Predicting Airbnb First Users' Bookings

Remzi Kizilboga

June, 2020

Problem Statement

We all might travel. Airbnb has always been an option for a traveler’s stay. However, it is important to know that where travelers will make their first bookings because having this information will help Airbnb to share more personalized content with their community and Airbnb hosts to make better arrangements for customers’ stays so that increase travelers’ satisfaction.This analysis explains the model which developed to predict Airbnb first bookings based on various factors.

Stakeholders

- Primary Stakeholders
  - Airbnb
  - Airbnb hosts
  - Travelers
- Secondary Stakeholders
  - Other booking companies like Vrbo, Booking.com, HomeAway.com, FlipKey.com

Data Acquisition and Wrangling

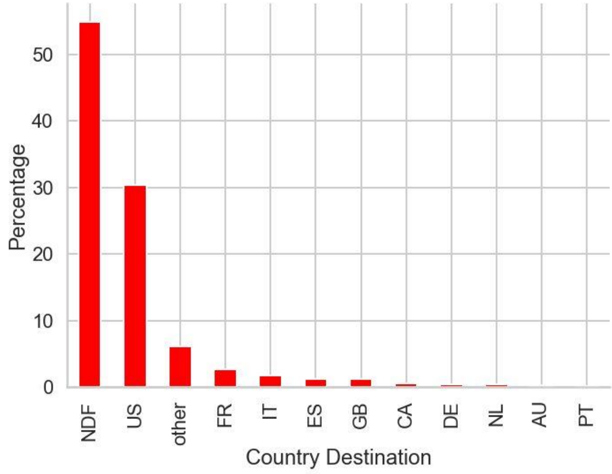
- https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings
-

```
In [54]: dfmergefinal.shape

Out[54]: (3340486, 30)

In [59]: dfmergefinal.country_destination.value_counts(dropna=False)

Out[59]: NDF      1833467
         US       1013036
         other     205497
         FR         90282
         IT         59932
         ES         41473
         GB         39540
         CA         19006
         DE         14512
         NL         12830
         AU          7434
         PT          3477
         Name: country_destination, dtype: int64
```

Data Storytelling

- What countries do users mostly want to travel to?
- Imbalanced classification

```
NDF      54.886235
US       30.326006
other     6.151710
FR        2.702661
IT        1.794110
ES        1.241526
GB        1.183660
CA        0.568959
DE        0.434428
NL        0.384076
AU        0.222542
PT        0.104087
Name: country_destination, dtype: float64
```



Data Storytelling

- What is the distribution of genders who are being active to book a stay?

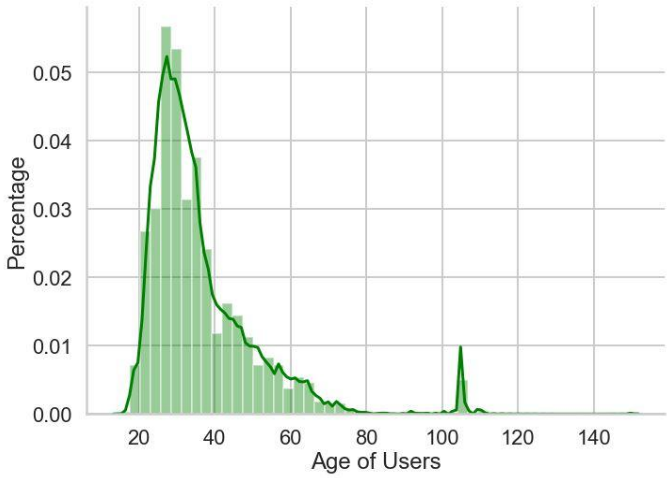
```
NaN       47.222590
FEMALE    29.443530
MALE      23.225363
OTHER      0.108517
Name: gender, dtype: float64
```

- What device do users mostly use for booking?

```
Mac Desktop           41.706386
Windows Desktop       30.372796
iPhone                10.797770
Other/Unknown          7.556445
iPad                   7.068642
Android Phone          1.136691
Android Tablet         0.734594
Desktop (Other)        0.606109
SmartPhone (Other)     0.020566
Name: first_device_type, dtype: float64
```

Data Storytelling

- What is the distribution of the age of the users?



Inferential Statistics

- p-value, t-test

```
In [19]: #calculate t value manually
         n0 = len(male_age)
         n1= len(female_age)
         std0 = male_age.std()
         std1= female_age.std()
         mean0 = mean_male
         mean1= mean_female
         sp = np.sqrt( ((n0-1)*(std0)**2 + (n1-1)*(std1)**2)/ (n0+n1-2) )
         t_ = (mean1 - mean0)/(sp * np.sqrt(1/n0 + 1/n1))
         print(t_)

         -7.118804078367921

In [20]: # Use 0.05 Significance level in two sample t-test
         t_val=((male_age_mean - female_age_mean)-0)/SE
         print(t_val)

         8.056237429570178

In [21]: #calculate p value manually
         p_value = (1 - t(n0 + n1 - 1).cdf(t_)) * 2
         p_value

Out[21]: 1.9999999999989102

In [22]: #calculate t and p values using scipy
         ttest_ind(male_age, female_age)

Out[22]: Ttest_indResult(statistic=7.118804078366678, pvalue=1.0897469575308629e-12)
```

Inferential Statistics

- p-value, bootstrapping

In [51]:
```
# # Shifting the Dataset so that the two groups have equal means

# First calculating the combined mean
combined_mean = np.mean(np.concatenate((male_bts, female_bts)))

# Generate the shifted dataset
male_shifted = male_bts - np.mean(male_bts) + combined_mean
female_shifted =  female_bts - np.mean(female_bts) + combined_mean
```

In [52]:
```
# Draw the bootstrap replicates from the shifted dataset
bs_replicates_male = draw_bs_reps(male_shifted, np.mean, size=1000)
bs_replicates_female = draw_bs_reps(female_shifted, np.mean, size=1000)
```

In [53]:
```
# Get the differences for the bootstrap simulated sample
bs_differences = bs_replicates_male - bs_replicates_female


# Get the observed difference from the actual dataset
obs_diff = np.mean(male_bts) - np.mean(female_bts)
obs_diff
```

Out[53]: 1089.7511982369979

In [54]:
```
# Calculate the p-value by comparing the bootstrap replicates against the observed difference of the means
# The fraction of values WITHIN bootstrap replicates array that meet a certain criteria against the obs_di
ff

p = np.sum(bs_differences >= obs_diff)/ len(bs_differences)
print('p-value =', p)
```

p-value = 0.0

Baseline Modeling

- Additional wrangling
- First parameter - 1507019 data points in total while 181 features
- Second parameter - “country destination” without NDF - Target variable
- Accuracy scores

```
warm_start=False)

In [96]: y_predict_test = clf.predict(Xtestlr)
         print("\n")
         print("[Test] Accuracy score (y_predict_test, ytestlr):",accuracy_score(y_predict_test, ytestlr))

         # Note the order in which the parameters must be passed
         # according to the documentation ... although there should be
         # no difference since it is a one-to-one comparison ...
         # ref: http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html#sklearn.metric
         s.accuracy_score
         print("\n")
         print("[Test] Accuracy score: (ytestlr, y_predict_test)",accuracy_score(ytestlr, y_predict_test))

         # also printout the training score
         y_predict_training = clf.predict(Xlr)
         print("\n")
         print("[Training] Accuracy score: (ylr, y_predict_training)",accuracy_score(ylr, y_predict_training))

[Test] Accuracy score (y_predict_test, ytestlr): 0.9564770208756354


[Test] Accuracy score: (ytestlr, y_predict_test) 0.9564770208756354


[Training] Accuracy score: (ylr, y_predict_training) 0.9562538621367518
```

Baseline Modeling

- Imbalanced classification
- AU, CA, GB were problematic

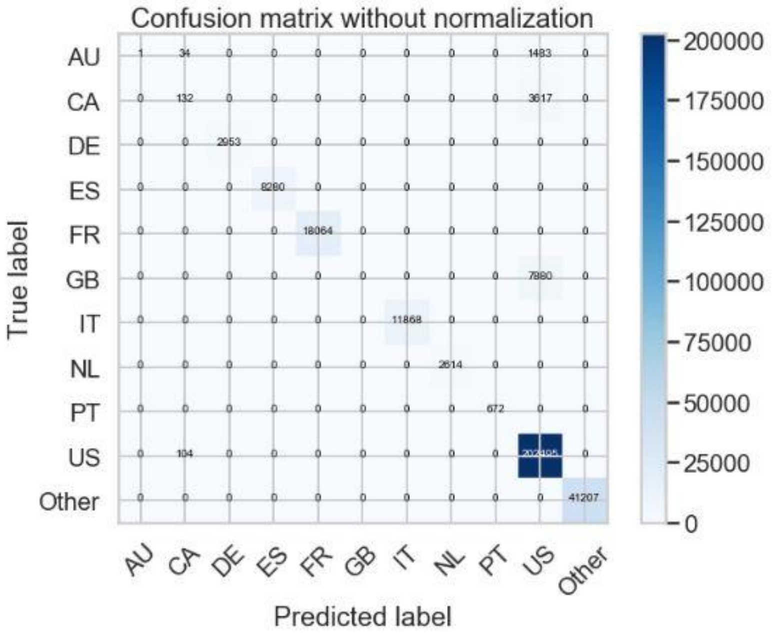
	precision	recall	f1-score	support
AU	0.54	0.00	0.00	5916
CA	0.46	0.03	0.06	15257
DE	1.00	1.00	1.00	11559
ES	1.00	1.00	1.00	33193
FR	1.00	1.00	1.00	72218
GB	0.00	0.00	0.00	31660
IT	1.00	1.00	1.00	48064
NL	1.00	1.00	1.00	10216
PT	1.00	1.00	1.00	2805
US	0.94	1.00	0.97	810437
other	1.00	1.00	1.00	164290
accuracy			0.96	1205615
macro avg	0.81	0.73	0.73	1205615
weighted avg	0.92	0.96	0.94	1205615

[Test Classification Report]

	precision	recall	f1-score	support
AU	1.00	0.00	0.00	1518
CA	0.49	0.04	0.07	3749
DE	1.00	1.00	1.00	2953
ES	1.00	1.00	1.00	8280
FR	1.00	1.00	1.00	18064
GB	0.00	0.00	0.00	7880
IT	1.00	1.00	1.00	11868
NL	1.00	1.00	1.00	2614
PT	1.00	1.00	1.00	672
US	0.94	1.00	0.97	202599
other	1.00	1.00	1.00	41207
accuracy			0.96	301404
macro avg	0.86	0.73	0.73	301404
weighted avg	0.93	0.96	0.94	301404

Baseline Modeling

- The confusion matrix results showed that most of the data points for Australia (AU), Canada (CA), and all data points for Great Britain (GB) were classified as the US



Conclusions and Future Work

- ❑ The destination countries rather than the US could be treated like “other” countries so that it could be possible to create a model for binary classification.
- ❑ Another classification model can be applied only for the countries excluding the US.
- ❑ The macro average scores were pretty low, and the weighted average scores are much higher. That tells us that the predictions were good on the larger classes and much poorer on the other ones. The model could be improved using another model like XGBoost.

Recommendations for the Clients

- ❑ The majority of users use Web over mobile devices for booking. It can be reasonable to say that it is better to invest more in Web applications.
- ❑ The majority of users prefer Apple devices (Macbook, iMac, iPhone, iPad) than other devices. It perfectly makes sense for businesses to invest in improving their systems in Apple devices.
- ❑ In addition, a huge majority of users mostly prefer using the basic signup method over Facebook and Google. The business might want to put more investment in posting ads on Facebook because the social media use for people between 20-40 is very high. https://www.statista.com/statistics/246221/share-of-us-internet-users-who-use-facebook-by-age-group/. This might probably increase the trends in bookings between and among friends.

Resources Used

- Itertools: https://docs.python.org/3/library/itertools.html
- Matplotlib: https://matplotlib.org/
- Pandas: https://pandas.pydata.org/
- Pickle: https://docs.python.org/3/library/pickle.html
- Seaborn: https://seaborn.pydata.org/
- Sci-kit Learn: https://scikit-learn.org/stable/