**The Prediction of Booking Destination of New User on Airbnb**

**Chien-Wei Lin, Po-Kang Huang**

**Abstract**

New users on Airbnb can book a 34,000+ cities across 190+ countries. By accurately prediction where a new user will book their first travel destination, Airbnb can share more information with their community, decrease the average time to first booking. This project aims to analyze Airbnb dataset and design a machine learning flow to do the prediction task. The task was assigned by Kaggle competition, which also provided training and test dataset for this project. We plan to learn the models and processes used by several winners in the competition, and try to combine all advantages of various machine learning methods.

**Methodology**

1. Dataset
2. Description of dataset

The dataset provided by Airbnb is a list of users’ information, including their demographics, web session records, and some statistics data. The whole dataset contains the following 5 files:

1. train-users and test-users

These two files respectively contain 213451 training samples and 62096 testing samples with 16 (country-destination, which is the label) and 15 (without label) properties, which are some basic information of some user data in their history records. The properties contain: id, date-account-created, date-first-booking, gender, age, signup-method, signup-flow, language, affiliate-channel, affiliate-provider, first-affiliate-tracked, signup- app, first-device-type, first-browser, country-destination, timestamp-first-active.

1. sessions

This file contains the web session log records for users. It contains 1048576 samples in total 74610 different users with 6 properties: user-id, action, action-type, action-detail, device-type, secs-elapsed.

1. countries

This file contains geometric statistics of destination countries in this dataset and their locations. There are total 10 destination countries, and each country contains 7 different properties, such as latitude, longitude, language, levenshtein distance, and distance from origin place.

1. age-gender-bkt

This file contains the summary statistics of users age group, gender, country of destina- tion. It consists 420 samples with 5 properties.

1. Observation and analysis of Data
2. Country Destination

Based on figure.1, we found that 64% of the users ended up booking nothing which indicated as NDF in *country\_destination*. Furthermore, there are 97.97% of those users who didn’t booked in the Airbnb didn’t have the record of *data\_first\_booking*. Therefore we turn to the users who had booked in Airbnb to ensure the effectiveness. The result is that all the users who didn’t book in Airbnb have NDF on the *country\_destination*. To predict the country destination, we want to classify the users who booked in the airbnb and those did not. As the previous observation, we found out a useful feature: *data\_first\_booking*

Fig. 1 percentage of the destination

1. Range of Age

From Fig.2, we can observe that most of the users who booked on Airbnb are in the range between 20 to 45 years old.

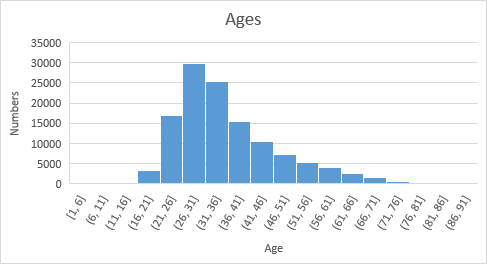


Fig.2 Age of users without missing data and outliers

1. Data pre-processing
2. Feature space

For all the features in our raw data, they are string form which is harder for computation. Moreover, many of our feature can be represented by numerical values. For example, *gender* contains male, female, other, and unknown in raw data. It will not be proper to assign 0, 1, 2, 3 to represent male, female, other, and unknown, respectively. Instead of using numerical value, we translate this feature from 1-dimension vector into 4-dimension vector. Each dimension will be 1 if it matches the description of the feature, and will be 0 otherwise. If User1 is male, then then User1 has the 4-dimension vector as Table1. For most of our feature, we translate each 1-dimension features into multiple dimension binary features

|  |  |  |  |
| --- | --- | --- | --- |
| Male | Female | Other | Unknown |
| 0 | 0 | 1 | 0 |

Table1

1. Age

We first represented the age with its number directly, which is a trivial result. We then found out that it was the age distribution which matters. From Fig.2, we assume that children and seniors have less probability to use the Airbnb to plan a trip. After discarding the data with and , we create 15-dimension vector, which counted the index between adjacent pair in . From Fig.2, we assume children and seniors has less probability to use Airbnb on planning a trip. Therefore, we focus on the range of age on 20 to 40.

1. Royalty of Using Apple Product

We first treated these three features, *first\_device\_type, first\_browser,* and *signup\_app*, separately, which didn’t improve our performance. After analysis, we noticed that they have some relations when it comes to Apple product, such as iPhone, iPad, Macbook, or Safari. As a result, we shrink these three feature into one feature. If one of the features contains the iOS, Mac, or other Apple related product, then we set this feature to 1, or 0 otherwise.

1. Learning Model
2. Model selection
3. Evaluation model
4. Binary Evaluation

For the binary classification

1. Multiclass Evaluation

For the multiclass evaluation, we follow the evaluation metric for the Kaggle competition, NDCG (Normalized discounted cumulative gain) @k where k=5. The NDCG evaluation is shown as:

whereis the relevance of the result at position , in our case

The term is the maximum possible ideal for a given set of queries. All calculations are relative values on the interval . Using this evaluation metric, we predict a list of 5 destination countries with highest probability, and compare, in order, the countries on the list with the ground truth country to calculate its evaluation score. The ground truth country is marked with relevance = 1, while the rest have relevance = 0. For example, if for a particular user with the prediction list and the ground truth country , the evaluation score becomes:

The maximum score for each user is 1 and minimum 0. The score is determined by the order of the prediction countries in the prediction list. If the correct prediction is at first position, the score will be 1, and the score decreases if the position of the correct country goes larger. This method is often used to measure effectiveness of web search engine algorithms or related applications, which avoid the advantages of arbitrarily giving the score 1 or 0.

1. Cross validation
2. Bagging
3. Description of model

**Result**

**Conclusion**

1. Performance
2. Future Directions
3. Learning from this project

**Reference**