Overview

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. The project will be based on supervised machine learning whereby the model will be trained using training data and prediction will be done on testing data.

Problem Statement

The Airbnb challenge is to predict in which country a new user will make his or her first booking. This is a classification problem whereby given training data, we train the model using one of the supervised machine learning model which can accurately predict the destination country for the testing data.

Datasets and Inputs

The dataset is used from Kaggle competition. Following are the datasets used as part of the project:

1. train_users.csv - the training set of users. The test_users.csv contains following features:

    - 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: Male/Female, but some records are marked as unknown.
    - Age: Age of the customer.
    - signup_method: Mechanism used for sign up with AirBnB.
    - 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: website
    - first_device_type: device type used for booking.
    - first_browser: Browser used
    - country_destination: this is the target variable we train the model and predict against the testing data.

2. sessions.csv - web sessions log for users. The sessions.csv contains following features:
    - user_id: to be joined with the column 'id' in users table
    - Action: Action performed on the website, could be lookup, search results
    - action_type: Action type, could be click, data displayed.
    - action_detail: Details of the action.
    - device_type: Device used, could be desktop/any mobile device
    - secs_elapsed: time spent on the action

3. countries.csv - summary statistics of destination countries in this dataset and their locations
    - country_destination - Country name
    - lat_destination – Latitude
    - lng_destination – Longitude
    - distance_km - Distance in km.
    - destination_km2 - Distance in km2.
    - destination_language - Language spoken in destination country
    - language_levenshtein_distance - Silimarlity in terms of language distance

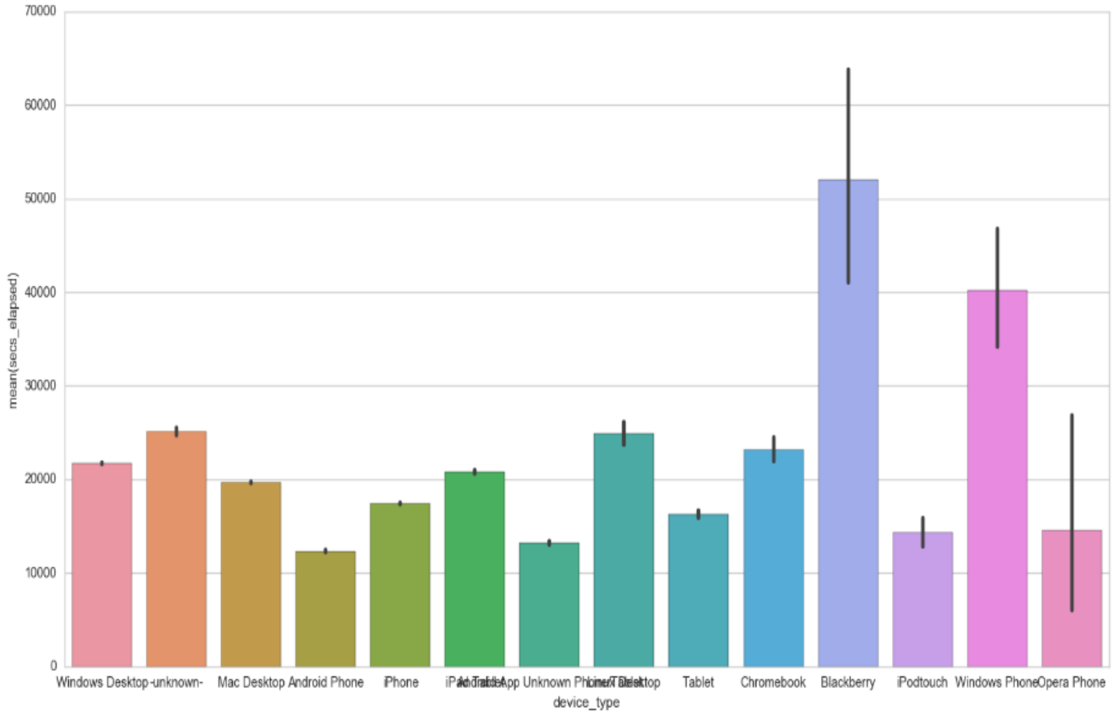
4. age_gender_bkts.csv - summary statistics of users' and it contains following set of features:
    - age group: Age group of people visiting the destination.
    - Gender: Gender of the customer.
    - destination country: Destination country
    - Year: Year of booking made.
    - population (in thousands) - Population.

Dataset Exploration

1. sessions.csv

The sessions csv file captures exhaustive list of actions performed by the user. These actions are clicking the various links in order to create wish list, searching for destination, modify the search criteria and other various actions. In addition to this, the CSV file provides time spent during each action on a particular device, which will help us in understanding intended action by the customer and the device used for making the booking. For example, if a customer spends lot of time in exploring the destination, which is more likely to make booking around that location/destination when compared with spending less time in exploring the destination.

Below barchart depicts data extracted from the sessions.csv.

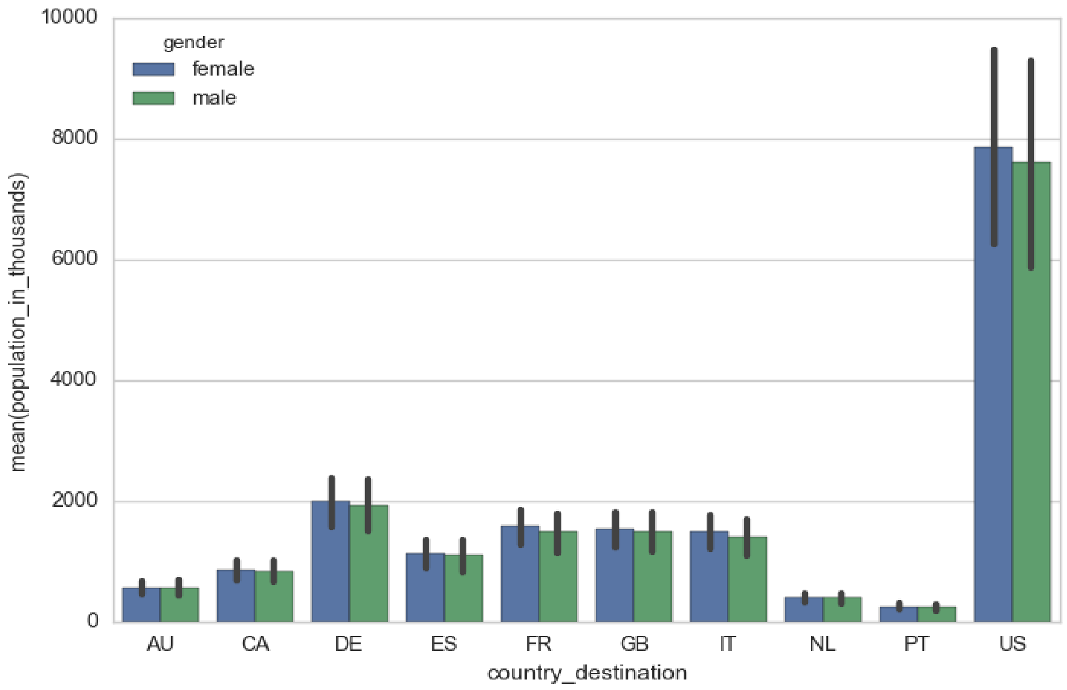


2. countries.csv

The countries csv file captures details about the destination, its location in terms of longitude and latitude. In addition to this it all also describes language spoken in the destination. This piece of information could be taken into account when the customer considers booking his holiday.

3. Age Gender buckets csv

This csv file contains valuable information such as range of age groups of people (male/female) visiting the destination country. In addition to it the file also contains information about population visiting the country for a given age group in a given year. This will help the model to learn the choices done by different age groups, male/females. Apart from this we can also deduce favourite destination by looking at the population count in each destination country.



Metrics

For evaluation process, I shall be using cross validation to validate the outcome of result. As there are many times of cross validation, I

shall be using k-fold cross validation. In case of k-cross validation, I shall split the input data in k-sets, in which k-1 subsets shall be used for training and the kth subset shall be used for testing. This process will run through all the k-sets in the training dataset.

Models

There are exists various classification models such as Decision trees, Support Vector Machines, K-NN, Neural networks and XGBoost. I prefer to use to use XGBoost, because of following reasons:

1. based on iterative model, whereby the model builds with basic set and learns from earlier trees.
2. can handle large amounts of data
3. xgboost is more flexible, i.e. has more customizable parameters
4. xgboost is faster

However I also would like to make use of some of the classification models and compare the prediction against each other.

Project Design

As it is clear that the project goal is to predict the country where the customer will make his first booking. Given the dataset, I prefer to follow these steps:

1. Data cleansing: This step is mandatory, because some of the data is either missing or incorrect. For instance, age of the customer cannot be 0 or more than 100. In addition to this, there are some features which we need to get rid of them.
2. Data Transforming & Normalization: Some of the data is not in preferred format, so we need to get the data in desired format. This will be done as part of this step. As we have noticed that some of the features are of categorical data type, so we need to apply some sort of normalization. I am planning to use One-Hot Encoding, wherein the n categorical options will be converted into n columns.
3. Feature Engineering: In this step we will be extracting information from various set of features and try to add more information to facilitate model in predicting the destination country. One way would be pick the information such as season/month of the year used by the customer to make the booking and also find out which resident country. Most of the time people living in cold countries prefer to holiday in warm country and other way around. This will be achieved as part of this step.
4. Finally, the new data constructed will be feed to the model and comparison will be done against the test data.

Benchmark

As part of this project, I would like to compare the performance of some of models (DecisionTrees, K-NN and XGBoost). The benchmark will be based on F1-score for each model. As we know that the dataset is used from Kaggle competition, the output/target results are not published. So I shall be solely using training data to train the model and also to predict.