Airbnb New User Booking Destination Prediction

Vibhuti Gupta
Department of Computer Science
Texas Tech University
Lubbock, TX 79415
Email:vibhuti.gupta@ttu.edu

Gantaphon Chalumporn
Department of Computer Science
Texas Tech University
Lubbock, TX 79415
Email:g.chalumporn@ttu.edu

Fang Jin
Department of Computer Science
Texas Tech University
Lubbock, TX 79415
Email: fang.jin@ttu.edu

***Abstract*—Airbnb is an online market place for renting apartments, houses, hotel rooms or vacation rentals for short time period. Users can book in 65000 cities across 190+ countries all over the world using this service.We have analyzed the airbnb dataset in this project to predict which country each new user will make the first booking or most likely to visit. We have used the dataset provided by Airbnb for a kaggle competition. The dataset consists of information about the new users like demographics, web session records and various other features.It also contains 12 different destination countries as the variable to predict.Firstly we will perform an exploratory analysis of the dataset to determine the statistics about various features such as age,gender,language,destination country distribution etc. and then feature selection for the most useful features responsible for prediction.After this we will train the appropriate classification model with already labelled destination countries and use it to predict for the test data. Finally we will evaluate our prediction model with various evaluation measures such as accuracy, precision, recall and AUC curve.**

***Index Terms*—Airbnb**

I. Introduction

Airbnb is one of the biggest online platform for short term lodging of houses, hotel rooms, vacation rentals etc. It is spanned across various countries with almost 3000000 lodging lists. It can be accessed through its website or mobile apps. Users have to register with their general information such as name,email address, telephone number etc. to book a lodging using airbnb. Also it facilitates users to determine various types of lodges by filtering through location, price etc. All this data generates an enormous amount of information which requires a data analytics approach to make the best use of it.

Not only promotions or competitive prices matter to get attention from new customers but also the right product to the right people. For a temporary rental residency company like Airbnb, their customers destinations are the key of their business. Where is the destination that a new registered user will chose to book with Airbnb? Which places can get the most attention from the new customers? These are one of the most crucial questions for a marketing team at Airbnb. Finding the right places which have high possibility to be booked by user will surely give a lot of benefits. Knowing the right product will be the most effective, once present it to the right person. The main key that must be defined for the better marketing is the line that links between the customer and the product. If the company provides the best services according to the customer's preferences,life style , likes, dislikes and other factors then market of the company will definitely increase.

No matter how good the products or services the company has, it would mean nothing if the customers never experience it yet. This is an important reason for gaining attention from new customers for Airbnb. Once the customers already experienced the products or services, they will acknowledge the existence of the product and once experienced the booking process they will be able to make the next one. First impression is another important factor to present the product. After customers get into the website and registered, they might not have a lot of time to spend on the website. Then, the products presented to them are the key to catch their attention and also cultivate their expectation about Airbnb.So new users spend more time on the website or in making their first booking by providing the places that are able to catch their attention.

Airbnb has the information filled by user from their first-time registration. The information of gender, age, places, language, devices, and affiliate are the example of information that are provided within the registration process.Some of these information is being used to determine the characteristics of a user which leads to the prediction of places for them which might be their first booking.

The main significance of this project is that Airbnb uses all the above factors to provide the best services to the customers. The prediction task in this project helps airbnb to increase its business since it can provide the destination country information beforehand so that users will be served well by airbnb.Airbnb can send the promotional messages to the new users which helps in promoting the hotels or houses. This project also reduces the time and effort to arrange the rental houses as we know in advance about the country of booking.

Some of the main goals of this project are as follows:

1) Determine the factors responsible to choose the travel destination booking of the user.

2) Predict the country booking for the new users using their demographic information and important features extracted.

3)Evaluate the prediction model using various measures such as accuracy, precision, recall etc. and analyze it with various factors.

II. Related Work

The Airbnb new user dataset is an open challenge problem on Kaggle [1]. This data set has been used by many researches with different approaches. As an example approach from Zhang, et al. [2], from their approach, after preprocessing some data they present multi-level of classification to classify the final destination country of the new users. In their first level they combine gbm, logistic, and polynomial classifier with voting process. And in the second level they used both SVM and logistic classifier to classify the rest classes. However, this work [2] didn't use the session data that is provided together with this data set which might be opportunity to gain the accuracy from this information. In the other hand the two levels approach seem to be very useful to classify the unbalanced data.

XGBoost or Extreme Gradient Boosting [3] is a scalable end-to-end tree boosting machine learning technique that very efficient and wildly use by data scientists. The example of the research that used XGBoost technique is a research proposed by Wang and Tan [4]. They used XGBoost and other methods to extract the features of BNP Paribas Cardif Claims Management which is one of another Kaggle competition. This methods is also used in scientific research [5] , which used XGBoost and another methods to create recurrent neural network to generate the molecule libraries for drug discovery. Their result shown that the Gradient Boosting Trees method was outperform all other methods.

III. Dataset

We will describe the dataset used for this project in this section. The dataset contains a list of users along with their demographics, web session records and some summary statistics. All the users in the dataset are from USA. It includes 5 csv files: train-users,test-users,sessions, countries,age-gender-bkts. The description of each of the file is as follows:

1) *train-users*: This file contains 171239 training examples with 16 properties

- id: userid
- gender
- date-account-created: date of account creation
- age
- date-first-booking
- timestamp-first-active
- signup-method
- signup-flow: the page a user came to signup from
- language: international language preference
- affiliate-channel: what kind of paid marketing
- affiliate-provider: where the marketing is
- first-affiliate-tracked
- signup-app
- first-device-type
- first-browser
- country-destination: target variable

2) *test-users*: This file contains 43673 items with 15 properties. Only country-destination variable is missing in this set. The training and testing set are split by dates. In the test set we will predict all the new users with first activities after 7/1/2014.

3)*sessions:* It includes the web session log for users. The sessions file contains 5600850 examples and 6 properties: user-id,action,action-type,action-detail, device-type,secs-elapsed.

4) *countries*: It contains statistics of destination countries and their locations i.e. latitude and longitude. It contains information about 10 countries. Destination country include 12 outcomes: US,FR,CA,GB,ES,IT,PT,NL,DE,AU,NDF and other. NDF means there wasn't any booking.

5) *age-gender-bkts*: It contains summary statistics of user's age group, gender, country of destination. It contains 420 examples with 5 properties.

IV. Methodologies

Our proposed approach includes data collection, exploratory analysis, data preprocessing, feature selection and classification. The first three steps are the data preparation tasks while the feature selection part includes retrieval of relevant features from the dataset.

The following part describes the above methods in detail. We will use the dataset mentioned in previous section from Kaggle. We have total of 214912 instances which will be divided in training and testing set in 80-20 ratio. Now firstly exploratory data analysis is performed in the whole dataset. It is used to determine the statistics and relationship between various features in the dataset. Also it provides an overview of the distribution of various features in the dataset.For example user's age, destination country might be related and range of ages might be related to destination country which provides the information of what age people prefer to go to which place. It will also help in determining the important factors responsible for destination country prediction.

We will preprocess the dataset to make it suitable for classification. We will preprocess the data by combining some of the data files based on some common attributes or group some attributes based on other attribute.We will look for any feature with lot of null values and correct the values with undefined format into proper format. We also need to normalize the values of various features so

as to prevent any discrepency in computation.Now we will apply the feature selection technique to determine the useful features out of the dataset.By looking into the distribution of various features in the dataset and relationship between them. We will determine what features are suitable for the target variable. Also the relationship of other variables with the target variable will provide an intuition of suitability of a feature. Feature selection technique will reduce the number of variables. We might use some existing feature selection techniques applicable for our data later.

Now we will apply machine learning classification algorithms for the prediction task. We will use the classification algorithm most appropriate for our data. We are planning to run two or three classification models on our data and compare their results. We will either use the multiclass classification or binary classification. We will use 10 fold cross validation for the validation of the model. Finally we will evaluate the prediction results with various measures. Accuracy, precision, recall are the common measures for the evaluation. We will compare the classification results of all models used using these measures.We will determine which method works well and the reason behind that. We will visualize the predicted countries in a map if time allows. Also we will visualize with various plots like scatter plots,bar plots etc. to get the relationship between various features.

V. RAW DATA VISUALIZATION

Before choosing classification methods, it is important to understand the characteristic of the data. The characteristics include structure, types, distributions, and cleanliness of the data. Once we understand all their characteristics, choosing to preprocess and classification methods will be more efficient. To help identify its properties, we visualized raw data and showed in this chapter. Each property is visualized and shown as follows:

A. country destination

This is the class data that we want to classify. It shows the destination that user booked. There are 12 possible classes : US (United States) , FR (France) , CA (Canada) , GB (Great Britain), ES (Spain), IT (Italy), PT (Poland), NL (Netherlands), DE (Germany) , AU (Australia) , NDF (No destination found) and other. The NDF status means there is no booking from that user. As shown in Figure 1, NDF is the most entity shown in our data, following by US.

B. gender

This attribute shows the gender of user. There are three possible factors of gender in this raw data unknown, female, male, and other. There is no NA values in this attribute. Figure 2 shows the quantity of each gender.

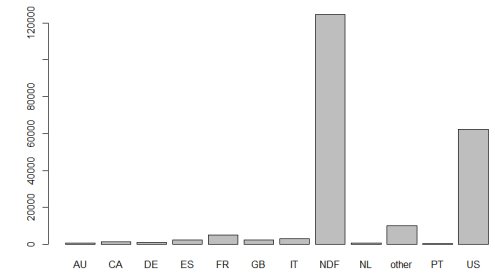

Figure 1. Histogram plot from raw data of country destination.

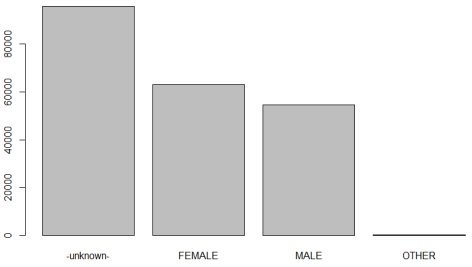

Figure 2. Histogram plot from raw data of gender of users.

C. age

This attribute shows the age of the user as an integer. In raw data, there are some mistake input which are birth years (eg. 1992) in the data. Figure 3, shows the scatter plot of raw age data. This attribute will be cleaned in the preprocessing phase. And there are 87990 of 213451 records that has no values in this attribute which is about 41 percent of all data.

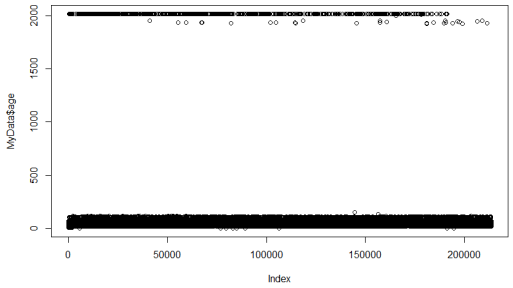

Figure 3. Scatter plot from raw data of age of users.

D. signup method

There are three possible methods that user able to used for signup basic, facebook, and google. Figure 4 shows the frequency of each method used by user.

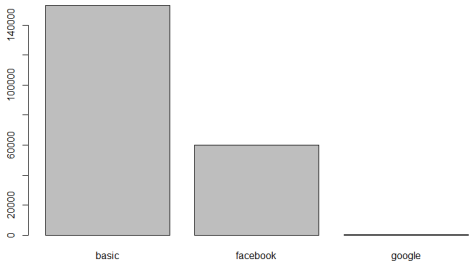

Figure 4. Histogram plot from raw data of signup methods.

E. signup flow

This data is an integer from range of 0 to 25. However, Airbnb doesnt provide any description about this attribute. The scatter plot of this attribute are shown in Figure 5.

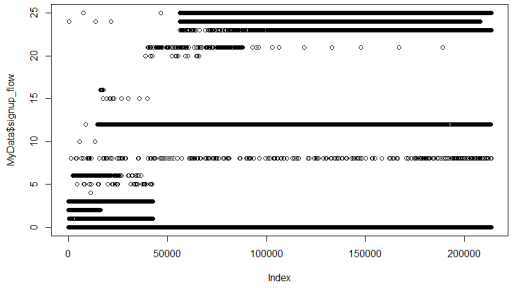

Figure 5. Scatter plot from raw data of signup flow.

F. language

Language is the attribute that determine the language preference preferred by user. There are 25 possible languages that user can choose. English is the most used language, selected by user. Exactly, there are 206314 of 213451 records which is more than 96 percent as illustrated in Figure 6.

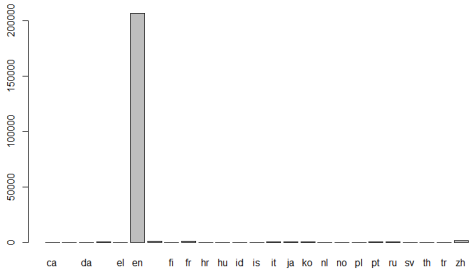

Figure 6. Histogram plot from raw data of language preference.

G. affiliate channel

There are eight possible affiliate channels for user to resister api, content, direct, other, remarketing, sem-brand, sem-non-brand, and seo. Figure 7 shows the histogram plot of all affiliate channels.

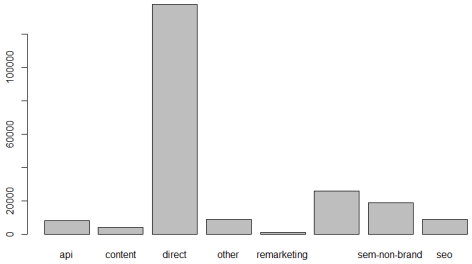

Figure 7. Histogram plot from raw data of affiliate channel.

H. affiliate provider

Registration is provide by different of 18 possible affiliate providers baidu, bing, craigslist, daum, direct, email-marketing, facebook, facebook-open-graph, google, gsp, meetup, naver, other, padmapper, vast, wayn, yahoo, and yandex. Figure 8 shows the histogram plot of all affiliate provider.

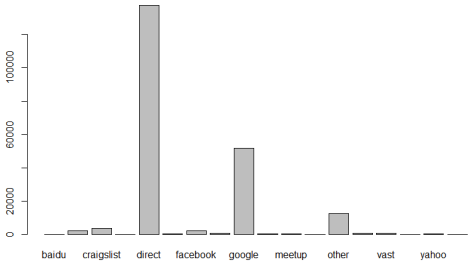

Figure 8. Histogram plot from raw data of affiliate provider.

I. first affiliate tracked

There are eight possible first affiliate tracked status linked, local ops, marketing, omg, product, tracked-other, and untracked. Figure 9 shows the histogram plot of first affiliate tracked.

J. signup app

User can use four possible application to do the signup process Android, iOS, Moweb (mobile-web) , and Web. Figure 10 shows the histogram plot of the application that user used for signup.

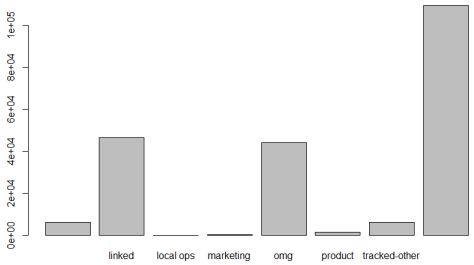

Figure 9. Histogram plot from raw data of first affiliate tracked.

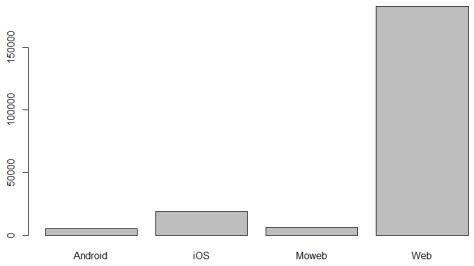

Figure 10. Histogram plot from raw data of affiliate provider.

K. *first device type*

From the raw data there are nine possible devices that user used to access Airbnb account which are Android Phone, Android Tablet, Desktop (Other, iPad, iPhone, Mac Desktop, Other/Unknown, SmartPhone (Other), and Windows Desktop. Figure 11 shows the histogram plot of the device that user first used with their account.

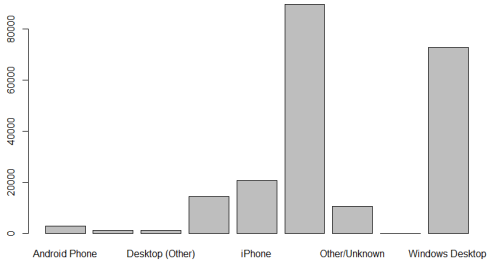

Figure 11. Histogram plot from raw data of affiliate provider.

L. *first browser*

From the record, users used 52 types of browser to access their Airbnb account. This attribute shows which browser that the user used for the first time. Figure 12 shows the histogram plot of the browser that user first used.

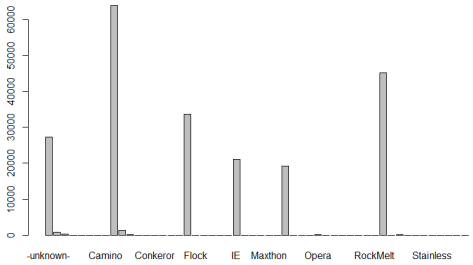

Figure 12. Histogram plot from raw data of affiliate provider.

VI. Data Preprocessing

After reviewed all attribute we did some data preprocessing with the data. This process includes reformatting, fixing mistake, extracting information, joining, and cleaning data. This chapter described the process that we apply to the data and visualized the processed result.

A. *Remove uninterested class*

From the country destination attribute that is the target class that we want to predict. There is a class NDF or No Destination Found class, which means user didnt booked Airbnb. So we removed all records that classified as NDF. The remained classes will be used as a result for our classification model. Figure 13, show the histogram plot of remaining data. There are 88908 left after remove NDF class.

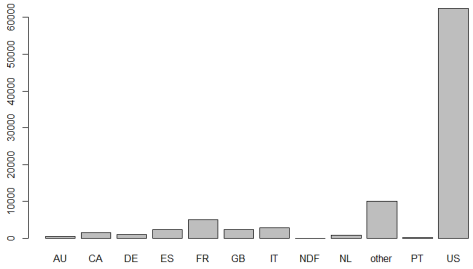

Figure 13. Histogram plot of destination after remove NDF.

B. *Fix mistake age data*

Some data in the age attribute are in the year format. So we recalculated age of user based on the date of last update of this data which is on December 15, 2015. In Figure 14 we plot the histogram of the age attribute. However, there are some unreasonable values so we filtered and keep the age range between 18 to 94. Then in Figure 15 show the divided age in to 10 ranges.

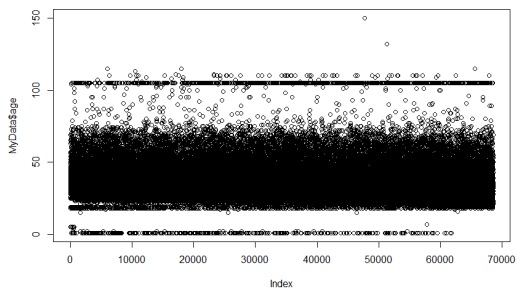

Figure 14. Scatter Plot of age after correct the mistake.

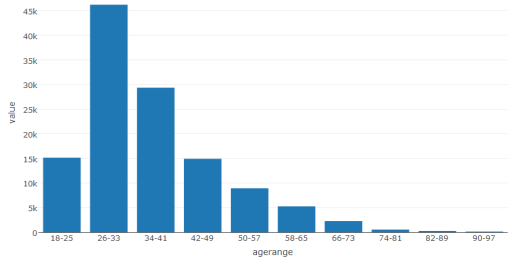

Figure 15. Histogram plot of age in range.

C. Adjust Data Format

Some attributes are not in the general format. For example, date account created, timestamp first active, and date first booking are date and time variables. However, none of them are in the same format, so we generalize them into the same format.

D. Join Data

From Kaggel, they also proved a dataset contain the session and time that user spend in each session. So we aggregate the amount of time that each user spend on Airbnb and join it with our dataset. Figure 16 illustrated the aggregated joined data with amount of time spend on Airbnb according to their destination and device types. Moreover, we created more features according to frequency of each session type that new user used.

E. Visualize Data

After the data was cleaned, we can better be visualized the attribute according to their class. As an example, gender attribute are plotted and shown in Figure 17.

VII. Proposed Methods

After preprocessing and cleaning the data, now we will use classification methods to classify the destination of Airbnbs new users. However, as shown in Fig. 14 the country destination in the data set contain a lot of US as the destination which makes the data unbalanced. To solve this problem, we decided to use two level classification as shown in figure 18. The classification are separated into two levels.

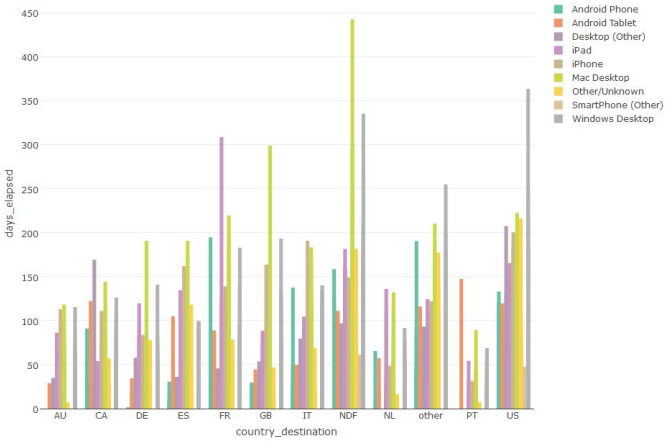

Figure 16. Histogram plot of time spend on Airbnb base on destination and device type.

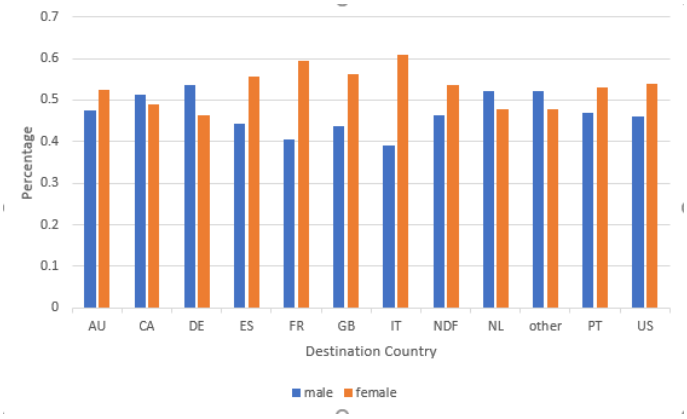

Figure 17. Histogram plot of gender base on destination.

The first level is a binary classifier, which will classify only US and non-US classes. Before putting the data into the first-level of classification,we labeled the data with US as the destination with 1 and other countries with 0. In this first-level, we used two classification methods, logistic regression and gbm model to be classified if the destination is US or not. We also performed 5-fold cross validation in each method to get more reliable models. We used averaging technique to combine two probability results from both the models. This combined probability will be compared with the set-up threshold, if the probability is higher than the threshold then the result from this first-level will be classified as US, otherwise it will be non-US. The non-US results from the first-level will be passed to the second-level classifier. The data of non-US countries are more balance than the previous stage that contain US. This second-level classifier is a multiclass classifier. It classifies other countries which are labeled as non-US into the real countries. We used extreme gradient boosting tree to get the classifier model for this second-level. We also used cross validation on this level to prevent creating an over fitting model. The final result from this second level

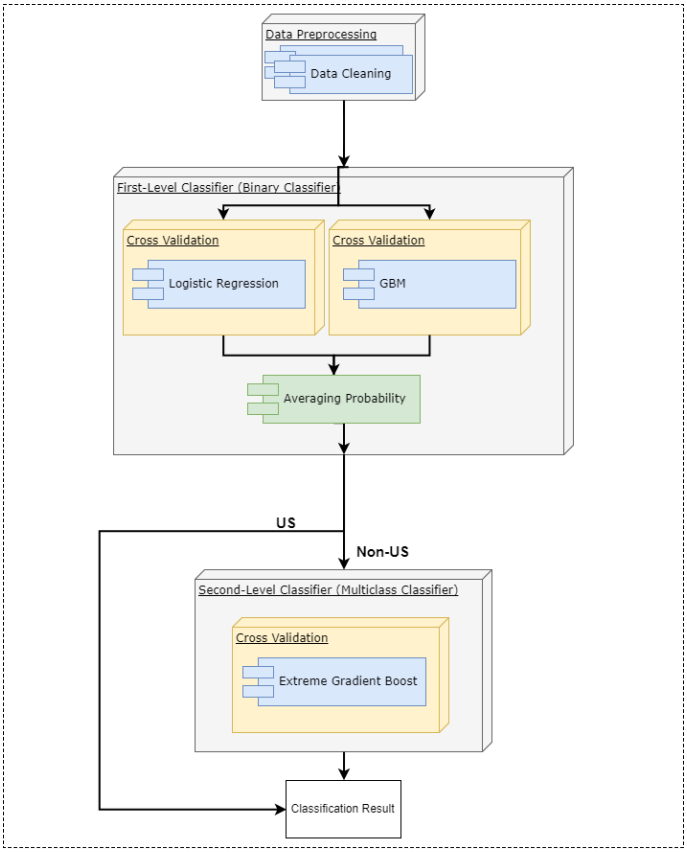

Figure 18. Work flow diagram of two levels classification.

will fulfill the final result in case that the destination is not US To measure the overall performance, we used 70 percent of data as a training set to create both first and second levels classifier. While another 30 percent will be used as a testing set, to measure the performance result which will be shown and discussed in the next section.

VIII. Experiment Results

In our proposed approach, we used a binary classifier to separate US users with Non US users. We use an ensemble of two classifiers .i.e. logistic regression and gbm, predicted the country destination as US or Non-US for each user and then take the average of both to be the final predicted probability. For the logistic regression, we determined the important features used for the classification as shown in Fig 19.

As we can see that session features are the most important ones for classification. The names for different actions and their descriptions are rather cryptic and it is hard to tell exactly what each action does, the 2 most valuable features are actions where one is called language multiselect and other is ajax google translate reviews. We can make a guess by seeing the importance values that these 2 features are closely tied to the confirmation of booking for destination as US or Non-US. The results for logistic regression are shown in Table 1.

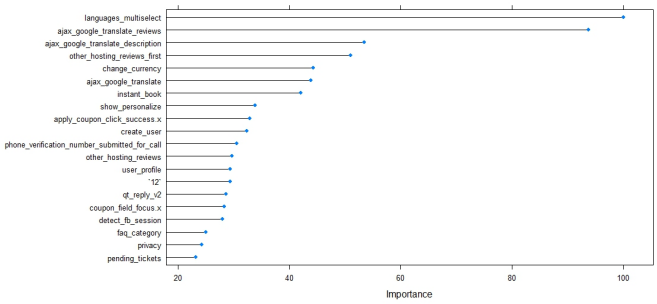

Figure 19. Variable importance plot for logistic regression.

Model	RMSE	RSquared	MAE
Logistic regression	0.45	0.029	0.4
gbm	0.446	0.054	0.4

Table I
Evaluation of Logistic and GBM classifier

ROC curve for the logistic regression and gbm part is shown in Fig 20. We achieved AUC value of 0.6351 for logistic regression which is not that good in this case. For the gbm model on the same data, the important features are related to session data but we achieved an auc value of 0.6367 in this case which is almost equivalent to logistic regression. Finally the average auc of both the models is 0.6352.

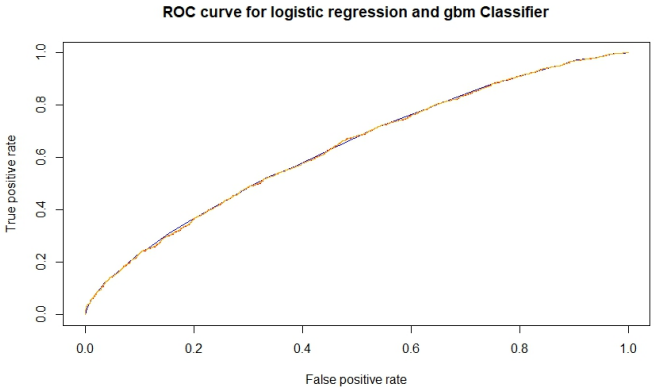

Figure 20. ROC curve for logistic regression and gbm.

After getting the average prediction probabilities, we build an xgboost model using the previous prediction values and other features in the training set. Since its a multiclass classification where we have to predict top 5 countries where a user will most probably book using Airbnb. So we used a technique to evaluate our classification. We determined the top 5 countries for each user using xgboost but the actual label is a single country destination with whom we need to compare. So we compared the actual country destination with all 5 predicted country destinations. If at least one of the predicted country matches the actual one, it is considered as correctly classified otherwise wrongly classified. We used this measure for each of the 5556 users and determined the accuracy.

We got an accuracy of 87.15 percent after applying xgboost classifier. We computed the important features required for model building as well as after model building. Important features before xgboost model building is shown in Fig 21. As we can see the important features in this case are age, month account, year account etc. It includes features from training set as well as from sessions data. This provides us information that the future booking depends upon the age of people too since younger and middle age people mostly book to travel across various country destinations. Also, date of account creation, gender, language and device type is also a factor for predicting the future bookings. Session information include the signup flow, affiliate channel etc. Fig 22. Shows the variable importance after xgboost model building. As we can see the most important country destinations include other, Australia, France, Canada etc. which shows that these country destinations are most likely be booked using Airbnb by the users.

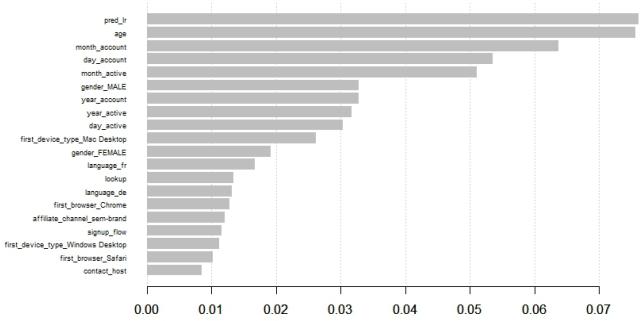

Figure 21. Variable importance before xgboost model building.

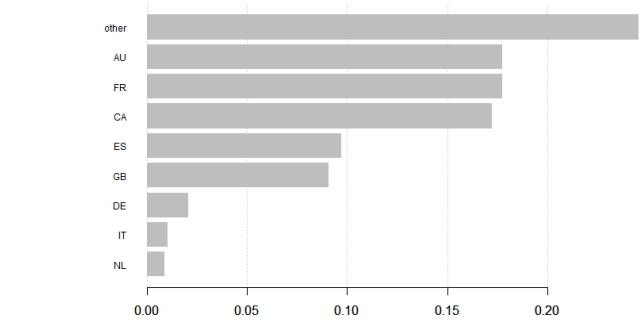

Figure 22. Variable importance after xgboost model building .

IX. Conclusion and Future work

In this paper we proposed an approach to predict top 5 country destinations that a user will book using Airbnb using session features, demographics and other factors etc. It provides us information about the factors affecting future bookings in Airbnb. As a future work, we have to experiment with different preprocessing procedures to improve the classification results. Also we will use sampling techniques to make this highly unbalanced data to balanced.

X. Conclusion and Future work

This is a collaborative work between us. We have equal contribution in each step i.e. data exploratory analysis, preprocessing, classification and paper writings. We have discussed and proposed the solution collaboratively.

References

[1] Kaggle, Airbnb New User Bookings, `https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings`

[2] Ke Zhang, Zhengren Pan, and Sichao Shi. *The Prediction of Booking Destination On Airbnb Dataset*. University of California, San Diego.

[3] Tianqi Chen and Carlos Guestrin, *Xgboost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco, California, USA, 2016.

[4] Wang, Peidong, and Ke Tan. *KAGGLE COMPETITION: BNP PARIBAS CARDIF CLAIMS MANAGEMENT*. Department of Computer Science and Engineering The Ohio State University, Columbus, OH.

[5] Segler Marwin H. S., Kogej Thierry, Tyrchan Christian, and Waller Mark P. *Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks*. ArXiv e-prints, 2017