Springboard DSC

CP1: Airbnb Final Report

Vikram Lucky

Dr.vlucky@gmail.com

26 Jan 2019

PROBLEM STATEMENT

The Problem this capstone project aims at solving is predicting to where a newly registered Airbnb user will book his/her first travel experience, By analyzing a variety of data such as user information, country information, user browsing session records, etc. To accomplish this goal, we applied several machine learning models that could accurately predict for which country a user would book his/her first travel experience, So that more personalized content can be recommended, and overall the booking process can be optimized.

THE CLIENT

The client in question is Airbnb, the world's biggest accommodation-sharing site, enabling people to short-term lodging including vacation rentals, apartment rentals, homestays, hostels etc, usually offering more affordable alternative when compared to standard hotel bookings. Guests get good value accommodations at a huge range of price points, from a few dollars a night to hundreds and often in prime locations where a normal hotel would cost infinitely more. There is more than a financial impetus though: many guests like living like locals and getting restaurant recommendations from people really in-the-know, while many hosts just enjoy meeting new people and showing off their home.

When user signs up for Airbnb, it is in Airbnb's best interest to show best deals in a country that the user intends on visiting. It would make sense to build a predictive model around the user's browsing data and give predictions based on it, that way more personalized content can be share with the user.

THE DATA

The dataset was obtained from Kaggle and is composed of several.csv files, which need to be cleaned and wrangled before they can be analyzed These files are:

- train_user.csv: The training dataset contains users information
- test_users.csv: The test set for competition submission (not used in this project)
- session.csv : Web sessions log for users
- countries: Summary statistics of destination countries
- age_gender_bkts.csv: Summary statistics of users demographic information

DATA WRANGLING

This section gives us an overview of the various data cleaning and data wrangling techniques applied on the datasets to make them more suitable for further analysis. Explanations of what we did with each of these files follow.

Training data: Training dataset contains 213,451 rows and 16 columns.

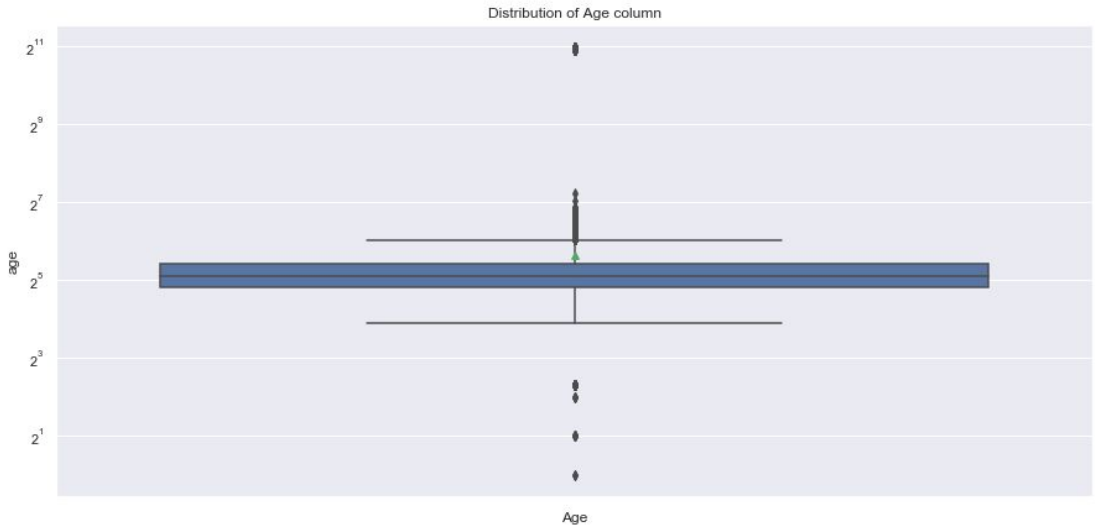
Column name and information about their data types follow.

1. **Id**: user id, unique for each user.
2. **date_account_created**: The date of account creation.
3. **timestamp_first_active**: Timestamp of the first activity, it can be earlier than date_account_created or date_first_booking as user can search before signing up.
4. **date_first_booking:** Date of first booking on Airbnb
5. **gender**
6. **age**
7. **signup_method**: How user signed up for example: using basic method, facebook or google
8. **signup_flow**: The page a user came to signup from
9. **language**: International language preference
10. **affiliate_channel**: Kind of paid marketing
11. **affiliate_provider**: Where the marketing
12. **first_affiliate_tracked**: What is the first marketing the user interacted with
13. **singup_app**
14. **first_device_type**

15. **first_browser**
16. **country_destination**: Target variable

Most of the columns in the dataset had either missing or invalid values.

- For instance, date_account_created and timestamp_first_active had an incorrect type. The type of these columns were converted to datetime.



- All the invalid and irregular fields mostly in age column shown in above boxplot and gender columns were converted into np.nan to give it the semantic, later columns with large number of NaN values >58% such as *Date first booking* were completely dropped, and the rest were imputed with sensible data relevant to the column.

Sessions data: sessions dataset contains 1053324 rows and 6 columns.

- This file contains information about web sessions log for users
- This file contains multiple np.nan values, but as this is not our training dataset, we left this out for now until we get to Machine learning part, if required we might extract some information from this dataset.

Age gender bucket: This dataset has 420 rows and 5 columns

- This dataset contains summary statistics of users age group, gender and country destination
- Contains no missing values

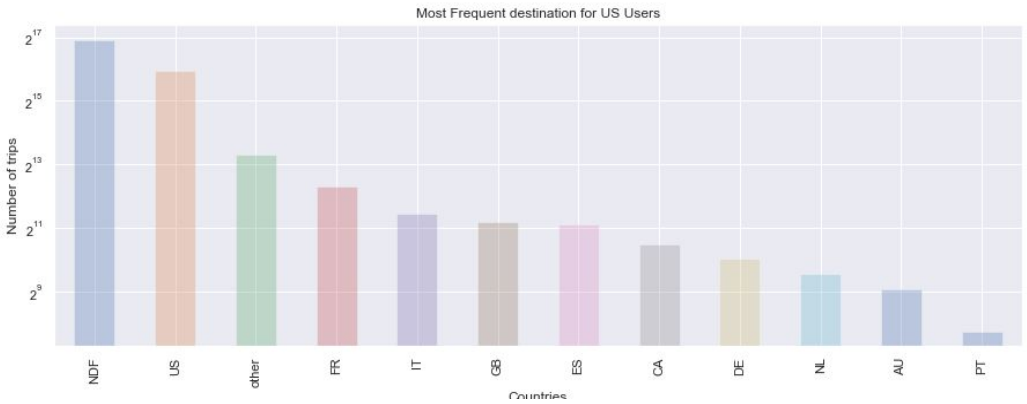
Countries: This dataset has 10 rows and 7 columns

- This dataset contains summary statistics of destination countries
- Contains no missing values

EXPLORATORY DATA ANALYSIS and INFERENTIAL STATISTICS

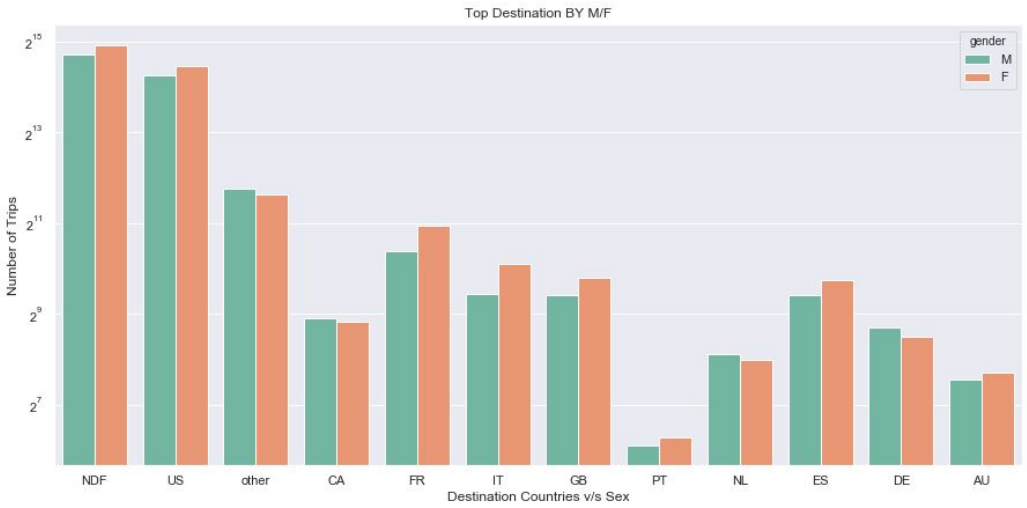
In this section we present a series of questions we posed as part of our storytelling.

We begin by looking into the distribution of choices among destinations.

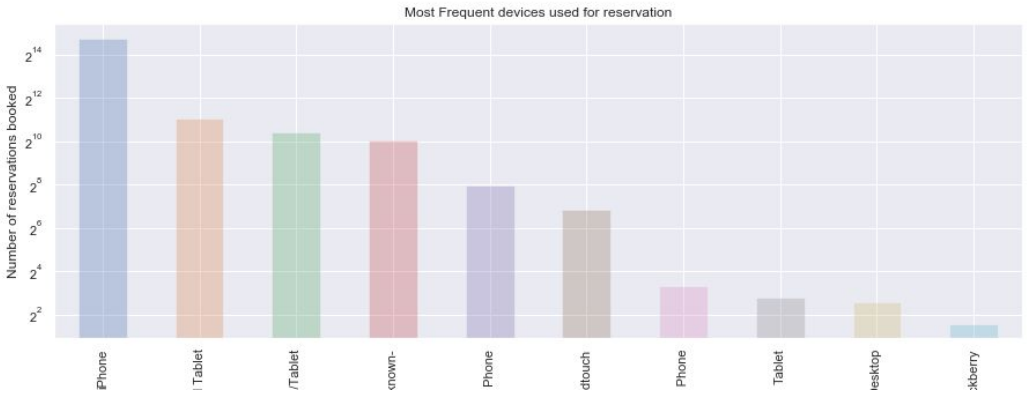


As shown in the Bar chart above, most of the user ended their search in NDF (mean no trips were booked), followed by US, and other (which represents trips were booked by not in our 12 category). Almost 60% of the users did not book their trips.

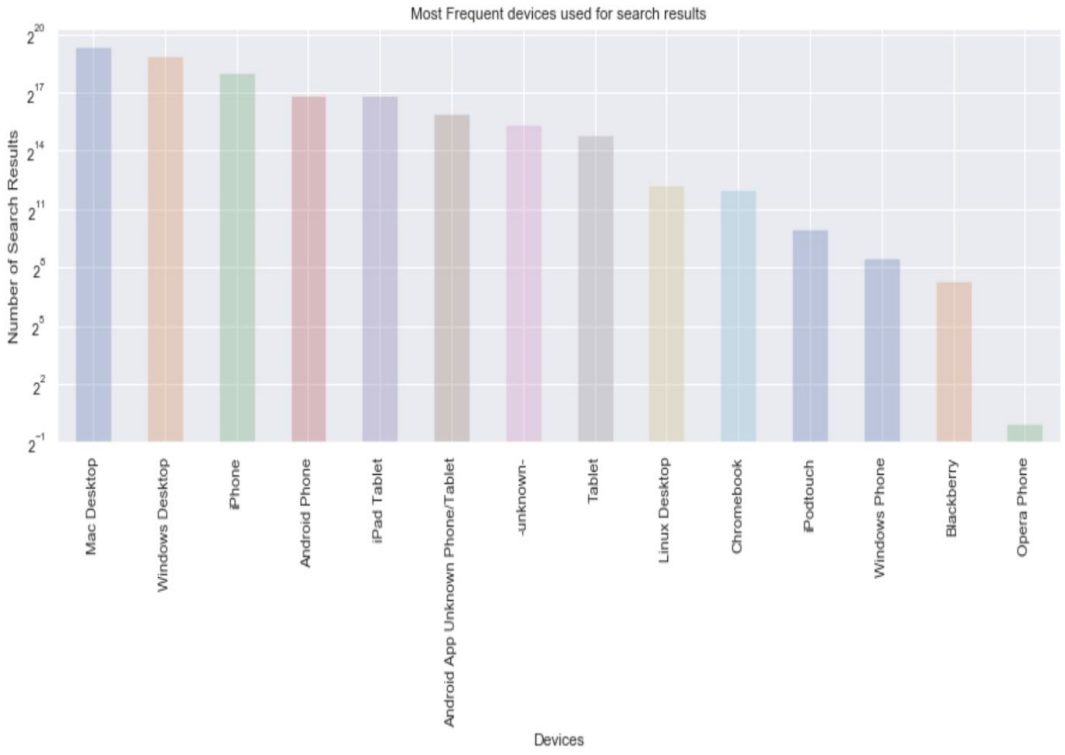
We now consider connections between various destinations and the age of users.



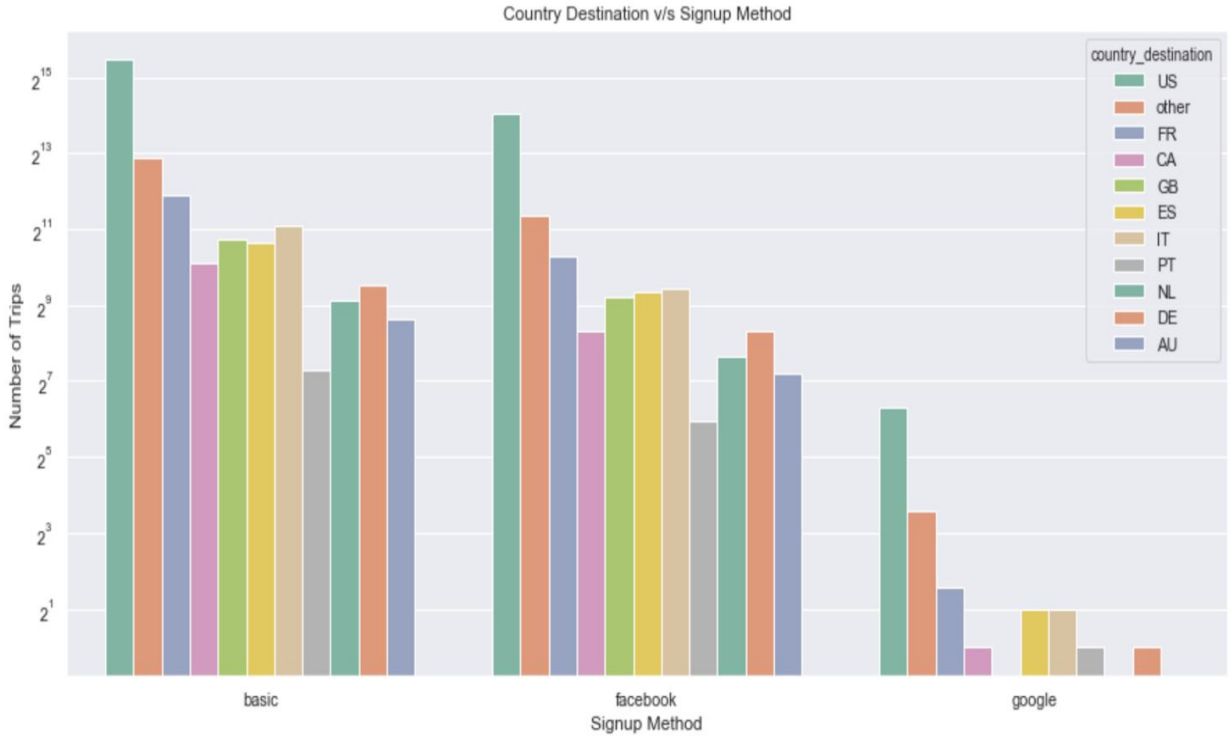
Graph above represents trips booked by each gender.



Above Bar chart represents number of most frequent devices used for **Reservation**. We clearly see Apple products come on top of the list for reservation. Surprisingly, neither MAC nor Windows desktops made it to the top of the list. But how search results. Below we will see what devices are used for search results (to browse destinations.)



Looks like users prefer to search and browse destination on bigger screens, as shown in the above bar graph, Mac desktops and Windows desktops are on top of the list followed by Iphone and Android phones. Another interesting fact according to above to bar graph, Android users who browse and search using their phones are less like to lock the deals, whereas IOS user are more likely to book their trips.



Above Chart represents the connection between signup method used and destinations. Looks like clients who signed up using the basic method and using Facebook have a slightly higher chance of booking a trip with Airbnb compared to users who signed up using Google.

Inferential Statistics

A Chi-Square significance test was performed to check the relationship between the signup method and the country of destination. For instance, is United States a top destination among users who sign up via Facebook?

	AU	CA	DE	ES	FR	GB	IT	NL	PT	US	other
signup_method											
basic	393	1105	737	1601	3767	1727	2147	560	154	45430	7430
facebook	146	322	323	646	1253	597	686	202	62	16867	2652

```
chi2, p, dof, expected = stats.chi2_contingency(cont_table)
```

```
print("Chi2: ", chi2)
print("p value: ", p)
```

```
Chi2:  48.529098330938574
p value:  4.967768526593509e-07
```

- There is a significant relationship between signup method and country destination
- The p-value obtained was less than our threshold [0.05], therefore we rejected our null hypothesis, which is [there is no significant relationship between signup method and destination country]

MACHINE LEARNING

Baseline Modeling:

We began predictive modeling by modeling the business problem as a multi-class classification problem, and chose to use Logistic Regression as the baseline algorithm.

- 1. Logistic regression with default parameters:
    - Not impressive performance model was highly biased towards majority (NDF) class.
- 2. Logistic regression with L1 & L2 Regularization technique and different values for the regularization parameter c = [0.001, 0.1, 1],
    - Model's performance was identical for L1 & L2. Trying multiple c values with L1 & L2 technique did not improve models performance as, model still remains highly biased towards the majority class.

```
-------------------------------------------------------------
Accuracy score w/ C:  1 0.6245080891998251
Classification Report C:  1
             precision    recall  f1-score   support

         AU       0.00      0.00      0.00       162
         CA       0.00      0.00      0.00       428
         DE       0.00      0.00      0.00       318
         ES       0.00      0.00      0.00       675
         FR       0.00      0.00      0.00      1507
         GB       0.00      0.00      0.00       697
         IT       0.00      0.00      0.00       851
        NDF       0.67      0.87      0.76     37363
         NL       0.00      0.00      0.00       229
         PT       0.00      0.00      0.00        65
         US       0.49      0.40      0.44     18713
      other       0.00      0.00      0.00      3028

avg / total       0.53      0.62      0.57     64036

-------------------------------------------------------------
```

Extended Analysis:

As baseline model did not perform well due to the high class imbalance of the dataset, we decided to consider the following problems:

- Original problem
- Dropped Majority class (‘NDF’)
- Binary classification problem (US destinations vs International destinations)

We then used the a more robust ML algorithm, Random Forest, with:

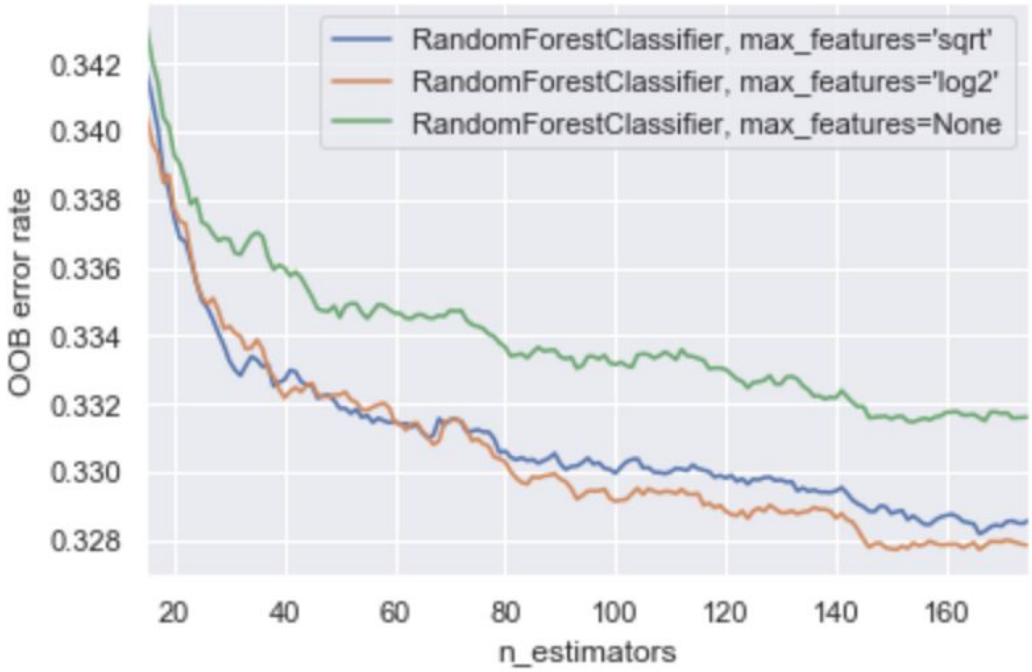
- Hyperparameter tuning & Grid search
- Cross validation
- SMOTE (Synthetic minority oversampling technique)
- RandomUnderSampler & Near Miss (Under Sampling technique)

We trained multiple models on 3 different categories.

- Original Problem (Multiclass w/ All the classes)
    - Tried Logistic Regression and Random Forest with all the techniques mentioned above.
    - Model's performance did not improve and it was identical to baseline models performance

- Dropped Majority class ('NDF')
    - After dropping 'NDF' class, 'US' class became majority class
    - Data was still highly biased towards majority class ('US'), thus model's performance was still the same

- Binary classification ('US' vs 'Rest')
    - We converted multiclass classification problem to binary classification problem
    - Model's performance did improve, the best score achieved was ~.70 using Random forest classifier with hyperparameters such as max_features: 'log2', n_estimators: 194. The graph below shows a summary of the search method we implemented.



CONCLUSION AND FUTURE WORK

In this project we got chance to explore and learn from Airbnb's data, and draw many insights from it, which might help to make business decisions. We trained multiple models, and they did not perform as expected, due to the dataset class imbalance. As part of future work we would like to consider if this problem can be overcome by supplementing existing data with more data, and by using more robust algorithms such as XGBOOST and Neural Networks.

RECOMMENDATIONS FOR CLIENT

- Desktop users book far more often than other device users, which suggests that users prefer to explore on smaller devices but when it comes to actual booking they prefer desktops. So, the UX on different devices must be tuned accordingly to achieve best results.
- Most of the Airbnb users in US tend to book within the US. Therefore, it makes most sense to give users more recommendations within the country.
- The majority of Airbnb's users are on Apple devices. Android users are the minority and do not book that often compare to Apple device users. So, it makes perfect business sense to invest more resources in improving the UX on Android devices So the number of booking using Android devices can be improved.