Predicting Airbnb User Booking Destinations

Lisa Oshita

California Polytechnic State University, San Luis Obispo

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
- To minimize computation time of the sessions data, 10% of the rows from each unique user were randomly sampled
- New sampled 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

- Year, month, day of the week, season features of dates
- Days between date account created and date of first booking
- Days between date first active and date of first booking
- Age
- 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...)
- Mean, median, standard deviation, minimum, maximum of seconds elapsed for each users web activity
- 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

Model Building: Feature Importance

include tables of feature importance?

Model Building: Oversampling

- Oversampling with replacement from countries under-represented in the data, and undersample from countries over-represetned in the data
- Synthetic Minority Oversampling Techniques (SMOTE)

Model Building: Results from Oversampling Techniques

- Same accuracy and NDCG scores as all previous models
- Both random forest and XGBoost were now able to make predictions for all booking destinations (though not well)

Model Building: Stacked Model

- visual describing the process

Model Building: Results of the Stacked Model

- Accuracy and NDCG scores were the same as all previous models
- include confusion matrix

Model Building: Stacked Model confusion matrix

- include confusion matrix here

Model Building: Final Models

- 5-fold cross-validation was performed for all three models on the entire data
- Accuracy and NDCG scores remained the same as all previous models
- Run times for the XGBoost, random forest and stacked models were 27.5 minutes, 28.7 minutes, and 3.24 hours, respectively (include this as a table?)

Discussion and Conclusions

- No effective strategy was found for improving model accuracy
- The stacked model did not perform better than the two base models because both base models could not make accurate predictions for the under-represented booking destinations

Next Steps?

- Stack more than just two models
- Build additional models to predict and impute missing values
- Find a way to incorporate the other 4 remaining data sets
- Parameter tuning