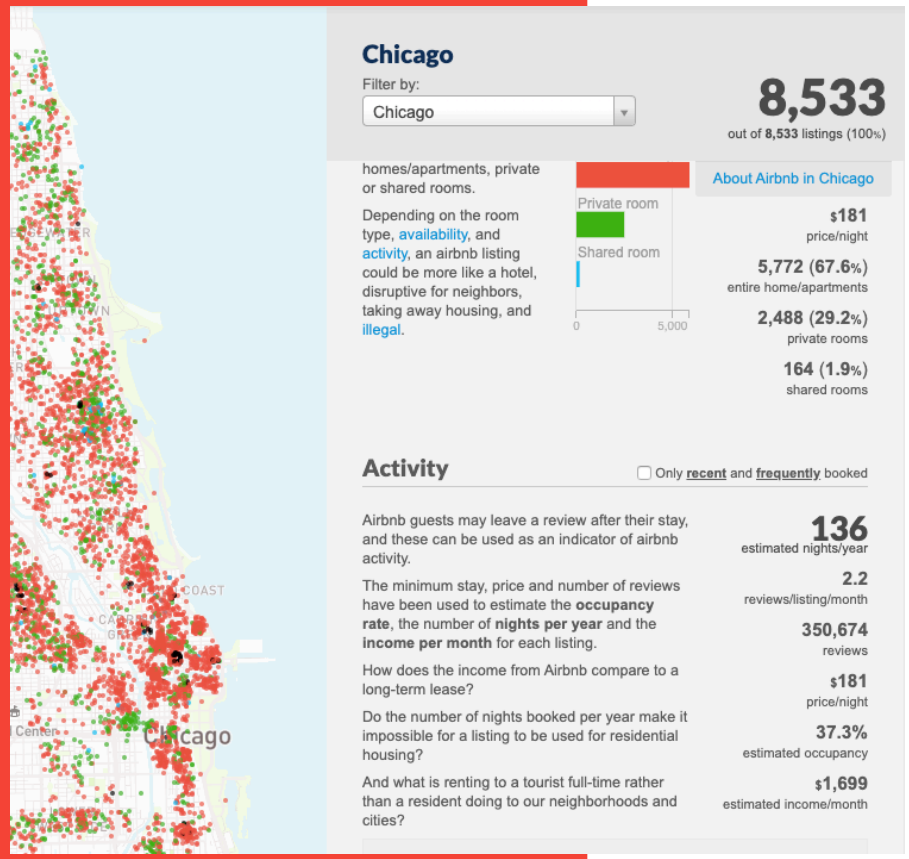
Predicting Airbnb Booking Activity: Chicago

Catherine Yang

6/8/2020

Background



Source: Inside Airbnb Chicago

Description:

Predicts the number of reviews per month that an Airbnb listing (or potential listing) in Chicago will receive, based on characteristics pertaining to its host, property, and booking process.

Motivation

- To provide hosts with more customized insights into attributes driving listing popularity and increased booking activity.
- Existing dashboards only provide aggregated, descriptive statistics without informing how booking activity and user engagement can be improved.

**Number of reviews per month is used as proxy for measuring booking activity but does not explicitly account for listing quality. Through EDA, it is noted that reviews scores are generally high, so it is assumed that listings with higher review count are in general favorable listings, since negative reviews would disincentivize guests from booking those listings.

Data

Dataset

- Chicago Airbnb listings as of 11/21/2019
- Source: Inside Airbnb

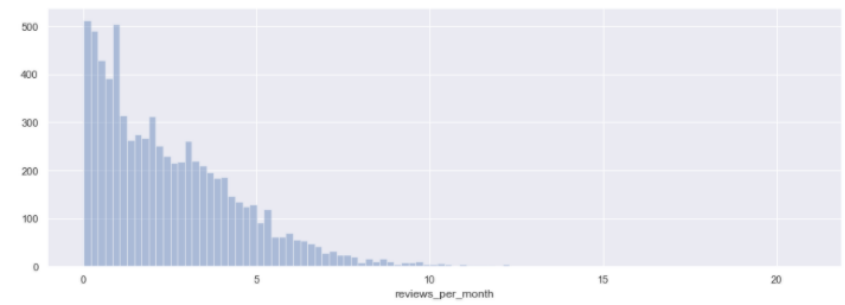
Features

- Host: number of years as host, response time, superhost status, number of listings, etc.
- Property: property type, room type, number of bedrooms / beds / bathrooms / guests accommodated, price, etc.
- Booking: max / min nights, instantly bookable, cancellation policy, etc.
- Performed feature engineering (e.g., binning, categorization) to arrive at appropriate features for modeling

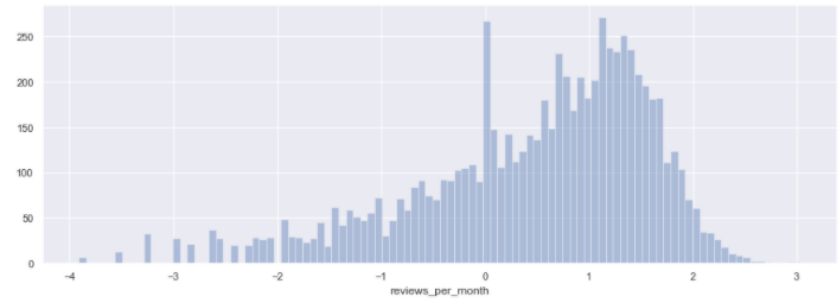
Target Variable

- Number of reviews per month

Model & Success Metric



Histogram of target variable



Histogram of log-transformed target variable

Model

- Log-transformed target variable
- One-hot encoded categorical features
- Standardized numeric predictors
- Ensembled a Random Forest Regressor, Gradient Boosted Tree Regressor, and XGBoost Regressor with specific set of weights

Success Metric

- Desired RMSE: < 0.5
  - Measure on average, how far the predicted number of reviews per month is from the actual
- Achieved: 0.767
  - There is margin for improvement!
- R2 metric was used for hyperparameter tuning

Insights

Random Forest

Feature	Importance
min_nights_cat	0.097191
host_since_years	0.090324
amenities_count	0.087168
price	0.078147
host_listings_count	0.069527
host_response_rate	0.064445
cleaning_fee	0.061138
host_is_superhost	0.039244
security_deposit	0.036792
guests_included_cat	0.027959

XGBoost

Feature	Importance
min_nights_cat	0.237894
room_type_Shared room	0.041442
cancellation_policy_strict	0.036493
host_response_time_within an hour	0.036378
host_is_superhost	0.026711
host_response_time_within a day	0.026492
host_response_rate	0.022769
room_type_Hotel room	0.018963
require_guest_phone_verification	0.018412
property_type_cat_House	0.016862

Model Observations

- Surprisingly, XGBoost utilized categorical variables as splitting criteria more than Random Forest or GBM.
- Minimum nights required was the most important feature for all models – an obvious result, since shorter stays allow more bookings per month, and thus, greater reviews per month.

Overall Conclusions

- Neighborhood location of listing didn't seem to be a top factor in the number of reviews, indicating that booking activity is generally commensurate with the number listings in an area.
- While number of reviews per month is correlated with obvious factors, e.g., host experience, there are a few important features that are more actionable in helping to improve booking activity and user engagement.

Thank You!

Class of 2016

Consultant, 2016-2019

MS in Analytics

Contact Info:

Catherine Yang

catherineyang@u.northwestern.edu