Airbnb New User Bookings

Rob Castellano, Zi Jin, Yannick Kimmel, Michael Winfield

Introduction
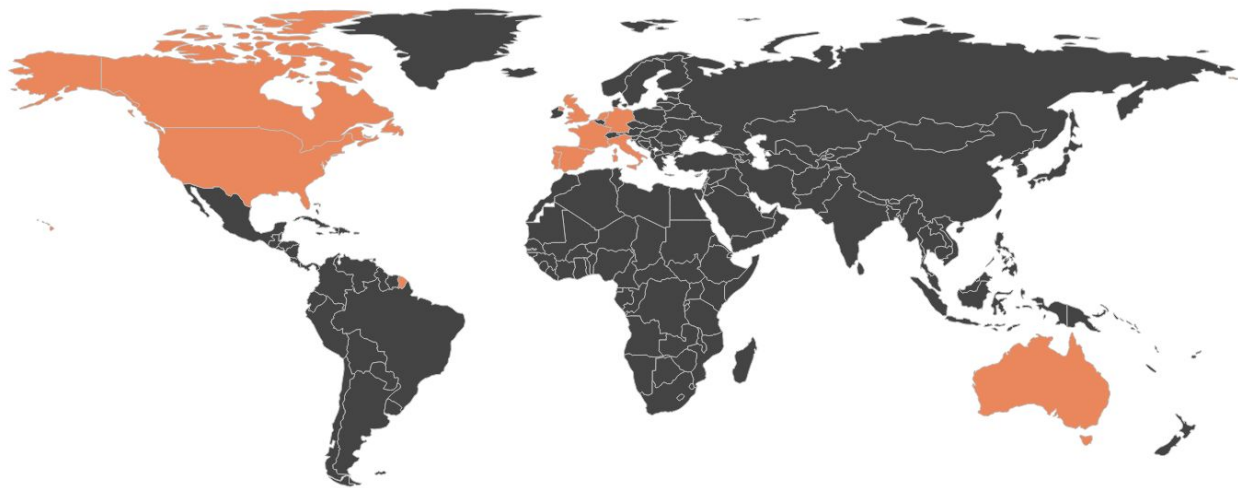
Goals

- Kaggle competition hosted by Airbnb, ending Feb 2016.

- Goal: Predict the country of a new user's first destination. This can include not booking (NDF).

- The competition allowed by the submission of five suggestions for each user.

- The competition was graded on normalized discounted cumulative gain (NDCG), which measures the performance of a recommendation system based on the relevance of the recommended entries.

Airbnb Kaggle Dataset

The Airbnb Kaggle dataset consisted of:

- **User information:** Unique ID, age, gender, web browser, avenue in which the user accessed AirBnB, country destination, timestamp of first activity, account created, and first booking.
- **Browser session data:** Unique ID, action type, and time elapsed.

- Training set: 200,000 users--Jan 2010 to Jun 2014
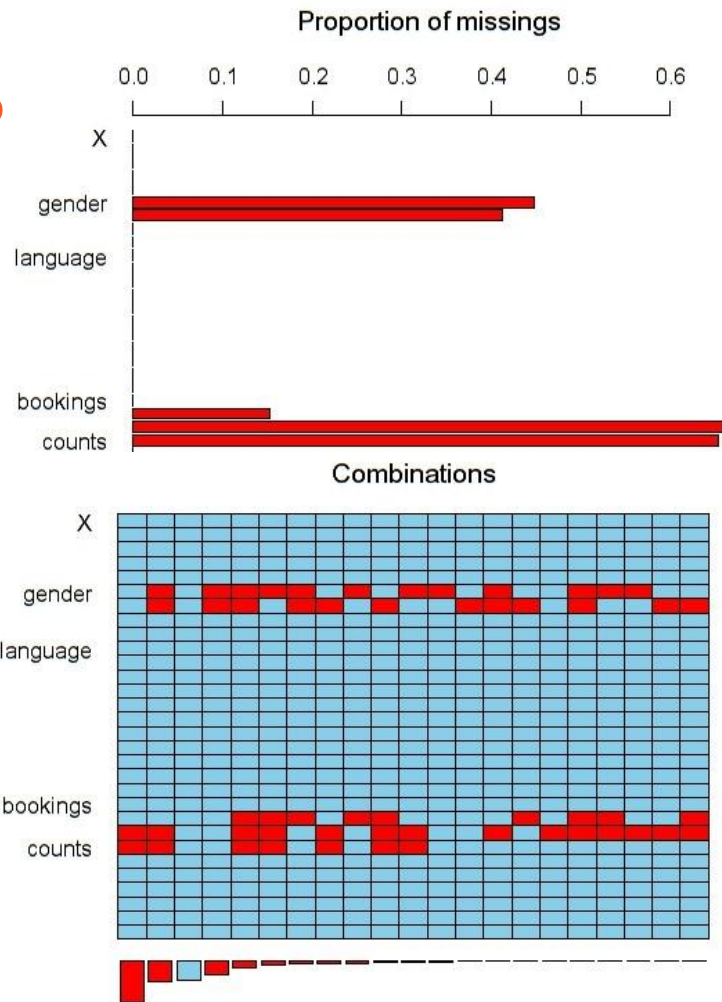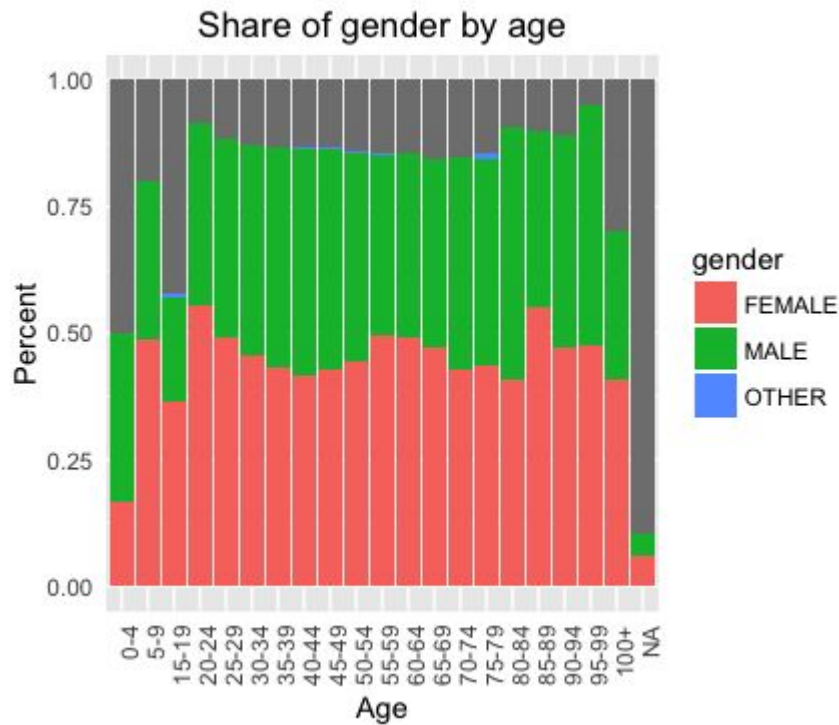 Test set: 60,000 users--July 2014 to Sep 2014

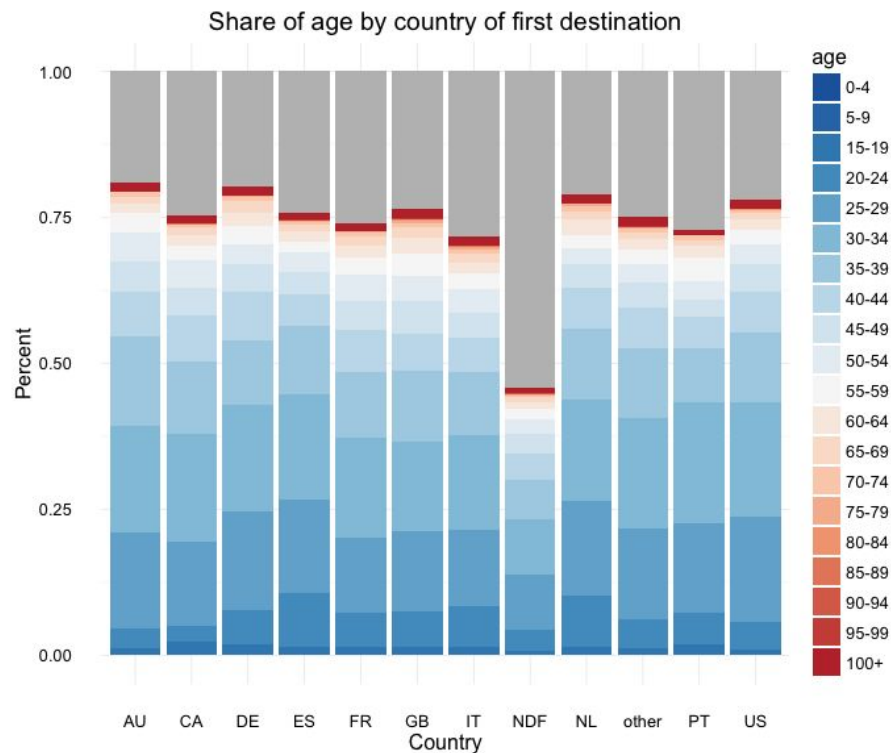# Airbnb User Booking Behavior

Distribution of country of first destination
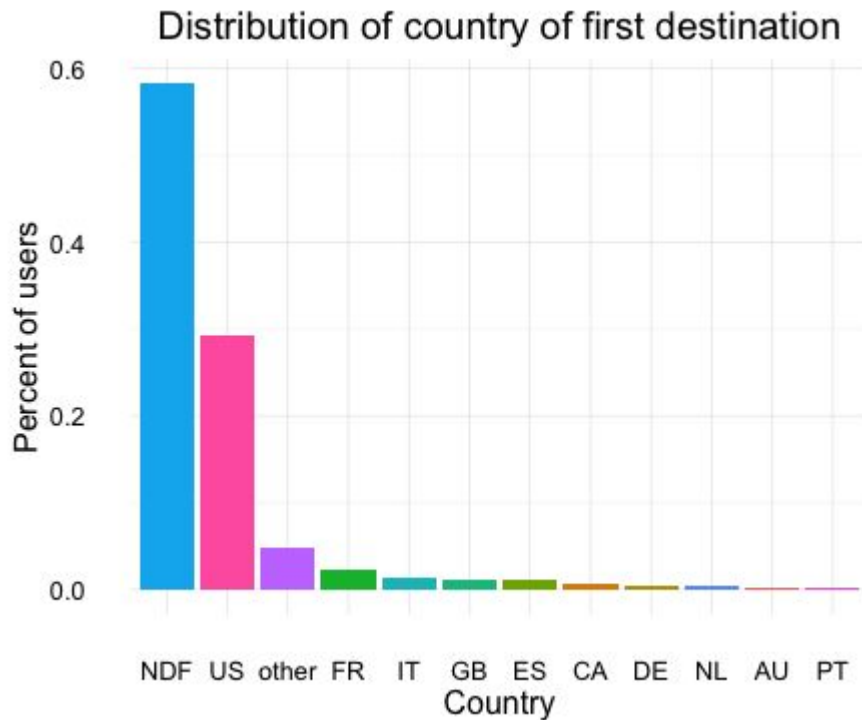
User demographics
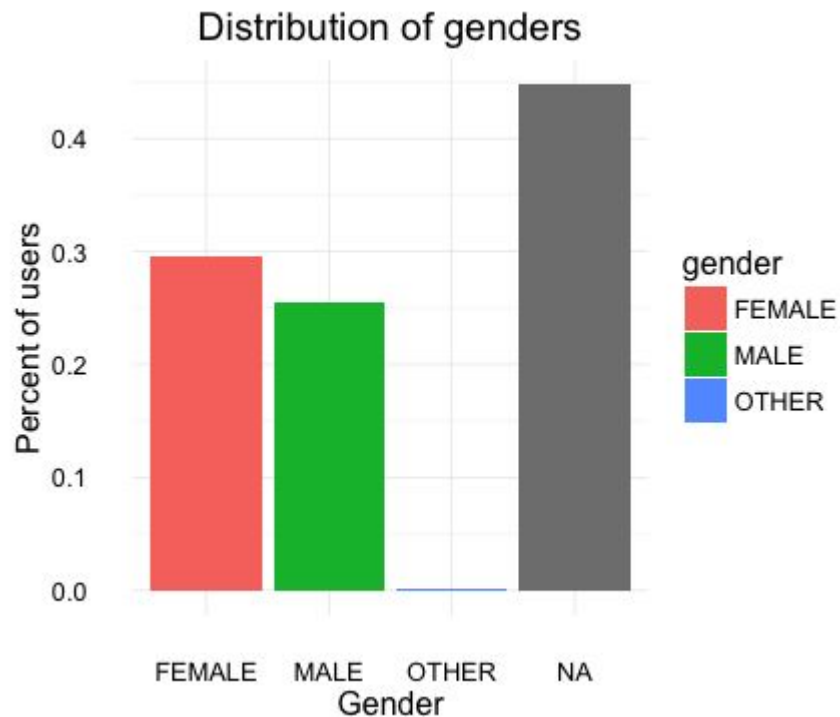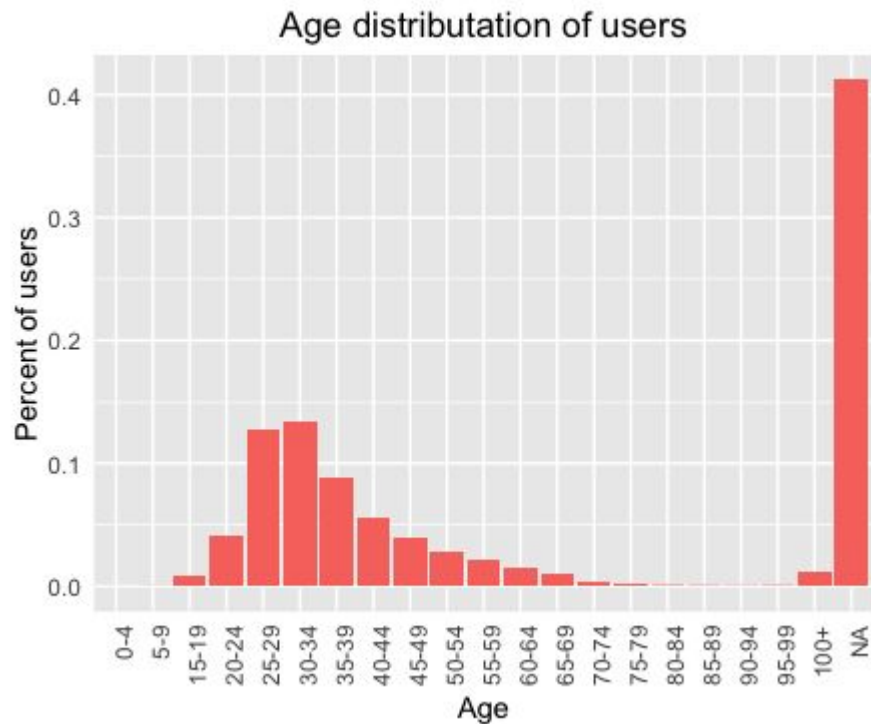
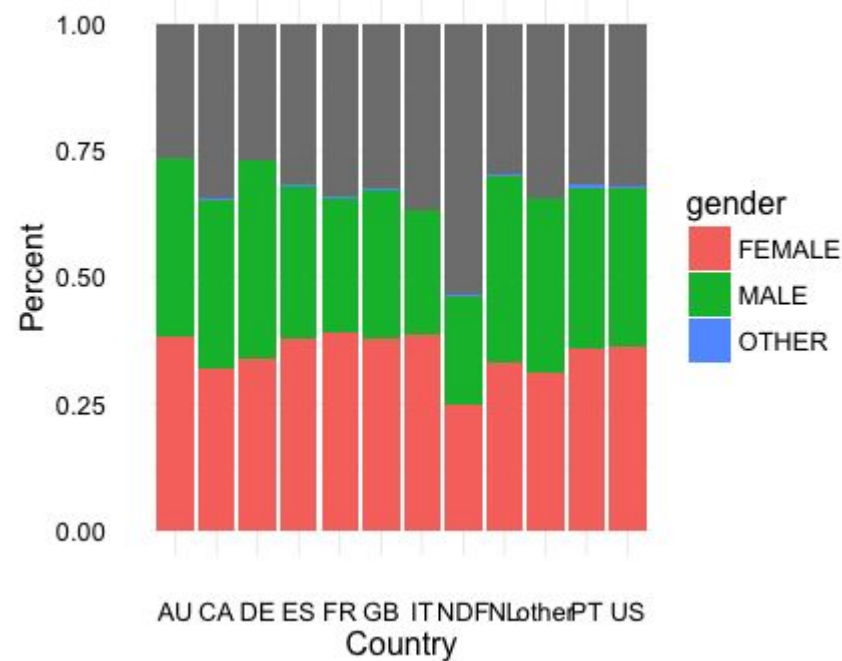# Age/Gender Missingness

Age & Gender on Country Destination


Share of age by country of first destination

Share of gender by country of first destination

Time variable feature engineering

- *We decided to engineer 3 features based on user booking behavior, specifically the time between the creation of Airbnb accounts, a user's first activity on the website, and their date of first booking.*

Distribution of time of booking

How long users wait to book, by country of first destination
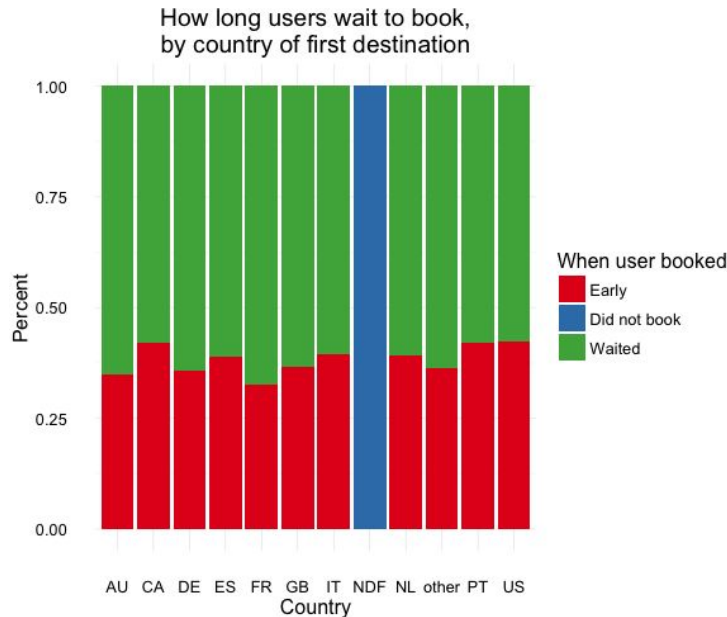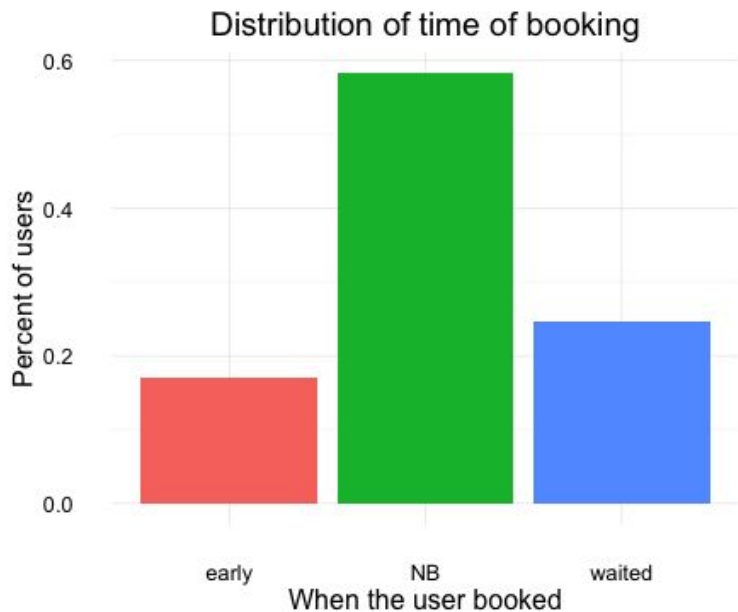
Stacking

- *Out-of-fold predictions of those three features were then added to the training dataset and test dataset through the process of stacking.*
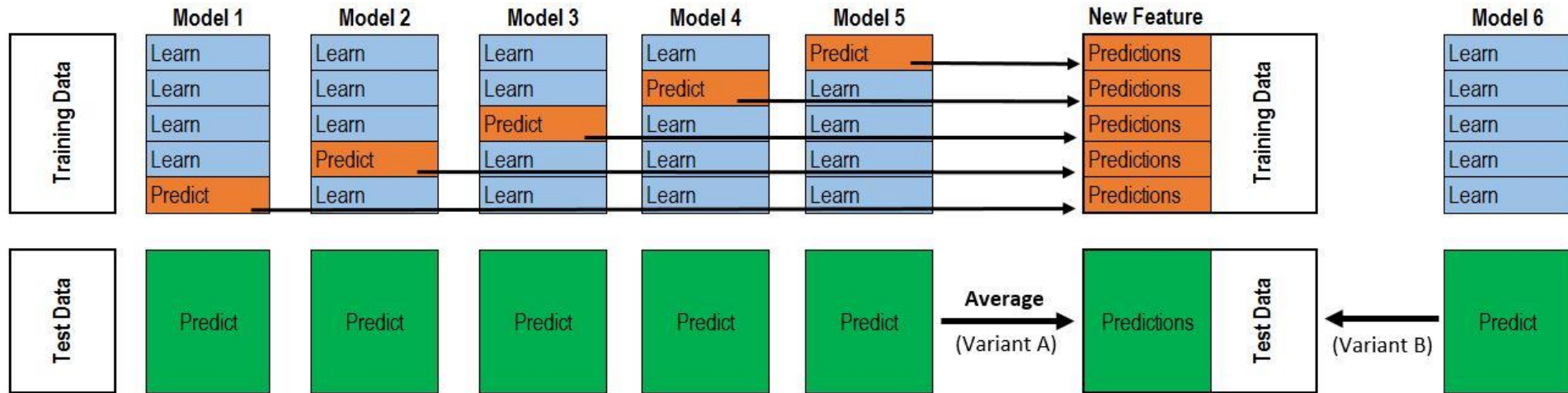
Workflow

Predicting Country Destination

Ran grid search cross-validation
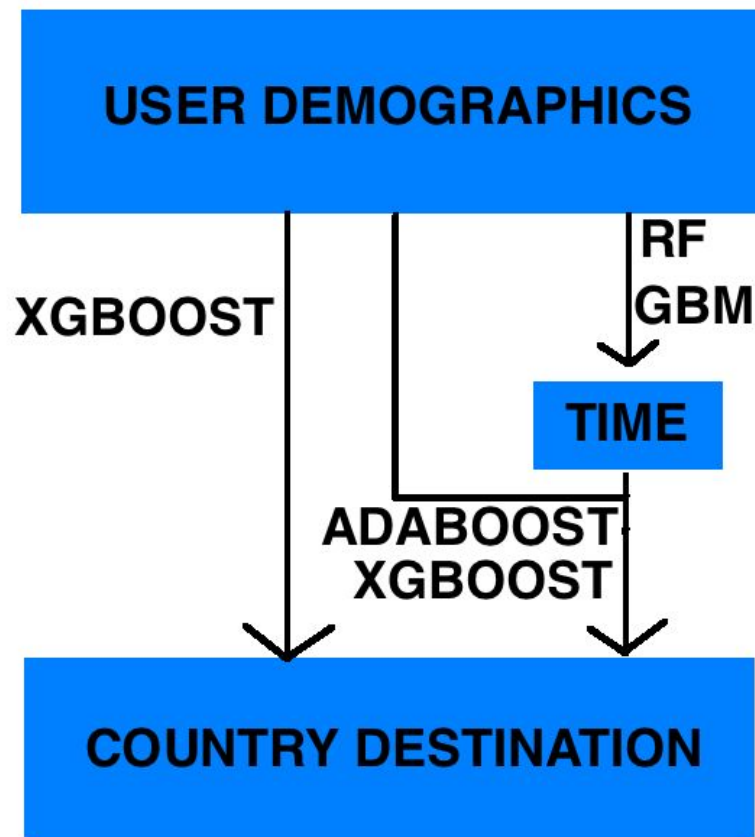
XGBOOST stacked and unstacked

ADABOOST stacked

Variables of importance

Conclusions

1. Performed exploratory data analysis on AirBnB new user information.
2. Wrangled and munged data in Python and R.
3. Used R for visualization and the creation of a Shiny App.
4. Feature engineered time-lag-based variables using Python and R.
5. Fit models (XGBoost/Random Forest/AdaBoost) using Python.
6. Performed predictions on users using XGBoost that ranked at 392 on Kaggle.

Recommendations to Airbnb

- Invest in collecting more demographic data to differentiate country destinations. A possible source includes Facebook (~¼ users enter through FB).
- Flag users who decline to enter age and gender; such users are more likely to browse without booking.
- Continuously collect browser session activity; such data was helpful for predictions. This data was available only for newer users.

Future Directions

Steps to improve our predictions:

- Optimize tuning parameters for XGBoost on the stacked dataset.
- Additional stacking: add country of destination predictions to dataset as features to improve predictions.
- Use multiple XGBoost models and ensemble them.