Airbnb Recruiting : New User Bookings

Where will a new guest book their first travel experience?

Nabil Abdellaoui
February 2016

The Challenge

User information

+ Age
+ Gender
+ Affiliate channel
+ Device and browser
+ Session log (clicks and times spent on Airbnb website)
- No contents
- No search queries
- No geographic or social details
- No dates or stay durations

Machine Learning

Learn relations

Classify users

Detect patterns

Produce insights

Predictions

Where will a new guest book their first travel experience?

Is it possible to predict it without looking at his search queries or viewed content ?

Who is this visitor ?

Intuitions

- A user who makes bookings rarely is more likely to travel far away, while a frequent traveler may be making shorter distances.

- Someone who connects from desktop, phone and tablet, and uses the website faster, is more likely to have business purposes rather than tourism.

- Someone who searches the website, comes back after one week, searches again for a few days, and takes many days to book, may be booking some unusual trip, far away dream destination?

- People of different ages will have different preferred destinations.

- Someone who is booking few days or weeks before Thanksgiving or Christmas is preparing some family trip.

I want to book a place in …

The sequence of actions taken by a user, and his browsing pace, are like a signature .

Graphologists pretend that they can deduce a psychological profile from handwriting patterns.

How about website browsing patterns ?

Feature Engineering (1)

- One Hot Encoding and some binary indicators : "age between 35 and 44" , "asian language" , "latin language" , "first device is tablet" , etc.

- Population in the same age/gender bucket (used the table age_gender_bkts to add one feature by country)

- Number of different devices that appear in the sessions log : users who use multiple devices are frequent travelers, maybe for business ?

- timeBeforeConfirmEmail and timeBeforeVerify : these two actions appear often in sessions logs, more than 5000 times, and appear in general only once (average frequancy in sessions < 1.2) they may help predict the time taken by the user to book after his first connection.

- Different counting of the actions : distinct different actions, total number of events logged, counts by action type and by action.

- Percentages of events by action type : could define a kind of user experience.

- Percentages of events by device type : could give some information about leisure vs. business kind of user.

- Time before actions : maybe some actions done at a certain point in time will provide useful information.

- Sequences of actions : after transforming “time elapsed” from numerical to categorical, I detected all sequences of 5 events with more than 100 occurrences both in train and test sets , then defined 734 binary features indicating if one of those sequences appears or not in the user log.

- Cluster number (one hot encoded) in a K-Means clustering with K=20.

Feature Engineering (2)

- Total times by user : In addition to the total elapsed time for a user in his session log, added 5 subtotals representing the following categories :

  A. Total of times when less than 60 seconds, which represent "live" actions.

  B. Total of 1 to 10 minutes actions, which are probably actions stopped at to read carefully or moved to another task before continuing.

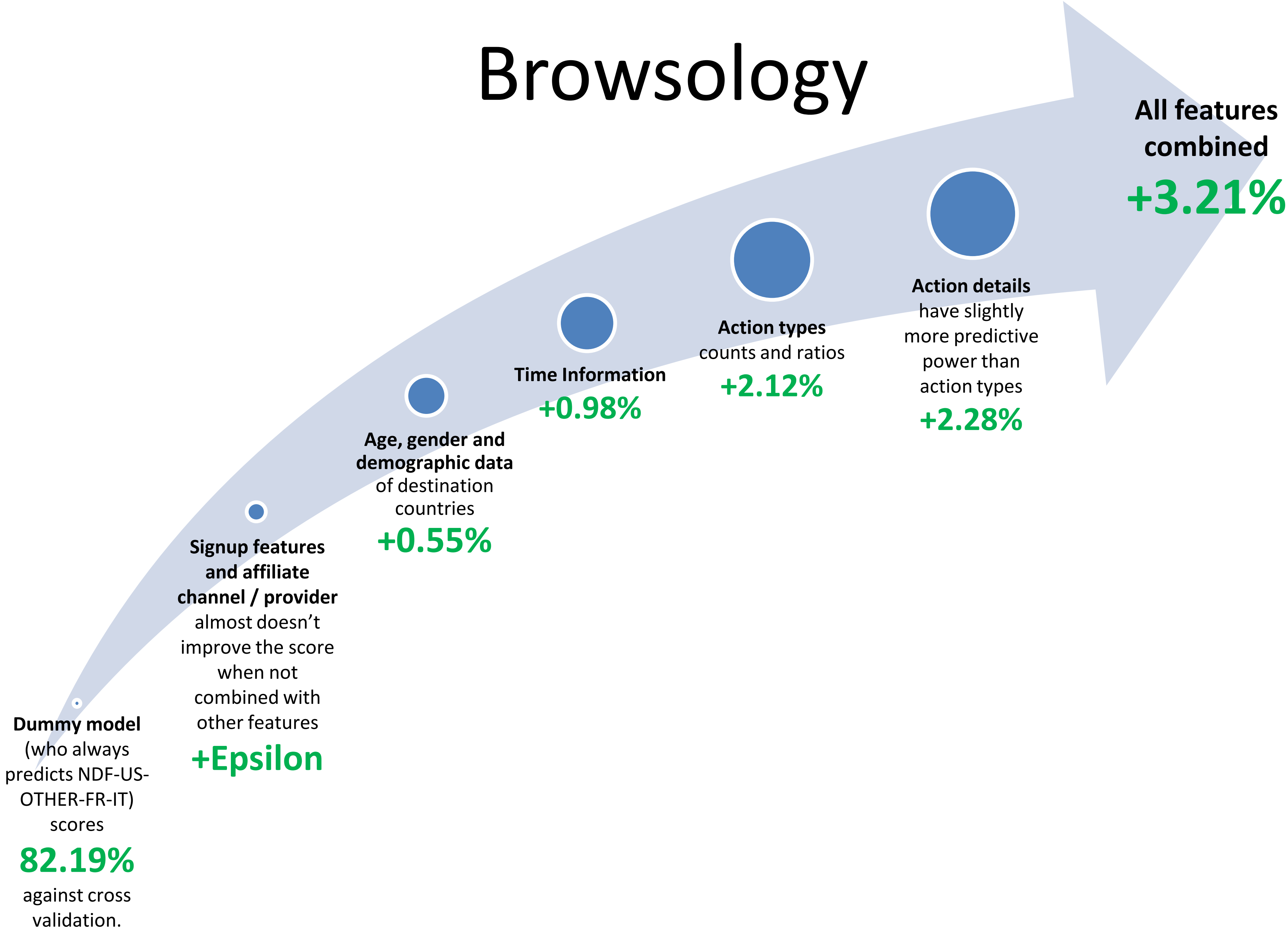
  C. Total of 10 minutes to 2 hours actions, the user must have left the page in these cases, but probably came back to follow the same process.

  D. Total of 2 to 24 hours actions, similar to C category but with a longer span.

  E. Total of times superior to one day, where the user might have came back an other day to start a different process.

- Based on previously defined thresholds, I computed the number of subsessions and their average duration and separation, plus some other statistics about the longest subsession of a user. (After trying models and analyzing these features, it seems that session data doesn't cover all of the user's actions, which probably makes these features erroneous.)

Browsology



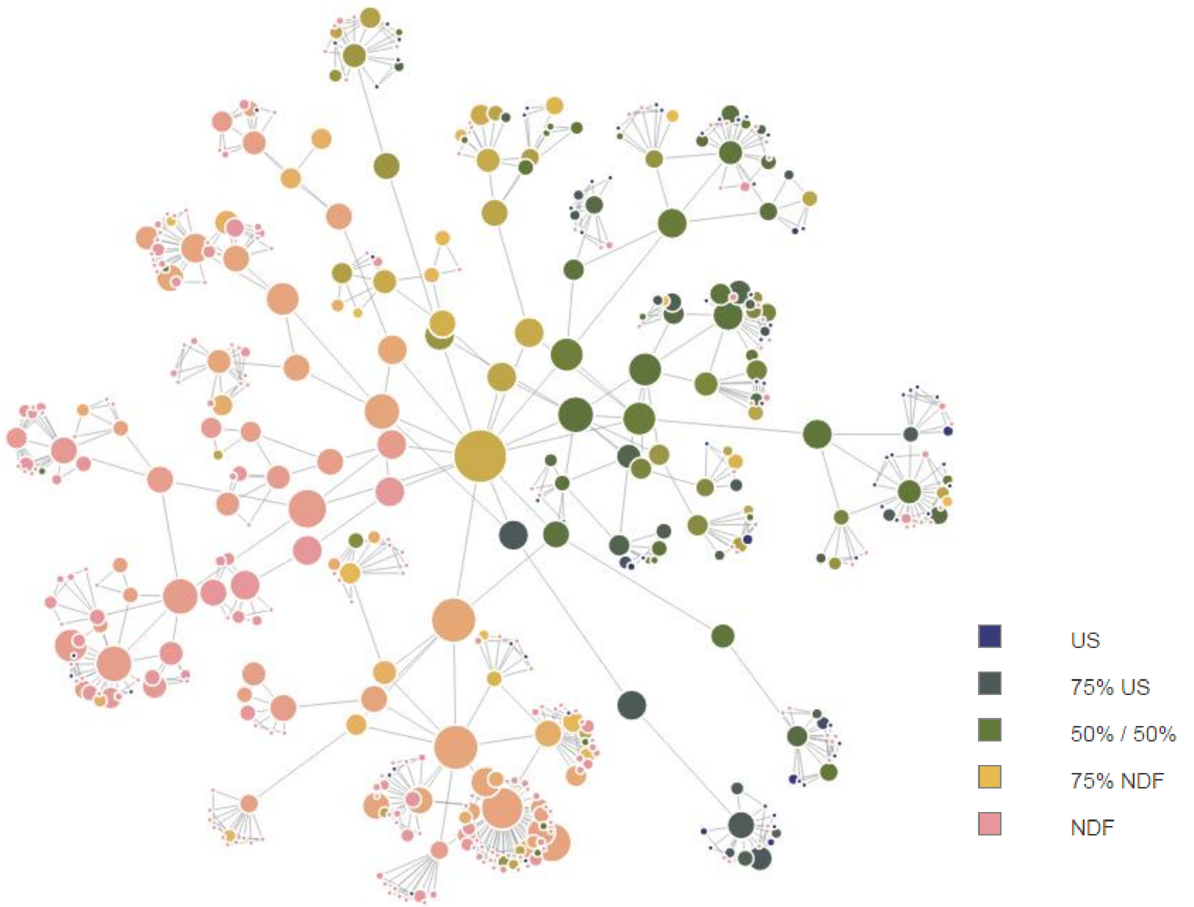
Dummy model (who always predicts NDF-US-OTHER-FR-IT) scores
82.19%
against cross validation.
Signup features and affiliate channel / provider almost doesn't improve the score when not combined with other features
+Epsilon
Age, gender and demographic data of destination countries
+0.55%
Time Information
+0.98%
Action types counts and ratios
+2.12%
Action details have slightly more predictive power than action types
+2.28%
All features combined
+3.21%

User visualization

- Clustering of users based on cosine similarity between sessions

- The centers of the clusters represent "prototypes" of users with significant differences when it comes to booking or not booking.

- **The clustering algorithm doesn't see the destination countries, the structure and separation of classes results only from user features.**



http://randombishop.github.io/aerograph-front/

Modeling and ensembling

-I tried XGBoost, H2O Deep Learning, Random Forests and Aerosolve.

- Features were selected by greedy forward search.

- After reaching single models ceiling, I tried submitting 5 different variations of XGBoost models per day and built a majority vote ensemble based on the three most different amongst the 50 best ones. This approach yielded a score of 0.88591 in private leaderboard, ranked #54.

#	Team Name	Score
1	Anupam Pandey	0.88697
2	Keiku	0.88682
3	Sandro	0.88670
4	SK	0.88659
5	Branden Murray	0.88657
6	SRK	0.88655
7	SkyLibrary	0.88653
8	lionfishy	0.88651
9	pxk	0.88651
10	renman	0.88648
35	Adhir Badul	0.88609
36	Bikash Agrawal ‡	0.88609
37	George	0.88608
-	**Randombishop**	**0.88608**
	My best single model is XGBoost with logloss minimization objective trained for 270 iterations (about 5 minutes of training)	
38	波波头一头	0.88607
39	Anonymous 12673	0.88607
40	#1 OVERFITTA!!!111oneoneone	0.88606

Learning from this challenge

- There is a weak signal in user features and session data, but it is worth studying as it produces significant information gain (around + 3%)

- For this particular dataset, single 5 minutes XGBoost performs almost (-0.1%) as well as complex multi layer stacked ensembles requiring many hours of training.

- Similarity based clustering and graph visualization provide an interesting way to explore the data and discover features' effects.

- Majority vote ensemble of three most different models produces interesting results, but was not enough to make significant improvements.

- I adopted a wrong model selection method, dropping from #8 to #54 on private leaderboard. A better selection strategy would have been to trust my local cross validation more, with best single model scoring 0.88608 and ranking #38. More generally, this challenge was a good exercise of hyper-fine-tuning, where 0. 01% improvements have to be found with very high resolution cross validation. I mistakenly dropped ensembles with neural networks and random forests because the improvement didn't show clearly by cross validation, and was never confirmed by LB score.

- If I had developed higher resolution cross validation, I would have done better feature selection and model ensembling.