Airbnb New User Bookings

Derek Halliwell

General Assembly, DAT SF 18

Question

- Background:
  - Airbnb gives users the ability to rent temporary lodging from other users
  - Rentals can vary from space on a couch to a very large house
  - Airbnb has over 2 million listings in over 35,000 cities and 190 countries
- Where will a new Airbnb user book their first travel experience?

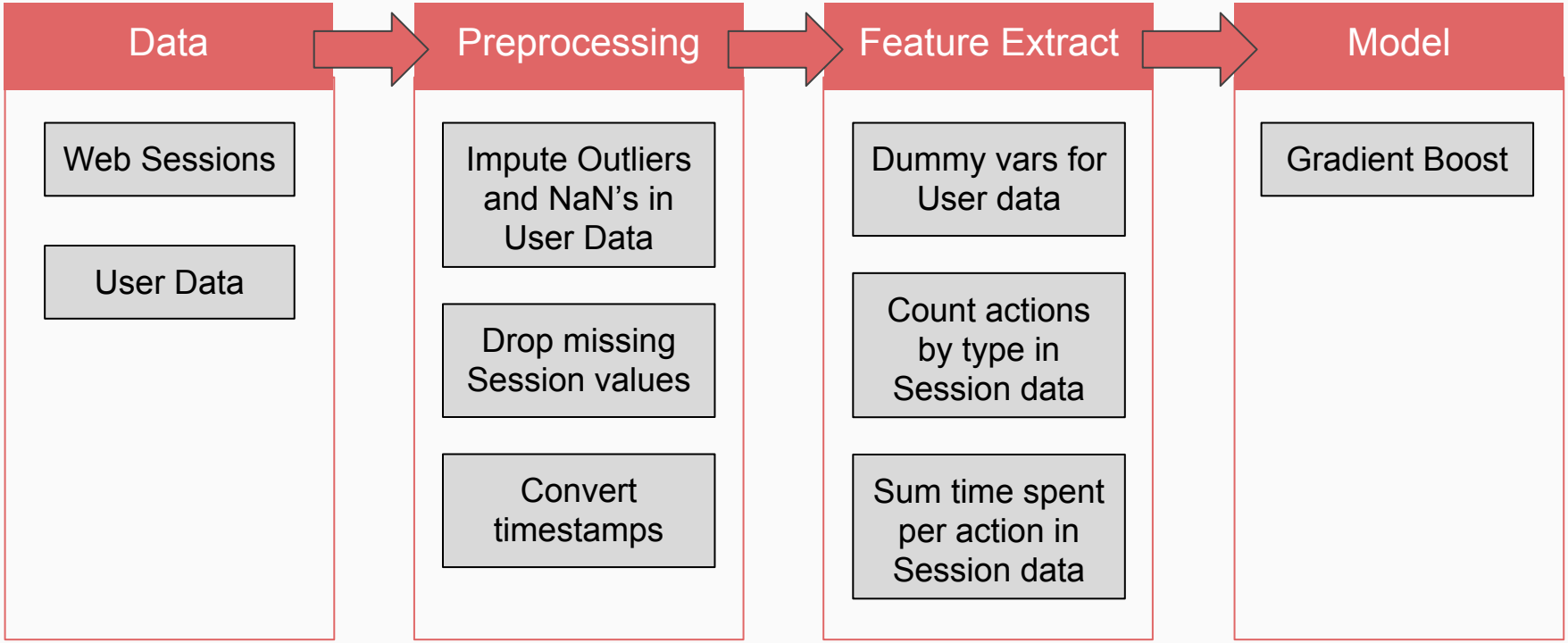
Data

- User Data: user demographics and marketing characteristics
- Session data: Detailed log of web session data (Includes data on clicks, different search results, etc.).
- All Users in dataset live in the U. S.
- A vast majority of this data is categorical

Sessions
user_id
action
action_type
action_detail
device_type
secs_elapsed

User Data
affiliate_channel
affiliate_provider
age
date_account_created
first_affiliate_tracked
first_browser
first_device_type
gender
id
language
signup_app
signup_flow
signup_method
timestamp_first_active

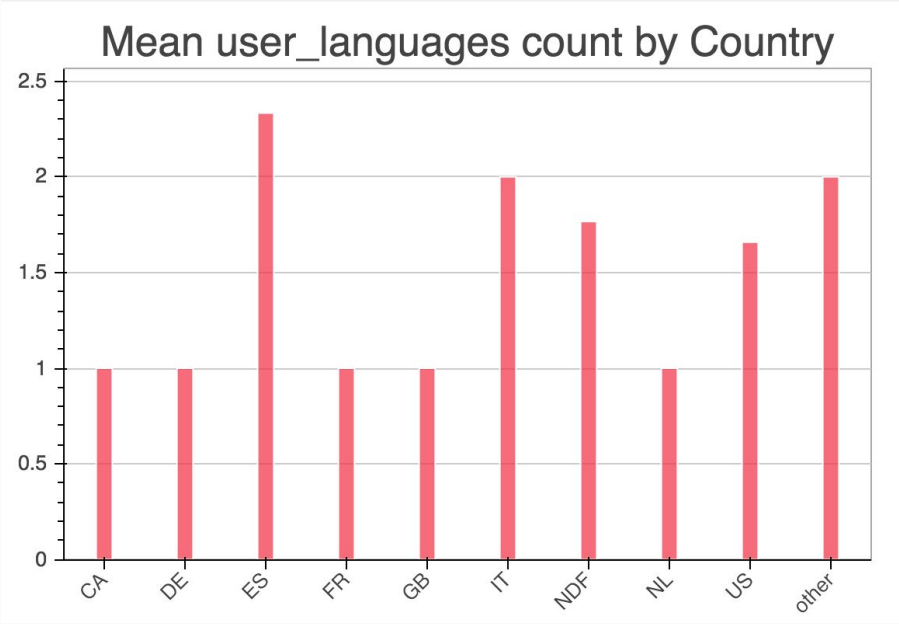
Summary

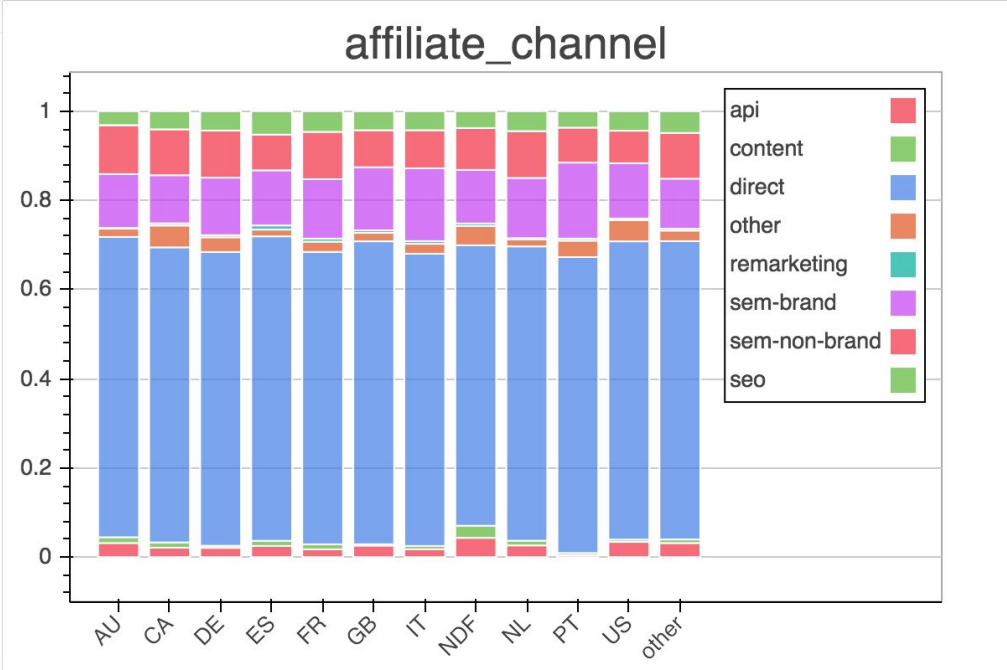

Data
Web Sessions
User Data
Preprocessing
Impute Outliers and NaN's in User Data
Drop missing Session values
Convert timestamps
Feature Extract
Dummy vars for User data
Count actions by type in Session data
Sum time spent per action in Session data
Model
Gradient Boost

Preprocessing

- **User Data**:
- gender: assigned NaN and ‘Other’ gender values as ‘Unknown’
- language: replaced with most common language value
- first_affiliate_tracked, first_browser: assigned as ‘Unknown’
- age: outliers/null values were replaced by average age by browser
- convert timestamp first active to DateTime64
- **Sessions Data**: dropped all NaN and ‘Unknown’ values.
- **Feature Creation**:
- Pivoted Sessions data to get count and total time spent for each action_detail and action_type
- E.g. how much time a user spent looking at search results, and number of clicks during that search

Visualizations





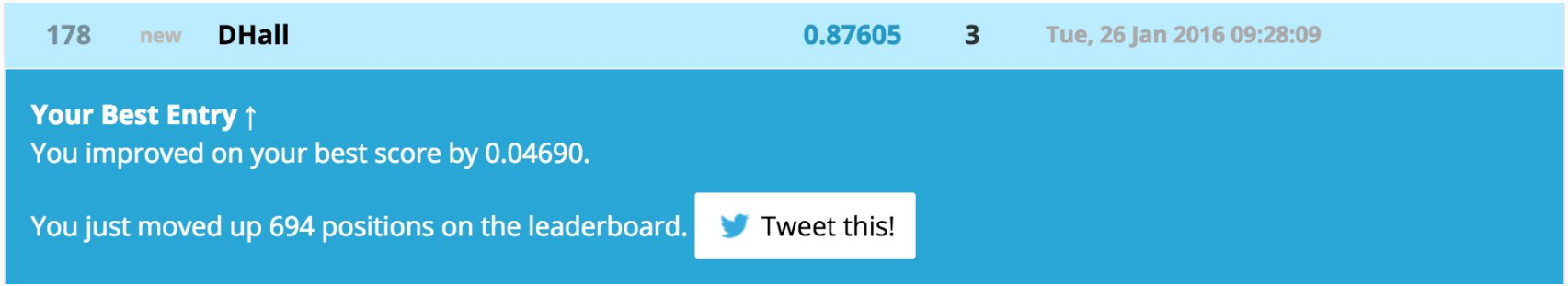
Hypothesis: Prediction accuracy will largely depend on how I extract features from the web session data.

Model

- Extreme Gradient Boosted Classifier (XGBClassifier)
  - Pros: Speed, protection against overfitting compared to sklearn's Gradient Boosted Classifier
  - Cons: "Black Box" algorithm, not very intuitive
- Methodology:
  - Cross validated model on training data to tune parameters
  - Tested final score by submitting to Kaggle

Results

- Kaggle's scoring: Normalized discounted cumulative gain (NDCG)
  - Score is calculated on 5 predicted countries per user, sorted by likelihood
  - Varies from 0.0 to 1.0
- My most accurate XGBoost model scored **0.87605**



Next Steps

- Improved feature extraction
- Fine tune model based on Kaggle’s scoring metric, not just cross validation score
- More submissions!