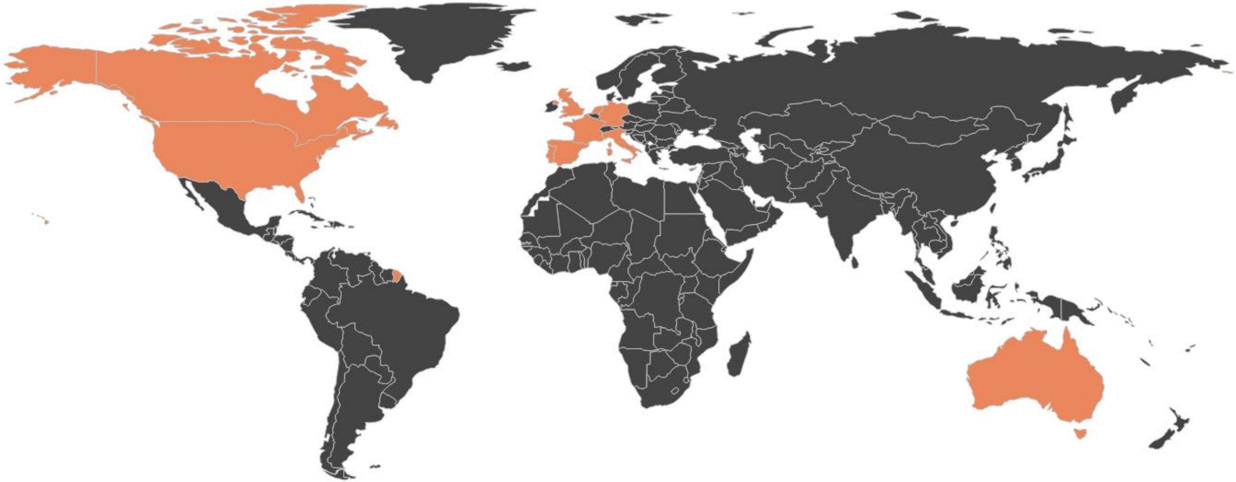
Airbnb New User Bookings

Rob Castellano, Zi Jin, Yannick Kimmel, Michael Winfield

Introduction

I. Goal of our project
II. The data
III. Insights into the data
IV. Our strategy
   A. Feature engineering
   B. Stacking
   C. Feature importance
V. Results
VI. Conclusions



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.

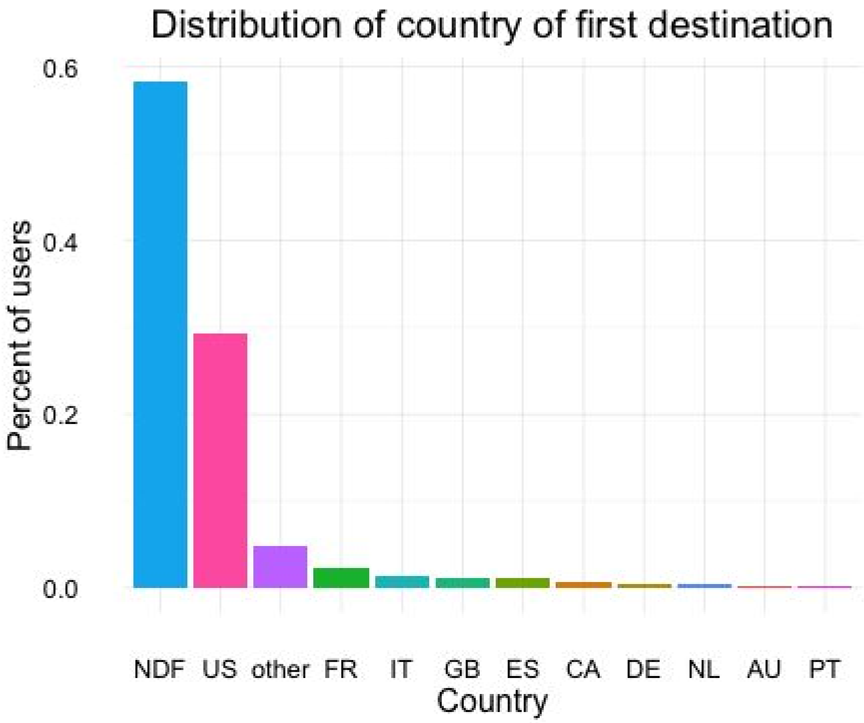
Airbnb Kaggle Dataset

The Airbnb Kaggle dataset consisted of:

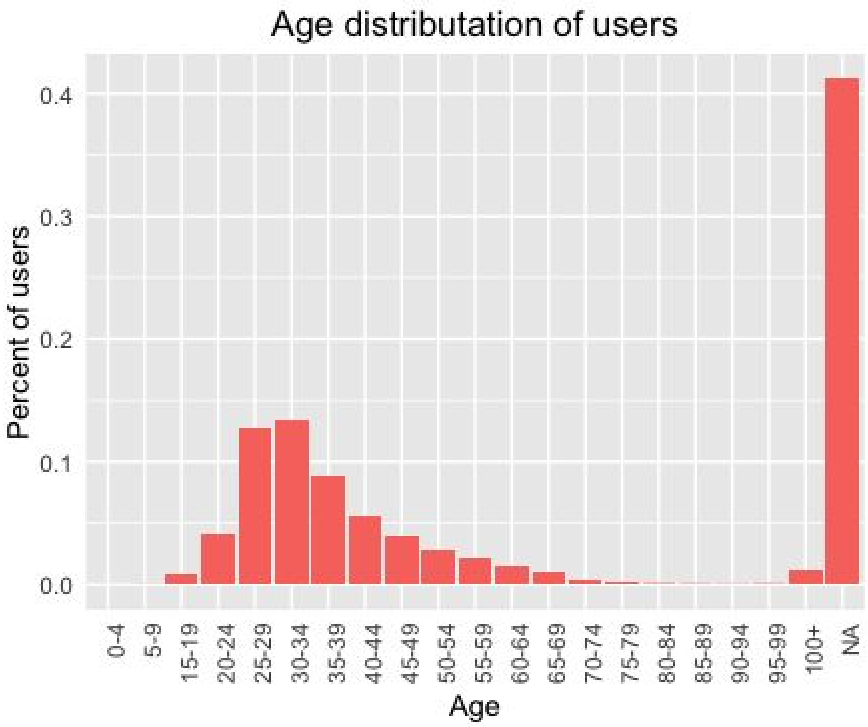
- **User information:** Unique ID, age, gender, web browser, avenue in which the user accessed AirBnB, country destination, timestamp of first activity, account created, and first booking.
- **Browser session data:** Unique ID, action type, and time elapsed.

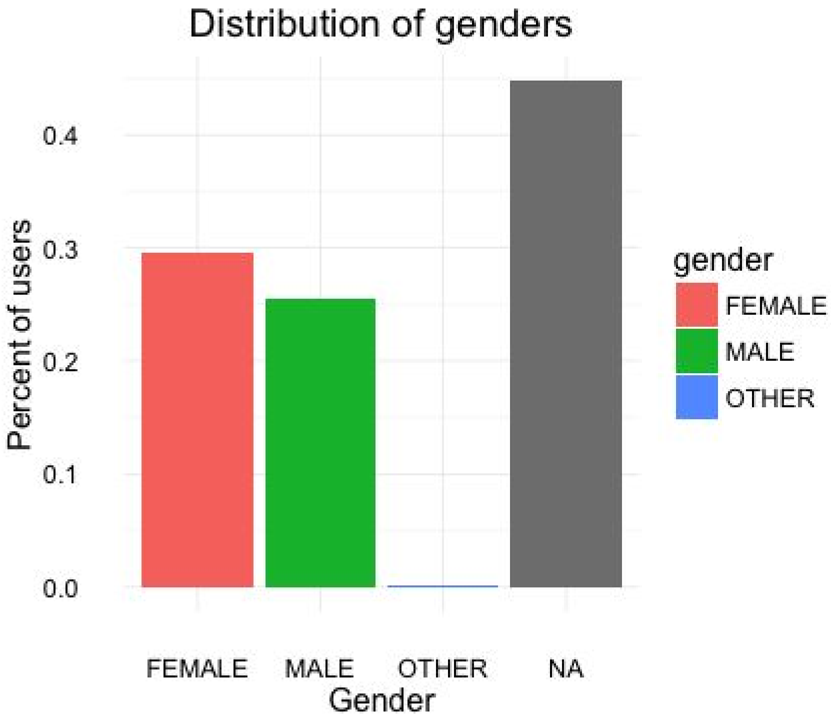
- Training set: 200,000 users--Jan 2010 to Jun 2014
  Test set: 60,000 users--July 2014 to Sep 2014

Airbnb User Booking Behavior

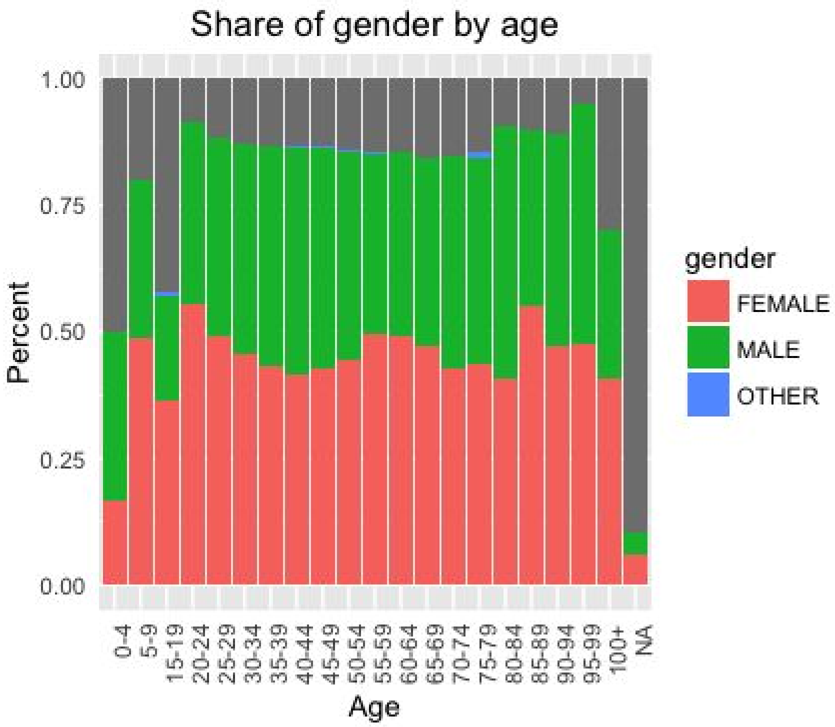


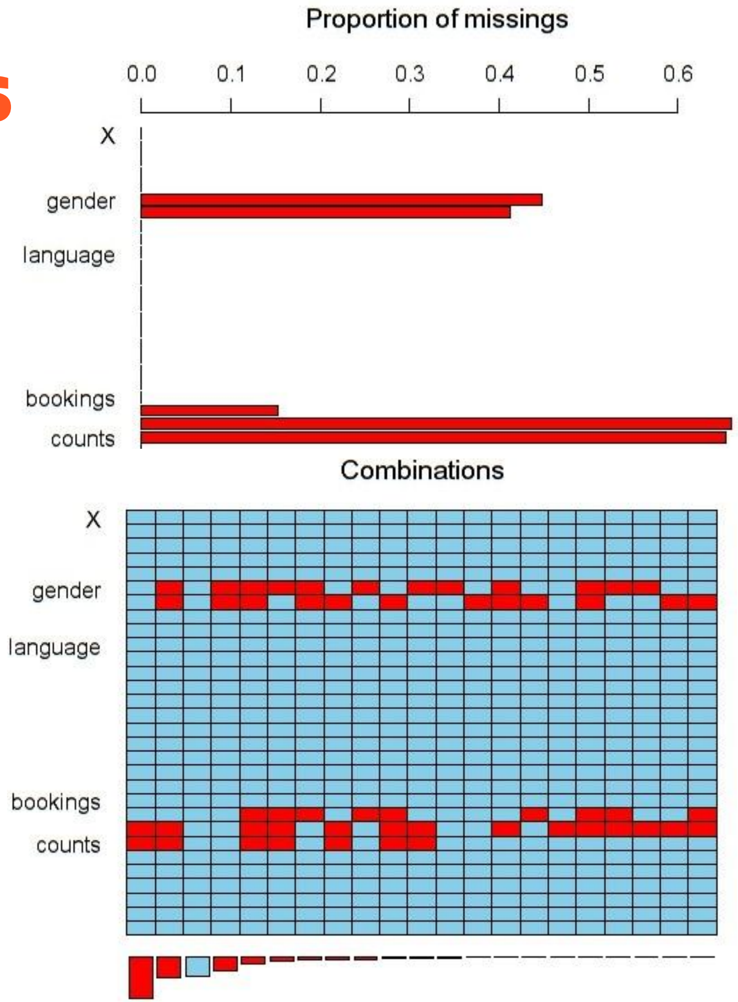
User demographics



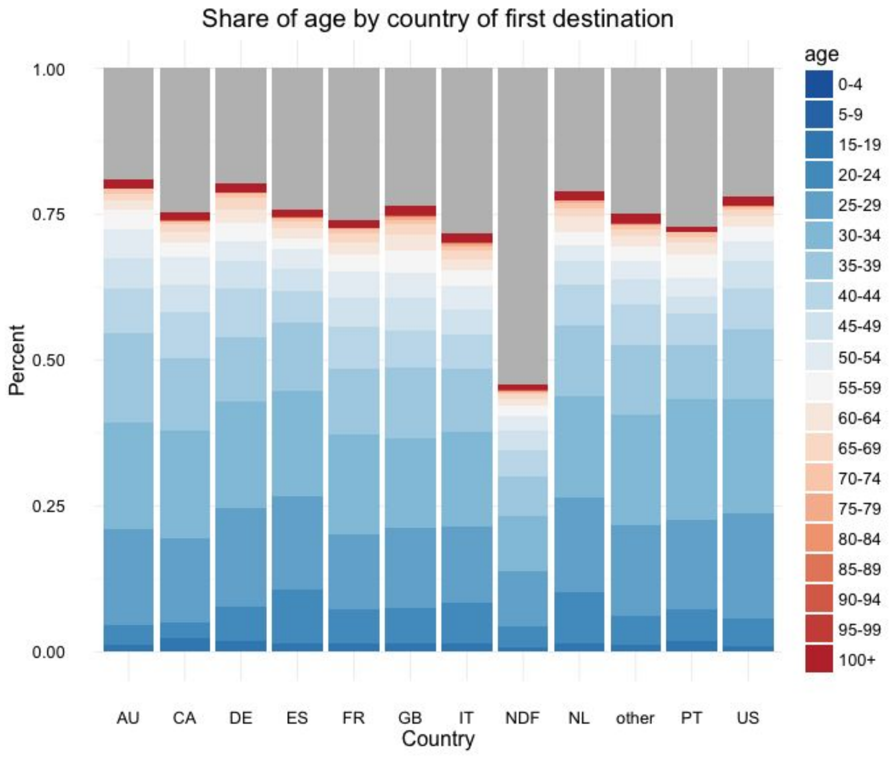


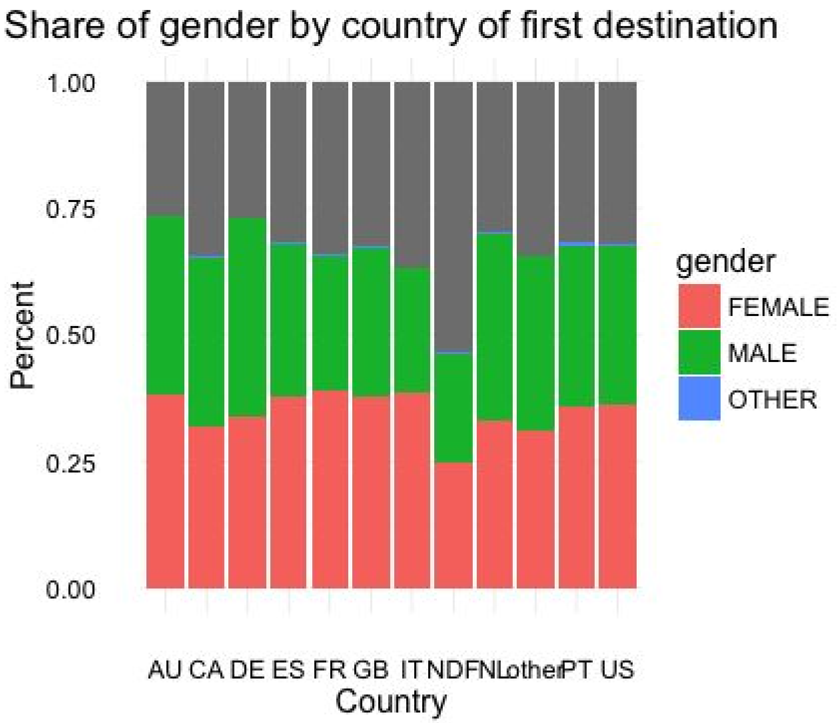
Age/Gender Missingness





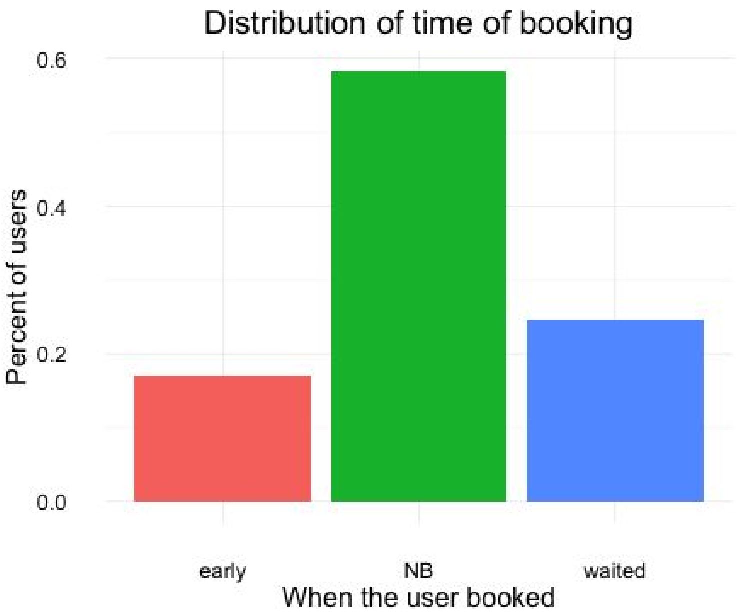
Age & Gender on Country Destination

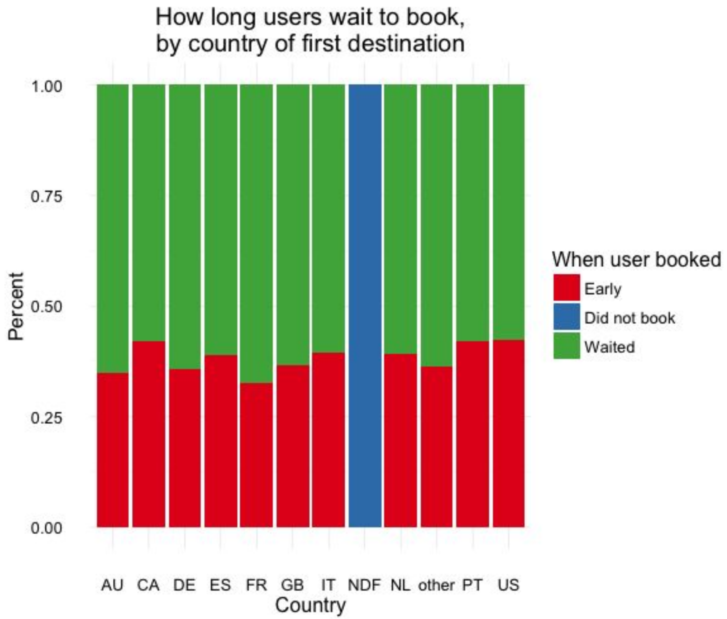




Time variable feature engineering

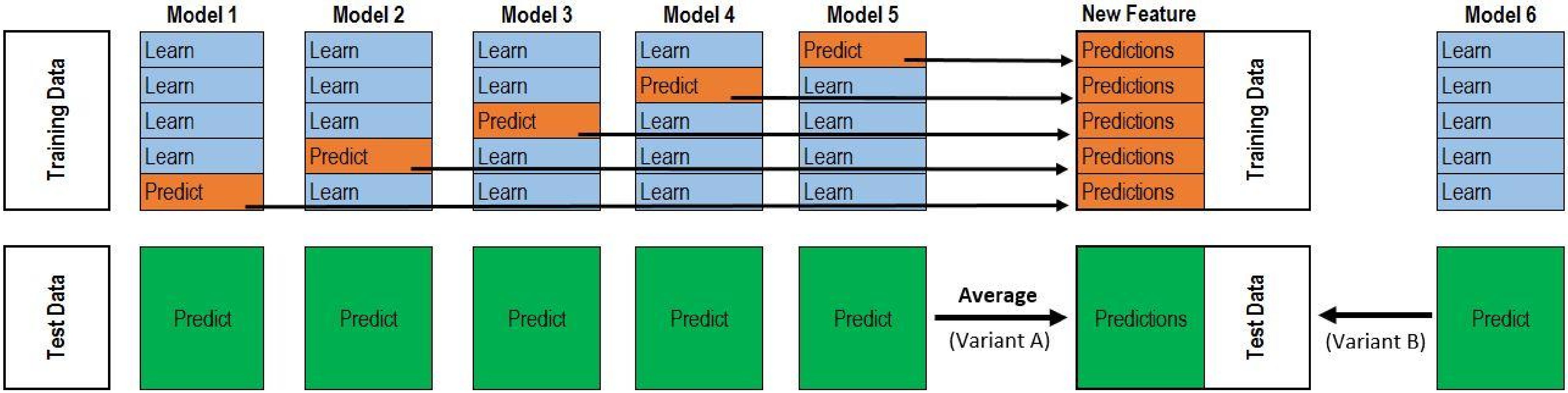
- *We decided to engineer 3 features based on user booking behavior, specifically the time between the creation of Airbnb accounts, a user's first activity on the website, and their date of first booking.*



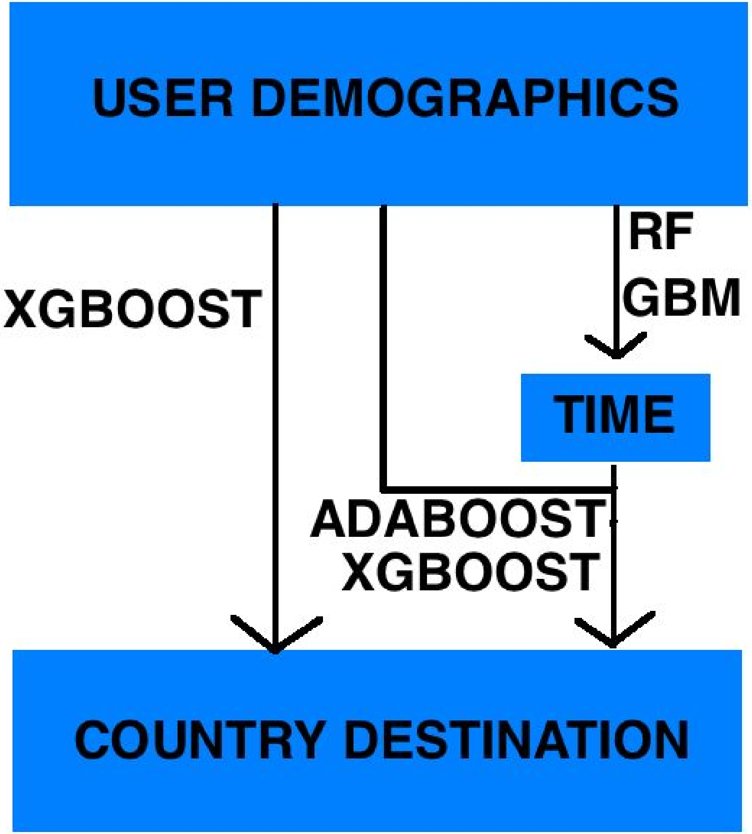


Stacking

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



Workflow


USER DEMOGRAPHICS
XGBOOST
RF
GBM
TIME
ADABOOST
XGBOOST
COUNTRY DESTINATION

Predicting Country Destination

- First choices either NDF or USA.

- Ran grid search cross-validation

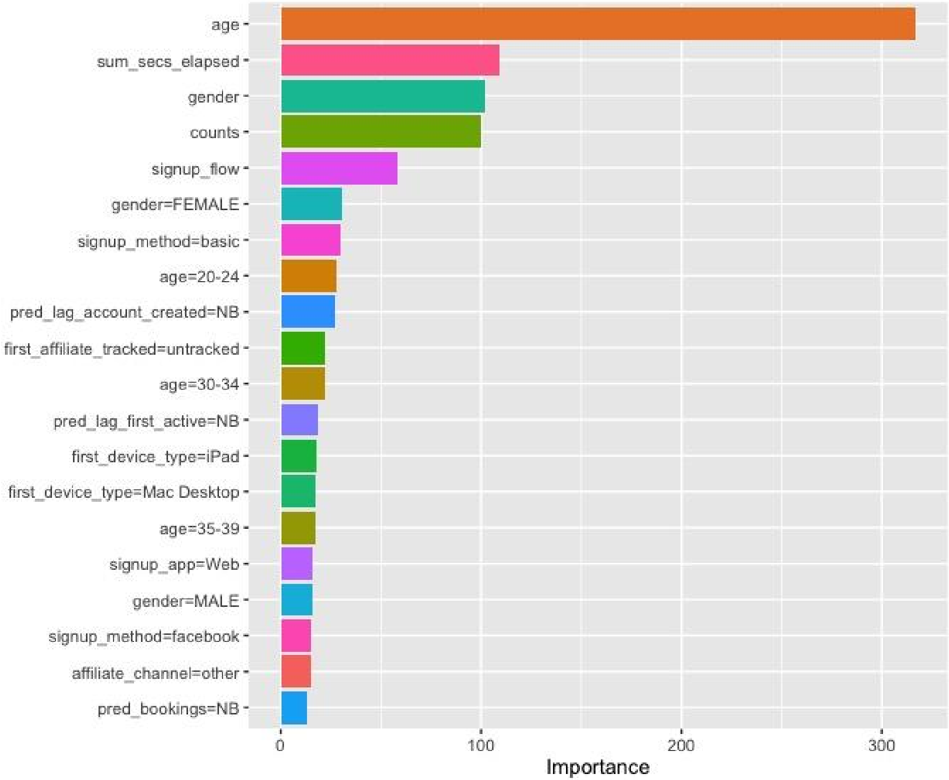
- Unstacked:
    - XGBoost -- Improved Kaggle ranking from #1165 to #390 with a score of 0.87036

- Stacked:
    - XGBoost -- Kaggle ranking of #1030 with score of 0.86332
    - AdaBoost -- Kaggle ranking of #1028 with score of 0.86445

Variables of importance in XGBoost

Unstacked Model

Stacked Model



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 390 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.