Kaggle Competition: Airbnb New User Bookings

Alper Halbutogullari, January 23, 2016

Overview

The problem

In this Kaggle Competition, Airbnb is challenging participants with the question: "Where will a new guest book their first travel experience?". Participants are required to predict in which country a new user will make his/her first booking.

Why is this important?

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

What is in it for Airbnb?

Participants who impress Airbnb with their answer (and an explanation of how they got there) will be considered for an interview for the opportunity to join Airbnb's Data Science and Analytics team.

Data

In this challenge, the participants are given a list of users along with their demographics, web session records, and some summary statistics. All the users in the given dataset are from the USA.

There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination/booking found), and 'other' (meaning there wasa booking, but is to a country not included in the list).

The data provided has the following files and fields:

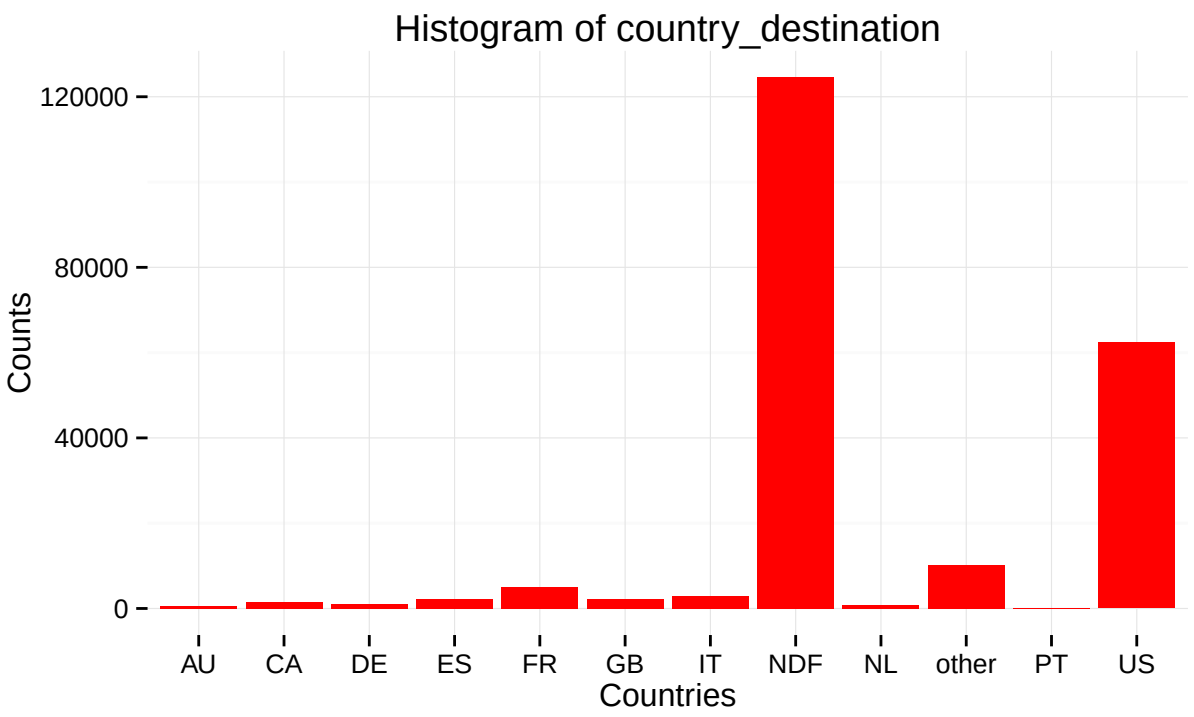
- train_users.csv - the training set of users
- test_users.csv - the test set of users
  - id: user id
  - date_account_created: the date of account creation
  - timestamp_first_active: timestamp of the first activity, note that it can be earlier than date_account_created or date_first_booking because a user can search before signing up
  - date_first_booking: date of first booking
  - gender
  - age
  - signup_method
  - signup_flow: the page a user came to signup up from
  - language: international language preference
  - affiliate_channel: what kind of paid marketing

  - affiliate_provider: where the marketing is e.g. google, craigslist, other
  - first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
  - signup_app
  - first_device_type
  - first_browser
  - country_destination: this is the target variable you are to predict
- sessions.csv - web sessions log for users
  - user_id: to be joined with the column 'id' in users table
  - action
  - action_type
  - action_detail
  - device_type
  - secs_elapsed
- countries.csv - summary statistics of destination countries in this dataset and their locations
- age_gender_bkts.csv - summary statistics of users' age group, gender, country of destination
- sample_submission.csv - correct format for submitting your predictions

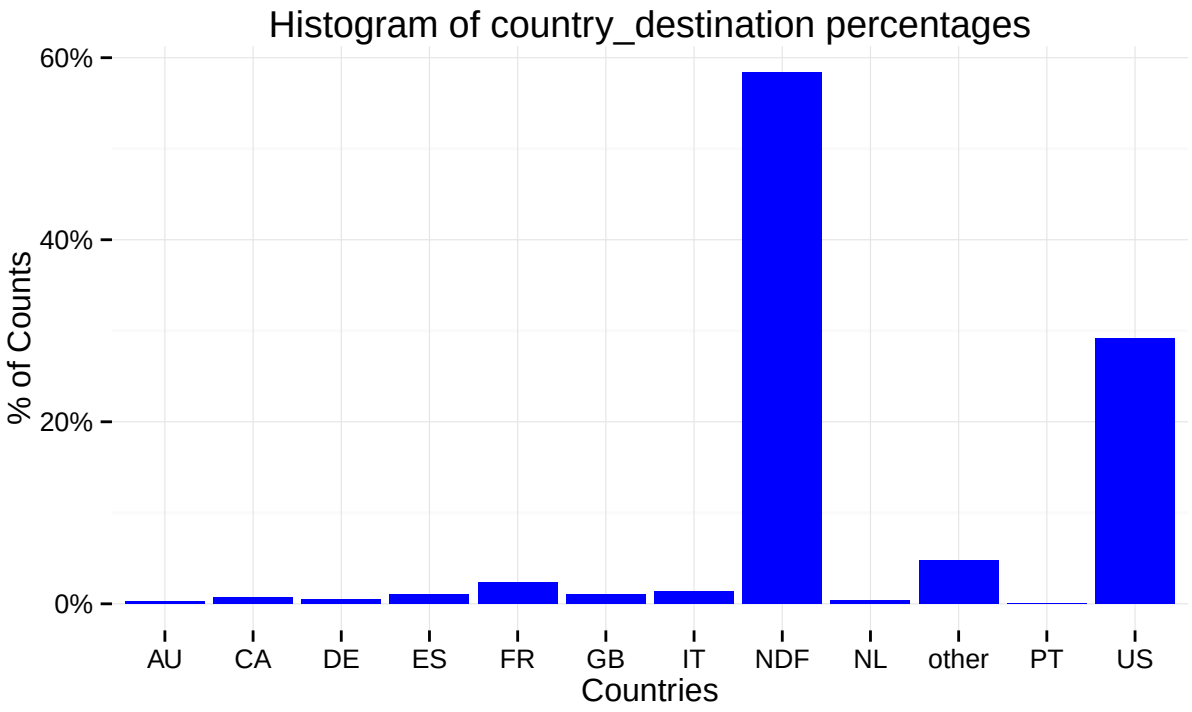
Exploratory Analysis

I first investigated the histogram of the target and some other fields:

Here is the histogram of the target variable "country_destination":



Here is the same data in percentages:



Seeing that there is a large bias towards NDF and US, I build my first few models based on that (more about this later).

I did many similar explorations on age, gender, language, actions, etc. There were some correlations between some of those and the target variables, but nothing generalized. Finally, I ended up building models on different subsets of them, as will be described below.

Model Building

In this competion, one is allowed to submit 5 guesses for each user. The scoring is based on NDCG (Normalized discounted cumulative gain) @k where k=5.

Initial Models

Based on initial analysis of the data (as can be seen in the previous section), I built first few models, just with the intention of testing the system.

I tried to set everything to NDF which received a score of 0.67909 (ranking ~900). I tried to randomly assign counries based on their relative frequencies, but that didn't result in a successful model.

Advanced Models

Then I tried to build models based on different factors: based on the distribution functions I observed, I saw some relation between the language of the browser used and the country visited, however models built on this didn't show much success rate.

Not succeeding much with above approaches, I decided to dive deeper: I build many different models using different approaches: Naive Base, GradientBoostingClassifier, XGBClassifier, etc. I used many different subsets of the features. I also tried many different combinations of those models: weighted average, max score, etc.

I tried a probabilistic approach for "age", "gender" and "language" separate, however I only saw incremental improvements. The major leaps came when I used the session information, which I first ignored.

I was mostly limited by the "5 submissions per day" limit. So I used Cross Validation to test my models locally before submitting.

After building close to 200 models and 60 submissions, I have improved my score from 0.67909, ranking ~900, to 0.88050, ranking 77 (as of this writing).

For the sake of completeness, here are the steps that I have followed while building models:

- Analysis and feature selection:
- test_users.csv - the test set of users
  - Plot the features to see their relevance to the target variable
  - Clean up the data by imputing the missing values
  - Run correlation analysis
  - Later use feedback from the results to add/remove some of the features
- Split the data into training, cross validation and test sets
- Model building:
  - Convert the features into a factor or numeric type if necessary
  - Select a suitable model (experiment if necessary)
  - Evaluate on cross-validation set
  - Change the model if necessary based on feedback (from cross-validation or from submission)
- Export the results into a file
- Submit on Kaggle website

For example, to get the model based on language: get the "language" and "country_destination" columns as features, fill in the missing values, use random forest to build a model and then later use this model to predict the "country_destination" for the provided test set. Output the predictions into a file and submit.

Conclusion

I have built many models and used many combinations to improve my initial results. I have improved the baseline by ~28%. The data was challenging because many of the fields were provided with no info, and some very important information that is available to the company, such as the country of the page being viewed by the user, was not provided. Also, I joined the competition 3 months late, which prevented me to do a lot of experiments, but still was able to climb up to the top 80 in rank in a matter of days.