Capstone Project #1 - Airbnb New User's Bookings
Milestone Report

Overview

According to Wikipedia, "Airbnb is an American online marketplace company based in San Francisco, California, United States. Airbnb offers arrangements for lodging, primarily homestays, or tourism experiences. The company does not own any of the real estate listings, nor does it host events; it acts as a broker, receiving commissions from each booking" (https://en.wikipedia.org/wiki/Airbnb).

As known, Airbnb has hosts all over the world. When planning to visit any places, people do not only look at the hotels for their stay, but they also search Airbnb places. There might be many factors that affect visitors' decisions of their stay. With this Capstone project, I am trying to predict a new user's first choice of stay in their travel destination. New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. 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.

Potential Stakeholders

This analysis is expected to make direct and indirect impact on several stakeholders. First of all, using the model developed at the end of this project, Airbnb can make arrangements on their booking system. Therefore, they will be able to connect potential visitors with Airbnb hosts in more effective ways. Second of all, this project might help the potential Airbnb hosts to make arrangements to better serve their visitors. For example, Airbnb hosts might want to decrease or increase the cost of stays on certain days. In addition, potential visitors might also be affected by this project because the arrangements made by Airbnb or Airbnb hosts might increase the visitors intention to book their stay. Other companies similar to Airbnb as well might want to adopt the model created in this project. Furthermore, the users' preferences of days for bookings and browser, device, and application selections might create an impact on various potential stakeholders, such as App Developers.

Data

The data used in this project has been already available to people who are interested in working on this project. It was posted within the scope of a Kaggle competition in 2015. Even though the data was quite straightforward, it was still requiring a data cleaning and wrangling process prior to analysis. The data has six data sets in the form of CSV files. Here are the data sets and their summaries:

1. train_users.csv - the training set of users
2. test_users.csv - the test set of users
    a. id: user id
    b. date_account_created: the date of account creation
    c. 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
    d. date_first_booking: date of first booking
    e. gender
    f. age
    g. signup_method
    h. signup_flow: the page a user came to signup up from
    i. language: international language preference
    j. affiliate_channel: what kind of paid marketing
    k. affiliate_provider: where the marketing is e.g. google, craigslist, other
    l. first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
    m. signup_app
    n. first_device_type
    o. first_browser
    p. country_destination: this is the target variable you are to predict
3. sessions.csv - web sessions log for users
    a. user_id: to be joined with the column 'id' in users table
    b. action
    c. action_type
    d. action_detail
    e. device_type
    f. secs_elapsed
4. countries.csv - summary statistics of destination countries in this dataset and their locations
5. age_gender_bkts.csv - summary statistics of users' age group, gender, country of destination
6. sample_submission.csv - correct format for submitting your predictions.

DATA WRANGLING

In order to do further analysis, I applied a few data wrangling and cleaning techniques. First of all, I explored each dataset separately. I applied the “replace” method to have consistent data

across datasets. For example, I replaced “unknown” values with “NaN” and “male” values with “MALE.”

Then, I started to merge datasets to have my main dataset for further analysis. First, I merge the train dataset with the countries dataset using the “MERGE” method on “country_destination.” Each time, I checked whether or not the datasets were merged correctly using the “shape” method on new dataframes.

Sometimes, I figured out that I needed to rename some features. For example, I renamed the “id” column with “user_id” column for correct merging. Therefore, secondly, I merged the first merged datasets with the “sessions” dataset using the “MERGE” method on “user_id.” Third, I merge the second merged datasets with the “age_gender_bkts” dataset using the “MERGE” method on the “country_destination” and “gender” columns, which was my final merging.

This section describes the various data cleaning and data wrangling methods applied on the Airbnb datasets to make it more suitable for further analysis. The following sections are divided based on the datasets provided.

Finally, I conduct descriptive analyses on the features of the final dataset because I wanted to better understand the final dataset in terms of categorical and non-categorical variables. During this process, I realized that there were some ages over 100 and less than 18, which is not allowed. I replaced the ages more than 150 with NaN values. At the last step of data wrangling and cleaning, I created an output dataset in the CSV format.

DATA STORYTELLING

In this part of my first capstone project, I will work on the output that I created after the data wrangling process to explore, visualize, and create stories based on the visualizations. I created visuals based on 13 questions in total.

1. **What countries do users mostly want to travel to?**

   About 55% of the users did not travel to anywhere while about 30% of the users travel to the U.S.

2. **What is the distribution of genders who are being active to book a stay?**

   The percentage of NaN values is about 47%. In addition, while 29% of the users are female, 22% of them are male.

3. **What is the distribution of the age of the users?**

   Most of the users are between 20-40 years old. There are only a few users in the age of 80-100. Also, there are some users over 100 age but I believe those users might not want to enter their original ages. Furthermore, there are some values lower than 18 and I don't think Airbnb will allow them to book on their website.

4. **What is the distribution of the language of the destination countries?**

   Majority of users used English when they interacted on the Airbnb website. This is not surprising because the destination country for the most of the users was the U.S.

5. **What device do users mostly use for booking?**

   About 42% of the users did their search and/or booking on a Mac Desktop while 30% of them used a Windows Desktop computer. In terms of smartphones, the users mostly used an iPhone (~11%).

6. **What methods do users mostly use to signup for booking a stay?**

   Majority of users signup through basic signup method while 22% of users prefer Facebook.

7. **What browser do users mostly prefer using for booking their stay?**

   74% of the users used one of Chrome, Safari, Firefox, and Mobile Safari as the browser.

8. **What are the most common actions while booking?**

   21.5% of users apply "show" action. I can see that users did "show" action after doing other actions. Therefore, it can be said that "show" is the result of any other actions.

9. **What days of a week do users mostly become active for booking a stay?**

   Users are mostly more active on Wednesdays and Thursdays and less active on weekends. Then, we can say that users will more probably book on weekdays than weekends.

10. **What are the most common applications that users mostly use to signup based on destination countries?**

Among non-defined countries and the US, web is the mostly-used signup method while the Moweb is the least one. There were some other apps as well, which are not even listed because they don't make too much impact.

11. **What is the most common signup method that users mostly use to signup based on destination countries?**

    In all countries, a basic signup method is the most preferred method. Here, I directly eliminated NDF values.

12. **What is the relationship between users' ages and destination countries?**

    I created boxplots which showed that medians are very close to each other for all destination countries. Users who are booking a stay in the US are the youngest ones compared to other countries. Also, users who are booking their stays in Great Britain, France, Italy and Netherland tend to be older. Users for Canada and Australia also tend to be younger.

13. **What days of a week that users mostly prefer for booking based on destination countries?**

    Here, I used countplots which showed that users who want to book a stay in the US are mostly active on Wednesdays and Thursdays while less active on weekends. I wanted to look at the results one more time without US because it seems there might some interesting results for other destination countries.

    When users select France as their destination country, they become less active on Monday and Friday while more active Wednesday and Sunday. They are less active on Friday for Great Britain as well. Users try to book a stay mostly on Monday for Italy and Spain. For countries like Netherlands and Portugal, they prefer Saturday.

APPLYING INFERENTIAL STATISTICS

In terms of inferential statistics, I used frequentist approach and bootstrap method for hypothesis testing. In order to do that, I used a part of the dataset, which included user id, country destination, gender, age, and elapsed time.

Users booked in the US are mostly female.I wanted to find out if there is a significant difference between males and females in terms of their ages. The null hypothesis was "there is no difference

between males and females who select their country destination as US in terms of their ages". The alternative hypothesis was "there is a difference between two groups."

The difference between the means of males and females was 0.21900494962098804 and the standard error was 0.027184520259685614. When calculating the t and p values manually, I found t values as -7.11880407836792, which showed that the sample mean was less than the hypothesized mean. Then, I applied a two-sample t-test at 0.05 alpha level. I found t-value as 8.056237429570178 and p-value as 1.999. After that, I calculated t and p values using scipy and found the values as statistic=7.118804078366678, pvalue=1.0897469575308629e-12. According to these results, the p-value was less than 0.05 and I rejected the null hypothesis and obtained the alternative hypothesis which showed that there was a significant difference between the means of males and females ages.

I also applied the same method to find out if there is a significant difference between males and females in terms of their elapsed time. Therefore, the null and alternative hypothesis were like following:

Ho: There is no significant difference between the means of females and males.
Ha: There is a significant difference between the means of females and males.

In the dataset, females were 55% of the sample while 45% were males. I checked the confidence interval for the elapsed time at 95% level, and found it 22456.68561000778, which indicated that the mean was above 22456 at 95% confidence interval. When manually calculated the t and p values, I found t-value -4.521679006836198 and p-value 1.999. When applied a two-sided test using scipy method, the results were statistic=4.521679006836163, pvalue=6.136141100266454e-06, which showed that there was significant difference between two groups. But those were the results when we assumed that there was no difference between variances of two groups. I wanted to look for the values when the variances were not assumed to be the same. And I found statistic=4.487927633810977, pvalue=7.1932208754725195e-06. Because the pvalue was less than 0.05, I rejected the null hypothesis which was "there was no difference in elapsed times between females and males." Instead, I obtained the alternative hypothesis which obviously showed that there was a significant difference between these two groups who booked in the US.

Finally, I used the bootstrap method and confidence interval. My Null and Alternative Hypothesis were as follows:

H0 : there is no difference in standard deviations between males and females

Ha : there is a difference in standard deviations between males and females

When I performed 10000 replicates immediately after setting the random seed to 47, I got the value 22456.2 here, which compares almost perfectly with the value 22456.6 obtained using the t-distribution confidence interval previously. I also found the difference of the standard deviation of the original sample 6339.8689770809. Also, the 95% confidence interval of the difference in standard deviation between two groups was [3723.4564025594013 , 8950.454220856525], which was not containing zero. Thus I rejected the null hypothesis. There was a significant difference in standard deviation (and thus variance) between males and females in terms of elapsed time.

I, furthermore, performed a bootstrapped hypothesis test at the 5% significance level (alpha = 0.05) to calculate the p-value of the observed difference between females and males.

For α= 0.05:

H0 : there is no difference in charges between males and females.
Ha : there is a difference in charges between males and females.

Here are the steps, I performed at this point:

- Shifting the Dataset so that the two groups have equal means
- calculating the combined mean
- Generating the shifted dataset for males and females.
- Getting the differences for the bootstrap simulated sample
- Getting the observed difference from the actual dataset (1089.7511982369979)
- Calculating the p-value by comparing the bootstrap replicates against the observed difference of the means (p-value=0.00)
    - Under the null hypothesis, we get a p-value of 0. Thus it is sufficiently unlikely that the null hypothesis is true and thus we reject the null. There is a difference in elapsed time between females and males.