Airbnb New User Bookings

Overview

As part of the existing Kaggle competitions, the team decided to develop the Data Mining group project for the competition called “Airbnb New User Bookings”.

According to the description and the inspected datasets, the problem context is a scenario fairly familiar to any person with minimum travel experience, the required outcome is well defined and the skills expected to be applied are relevant for the Data Mining module.

The team **LlanfairPG** is integrated by five students, one from the MSc Computer Science program and four from the MSc Data Science program. All with multiple and complementary backgrounds. Team member details are available in the table below.

Team	
Alun Meredith	am5e15@soton.ac.uk
Manuel Llamas	mlf1g15@soton.ac.uk
Miguel Ballesteros	mabm1e15@soton.ac.uk
Nicola Vitale	nv5g15@soton.ac.uk
Olivia Wilson	oew1v07@soton.ac.uk

Propose

As described in the Kaggle page, the Airbnb new user bookings competence consist of predicting the country where a new user will book its first travel experience. The Airbnb service allows booking accommodation in more than 190 countries from which there is a ten elements subset. Such subset consist of the most common destination countries to be taken into account for the prediction, along with a special case when the user leaves the website without a transaction (NDF). It is known that the site users are all from the USA.

Kaggle URL
https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings

The available datasets contain basic user data mainly with demographic fields, some application specific properties and user activity while browsing for accommodation options. Additionally, there is a test set and a submission sample that provide clarity on how the final outcome should look like.

With the performed data inspection and with the sample outcome, there is enough evidence that suggest to follow a data analysis with the aim to find patterns in the user profile and its browsing activity. At first sight it will be necessary to identify the most relevant features suggesting a destination country like language, device type, referrer and session flow.

The data shows that users not finalizing a booking (NDF) are about 60% of the whole dataset, while those having the US as destination are nearly 30% leaving just a proportion close to 10% to international destinations.

Preliminary Analysis

During a brainstorm we discussed ideas about where the device type and the session flow may suggest when the user is hesitating on choosing an option and therefore this may evidence a research process. With the same logic, if a user is accessing from a mobile device probably is quickly looking for accommodation options for familiar and most likely local alternatives. Potentially, for the first scenario the user is looking for a remote and unknown option (a country different than the US) while the second scenario it suggest that the user is familiarized with the destination and therefore its most likely destination is local.

In the same brainstorm there were suggestions about trying to find relationships between the browser language in combination with other factors signaling that the user is from a different culture and may have a destination different to the US. In the same way, one of the proposed approaches was to assume that the user is intending to travel inside the US unless the data point in a different direction.

At the moment, the considered strategies are focused in implementing a classification tree with nodes clustering similar features and then weighting all outcomes. Also, with the existing data it seems possible to create new features based on the existing ones (feature engineering). Curiously, the countries dataset has an interesting field that measures the language distance to the English language (language_levenshtein_distance) and may have some relevance when a user is choosing a new and unknown destination.

As soon as the exploratory efforts are complete and the strategy is defined in more detail, the team will decide which language use to implement the final solution. At the moment most team members are confident in R and Python.