airbnb Classification

Patrick Nieto

Where will a new guest book their first travel experience?

Why?

The Airbnb dataset is the real-world data.

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries.

Accurate predictions:

1. Airbnb can share more personalized content with their community,
2. Decrease the average time to first booking, and
3. Better forecast demand.

FR

CA

US

GB

other

ES

NDF

IT

AU

PT

DE

NL

The Data

The dataset contains a list of users along with their demographics, web session records, and some summary statistics.

Sessions
10,000,000

Users		
ID	Dates	Age
Gender	Signup	Device
Browser	Language	App

200,000

Age Buckets
240

Countries
12

Step 1: Categorical Encoding

Replace the categorical fields in the dataset with multiple columns representing one value from each column.

ID	Gender
1	Male
2	Female
3	Not Specified
4	Not Specified
5	Female

ID	Male	Female	Not Specified
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1
5	0	1	0

Step 2: Feature Extraction

Breakdown single features into multiple sub-features in order to create as many factors as possible.

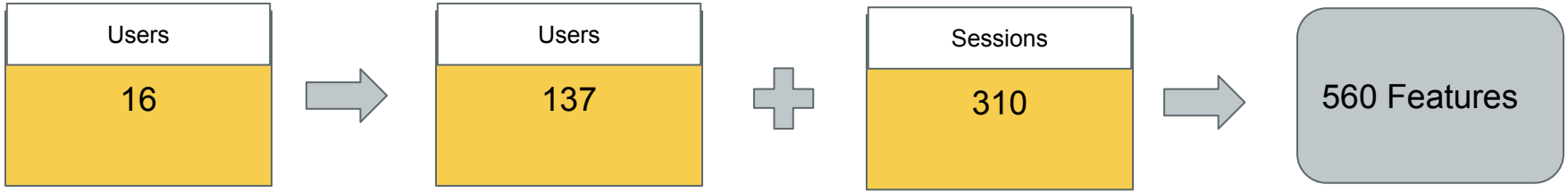
- Convert dates to hours, weekdays, months, quarters, and years
- Calculate lag-time

Step 3: Adding External Data

Expanding existing dataset by adding new data points for a given record.

Sessions Data = Web session logs of 10,000 different users

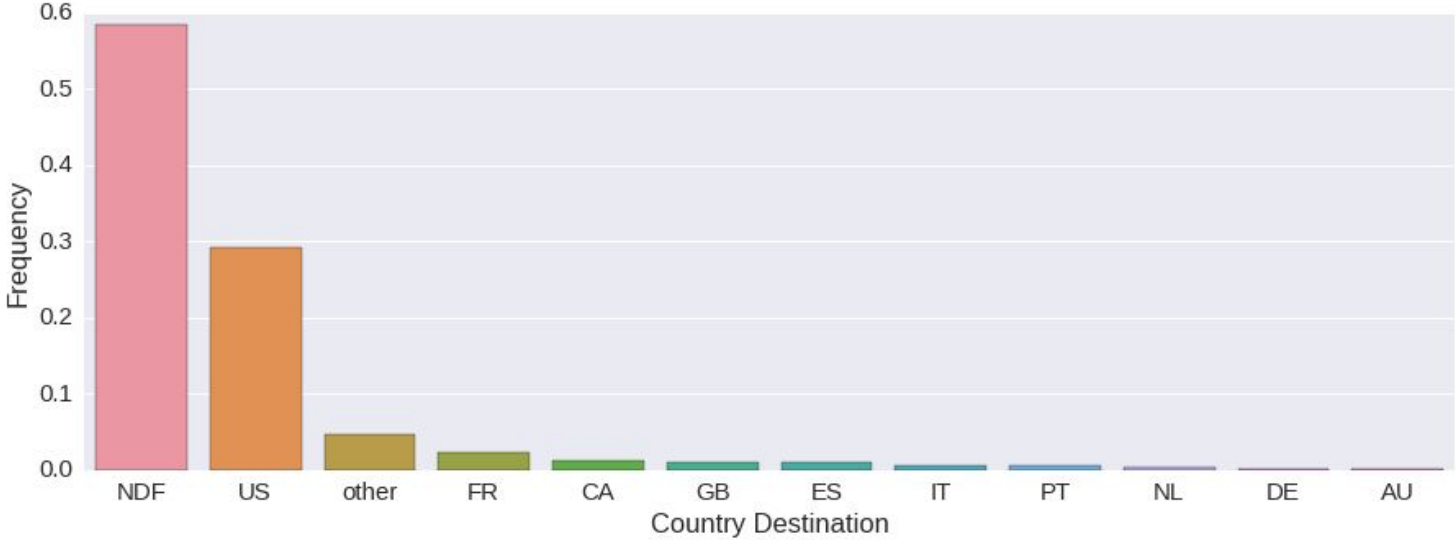
- Seconds elapsed for every action taken on every device
  - Eg. clicked on a listing, updated a wish list, ran a search etc.



Distribution of Classes

Because this is data comes straight from Airbnb, class distribution reflected real outcomes.

- 85% of users stayed within the US or did not travel at all.



Down Sampling

Random Forest (test set)

	precision	recall	f1-score	support
AU	0.00	0.00	0.00	30
CA	0.00	0.00	0.00	88
DE	0.00	0.00	0.00	50
ES	0.00	0.00	0.00	141
FR	0.11	0.01	0.01	287
GB	0.00	0.00	0.00	146
IT	0.00	0.00	0.00	196
NDF	0.72	0.89	0.80	9008
NL	0.00	0.00	0.00	49
PT	0.00	0.00	0.00	17
US	0.52	0.46	0.49	4019
other	0.03	0.00	0.00	731
avg / total	0.58	0.67	0.62	14762

	precision	recall	f1-score	support
AU + NL	0.46	0.56	0.51	80
CA	0.42	0.62	0.51	80
DE	0.61	0.81	0.70	80
ES + PT	0.43	0.47	0.45	80
FR	0.30	0.30	0.30	80
GB	0.55	0.44	0.49	80
IT	0.44	0.38	0.41	80
NDF	0.34	0.38	0.36	80
US	0.23	0.14	0.17	80
other	0.39	0.20	0.26	80
avg / total	0.42	0.43	0.41	800

Random Forest Classifier

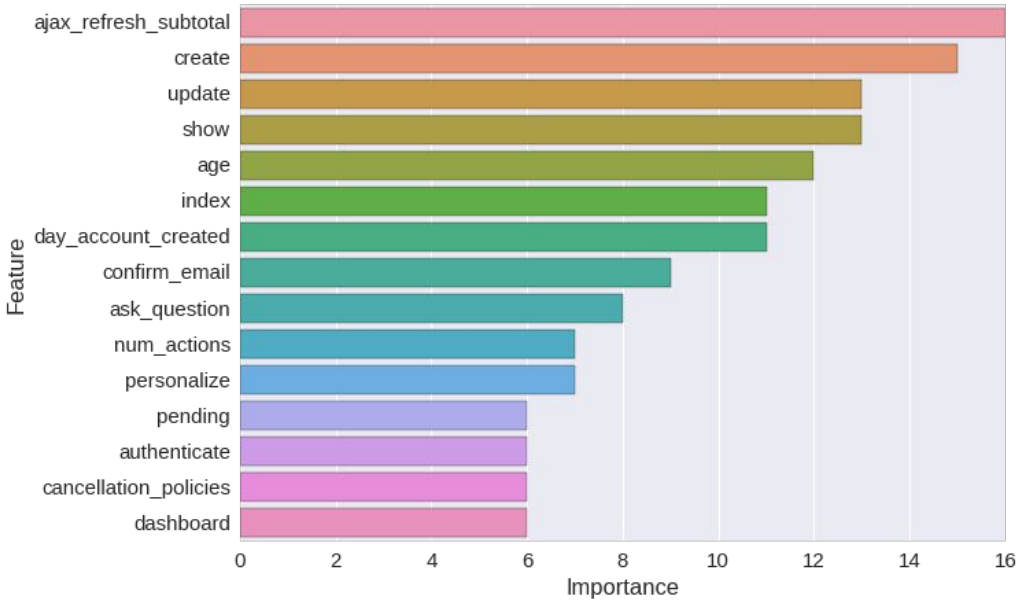
Reporting set

Original
57%

Feature
Extraction
61%

Final
64%

Conclusion



- The distributions for every country on age would allow airbnb to target certain age groups and send personalized content.
- A significant amount of importance is placed on the web session logs

Take-aways

1. Investing time looking for ways to add new and useful data to your existing dataset.
2. Understanding the distribution of your labels

Next Steps

1. Incorporating either logistic regression or nearest neighbors to form better insights and conclusions about features
2. Two step classification
3. State destinations
4. Parameter tuning (grid search)

Thank you!