DS 310 Machine Learning

Project #3: Airbnb New User Bookings

Fall 2020 / Amulya Yadav

Task: 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.

In this project, Airbnb challenges you to predict in which country a new user will make his or her first booking.

Dataset: You can download the dataset on Canvas.

In this challenge, you are given a list of users along with their demographics, web session records, and some summary statistics. You are asked 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 dataset consists of the following items, after unzipping:

- `train_users.csv`: the training set of users
- `test_users.csv`: a the test set of users
  - 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

  - age
  - signup_method
  - 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
  - first_device_type
  - first_browser
  - countr_destination: this is the target variable you are to predict
- `sessions.csv`: web sessions log for users
  - user_id: to be joined with the column 'id' in users table
  - action
  - action_type
  - action_detail
  - device_type
  - secs_elapsed
- `countries.csv`: summary statistics of destination countries in this dataset and their locations
- `age_gender_bkts.csv`: summary statistics of users' age group, gender, country of destination
- `sample_submission.csv`: correct format for submitting your predictions

The training and test sets are split by dates. In the test set, you will predict all the new users with first activities after 7/1/2014. In the sessions dataset, the data only dates back to 1/1/2014, while the users dataset dates back to 2010.

Evaluation Metric

The evaluation metric for this competition is $NDCG$ (Normalized discounted cumulative gain) where $k = 5$. $NDCG$ is calculated as:

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

where 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.

For each new user, you are to make a maximum of 5 predictions on the country of the first booking. The ground truth country is marked with relevance = 1, while the rest have relevance = 0.

For example, if for a particular user the destination is FR, then the predictions become:

$$\text{[FR] gives a } NDCG = \frac{2^1-1}{\log_2(1+1)} = 1.0,$$
$$\text{[US, FR] gives a } DCG = \frac{2^0-1}{\log_2(1+1)} + \frac{2^1-1}{\log_2(2+1)} = 0.6309.$$

Kaggle

After teams generate the prediction results for the test data, you can post it to the Kaggle to get the feedback on the accuracy of their model:

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

Deliverables: Each team needs to accomplish the following tasks:

1. Kaggle Leaderboard: At the end, we take the best ranked entry (performance) of each team's submission against `test.csv`, in the leaderboard as the team's final submission result.

2. Jupyter notebook: Each team posts its final model in Jupyter notebook, named as `submission.ipynb`.

3. Project Report: Project report in PDF that contains all the details of the major steps of the project.

Grading Rubric (100 points)

- 80 points: performance (65 points) and ranking (15 points) in the Kaggle Leaderboard, using each team's best `submission.csv` file. 50 points for baseline performance and 15 points for better performance.
- 10 points: submitted Jupyter notebook, `submission.ipynb`.
- 10 points: quality of report.

Grading Exceptions

The goal of this project is to encourage you to emerge yourself in a real-life data science process to learn the materials better in a hands-on fashion. **DO NOT CHEAT IN ANY WAYS.** If you do not attempt to get your hands dirty in this project, you will not learn much.

All team members need to participate in the project and contribute as much as he/she could. Each is expected to bring different skills to the table, i.e, some are good in programming, while others in analytics, etc. Find out what you can contribute within the team, and play the team work. While both members get the same points by default, I will intervene in an exceptional situation (e.g., one member does not participate at all).