McNulty MVP

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

Project and data source: https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings

MVP Summary:

The MVP is to generate a prediction for the AirBnB Kaggle challenge.

- **Domain:** The question is, Where will a new guest book their first travel experience? The AirBnB data set I'll work with comes from a Kaggle competition. The idea is to use user data to predict which of 10 countries a new user is likely to pick as their first destination.
- **Data:** The data are stored in five disparate .csv tables:
    - Age Gender Brackets (12 kb)
    - Countries (<1 kb)
    - Sessions (632 MB)
    - Train Users (25 MB)
    - Test Users (7 MB, same format as Train Users, except target column, `country_destination`).

Table	List of columns
`age_gender_bkts`	(`age_bucket`, `country_destination`, `gender`, `population_in_thousands`, `year`)
`countries`	(`country_destination`, `lat_destination`, `lng_destination`, `distance_km`, `destination_km2`, `destination_language`, `language_levenshtein_distance`)
`sessions`	(`user_id`, `action`, `action_type`, `action_detail`, `device_type`, `secs_elapsed`)
`test_users`	(`id`, `date_account_created`, `timestamp_first_active`, `date_first_booking`, `gender`, `age`, `signup_method`, `signup_flow`, `language`, `affiliate_channel`, `affiliate_provider`, `first_affiliate_tracked`, `signup_app`, `first_device_type`, `first_browser`)

Train users has one row per user `id` and includes the target column `country_destination`.

`sessions` contains web session data for about 135k users, and nearly all of those users can be matched to either `test_users` or `train_users`.

However, the reverse is not true: while nearly all `test_users` are also represented in `sessions`, only about 1/3 of `train_users` have a match in `sessions`.

I'll explore all the data that can be matched through a `join` in order to build this model.

- **Known unknowns:**
  - There are two columns in `sessions` that seem difficult to sort out: `action` (359 unique categories), and `action_detail` (155 unique categories). Not only are these a lot of categories, but they're also not defined by anything other than their naming, which is often not enough information to grasp what's going on.
  - There are many ML models to use for this. Most of the ML tools I've seen used for this challenge I've not seen before.