Home.PI: A Friendly Pricing Intelligence for New Airbnb Hosts

Yiming Xu yimingxu@mit.edu

Rong Liu rongl@mit.edu

Shu Xu shuxu@mit.edu

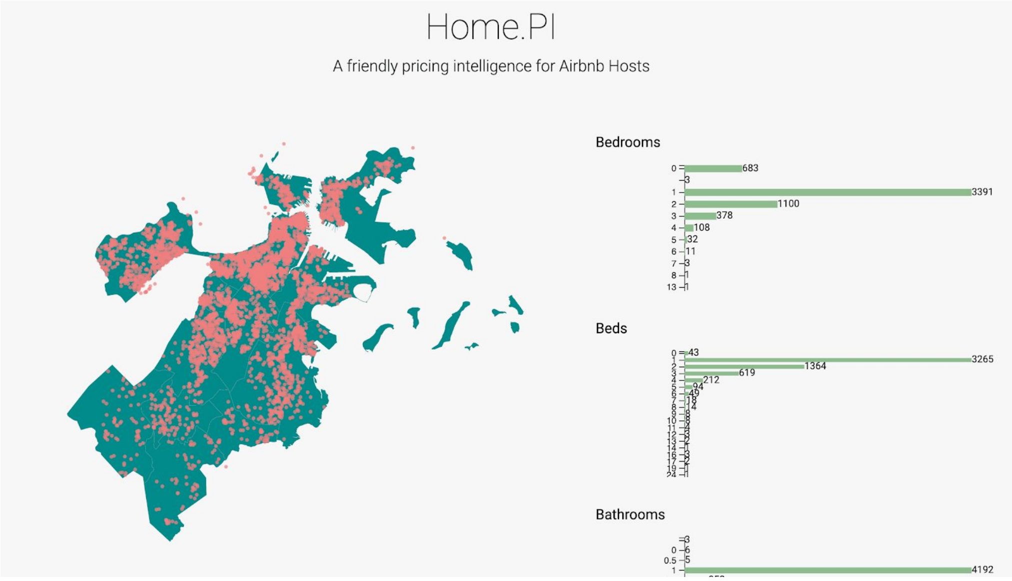
Project Progress

The completed tasks:

- Static visualization of the overall market
    - Map plot with listings location
    - Bar plots for numbers of bedrooms, beds, bathrooms, etc.
- Data integration and preprocessing
    - Integrate data of Boston and Cambridge
    - Compile calendar data pulled on different dates
    - Impute missing data (e.g. security deposit, cleaning fee, etc.)
    - Feature engineering (day and month) to study seasonality
- EDA to examine the relation between pricing and seasonality
- Baseline Pricing Recommendation Model
    - We build a classifier that uses listing features, booking date and price to predict Pr(booked) -- whether this particular listing will be booked on this date. Next, we use this classifier to construct an Optimization model that maximizes Earnings = (Price * Pr(booked)).
    - To improve classifier accuracy, we have experimented with Logistic Regression and KNN (both tuned through cross-validation), and LDA/QDA.

The finished deliverables:

- Preliminary website with static visualization and layout design



Home.PI

A friendly pricing intelligence for Airbnb Hosts

Preliminary Result

- The best booking status prediction model so far is KNN CV, but it is not very helpful in producing booking probability, so it is not useful to optimize earnings.
- Logistic Regression CV and LDA both achieve 0.67 accuracy on test set and are useful in optimizing earnings, but they both tend to raise the pricing by too much.

(For detailed information, please visit our jupyter notebook: https://github.com/YimingXu1213/AirbnbPricing/blob/master/Model/midterm_EDA_baselineModel.ipynb)

Next Steps

- Generate interactive visualization:
  Users should be able to specify key features (e.g. location, property type, number of bedrooms, etc.) about their listings, or feature combinations of their interest, and the distribution visualizations of the pricing and these features should update accordingly.
- Generate word cloud plots
  We want to extract keywords from reviews of similar listings and generate a word cloud plot, to inform users what their customers care about and look forward to in an Airbnb experience.
- Improve Pricing Recommendation Model
  (1) Feature Engineering: We will further extract useful information from the dataset and tailor it for our model.
  (2) Model Building: We will take a look at decision tree and different ensemble methods, such as bagging, random forest, boosting and even neural networks to improve model performance.
  (3) Optimization Process: So far, we only allow users to search for the daily price that maximize the expected return. In the following steps, we might want to extend it to include other fees such as cleaning and security deposit. In addition, we might also allow users to set their own constraints on the daily price as well as probability.
- Integrating pricing model into the HTML website
  We want this section to be interactive, as well. Users can input booking information and listing features. The website should return a series of pricing recommendations: a point recommendation that maximizes expected return, an interval recommendation that accommodates different pricing strategies, and suggestions such as ‘adding a TV will likely improve the daily earnings by $30’.

Potential Problem

- Calendar data refer to the information of whether listings are booked on certain days. It is crucial in studying seasonality and predicting booking status in general. However, we only have calendar data of future dates (from Inside Airbnb). Fortunately, we have future calendar datasets pulled roughly once a month, so we take one month’s data from each dataset to compile our final calendar data. On dates further away from the pull dates, we assume that the listings are less likely to be booked, so we incorporate ‘booking date - calendar data pull date’ as a predictor variable to account for the impact.

Updated timeline

Timeline is not changed by much, except that each person’s responsibility is clearer.

Tasks	Person	Deadline
1. Data Integration and Cleaning	Rong	Midterm Report(11/04)
2. Static Visualization	Shu	
3. EDA and Baseline Model	Yiming	
1. Interactive Visualization	Shu	Final Report(12/04)
2. Review Word Cloud	Rong	
3. Optimized Model	Yiming	
4. HTML Deliverable	Shu	
5. Poster and Report	All	
1. Presentation	All	Presentation(12/11)

**rows in shade are tasks already completed.*

Questions

- How should we evaluate our visualizations?