6/21/2020

Anu Kashyap

Springboard

Airbnb New User Booking

Springboard Capstone Project # 1

Contents

[Background 2](#_Toc43305765)

[Problem Statement 2](#_Toc43305766)

[Objective 2](#_Toc43305767)

[Kaggle Data 2](#_Toc43305768)

[Data Wrangling 3](#_Toc43305769)

[Exploratory Data Analysis 3](#_Toc43305770)

[1. Destination Country and Gender 3](#_Toc43305771)

[2. Destination Country and Signup Method/Language 3](#_Toc43305772)

[3. Destination Country and First Device Type 4](#_Toc43305773)

[4. Destination Country and First Browser 5](#_Toc43305774)

[5. Destination country and Affiliate Channel 5](#_Toc43305775)

[6. Destination Country and Affiliate Provider 6](#_Toc43305776)

[7. Destination Country and Age 6](#_Toc43305777)

[8. Destination Country and Days Since First Booking 7](#_Toc43305778)

[9. Destination Country and Seconds per Session 8](#_Toc43305779)

[Independent Variables 8](#_Toc43305780)

[Comparison Between Prediction Models 9](#_Toc43305781)

[Variable Importance Using XGBoost Classification 10](#_Toc43305782)

[Insights 11](#_Toc43305783)

# Background

Airbnb is an American online marketplace company based in San Francisco, California, United States.

Airbnb offer arrangement for lodging, primarily homestays, or tourism experiences

Instead of waking to overlooked "Do not disturb" signs, Airbnb travelers find themselves rising with the birds in a whimsical treehouse, having their morning coffee on the deck of a houseboat, or cooking a shared regional breakfast with their hosts

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries.

# Problem Statement

In which country will a new guest book their first travel experience?

# Objective

The objective is to help Airbnb understand the following:

1. Better forecast demand
2. Share customized content with client
3. Decrease the average time to first booking

# Kaggle Data

In this challenge, Kaggle has given a list of users along with their demographics, web session records, and some summary statistics.

All the users in this dataset are from the USA

There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other’.

'NDF' is different from 'other' because 'other' means there was a booking, but it is to a country not included in the list, while 'NDF' means there wasn't a booking.

The training and test sets are split by dates. In the test set, we need predict all the new users with first activities after 7/1/2014

The sessions dataset, the data only dates back to 1/1/2014, while the ‘users’ dataset dates back to 2010.

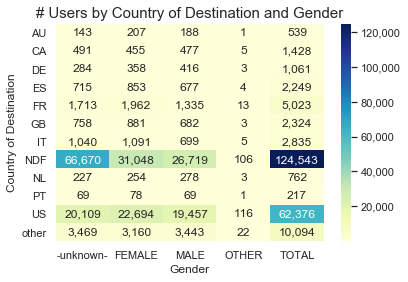
# Data Wrangling

1. Concatenate train and test data so that data cleaning performed together.
2. Drop ‘date\_first\_booking’ column which is entirely missing in test data.
3. Replace unknown values in ‘gender’ and ‘f'irst\_browser’ columns to NaN.
4. ‘Age’ column will only have values between 18-100 years, therefore, other values will be set to NaN.
5. Date time data of 'time\_first\_active' and 'timestamp\_first\_active' to be split into day, month and year columns.
6. Drop ‘date\_account\_created’ and ‘timestamp\_first\_active’ columns after completing step 5.
7. Remove leading and trailing spaces from ‘language’ column.

# Exploratory Data Analysis

## Destination Country and Gender

* 1. Each country has similar distribution of users in Male/Female categories
  2. Using Chi-Square test, there’s a statistically significant association between country of destination and gender

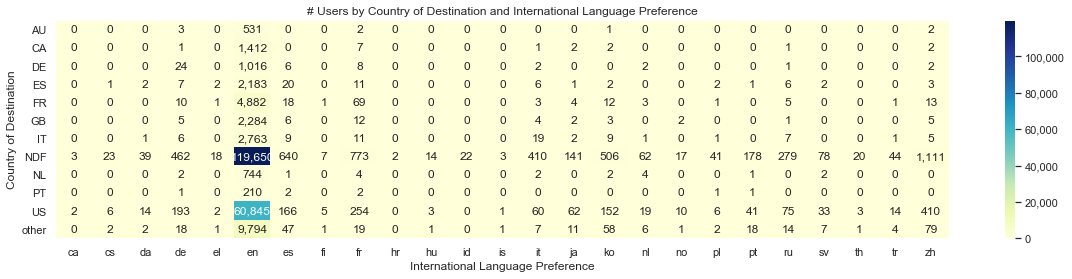


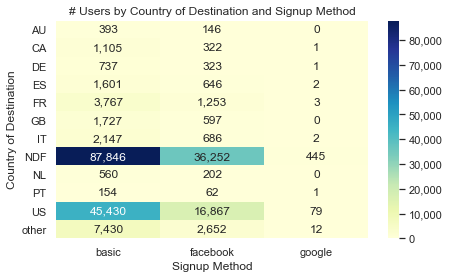
## Destination Country and Signup Method/Language

‘English’ is the international language of preference for almost all users

Most users signup either through their own website or Facebook

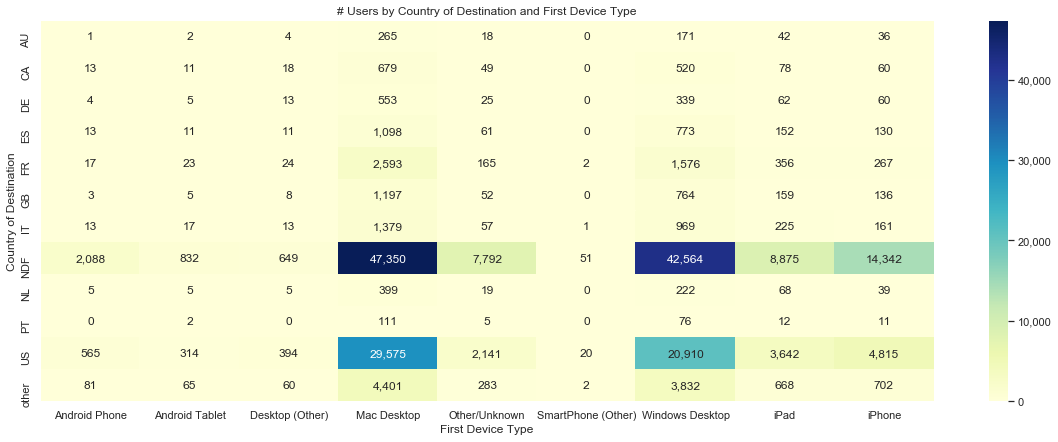
* Using Chi-Square test, there is a statistically significant association between:
* Country of destination and international language preference
* Country of destination and signup method





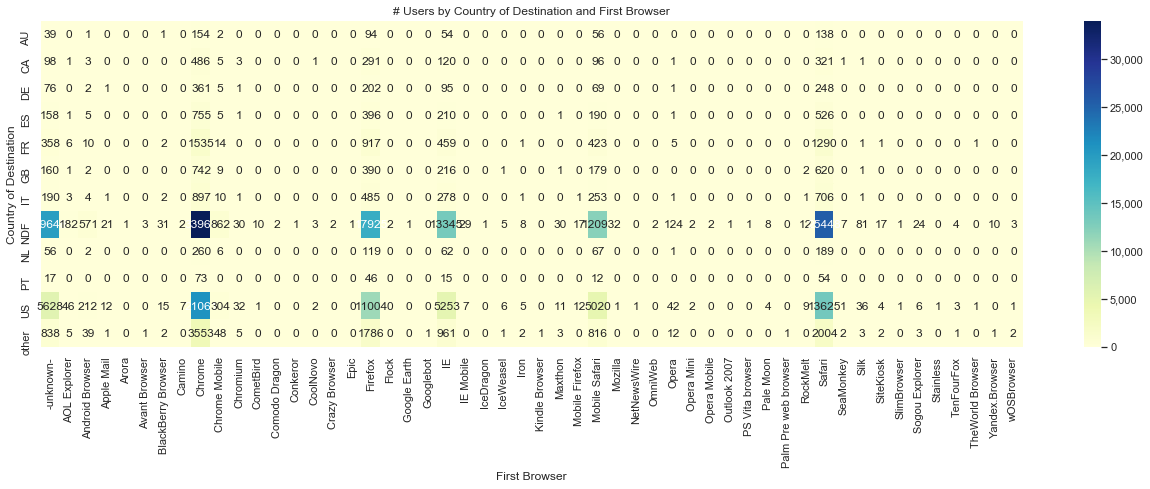
## Destination Country and First Device Type

* Most users are either Mac Desktop users or Windows Desktop users
* Using Chi-Square test, there’s a statistically significant association between country of destination and first device type



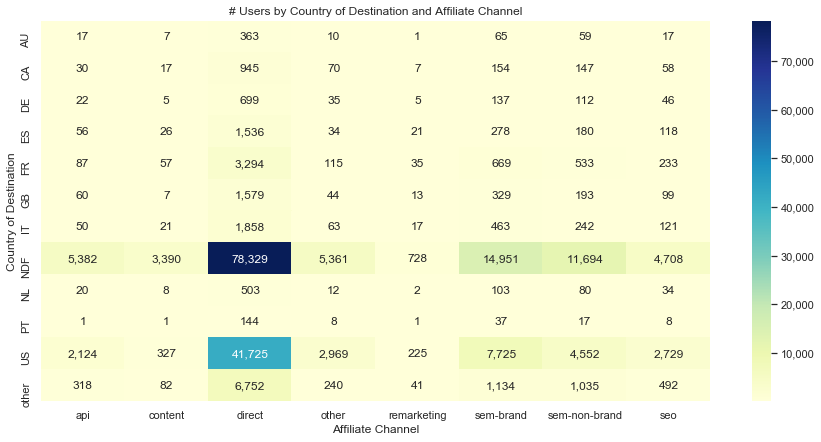
## Destination Country and First Browser

* Chrome, Firefox, IE, Safari and most popular first browsers used by user



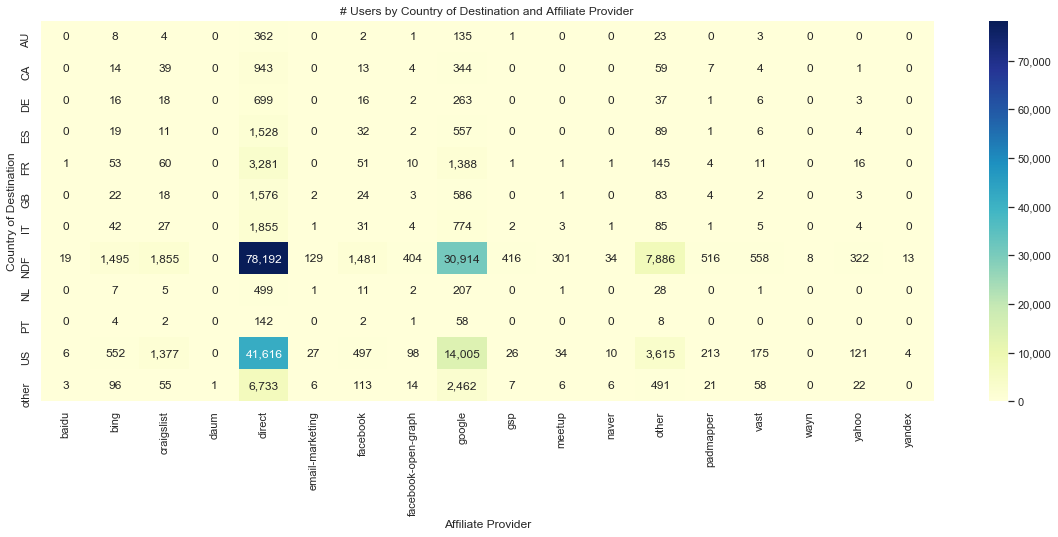
## Destination country and Affiliate Channel

* The most popular paid marketing channels are direct, sem-brand and sem-non-brand



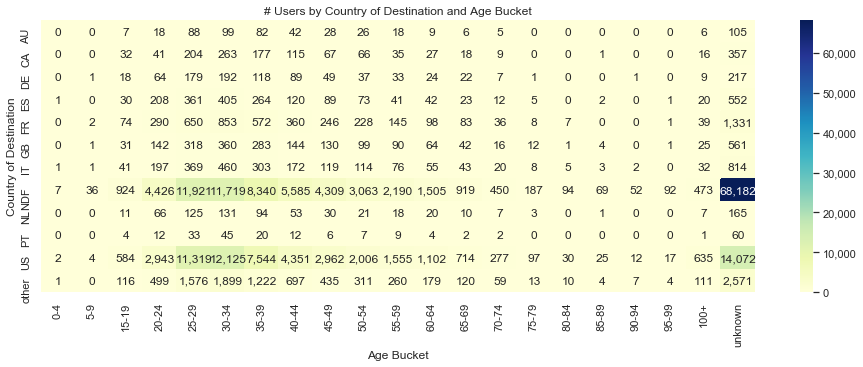
## Destination Country and Affiliate Provider

* The most popular affiliate providers are direct and google



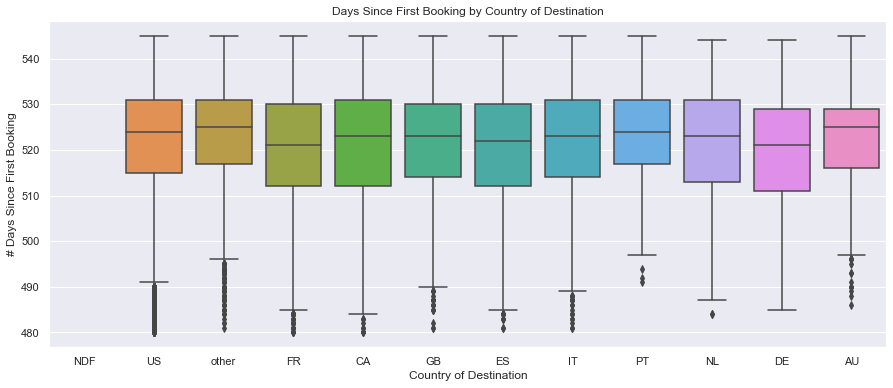
## Destination Country and Age

* Most users with destination country as US, are between ages 25-39 years

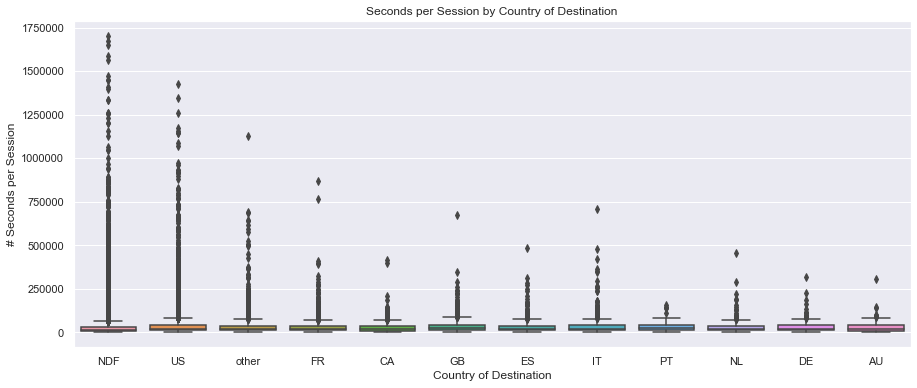


## Destination Country and Days Since First Booking

As per the ANOVA test, The F-statistic= 38.9 and the p-value < 0.05 which indicates that there is a statistically significant association between country of destination and days since first booking but this test may not be very reliable because it may violate some of the assumptions of ANOVA



## Destination Country and Seconds per Session



# Independent Variables

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Variable name** | **Description** | **Type** |
| 1 | gender | Gender of the user | categorical |
| 2 | age | Age (in years) of the user | numeric |
| 3 | signup\_method | Signup method used by user e.g. basic, facebook or google | categorical |
| 4 | Signup\_flow | the page a user came to signup up from | categorical |
| 5 | language | international language preference | categorical |
| 6 | affiliate\_channel | type of paid marketing | categorical |
| 7 | affiliate\_provider | where the marketing is e.g. google, craigslist, other | categorical |
| 8 | first\_affiliate\_tracked | whats the first marketing the user interacted with before the signing up | categorical |
| 9 | Signup\_app | app through which the user signed up | categorical |
| 10 | first\_device\_type | the first device type used by user e.g. phone, tablet, desktop | categorical |
| 11 | first\_browser | the first browser used by user | categorical |
| 12 | dac\_year, dac\_month, dac\_day | year, month and date when the account was created by user | numeric |
| 13 | tfa\_year, tfa\_month, tfa\_day | year, month and date when the user was first active | numeric |
| 14 | cnt\_action | # of actions by user | numeric |
| 15 | cnt\_uniq\_action\_type | # of unique action types by user | numeric |
| 16 | cnt\_uniq\_dev\_type | # of device types by user | numeric |
| 17 | secs\_per\_session | average # of seconds elapsed per session by user | numeric |

# Comparison Between Prediction Models

|  |  |  |  |
| --- | --- | --- | --- |
|  | **XGBoost Classification** | **Multi-Class Logistic Regression** | **Random Forest Classification** |
| Dependent variable | Predicts destination country out of 12 countries | Predicts destination country out of 12 countries | Predicts destination country out of 12 countries |
| Hyperparameter Tuning Method | Gridsearch | Not applicable | Gridsearch |
| Optimal Hyperparameters | max\_depth = 6  learning\_rate = 0.1  n\_estimators = 70  objective = 'multi:softprob' | penalty = ‘l2’  multi\_class = ‘multinomial’  solver = ‘lbfgs’  C = 0.1 | criterion = ‘entropy’  max\_depth = 15  max\_features = ‘sqrt’  n\_estimators = 150 |
| Accuracy (Train set) | 65.4% | 58.4% | 68.5% |
| Accuracy (Test set) | 64.2% | 58.1% | 49.1% |
| Accuracy (Hold out set on Kaggle) | 85.7% | Not calculated | Not calculated |

# Variable Importance

|  |  |  |
| --- | --- | --- |
| **Variable name** | **Importance** | **Description** |
| gender | 14.7% | gender of the user |
| first\_browser | 12.5% | the first browser used by user |
| signup\_method | 12.1% | signup method used by user e.g. basic, facebook or google |
| age | 10.5% | age (in years) of the user |
| first\_affiliate\_tracked | 8.1% | whats the first marketing the user interacted with before the signing up |
| affiliate\_channel | 8.0% | type of paid marketing |
| affiliate\_provider | 7.1% | where the marketing is e.g. google, craigslist, other |
| tfa\_year | 6.3% | year, month and date when the user was first active |
| first\_device\_type | 5.3% | the first device type used by user e.g. phone, tablet, desktop |
| cnt\_uniq\_action\_type | 5.1% | # of unique action types by user |
| signup\_app | 3.9% | app through which the user signed up |
| signup\_flow | 1.6% | the page a user came to signup up from |
| dac\_year | 1.2% | year when the account was created by user |
| tfa\_month | 0.7% | month when the user was first active |
| secs\_per\_session | 0.6% | average # of seconds elapsed per session by user |
| cnt\_uniq\_dev\_type | 0.6% | # of device types by user |
| cnt\_action | 0.6% | # of actions by user |
| dac\_month | 0.5% | month when the account was created by user |
| tfa\_day | 0.4% | day of the month when the user was first active |
| dac\_day | 0.4% | day of the month when the account was created by user |

# Insights

1. Out of the three classification models used, XGBoost, Multi-class Logistic Regression, Random Forest, XGBoost classification model gives the best accuracy at 85.7% on the Kaggle test set or hold out set
2. XGboost hyperparameters were determined using Gridsearch and the best hyperparameters are:
   1. max\_depth = 6
   2. learning\_rate = 0.1
   3. n\_estimators = 70
   4. objective = 'multi:softprob'
3. The top 5 variables as per the XGBoost variable importance are:
   1. Gender
   2. First\_Browser
   3. Signup\_Method
   4. Age
   5. First\_Affiliate\_Tracked