Airbnb Challenge: Where will a new guest book their first travel experience?

Pecoraro Cyril
Jaume Guillaume
Grisard malo

GitHub of the project

Abstract

This project will be developped within the framework of the course An Introduction to Data Science. It is inspired by a Kaggle challenge proposed in February 2016 via their website. It aims at determining where a new guest will book their first travel experience. By accurately predicting this information, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

Description

We are given a list of users along with their demographics, web session records, and some summary statistics. We will try to predict which country a new user's first booking destination will be. All the users in this 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 found), and 'other'. Please note that 'NDF' is different from 'other' because 'other' means there was a booking, but is to a country not included in the list, while 'NDF' means there wasn't a booking.

The training and test sets are split by dates. In the test set, we will predict all the new users with first activities after 12/05/15. In the sessions dataset, the data only dates back to 1/1/2014, while the users dataset dates back to 2010.

The evaluation metric to determine the accuracy of the model is NDCG (Normalized discounted cumulative gain). The destination country predictions will be ordered such that the most probable destination country goes first.

Data

The data are real and from Airbnb. For a complete description, see the file *Data.md*

Feasibility

The features used are all non intrusive and completely anonymized. A possible problem would be that the features used are not relevant to predict the first destination of a new AirBnb customer.

Deliverables

A Jupyter Notebook and the data used for the analysis will be handout. This project will be presented in January 2016.

Timeline

Preprocessing and Exploration - 3 Weeks

During this phase we will expore the different files and try to see what is interesting in it. There will be some cleaning in order to only retain meaningfull informaiton

Machine Learning - 3 Weeks

We will implement a model that will be able to obtain the maximum score at the NDCG (Normalized discounted cumulative gain). The NDCG is given by the following formula:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

where *k* = 5, *rel_i* is the relevance of the result at position *i*, *IDCG_k* is the maximum possible (ideal) *DCG* for a given set of queries. All *NDCG* calculations are relative values on the interval 0.0 to 1.0.

Vizualization - 2 Weeks

In the ideal case, an interactive visualization map of each destination country with a user type will be presented.