Machine Learning Engineer Nanodegree

Capstone Project Proposal

Airbnb New User Booking Dataset ¶

Saumitra Rawat
Jan 3rd, 2019

Domain Background

Airbnb, which is an online marketplace where people list, discover, and book accomodations around the world. It has collected various datapoints about users. This data about the usage patterns of its present user base can be utilized to predict patterns about its future users to provide them with customized suggestions to serve Airbnb's customers better. Airbnb had posted this on Kaggle as a Recruitment Challenge. Using user data effectively can help organizations increase metrics such as sales, user experience, customer retention and customer satisfaction. Machine Learning techniques can help organizations attain useful predictions using these data. The motivation for pursuing this project is to understand how to work on real world datasets and challenges that companies like Airbnb consider to be important and valuable for their companies and learn to provide similar value for organizations that I work with in the future.

https://medium.com/airbnb-engineering/learning-market-dynamics-for-optimal-pricing-97cffbcc53e3 (https://medium.com/airbnb-engineering/learning-market-dynamics-for-optimal-pricing-97cffbcc53e3)

Problem Statement

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.

Using the data from Airbnb New User Bookings (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings) dataset, the challenge is to predict the destination of choice for the users' first booking. This data includes demographics of users and their session data. The model will utilize these demographics and session data to make models that can predict the destinations.

In this project, I plan to use Machine Learning Techniques to predict in which country a new user will make their first booking on Airbnb. This project will involve data cleaning, data exploration using visualizations, and testing various algorithms for classification for the same.

Datasets and Inputs

The dataset is composed of 5 CSV files. It has been obtained from a Kaggle Competition provided by Airbnb. [link] (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data)

The most important file is the `train_users` file which has 16 columns containing user id, dates of account creation, first booking date, gender, age, signup method, signup app, destination etc along with the target variable `country_destination` and has 213451 rows. The `test_users` is similar to the previous file discussed but does not have our target variable and we have to use these to predict the destination and has 62096 rows. We have a good amount of data to work with to produce meaningful models.

The other three files contain web session logs (`sessions.csv`) for the users, summary statistics of destination countries (`countries`) and summary statistics of about the users age group, gender, etc. (`age_gender_bkts.csv`)

File descriptions

- **train_users.csv** - the training set of users
- **test_users.csv** - 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
  - country_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

As of now I am not plannning to split training data further into custom-training and cross-validataion set. Once the solution is ready we can first calculate the accuracy of model and then re-evaluate and test the same to check whether accuracy of the model is getting increased or not.

Solution Statement

The solution will largely utilize the fact that similarites in user demographics is likely to be correlated to the choices made by the users on the platform. This will be helpful for us to test supervised learning models to predict the behavious of new users. I will use the first 15 columns of the `users data` as input to these models and the `country_destination` as the target.

I will then test various models such as SVM, Decision Trees, Random Forest etc. we have learned in this course along with techniques such Grid-SearchCV to optimize and other models such as XGBoost which are used effectively in competitive environments such Kaggle.

Benchmark Model

To determine a baseline benchmark, we will find the metric value obtained by predicting the 5 most common outcomes [NDF, US, OTHER, FR, IT] against the train and test datasets.

Along with that, the goal will be to place our final model in the top 20% of the leaderboard on Kaggle using the evaluation metric describe below

Evaluation Metrics

Since this is a Kaggle Challenge, we already have an evaluation metric, that is the NDCG (Normalized Discounted Cumulative Gain)

For each new user, we 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.

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

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

where rel_i is the relevance of the result at position i and $k = 5$.

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

[FR] gives a $NDCG = \frac{2^1-1}{log_2(1+1)} = 1.0$

[US, FR] gives a $DCG = \frac{2^0-1}{log_2(1+1)} + \frac{2^1-1}{log_2(2+1)} = \frac{1}{1.58496} = 0.6309$

Project Design

The project will be composed of the following steps:

- *Data Exploration and Pre-processing*:

    1. Visualizing the dataset.
    2. Detect outliers.
    3. Remove null values.
    4. Cleaning the dataset.
    5. Check relevance of every column to the target column.
    6. Cluster the dataset using unsupervised techniques to see if we can engineer new features.
    7. Replacing unknown and missing values with '-1'.
    8. Replacing continuous values with interval type values(Like age interval).
    9. Computing season using dates and using them instead of dates.
    10. Converting some features into One-hot-encoding features.

- *Training*: Consider multiple supervised ML models and select the best one, use techniques such as cross validation, and optimizing using GridSearchCV for hyperparameter optimization.

- *Testing and Optimizing*: Optimizing the model offline, using the trained models to test on Kaggle and improve the rank on kaggle.