

Airbnb: Predicting Loyalty

Louise Y. Lai, Erica Ram

Microsoft Research Data Science Summer School

Research Question

How can we predict the loyalty of hosts and guests on Airbnb?

- How can loyalty be quantified?
- Are there certain characteristics of hosts and guests that can predict whether they are loyal or not?
- How does host loyalty interplay with guest loyalty?

Introduction

Airbnb has two million rooms available - nearly three times more than that of the largest hotel chain in the world.

Airbnb faces unique challenges because unlike traditional hotels, this sharing-economy platform relies on the return of both their hosts and guests.

Predicting customer loyalty, both on the host side and guest side, is arguably more important to Airbnb than to traditional hotels.

Methods

- Data sets provided by InsideAirbnb.
 - Listing information consists of host and listing information for listings in New York City between January 2015 and July 2016.
 - Reviews data includes all reviews for the United States on Airbnb's website as of May 2016.
- Host and guest loyalty were predicted by identifying predictor features for both hosts and guests and modeling using decision trees.
- Analysis conducted using the R programming language.

Discussion

Typically, **how frequently** and **how recently** a person uses a platform is predictive of their likelihood to return. This is the **recency/frequency** theory (Fader et al.) Despite running analysis on a wide range of variables such as amenities, availability, prices and even sentiment analysis on reviews, recency and frequency trumped all other variables in their predictive power.

Host Loyalty:

- Those who are listed most often and most recently, are more likely to be loyal.
- Combining recency/frequency features and other subsets, such as amenities, gave a slight improvement to the AUC and provided interesting insights.

Guest Loyalty:

- This was best predicted using recency/frequency features, however the **interplay** features of the listing from the last review do add value to the model.
- Guests who stay with highly-rated hosts are more likely to return. The prediction ranking increases when looking at guests for whom more data exists.

Limitations:

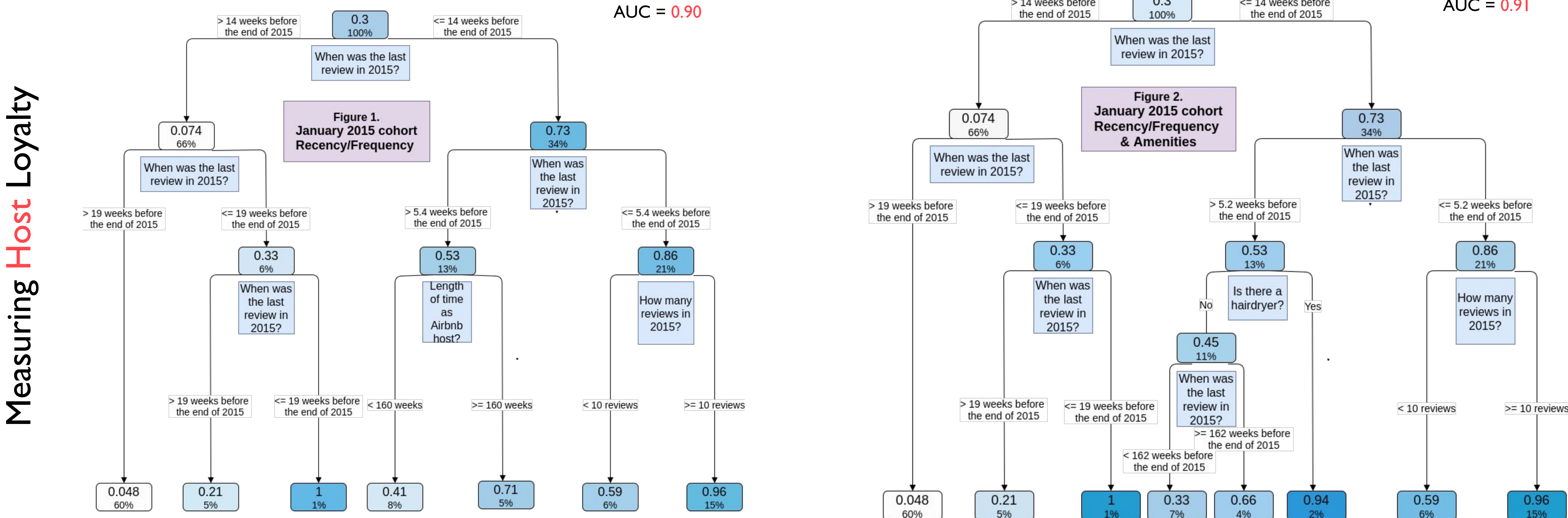
- Usage of scraped data was limiting because it is essentially a snapshot of publicly available data and not a comprehensive log of actual transactions between guests and hosts.
- Lack of transactional data led to the usage of reviews as a proxy for guest stays but estimates have shown that less than half of guests leave reviews.
- Certain features, such as amenities, aren't found in earlier datasets.

References

1. Fader, Peter S., and Bruce GS Hardie (2009). "Probability models for customer-base analysis." *Journal of interactive marketing*, 23.1, 61-69.

We thank our mentors Shawndra Hill, Amit Sharma, Sébastien Lahaie, Fernando Diaz, and Christopher Riederer for their guidance on this project as part of the 2016 Microsoft Research Data Science Summer School. Special thanks to Murray Cox, the founder of InsideAirbnb website, for making the datasets easily accessible to the public.

Results



The goal of this classification tree is to identify which hosts will return to the Airbnb platform. The algorithm does that by classifying hosts into buckets based on recency, frequency and other variables.

15% of all hosts fall into this bucket, and our model is saying these hosts are very likely to return to Airbnb next year a host again. These are our 'loyal' hosts!

The bottom row gives us the results of the classification model. Here, the model has classified 60% of all hosts into this bucket and they are very unlikely to return to Airbnb.

Measuring Guest Loyalty

