Airbnb New User Bookings

Derek Halliwell

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?

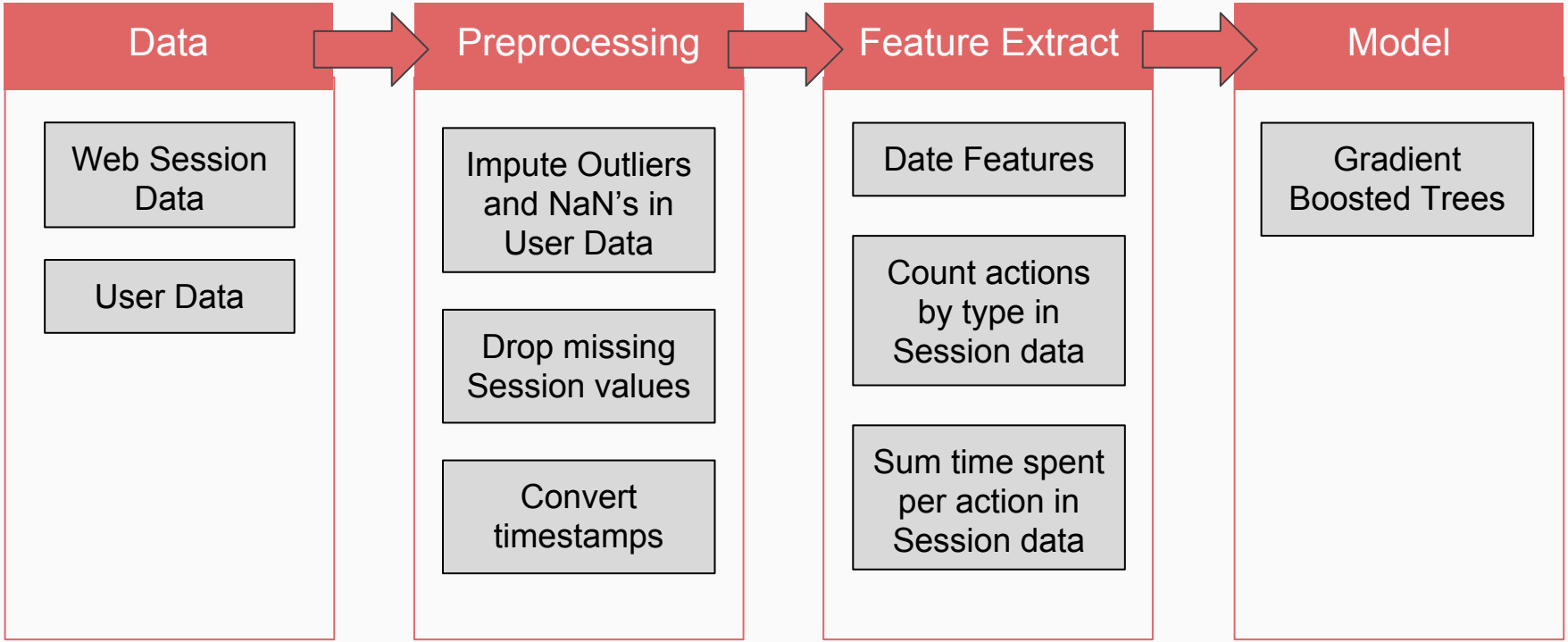
Data

- **User Data**:
  - User demographics and marketing characteristics
- **Session data**:
  - Detailed log of web browsing behavior (searches, emails, booking requests, etc.) including time spent on all actions
- All Users in dataset live in the U.S.
- Most of this data is categorical
- Target Variable: Country Destination

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 Session Data
User Data
Preprocessing
Impute Outliers and NaN's in User Data
Drop missing Session values
Convert timestamps
Feature Extract
Date Features
Count actions by type in Session data
Sum time spent per action in Session data
Model
Gradient Boosted Trees

Preprocessing

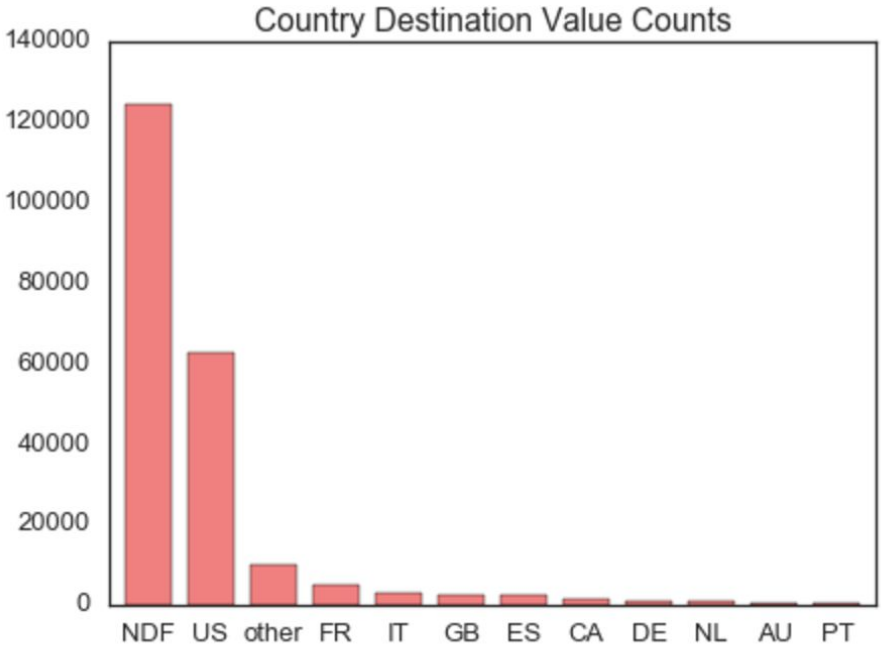
User Data:

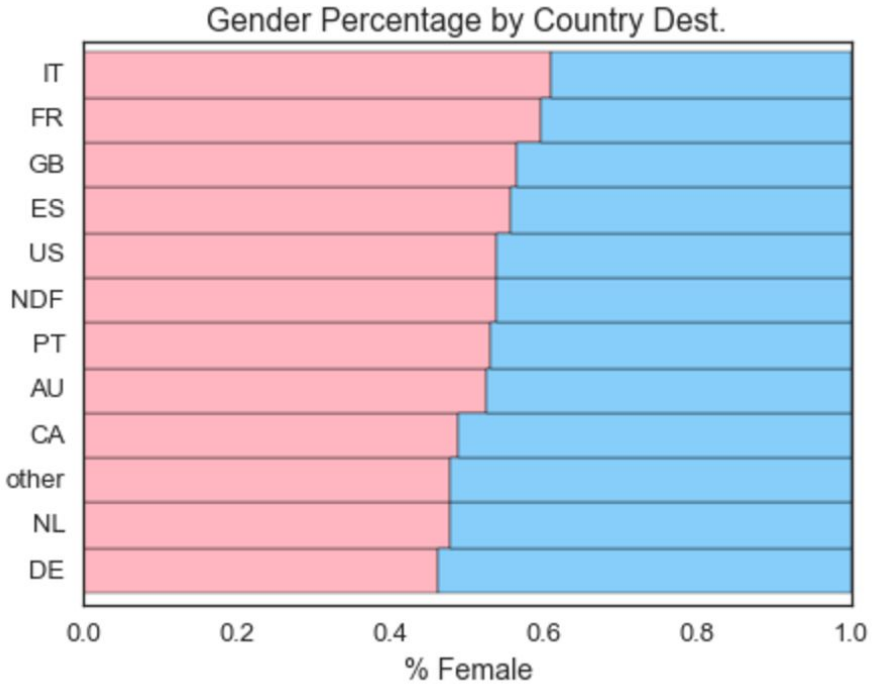
- gender: assigned null and ‘Other’ gender values as ‘Unknown’
- language, first_affiliate_tracked: replaced with most common value
- first_browser: assigned as ‘Unknown’
- age: outliers/null values were imputed with regression trees

Web Session Data:

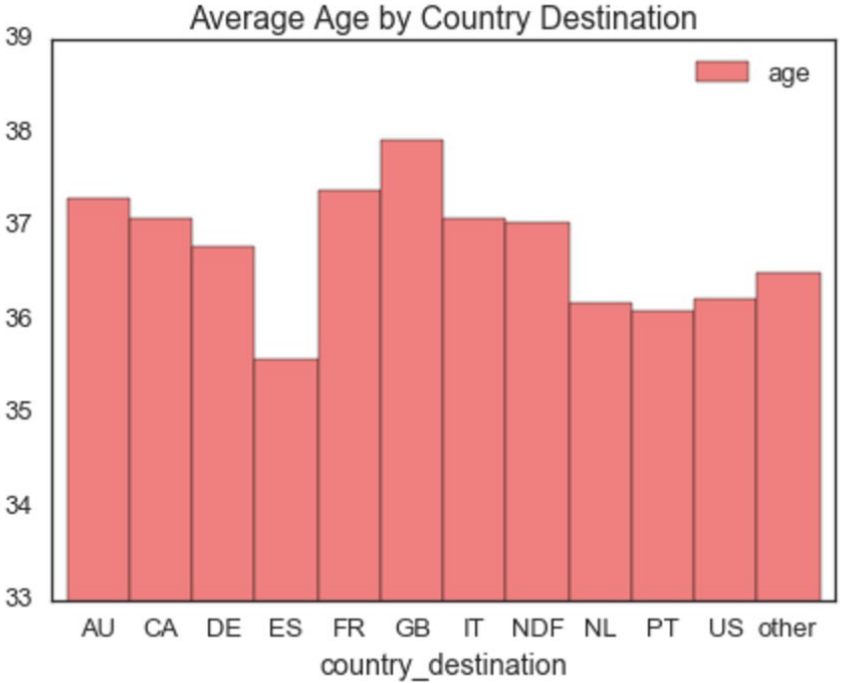
- Dropped all null and Unknown values.

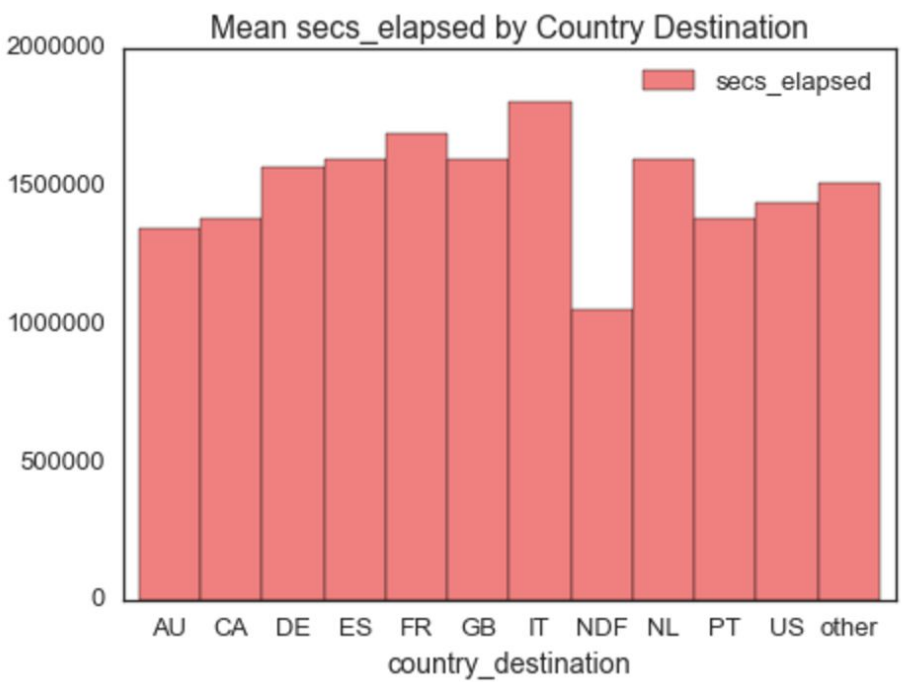
Data Exploration



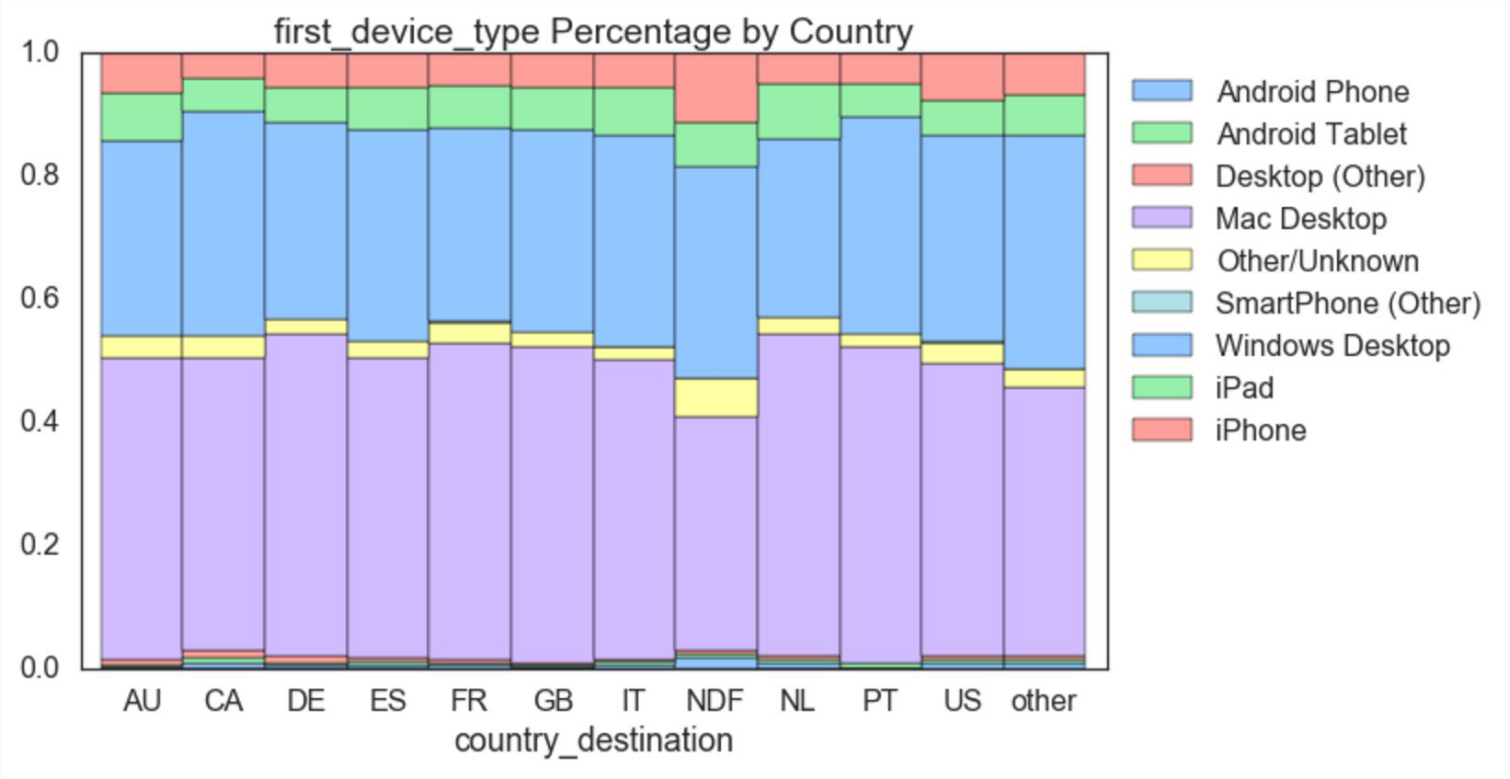


Data Exploration

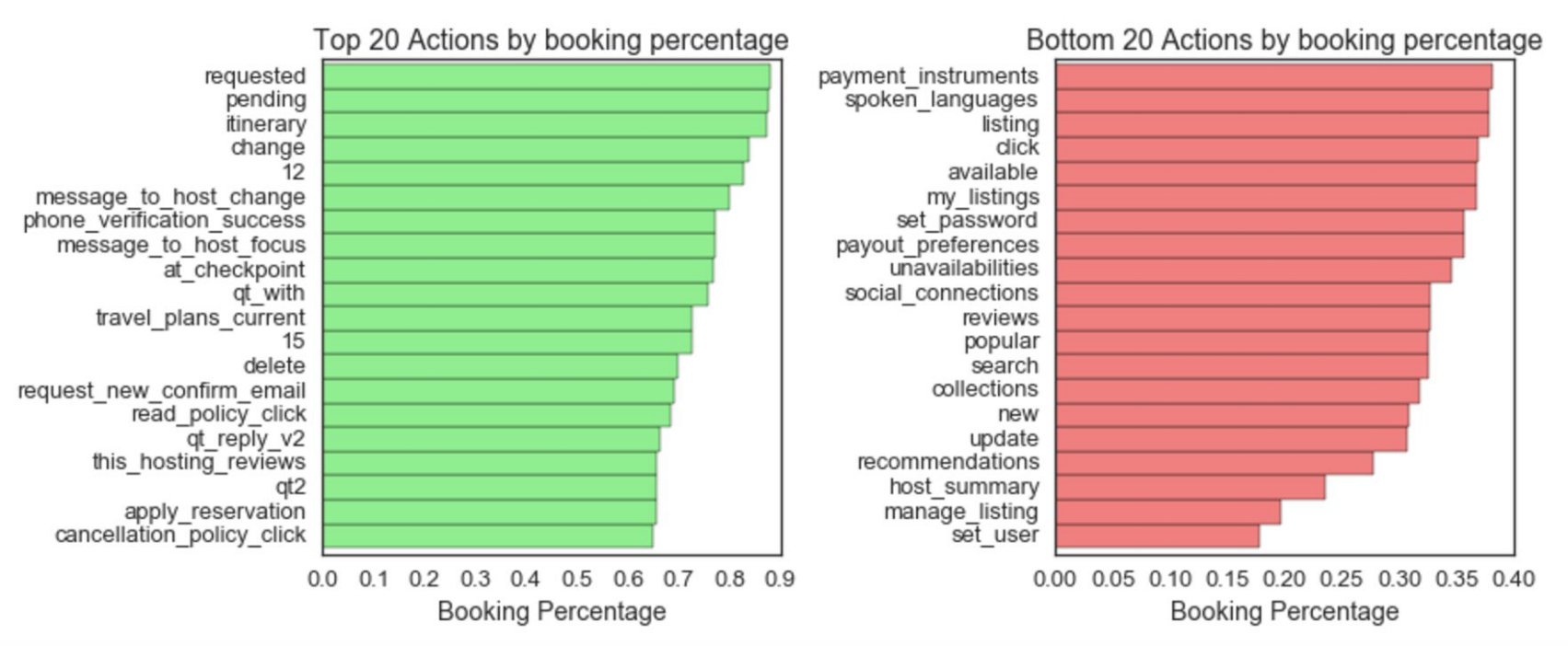




Data Exploration



Data Exploration

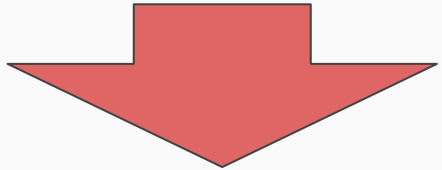


Hypothesis: Prediction accuracy will depend on how I extract features from the web session data.

Feature Extraction

- 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/views during that search

	user_id	action	action_type	action_detail	device_type	secs_elapsed
1	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	67753.000
3	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	22141.000
5	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	7703.000
7	d1mm9tcy42	personalize	data	wishlist_content_update	Windows Desktop	831.000
8	d1mm9tcy42	index	view	view_search_results	Windows Desktop	20842.000



	len account_notification_settings view	len apply_coupon submit	...	sum view_search_results click	sum view_search_results view
user_id					
00023iyk9l	0	0	...	22079	32712
0010k6l0om	0	0	...	45844	30107

Model

- Two models: Users with Web Session Data and Users without.
- Extreme Gradient Boosted Classifier (XGBClassifier)
  - Pros: Speed, protection against overfitting
  - Cons: “Black Box” algorithm, not very intuitive
- Methodology:
  - Cross validated model on training data to tune parameters

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 **0.88128**, resulting in a 291st place finish out of ~1,500 competitors.