Bayes is good resistance to noise compared to the former 2.

the result of Naive Bayes is 68%. Report

1.4 summaries

2. FUTURE PLAN