Travel Destination Prediction for Airbnb

Firstname Lastname

Abstract

In this project, I have built a proposed model to predict the Airbnb's new user's booking destination country based on their features such as age, gender, demographics and session data, language etc. I gathered data from Kaggle competition. Firstly, I did comprehensive analysis on the dataset, tried to explore most features and collected all features I thought that can be useful. This is how I did feature selection strategy based on the observations on dataset. Then, I described and interpreted the prediction task and the evaluation method and built a reasonable model for the prediction task. I used SVM (Support Vector Machine) and Neural Network to predict the top most visited countries from list of countries. I squabble and manipulated data in Python and R in my proposed model. Finally, I received the accuracy of 90% for the 5 top most countries that a user can visit.

Keywords: Airbnb, Prediction, Random Forest, Neural Network, Travel, Destination, Country, Language, New user, Booking.

1. Introduction

Airbnb is one of the best option for a traveler to save their expenses on hotels. Airbnb is a trusted marketplace that allow people to enlist, explore, and book some unique places/rooms/house all over the world. Moreover, Airbnb is becoming the visiting place for travelers of worldwide. With its attendance, in 34,000+ cities in 190+ countries, the users no longer should be out of choice about needing to stay in costly hotels. Users can use either web application or the iOS/Android applications. The users can have facility to browse through lists from many countries and book their accommodation with a single click in the applications.

Using few machine learning algorithms, we can design such systems that can be mapped to specific users' need to allow them for more personalized product/content that reduce the total average time of users to broadcast the list of countries according to the users' best option out of many countries to be travelled by them for the first time booking with better accuracy. This leads us to predict from many listed countries, how new user will make his or her first booking based on browser's session activity and user's demographic information such as user's browsing through certain country's listings, worldwide languages spoken by the users, country specific seasonality et cetera. I am trying to define potential features from the dataset and see how they are correlated to the countries that a proposed model can predict user's first booking effectively. Using Random Forest and Neural Network Algorithms, I am trying to make some equations between the features that how they are used to predict new users' first time booking for Airbnb.

Background

Using enormous dataset given by Airbnb to Kaggle, we can check how effectively augment user experience and rectify total bookings by utilizing big dataset remains a question to answer for us. That question can be resolved after going through many research papers. In [3], the author Narayanan Ramamirtham has concluded that the Random Forest classifier algorithm is superior to the Decision Tree model. In [3], the Random Forest Classifier model had an 87% accuracy compared to 79% for the decision tree model. Also, he had expected that pruning the dataset given to the Random Forest Classifier will improve the results. This proves that the classifier will not over-fit the data since it generates multiple decision trees and chooses the best one, even if we had a redundant feature, there won't be any degradation in the prediction accuracy. In [4], the author Arvind Rao has achieved 87.5% accuracy by using single level classifier Neural Network. In [5], the author Yingzhi Wu has received the accuracy of 91% using Nearest Neighbor classification model.

With the same purpose, I have used Neural Network classifier algorithm to improve the accuracy and compared it to the existing model that have used Random Forest algorithm. I did data pre-processing by replacing null or missing attribute values with the most common value relative to each attribute values.

2. Dataset

We have got datasets of Airbnb from the Kaggle competition. Airbnb has given 5 datasets to help with the prediction of travel destination as a list of 5 top most visited countries to be traveled by first time user in Airbnb.

(1) Train_users: This dataset contains data on 213,451 users of the Airbnb. It contains user id, date account created, date first booking, gender, age, country destination etc.

(2) Test_users: This dataset contains 62,096 users on which the prediction will be performed. It contains same data fields as Train_users.

(3) Session: This dataset contains device type, action, action type, user id and the number of seconds between actions were recorded etc.

(4) Age_gender_bkts: This dataset contains user's decided country information in groups age of 5 years difference i.e. range of age with 5 years gap in between, travel destination as a country and it shows information about each group's gender. It comprises of 5 properties and 420 examples. I see that users over age 30 has travelled more compared to users under age 30.

(5) Countries: This dataset contains a summary of the different country destinations and various data on those countries' locations and which common language is spoken in relative country.

In these datasets, we have eleven unique countries (CA, DE, AU, FR, IT, GB, PT, US, NL, ES, NDF). Here, NDF means no destination has found. We found that more than 50% of the data has No Destination Found (NDF) which means user have not been at any travel destination yet. It was interesting and challenging part that user have browse destinations but has not made any actual trip.

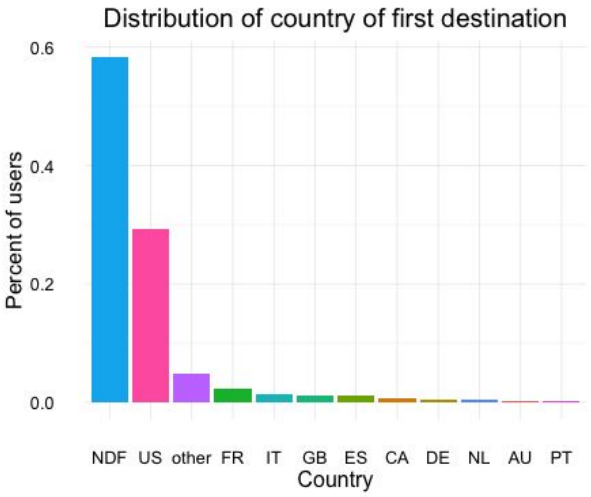

Figure 1: distribution of country of first destination

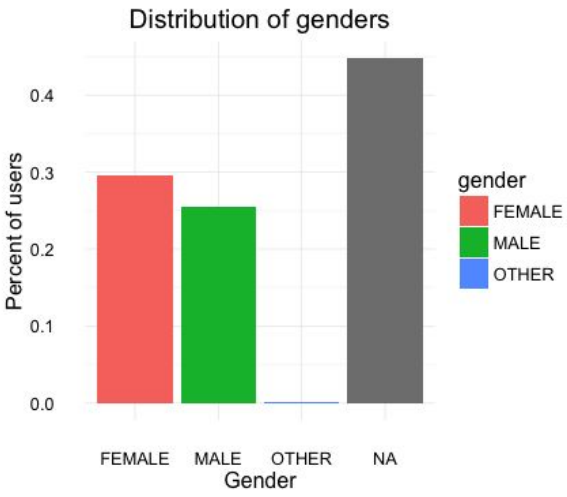

Figure 2: distribution of gender

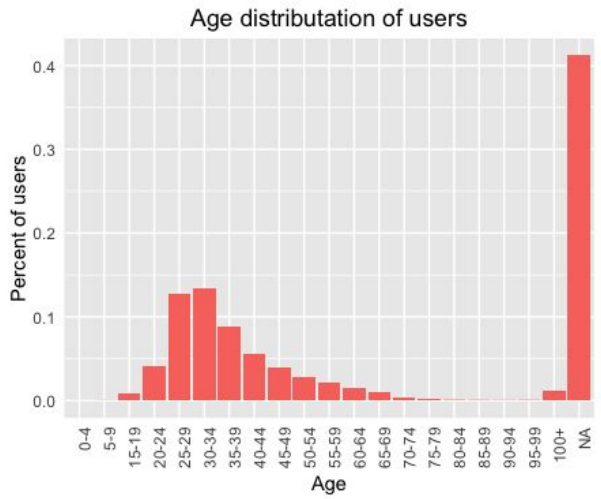

Figure 3: distribution of age

3. Data cleaning and Pre-processing

First of all, to fill in the missing numerical values, we have used median method for continuous feature. In particular field, if numerical data is missing, the mean of that field will be replaced in missing or null location.

For categorical feature, fill the null or missing attribute values using approach which is listed below:

1. Use the most common value relative to each user and action
2. Use the most common value relative to each action
3. Use the value 'missing'

We found that 99,152 users have not selected any destination. We have focused on users who have selected destination as US. Because first we want to find US and non-US label.
Following are the images of analysis of the dataset that can be useful to estimate which type of dataset I need to clean.

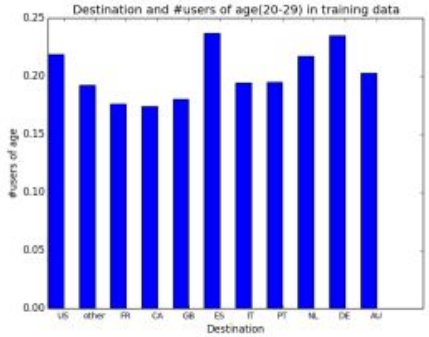

Figure 4: Ratio of users whose age is 20-29 of different destinations

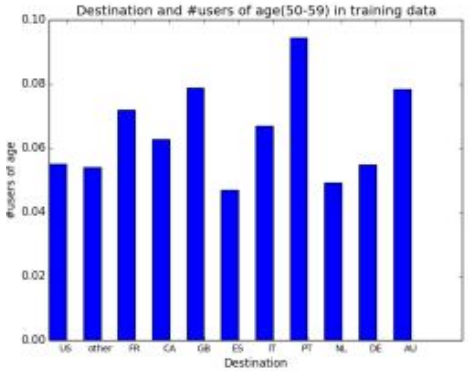

Figure 5: Ratio of users whose age is 50-59 of different destinations

4. Feature Selection and Extraction

Based on the analysis of pre-possessed data, we have considered following main set of the features.

1) Session data
2) Gender
3) Age
4) Seconds elapsed (session)
5) Language
6) Country

After selecting these main features, we have generated a summary data frame using this information. Also, we merged all these features of summary data frame into training and testing datasets in which training and testing datasets are equal to users that are not in previous training and testing datasets.

Since there are many users that have not appeared in the sessions data frame, all of their values are NaN. We sorted out these null values using similar approach that we applied earlier on categorical and continuous feature to pre-process the datasets. Further, we grouped all features as either continuous or categorical features. We created dummies of each value of a categorical feature to provide input of categorical feature to our proposed model. Then, we normalized the continuous features to improve execution process faster. We split these all data frame into training and testing dataset.

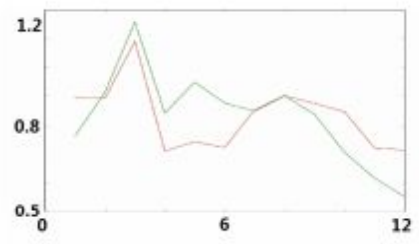

Figure 6: Relation between booking numbers and month

In figure 1, X-axis is month and Y-axis is number of bookings we found that months have good influence on the number of the bookings so it can be selected as a main feature. We also did analysis between languages and the user destination preference, gender, home country etc.

5. Proposed Method

Using given data, we are predicting user's first travel destination. We will consider gender, age, session, seconds elapsed, language, country et cetera input to our proposed model.

This problem is considered as a multi class classification problem. We have classified output to the one of the eleven countries (CA, DE, AU, FR, IT, GB, PT, US, NL, ES) or NDF, where NDF means model is unable to predict class for that user. The given data is unbalanced because most of the data records has US related information. So, we need to try different models to get different results and identify the best results. As suggested in [2] for the similar project, two level classifier performs well. In [2], the author Ke Zheng has selected Support Vector Machine (SVM) to classify the data. In [4], author Arvind Rao has constructed a pipeline for multilevel Random Forest classifier. In [4], after evaluating on the test set the result was far worse than the Neural Network classifier. This might be due to guessing only a top most destination prediction and not the top most 5 destination predictions. Consequently, our final classifier was to use a Neural Network algorithm in our proposed model. Therefore, we have used Neural Network model for a travel destination prediction to improve the accuracy from previous models.

We used 3 hidden layers with 50 hidden nodes and in input layer 186 input nodes in each layer ending with a 12-way softmax classification layer. Each class was represented with a one-hot encoding vector of length twelve. Following image show steps taken to implement Neural Network Algorithm.

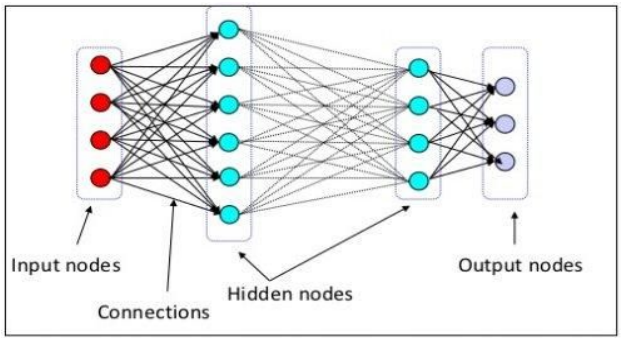

Figure: 7 Neural Network Algorithm

6. Evaluation Model and Results

Neural Networks are sensitive to class imbalance, because during the learning process, if the class is seen less often, the weights of the neural network will not update to account for that class. To overcome with this problem, we used epoch which is used to measure the number of times all of the training vectors are used once to update the weights. We have used total 3000 epochs with test-loss which shows how many error is there in our coding. If there is less number of test-loss, there are more chances to achieve better accuracy. We have used Normalized Discounted Cumulative Gain (NDCG) evaluation method based on Kaggle competition. NDCG is used for giving rank to the most visited countries.

We achieved accuracy of 87.5% in predicting the top 5 most countries to visit by new user of Airbnb using proposed model.

7. Conclusion and Future Work

In this project we have presented model which gives accuracy of almost 90%. Main reason behind this performance of this model is approach we have used for data pre-processing, data cleaning and feature extraction.

In future, I am planning to implement different models using ensemble learning algorithms which can overcome disadvantages of individual models.

8. References

[1] Hugo Ulfsson "Predicting Airbnb user's desired travel destinations", Stockholm, 2017. http://www.diva-portal.org/smash/get/diva2:1108334/FULLTEXT01.pdf

[2] Ke Zheng, Zhengren Pan, Sichao Shi “The prediction of booking destination on Airbnb dataset”, *UC San Diego*. *http://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/038.pdf*

[3] Srinivas Avireddy, Sathya Narayanan Ramamirtham, Sridhar Srinivasa Subramanian “Predicting Airbnb user destination using user demographic and session information”, *UC San Diego* https://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/040.pdf

[4] Arvind Rao “Predicting on Airbnb” https://cseweb.ucsd.edu/classes/wi17/cse258-a/reports/a065.pdf

[5] Yingzhi Wu “New User Booking Prediction for Airbnb Historical Data” https://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/045.pdf

[6] Kaggle competition Airbnb New User Bookings. https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

[7] Michael Winfield “Predicting a New User's First Travel Destination on AirBnB”, NYC Data Science. https://nycdatascience.com/blog/student-works/predicting-new-users-first-travel-destination-airbnb-capstone-project/