Machine Learning Engineer Nanodegree

Capstone Project Proposal

Airbnb New User Booking Dataset

Pin-Yi Tseng October 13th, 2018

Domain Background

The distance between countries are becoming shorter due to faster transportation like airplanes and trains; on the other hand, thanks for broadband universal service, people tend to share their daily life, experiences, special events, etc. on Facebook or Instagram. Instead of joining in traveling agency, more and more people get their bag, pick up destinations and take off their adventures by themselves or with a small group of their friends. Companies like Airbnb, Booking.com, Agoda, trivago, and so on booming nowadays. People tend to continuously use their familiar and trust website when booking their hostels, especially when they become VIPs after many purchases. Therefore, the way to correctly predict the destination, providing related service or recommendation and attract people when they make their first consumption is the critical point for getting long-term customers.

Airbnb has become a global platform that connects travelers and hosts from over 34,000 cities. As such, it has collected a diverse set of dataset about users which can be utilized to predict patterns about its future users and provide them with customized suggestions to serve Airbnb's customers Airbnb had posted this on Kaggle as a Recruitment Challenge. Using user data could help organizations increase metrics such as sales, user experience, customer retention, and customer satisfaction. Apply Machine Learning methodology can help the organization to reveal the mutual effect between different events And furthermore, making the prediction 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.

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`)

train_users/test_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

Solution Statement

The solution will largely utilize the fact that similarities in user demographics are 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 behaviors 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

The given dataset is a typical supervised learning problem and so far SVM is the algorithm I used mot of time. So I will pick Support Vector Machine (SVM) as a benchmark and try to beat the benchmark with hyperparameter turning. We will also try tree type models and ensemble methods if the hyperparameter tuning does not improve the score.

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*: Visualizing the dataset, detect outliers, remove null values, cleaning the dataset, check relevance of every column to the target column, cluster the dataset using unsupervised techniques to see if we can engineer new features, splitting training dataset into training and testing sets etc.
- *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*: Fine tune the selected algorithm to increase performance without overfitting and test the model on testing dataset.