Predicting Airbnb User Booking Destinations

Lisa Oshita
Dr. Glanz

May 13, 2018

About the Competition

- Recruiting competition hosted on Kaggle from November 2015 to February 2016
- Task: build a model to predict where new Airbnb users will book their first destinations
- 12 possible destinations to predict: Australia, Canada, France, Germany, Italy, Netherlands, Portugal, Spain, United Kingdom, US, other and no destination found (NDF)

About the Data

- train_users: 213,415 observations and 16 rows, contains information about users from 2010 to 2014
- sessions: 1,048,575 rows and 12,994 unique users, contains information about web session activity for each user
- 10% of rows from each unique user were randomly sampled
- New sampled sessions data contained 104,424 rows

Booking Destinations: extremely imbalanced classes

Destination	Percentage of the data (%)
NDF	58.35
US	29.22
other	4.73
FR	2.35
IT	1.33
GB	1.09
ES	1.05
CA	0.67
DE	0.50
NL	0.36
AU	0.25
PT	0.10

Table: Percentage of data each destination accounts for

Models

- Extreme Gradient Boosting (XGBoost)
- Random forest
- Stacked model

Feature Engineering

- Date features
- Age and gender
- Count features created from the sessions data (314 features: number of times a user viewed recent reservations, number of times a user viewed similar listings...)
- Summary statistics of seconds elapsed for each user's web session
- After all feature engineering and one-hot encoding, there were a total of 596 features for use in the model

Model Building

- Full data was split into training and test sets
- 5-fold cross validation with both the XGBoost and Random forest achieved 87% classification accuracy and NDCG score of 0.92, but only made predictions for NDF and the US
- Both models were fit to just the top 200 most important features and cross-validation was again performed - both achieved same results as previously, but computation time decreased