Airbnb First Destination Country Classification Project Milestone Report

Project Scope

Project Objective:

The capstone project aims to predict which country will be new users' first booking destination so that more relevant content will be displayed on the users' interface. The analysis could help to deliver more customized messages to the users across all the touchpoints to increase the booking conversion probability of the new users.

For the analysis, user demographic data, web session records and some summary statistics about the destination countries will be used. The dataset could be downloaded from the link below:

https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

Approach for solving the problem:

The whole project will have four main stages:

- Data Cleaning (Clean up all the fields and merge all different data sources into one single analytical dataset)
- Data Exploratory Analysis (Understand the distribution of all variables, identify some of the constraints of the analysis due to the limitation of the data)
- Modeling: This is a classification problem. The methodologies I will try out will include (but limit to) logistics regression, Random Forest, Ada Boost, Gradient Boost and Xg Boost.
- Validation: Leverage the model to predict the testing dataset. The metrics to evaluate the model would be precision, recall, F1 score, ROC AUC and also the accuracy score.
- Presentation: Present insights from the model (eg: important online factors that could impact new users' decision on choosing their first destination country)

Deliverables: Scripts for each stage of the project and slides for presenting the methodologies, whole workflow and some additional insights out of the project.

Data cleaning

Raw Data Set:

5 raw datasets:

- 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
- 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
- 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
- sample_submission.csv - correct format for submitting your predictions

Data Cleaning Step:

- Group count of session by user id as count of activity count and sum of “secs elapsed” by user id as total time spent with airbnb website
- Merge user training data with the session data by user_id
- 58% of users have not made a booking yet. So I created the trip_booking_flag field so that before getting into the multi class classification problem to predict the first destination country, I can start with binary classification problem.
- There are some users with unreasonable ages that are lower than 18 or higher than 122. For the users with age that is lower than 18, the age is computed as 18. For the users with age that is higher than 122, the age has been computed as 122.
- Since there are some null values in the age. To resolve the issue, I have cut the age into buckets based on quantiles and null values are categorized into the unknown bucket.
- Since there are some null values in total time spent with airbnb website and total activity count fields, I have cut the fields into bucket based on quantile and the null values are categorized into the unknown bucket.
- For “total time spent” and “activity count” fields, I have also filled the null value with zeros to keep the continuous variables.

Feature Engineering: Creating extra features based on the existing features

- For the account creation date, four more fields have been created: year, month, day and day of week.
- For the first active date, four more fields have been created: year, month, day and day of week.
- For the first active date, four more fields have been created: year, month, day and day of week.
- In the user dataset, there are two timestamp fields: account creation date and first active timestamp. By leveraging these two timestamp fields, two numerical fields have been created: number of days since the account creation as of latest day and number of days since the first activity as of latest day.
- Since there are missing values in the dataset for user age, total time spent (in seconds) and also the session count. The buckets have been created for these fields by cutting the data into quantiles. For the missing values, a bucket called “Unknown” has been created for each of the fields.

Final Cleaned Dataset:

Below are the fields in the final cleaned dataset:

```
Index(['id', 'date_account_created', 'timestamp_first_active_cleaned',
       'gender', 'signup_method', 'signup_flow', 'language',
       'affiliate_channel', 'affiliate_provider', 'first_affiliate_tracked',
       'signup_app', 'first_device_type', 'first_browser',
       'country_destination', 'age_computed',
       'Account_creation_before_booking_flag', 'Total time spent (in seconds)',
       'number_of_active_day_as_of_latest_date',
       'number_of_days_since_account_creation_as_of_latest_date',
       'session count', 'trip_booking_flag', 'Account_creation_date_month',
       'Account_creation_date_year', 'Account_creation_date_day',
       'Account_creation_date_day_of_week', 'first_active_date_month',
       'first_active_date_day', 'first_active_date_year',
       'first_active_date_dayofweek', 'age_bucket',
       'Total time spent (in seconds)_fill_null_zero',
       'session count_fill_null_zero', 'Total time spent (in seconds)_bucket',
       'session count_bucket'],
      dtype='object')
```

Exploratory Data Analysis Part One - Variable Descriptive Analysis

User Demographic:

- **What do the users in the dataset look like?**
    - 213,451 users in total from US.
    - 45% of users have unknown gender.
    - 59% of users have unknown age.

User Acquisition Channel:

- **Where do the users come from?**
    - Most users are either coming to the website by typing in URL directly or coming from Google.

User Behaviors:

- **How did the users come to the website for the first time? How long have they been browsing on the website?**
    - Majority of users signed up through desktop.
    - Mac is more popular among users than any other device.
    - As of 7/1/2015, user age ranges from 1 year to a bit over 6 years. Most of users are 1 to 2.5 years old (based on their first activity on the site).

Booking History:

- **How many users have made the booking? What destination are more popular? How long did it take users to convert after getting on the site? How does the user growth look for the website?**
    - Only 42% of users have made a booking.
    - Most users like to travel within US. Europe is the second place they like to go.
    - Majority of users (75%) that have booked their first destination within 30 days since they got on the website for the first time.
    - Between 2013 and 2014, number of users on Airbnb has grown dramatically.
    - Take 2013 as an example to take a look at the new user acquisition trend, you will see there's seasonality when it comes to the user acquisition. October is peak month for user acquisition while the beginning of the year is the off-peak season for user acquisition for Airbnb.

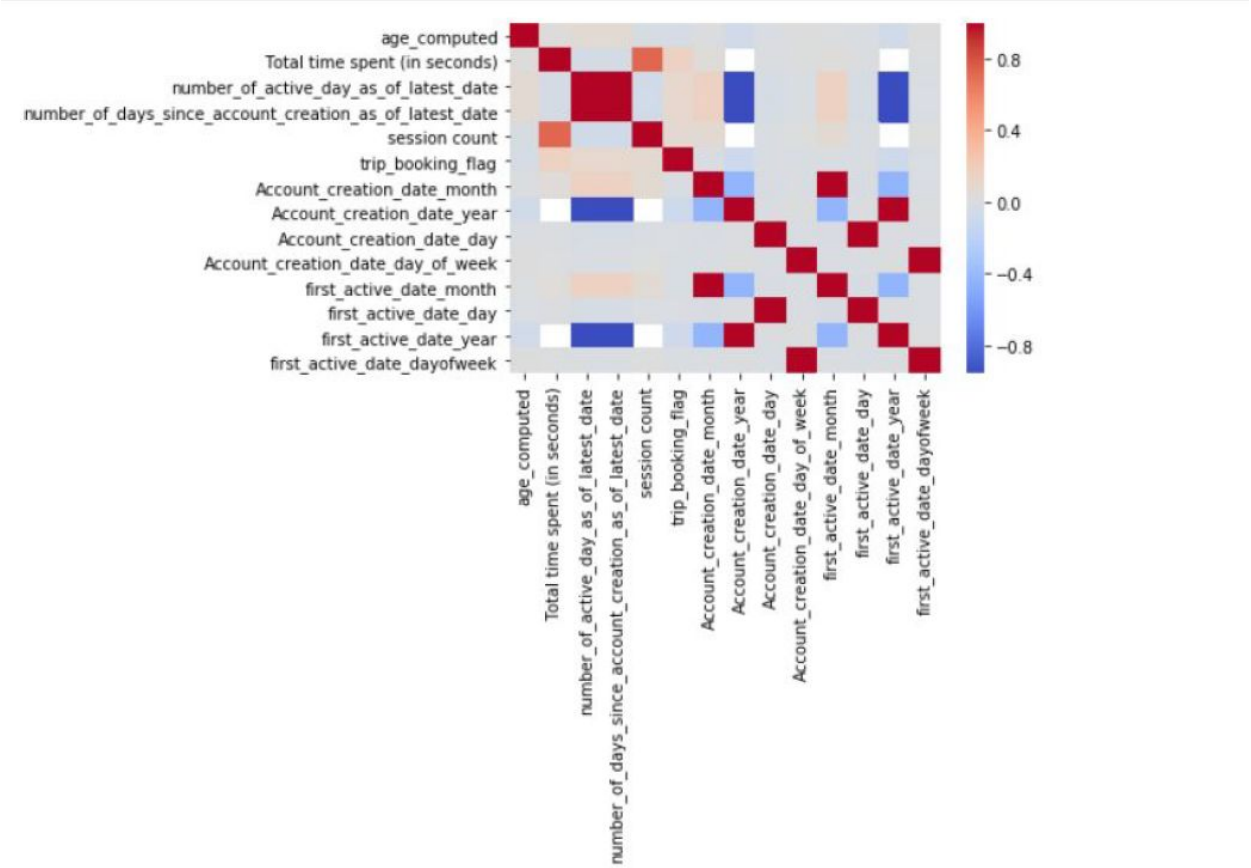
Correlation Exploration:

- **Is there any correlation between user acquisition channel and their booking behavior? Is it true the more time users have spent with website the more like users are to make a booking?**
    - Users coming from Web are more likely to make a booking.
    - Users signing up using Mac are more likely to make a booking.
    - Users that have made a booking have longer account age on average than users that have not made a booking as of 7/1/2015.
    - Users that have made a booking have done more activities and spent more time with the website on average than users that have not made a booking.

Exploratory Data Analysis Part Two - Variable Statistical Learning

Is there any strong correlation between the label (trip booking flag) and the features?

By looking at the heatmap below, the feature that is most correlated to the trip_booking_flag is the total time spent (in seconds).



Are there strong correlations between pairs of independent variables or *between an independent and a dependent variable?*

By looking at the data, below are the three question/hypothesis I have regarding the relationship between the variables:

- Bookers has stayed with the airbnb website longer than the non bookers (based on the first date active)
- Bookers has more activities on the airbnb website than non bookers (based on the session count)
- Age and gender could be important factors to determine which country users booked

What are the most appropriate tests to use to analyse these relationships?

For each of the hypothesis above, I have come up with the test below:

- Hypothesis 1: Bookers has stayed with the airbnb website longer than the non-bookers (based on the first date active)

  Ho: There's no difference between bookers and non-bookers on their number of days since they got active on the website.
  H1: Bookers stay longer with the website than non-bookers.

  For hypothesis 1, I have used z test to compare the average number of days since the first active date between the bookers and non bookers

  The conclusion is the high z test statistic indicates that the p value is very low and there's signification difference between bookers and non-bookers in number of active days.

- Hypothesis 2: Bookers has more activities on the airbnb website than non-bookers (based on the session count)

  Ho: There's no difference between bookers and non-bookers on their web activities.
  H1: Bookers have more web activities on airbnb than non-bookers on average.

  For hypothesis 2, I have also used z test to compare the average session count between bookers and non-bookers.

  The conclusion is bookers are on average more active than non-bookers on the airbnb website.

- Hypothesis 3: Age and gender could be important factors to predict which country users booked

  For hypothesis 3, I have used data visualization, chi square test and One Way Anova test to determine the interaction between the label and features like user age and user gender.

  By doing the Chi-square test and leveraging data visualization below,, we can reach the conclusion that there's a significant relationship between gender and booking destination. But the relationship is not that strong given Cramer's V is only 0.0635.