AirBnB New User Bookings

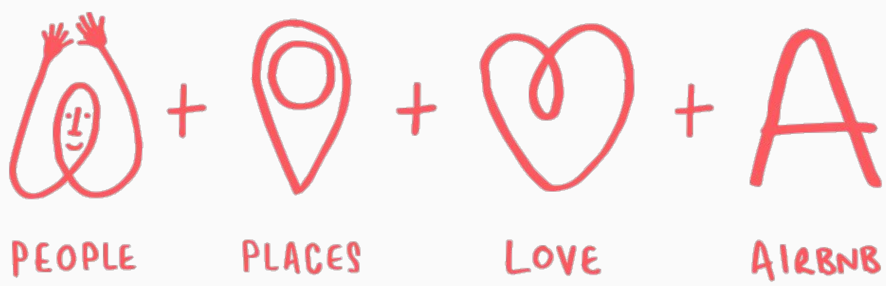
Kaggle Competition, January 2016

Ruben Bloom (rmb042@gmail.com)

Code and Detailed Technical Report: **https://github.com/darkruby501/AirBnBRecruitingComp**

Overview

1. Intro: The Challenge
2. The Data
3. Data Exploration
4. Predictive Modelling
5. Data Insights
6. Recommendations

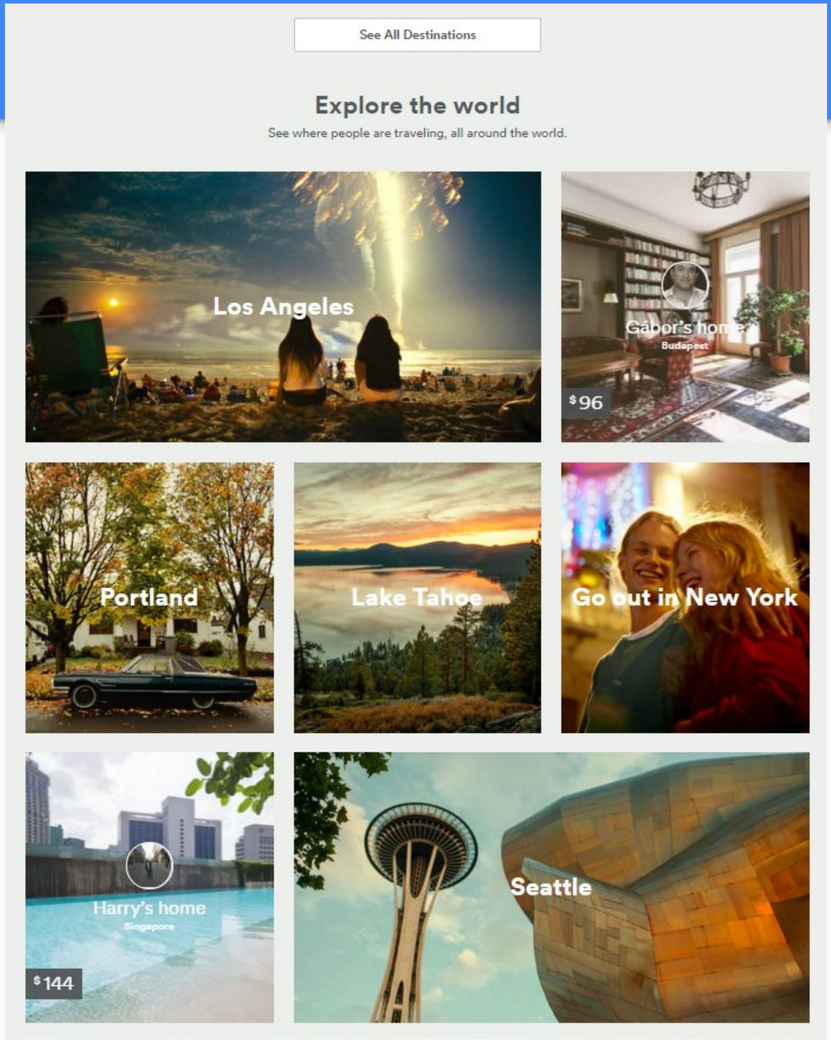


Intro: The Challenge

AirBnB wishes to increase bookings and improve user experience through customisation of content, e.g. personalised landing pages with suggested destinations (right).

AirBnB has challenged the Kaggle community to predict where a new user will book using demographic and web session data.

There are 12 possible destinations and participants must provide a list of the five most likely destinations for a user, in order.



AirBnB Homepage: Suggested Destinations

The Data

The following data was provided.

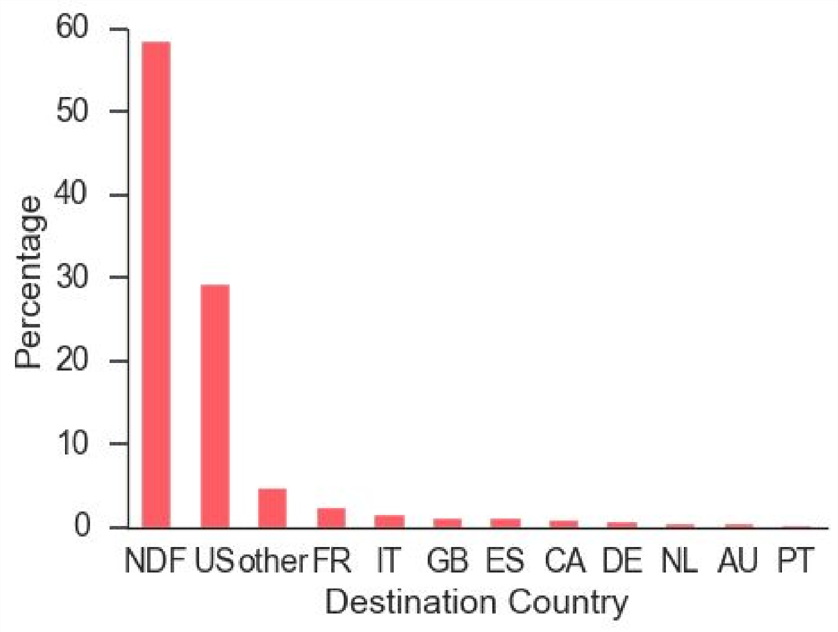
Demographic: age, gender, sign-up method, device used, browser used, etc.

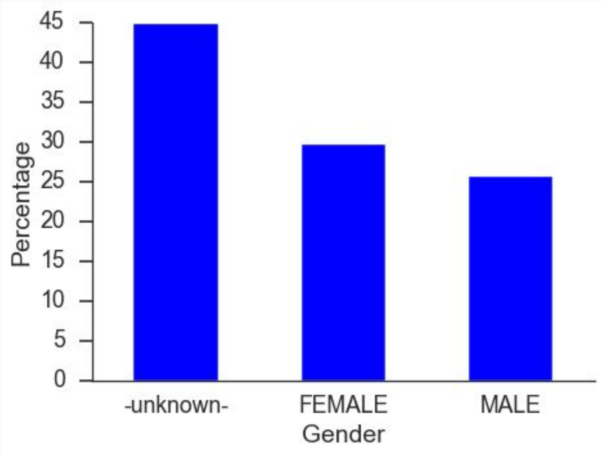
Web Sessions: sequences of actions taken on the website, e.g. click, view results, request booking, verify email.

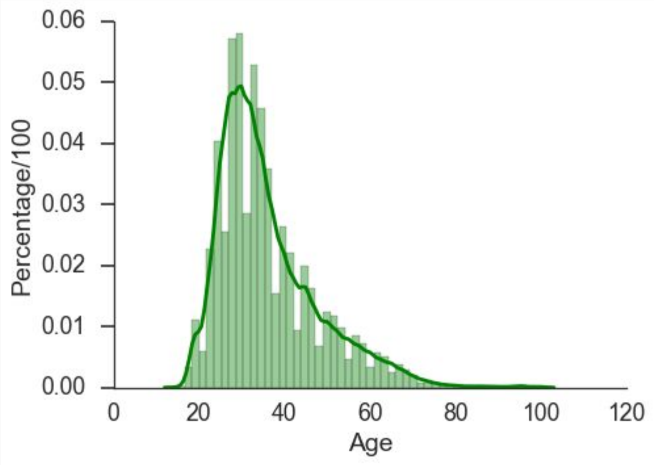
To use this data, appropriate processing was performed: cleaning missing and invalid values, joining records, scaling and normalisation, creation of usable features from unstructured session data.

Data Exploration

Some quick data visualization.







Most new users do not book! Most of those who do book, book domestic.

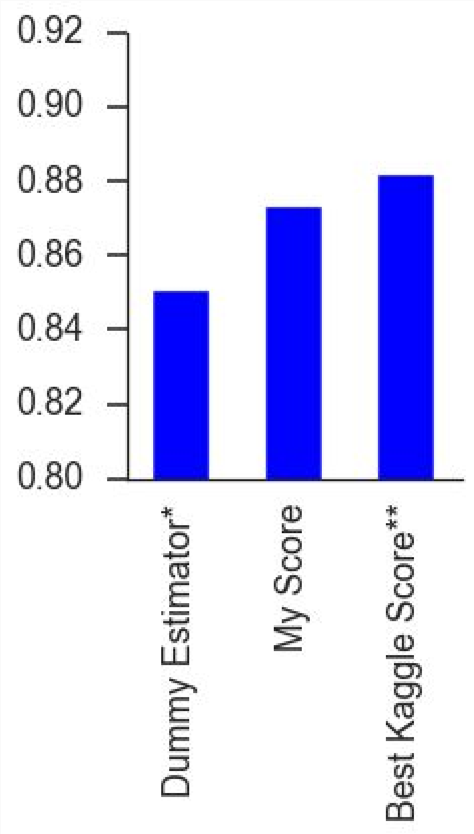
Extreme class imbalance is clear.

Predictive Modelling

Machine Learning Models attempted for this challenge: Logistic Regression, Random Forest, XGBoost.

Dummy Estimator*:	0.850
My Current Score:	0.873 (top 15%)
Kaggle Current Best Score**:	0.882

Are these scores good? No. The measure used is NDCG†, but equivalent accuracy is only ~70%. No one is outperforming a dummy estimator by very much.



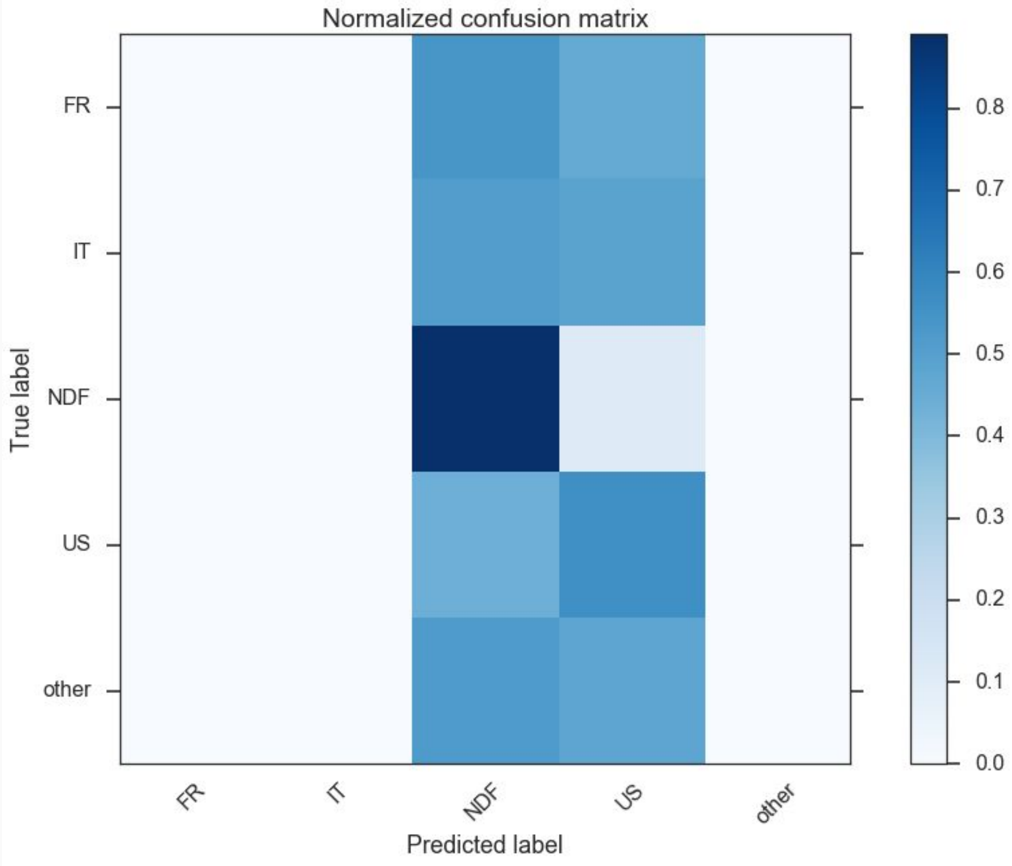
*Predicts solely based on class frequencies.
**As of 1/18/16, competition closes on 2/11/16
† Normalised Discounted Cumulative Gain

Predictive Modelling: Classification Report

What's going on?

Looking at the Confusion Matrix, all predictions are for US or NDF. Other countries are never predicted. Accuracy for US vs NDF is also poor, 44% of US bookings are misclassified as no booking.

1) Why can't the classifier distinguish non-US countries?
2) How does it tell the difference between booking and no booking?



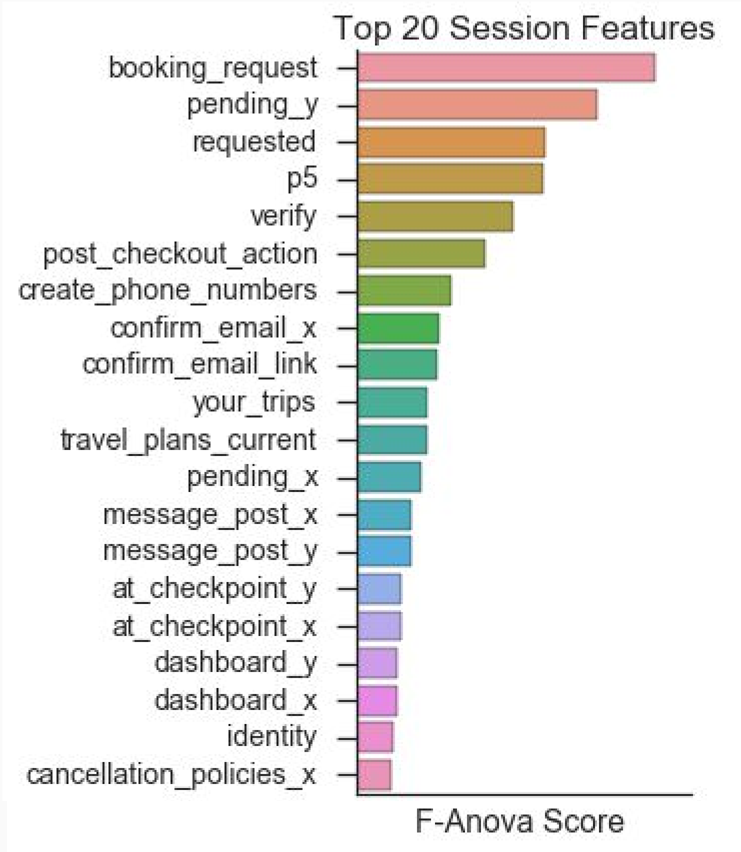
NDF = "No Destination Found", i.e. no booking

Data Insights: Feature

Digging in, we find that the most informative features include:

- Presence of “Booking Request”, “Travel Plans Current”, “Pending”, etc. in web session log.
- Sign up methods, e.g. direct vs Facebook
- Filled in profile, i.e. not missing age and gender
- Desktop vs Mobile (mobile conversion is lower)
- Mac Desktop (more likely to book)

These features would not differentiate bookings between different countries.



Differences between Booked and Didn't

Furthermore, the presence of "Booking Request" and "Travel Plans Current" are not conclusive.

- 13% of users who have make a Booking Request do not book.
- 33% of users who *do not have* a Booking Request make bookings (how?).

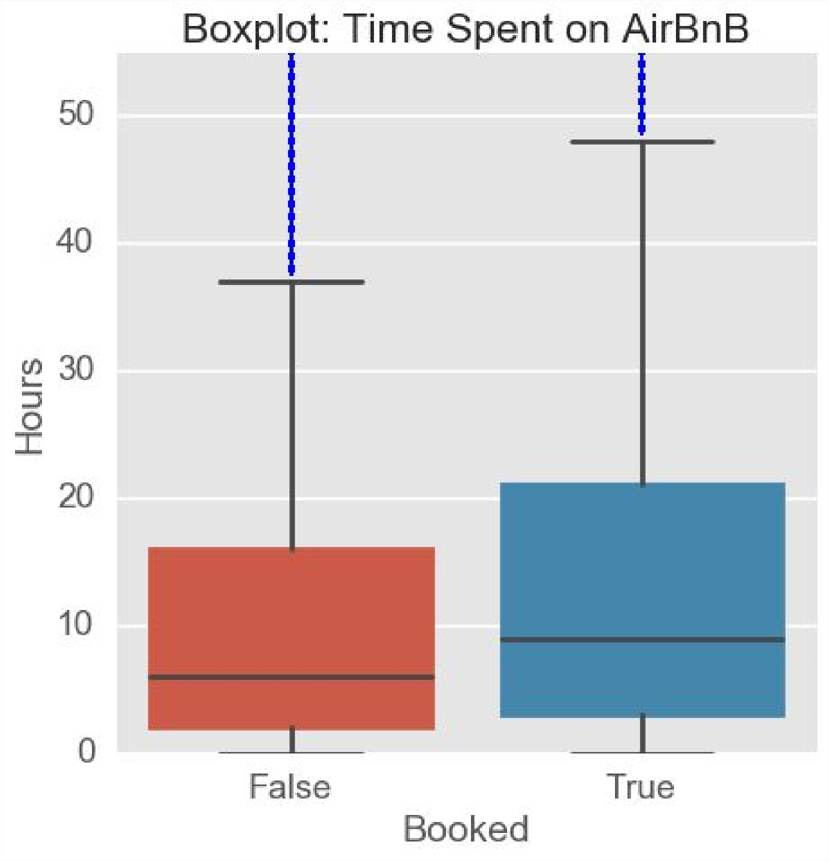
Other strongly correlated features are not conclusive!

There is significant attrition even in the lower funnel!

More Differences

Time spent on AirBnB.com shows a similar story:

- On average, users who book spend longer on the site.
- But the distributions overlap heavily.
- Therefore time spent on the site is weak evidence.

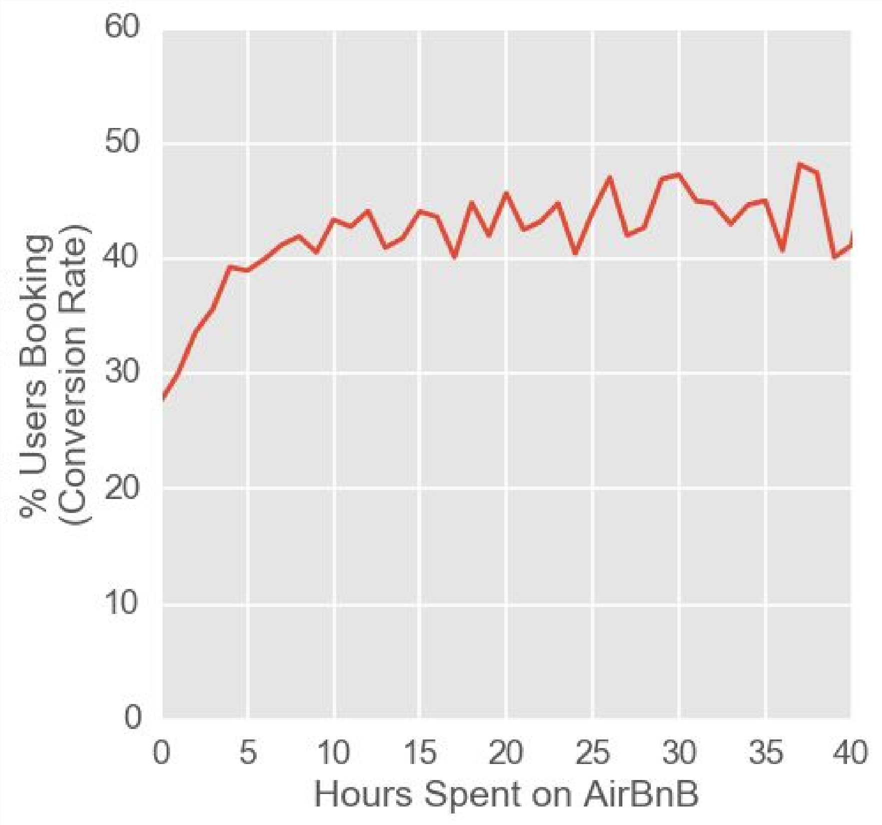


More Differences

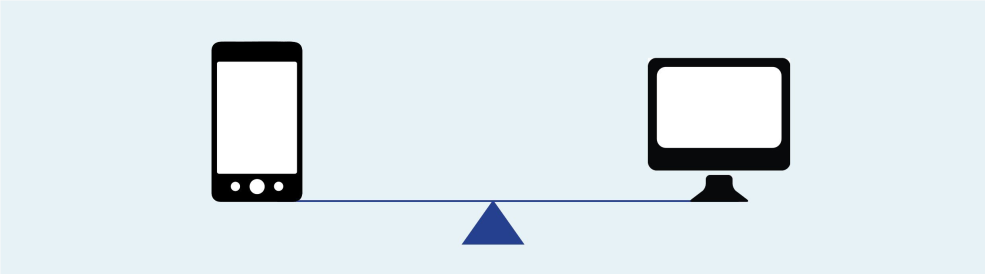
Equivalently:

After the first 10 hours, more time spent on the website does not increase the likelihood of booking.

The features in the dataset do not strongly separate people who book from those who do not. We should not expect high accuracy from a predictive model.



A Word on Data Quality



The data is not perfect!

Plagued by the “many screens problem” which makes perfect tracking impossible, e.g.

- Browse on work computer, book on home computer
- Browse on mobile, book on desktop
- Husband browses, wife books

Much of the data will be misleading. This is noise the machine learning model cannot beat.

Data Insights: Technical Conclusions

Back to the original goal.

We have many reasons to conclude that the data provided is not sufficient to accurately predict destination country. In addition to the above:

Technical Reasons
- The model doesn’t get better with more examples.
- More complicated ML models do not perform better.
- Techniques for handling class imbalance such sampling techniques do not help.

Domain Understanding
- We do not expect age, gender, device preferences, and browsing behavior to distinguish people who travel to Italy to those who travel to France, and and so on!

Final Recommendations

- To meet the original goal of predicting destination countries better than class priors, a richer dataset is needed.

- With the current data, it is possible to predict whether users will Book or Not Book with some accuracy. Currently all users are grouped together, but more useful predictions based on Sales Funnel position may be possible.
- For example, create a project to:
  - Predict which users will bounce within the first hour of browsing.
  - Predict which users who have requested bookings will not convert.
  - Target with appropriate interventions.
- AirBnB should attempt to quantify the "multiple screens problem"