Predicting in Which Country a New User Will Make His or Her First Booking

Problem Statement and Potential Stakeholders

Instead of waking to overlooked "Do not disturb" signs, Airbnb travelers find themselves rising with the birds in a whimsical treehouse, having their morning coffee on the deck of a houseboat, or cooking a shared regional breakfast with their hosts. New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. However, it is not clear that where a new user of Airbnb will make a booking activity.

It is clear that 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. It is also clear that providing more personalized content for Airbnb hosts will help them to make better arrangements which will increase travelers' satisfaction.

Airbnb hosts will get better review results and feedback. Travellers and other booking companies like Vrbo, Booking.com, HomeAway.com, FlipKey.com, etc. can use the model I will develop to adjust the search algorithms on their website and better serve the viewers.

I am planning to use a supervised learning model to predict where a new user will book their first travel experience. Among supervised learning models, I will do classification rather than regression because I will not aim to find a continuous variable but a categorical variable. Furthermore, the analysis should be multi-class classification because there are more than 2 categories so it cannot be binary. The target variable will be country destination and there are 10 countries listed in the dataset.

Using the model developed here, Airbnb hosts can better serve travellers and get better review results and feedback. Travellers and other booking companies like Vrbo, Booking.com, HomeAway.com, FlipKey.com, etc. can use the model to make their own arrangements and even configure and specify their search algorithms.

Data and Deliverables

The datasets are available on **https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data**. All datasets in the project are provided in CSV format. Four of six datasets are relevant to sessions, countries, age-gender, and train datasets. I will first clean and merge these four datasets to create a final

dataset for my analysis. These datasets were the most updated versions which were updated in 2015 for a competition. One of the datasets is a sample for the submissions and the last one is a test dataset so that I will not include both in the merging and cleaning processes.

Datasets

The following tables show the variables, data types, and variable types.

Sessions

Variables	Data Type	Variable Type
User ID	Object	Label
Action	Object	Categorical
Action Type	Object	Categorical
Action Detail	Object	Categorical
Device Type	Object	Categorical
Seconds Elapsed	Float	Continuous

Countries

Data Points	Data Type	Variable Type
Country Destination	Object	Categorical
Latitude	Float	Continuous
Longitude	Float	Continuous
Distance(km)	Float	Continuous
Destination (km2)	Float	Continuous
Destination Language	Object	Categorical
Language_Levenshtein_Distance	Float	Continuous

Age_Gender_BKTS

Data Points	Data Type	Variable Type
Age Bucket	Object	Categorical
Country Destination	Object	Categorical
Gender	Object	Categorical
Population in Thousands	Float	Continuous
Year	Float	Categorical

Train Users

Data Points	Data Type	Variable Type
Id	Object	Label
Date Account Created	Object	Time Series
Timestamp First Activity	Integer	Categorical
Date First Booking	Float	Time Series
Gender	Object	Categorical
Age	Float	Continuous
Signup Method	Object	Categorical
Signup Flow	Integer	Continuous
Language	Float	Categorical
Affiliate Channel	Object	Categorical
Affiliate Provider	Object	Categorical
First Affiliate Tracked	Object	Categorical
Signup App	Object	Categorical
First Device Type	Object	Categorical
First Browser	Object	Categorical
Country Destination	Object	Categorical