Springboard Capstone Paper

AirBnB New User Bookings

This paper contains the details of the problem taken up on AirBnb User data and solution plus recommendation on to the company on the basis of the findings.

Kshitiz Khatri

2/7/2016

Table of Contents

Introduction:

Imagine someone walks into her favorite restaurant and soon after ordering her food; it's on her table with chef's special customization to it. How will someone feel? It will be an awesome feeling to be treated as special among so many people visiting the restaurant.

AirBnB is among those companies in the world which is thriving to achieve this kind of feat by predicting which destination their next user will chose for booking an accommodation.

I have more than 250,000 users' data set for AirBnB (training and test) to use and to ultimately predict the destination country for a user. This data set contains the users' demographic details along with their usage and sessions history. There are details of the affiliate channels, affiliate providers, device type, and browser type; in total 16 different variables per user. I have put down the details of the data in the following pages.

The aim of this study is to explore the data thoroughly and apply the techniques learned throughout the workshop for past month and try to extract useful information which can be used further to model a solution to the problem stated above.

Details of the data set:

I have in total 6 files with the following description –

1. **Countries.csv**

It contains the summary statistics of the destination countries like longitude, latitude, area stated as distance in km, destination language and language levenshtein distance.

2. **Age_gender_bkts.csv**

This data set contains the age bucket for the user, country destination chosen, gender of the user, population in thousands for the city visited in that country and the year of booking.

3. **Test and Train Users**

Training and test data set have 16 different variables with details as follows:

- User ID
- Date of account creation (date_account_created)
- Timestamp of the first activity. It can be before than the account creation or booking, as a user might have searched even before creating an account. (timestamp_first_active)
- Date of first booking. (date_first_booking)
- Gender
- Age
- Signup method (signup_method)
- Signup_flow, the page user came to sign up from (signup_flow)
- Language, international language preference (language)
- Affiliate channel, the kind of paid marketing (affiliate_channel)

- Affiliate provider, where the marketing is, like google, craiglist, facebook, direct etc. (affiliate_provider)
- First affiliate tracked, it is the first marketing user interacted with before signing up (first_affiliate_tracked)
- Signup app (signup_app)
- First device type (first_device_type)
- First browser (first_browser)
- Country destination (country_destination)

4. **Sessions.csv**

It has the following variables related to the user activity for every session:

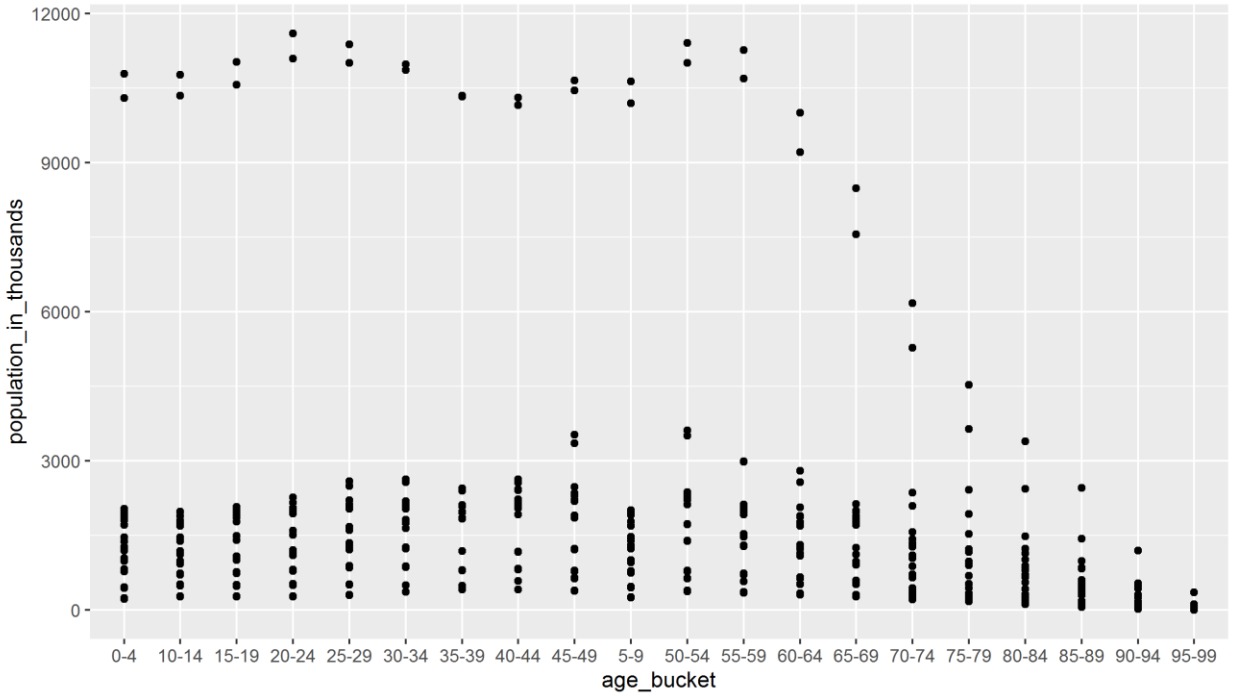
- User ID
- Action (action)
- Action type (action_type)
- Action detail (action_detail)
- Device type (device_type)
- Seconds elapsed, it is the time spend in every session by a user (secs_elapsed)

5. **Sample_submission_NDF.csv**

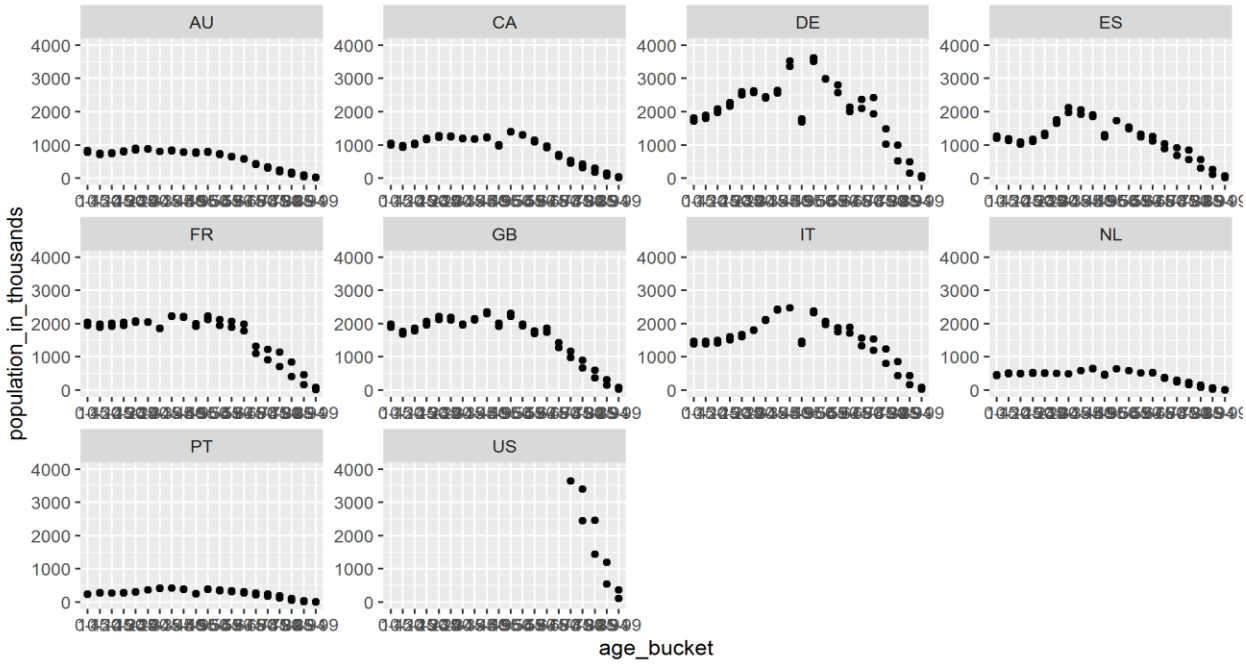
It contains the submission format for the output file generated through the model.

Exploratory Data Analysis

I will start with my findings from the age_gender_bkts data set. Below is the plot between age bucket and population in thousands for the city visited in the destination country.

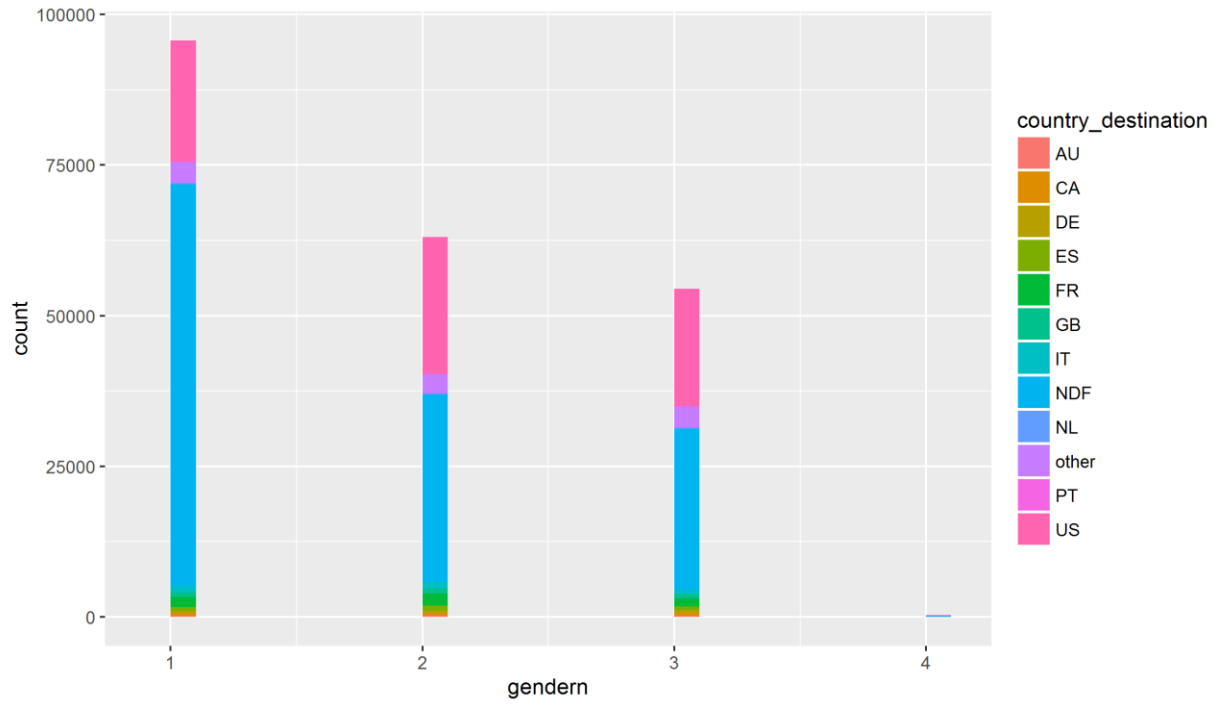


This picture contains the data for all the countries. The clear outlier is USA with very high relative population for its cities. Out of curiosity, I faceted the graph with destination countries for getting an insight country wise.



The picture in the previous page for population in thousands vs age group shows that in general, the places visited by the younger age groups are more populated than the older age groups.

I picked up the training data set and try to plot some relationships which might be useful in devising features.

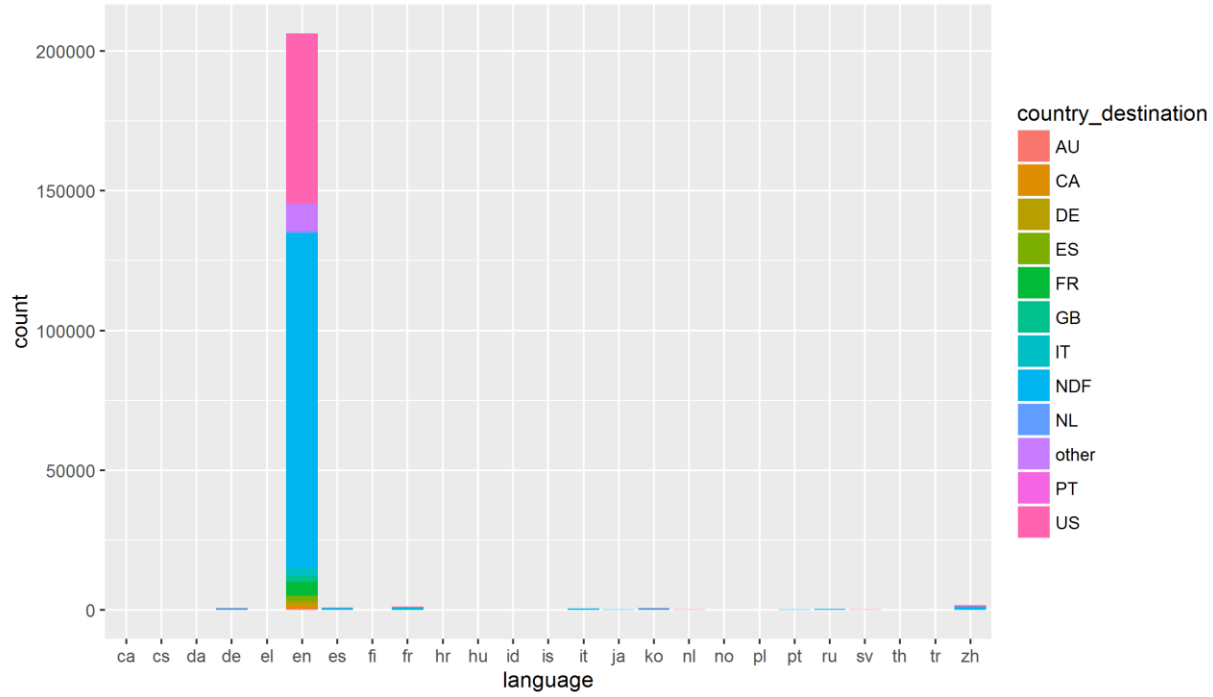


Picture pasted above represents the count of users by gender. The colors represent the proportion of the destination country chosen by the users for each gender. Following are the observations from each bar:

1. People with unknown gender have the highest number in the gender category.
2. The proportion of the no booking (NDF) is also highest in the unknown gender category.
3. Proportion of no bookings is almost same for male and female category.
4. Lastly, number of people in the ‘others’ gender category is negligible as compared to other categories.

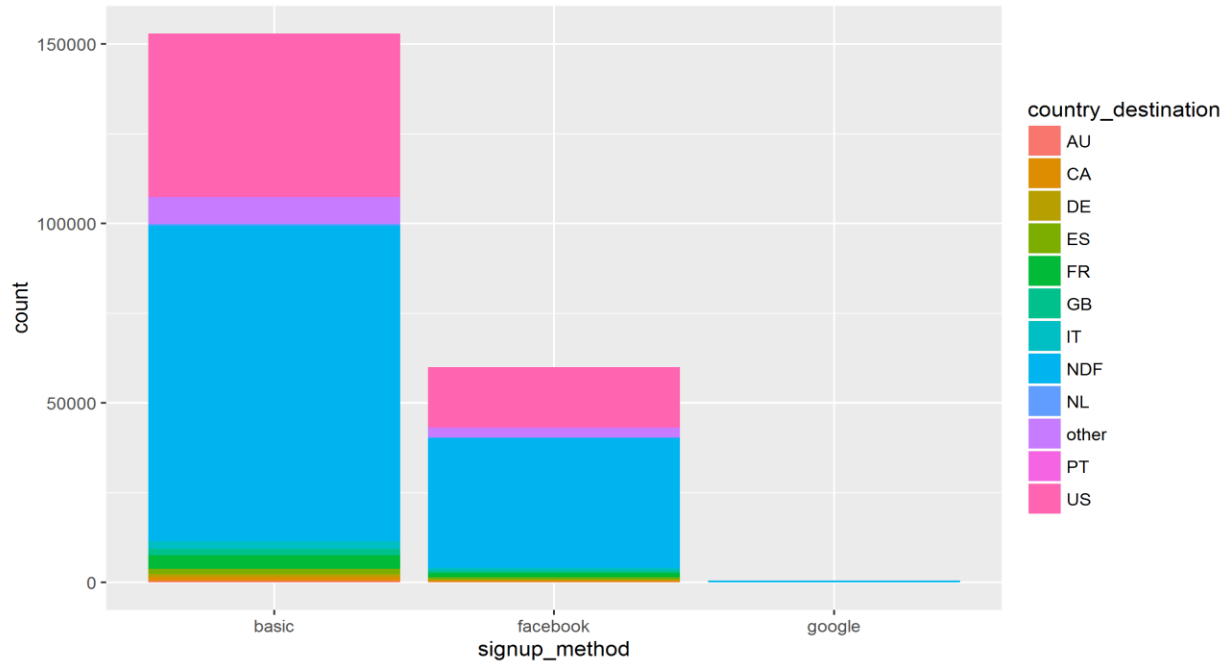
Following up next are the bar graphs for all the categorical variables with the count of users. The colors on each bar represent the destination country count proportion just like the graph presented above.

1. Count of users with the language preference



One clear observation from this graph is that almost all the users have a language preference of English.

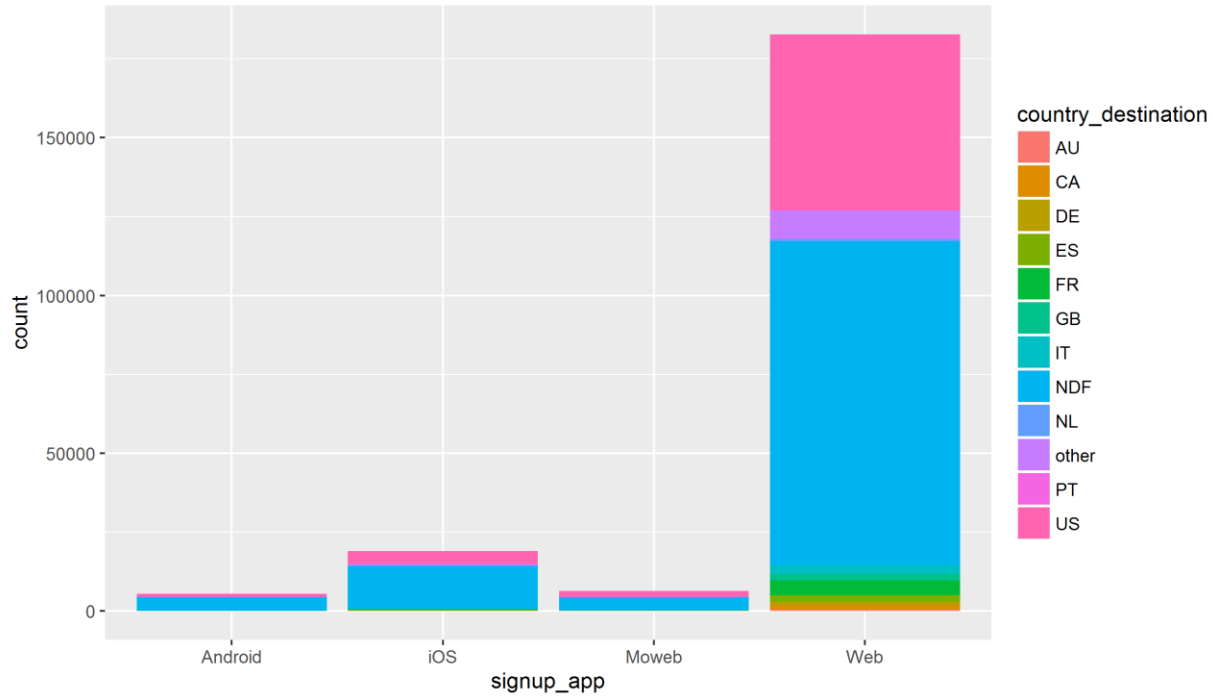
2. Count of users for different signup methods



There is couple of observations from the graph presented on the previous page:

- 2 out of three signup methods are clearly dominating that is basic and Facebook, Google is barely used as a signup method comparatively.
- Proportion of no bookings is relatively higher for Facebook signup method than the basic.

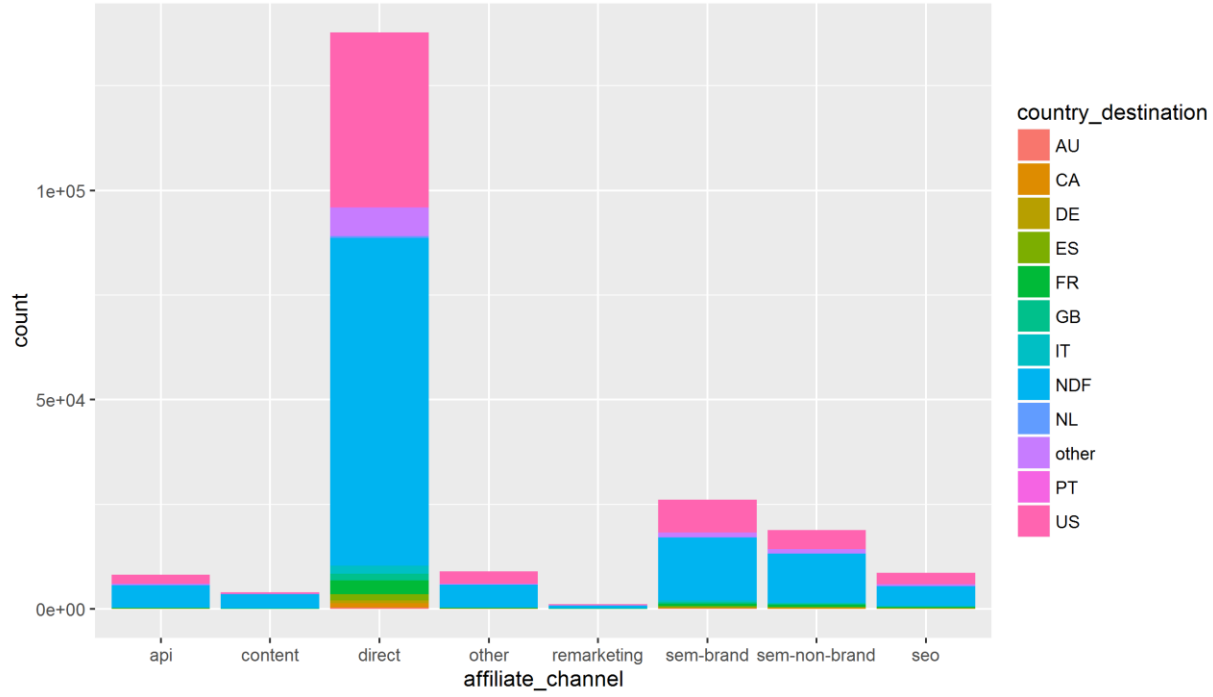
3. Count of users for different signup applications.



Observations:

- Highest number of users in the graph are using web as the signup application, followed by the iOS and then Android and Mobile Web share almost same number of user counts.
- Web has relatively lower proportion for no bookings.
- USA is consistent with its highest number of bookings as the destination country.

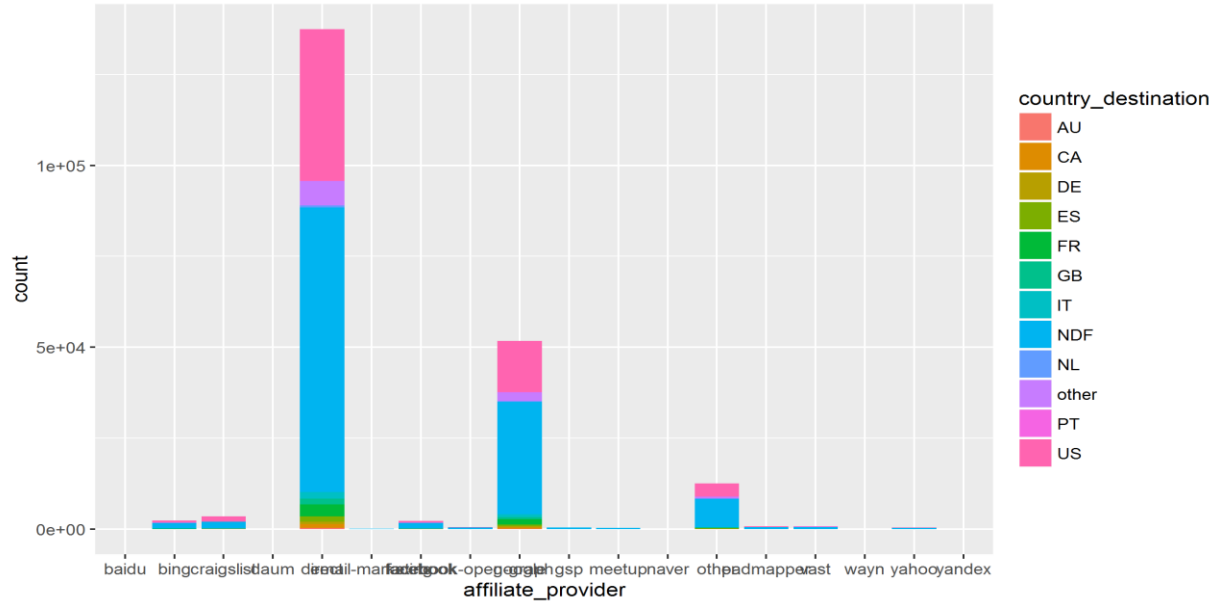
4. Count of users for different affiliate channels



Observations:

- There is a clear dominance of direct marketing as the affiliate channel followed by sem-brand, sem non-brand and seo and api with almost same number of user count.

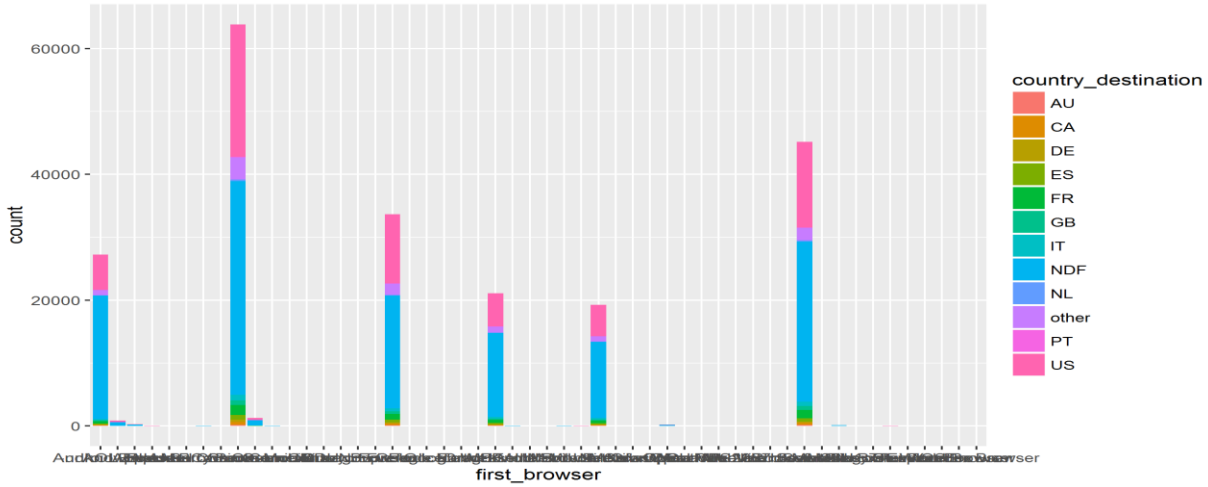
5. Count of users for different affiliate provider



Observations:

- Here also, direct marketing has helped AirBnB the most with the customer acquisition and surprisingly google is second highest which didn't performed well as a signup method.
- Only three affiliate providers are significant out of all the 18 affiliate providers.

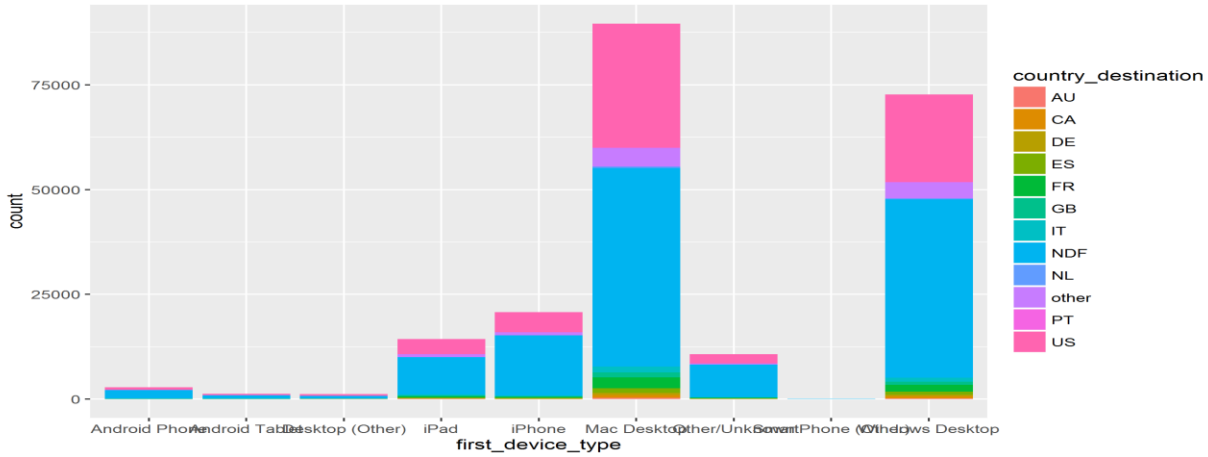
6. Count of users by different browser



Observation:

- Since there are many choices of browsers, the x axis is smudged, but it is also clear that there are only few browsers used by users.

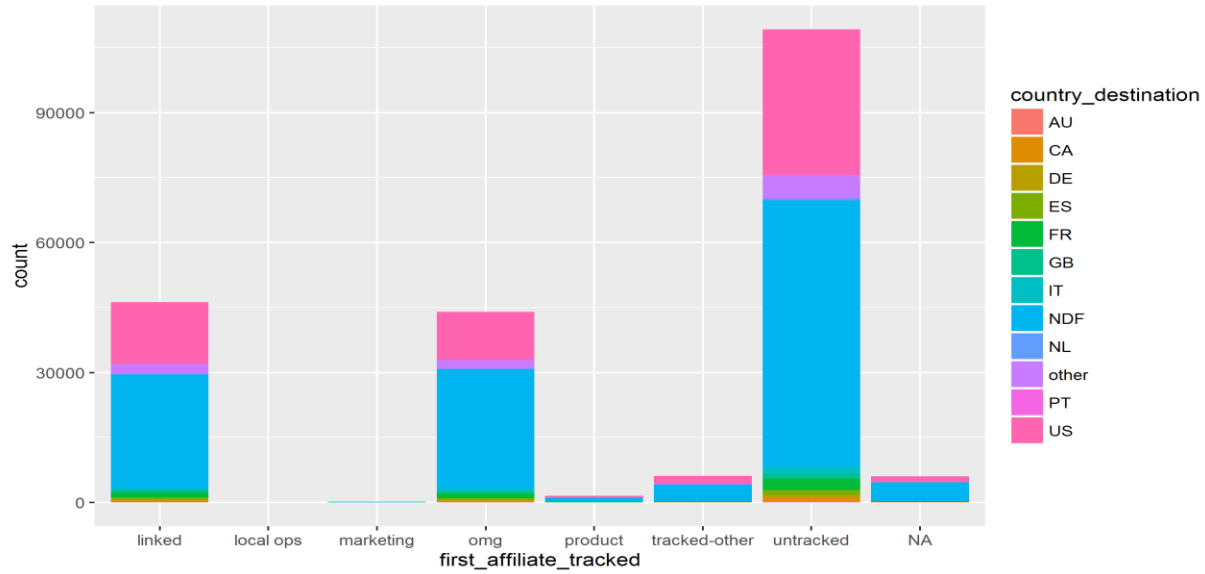
7. Count of users by different devices used



Observation:

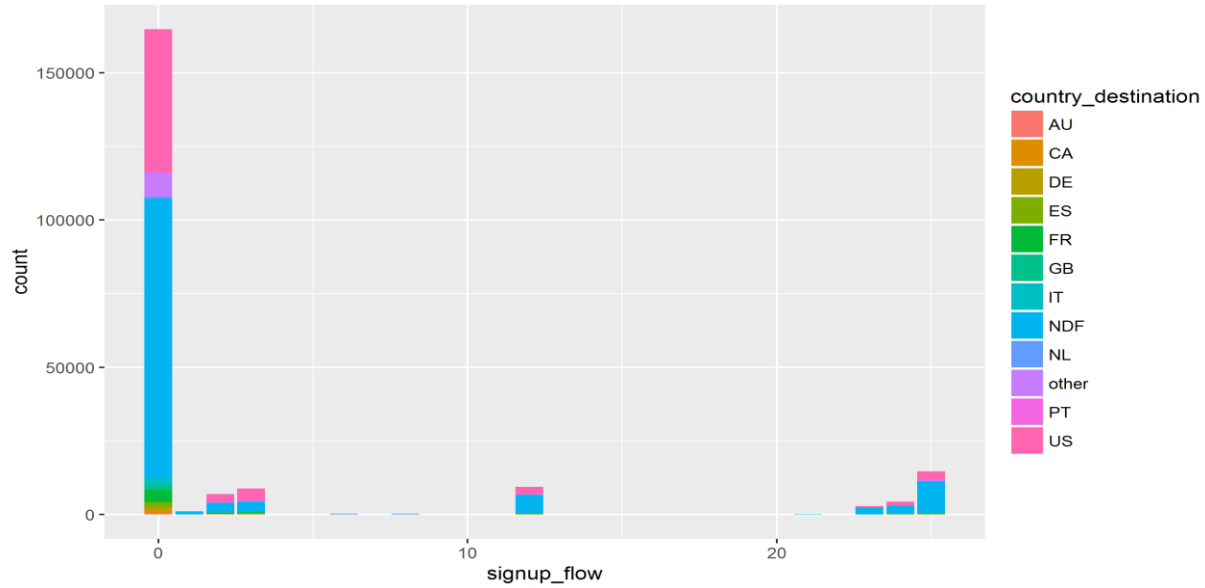
From second graph on the last page, it is clear that desktop is still the most used device among the users and the second most used device are apple products (desktop, iPhone, iPad).

8. Count of users by different first affiliate tracked



Maximum users out of the total who signed up remain untracked, although two most significant affiliates tracked are linked and omg.

9. Count of users by different signup flow



Observation:

Count of users by signup flow is quite dominated by the signup flow 0. Symbol 0 represents the page a user came to signing up from. It means that most of the users came through the same page to sign up on the website.

General Observations:

1. Correlation among all the groups (variable and the destination country) for which graphs has been plotted turned out to be very high with a very low p value.
2. Correlation among these variables also turned out to be very high, while I was checking for multi-collinearity.
3. If significant data is missing for a particular variable, it is best to remove that variable from the analysis. Though in my case, I had to keep the age variable as it is intuitive also that there should be some relation between age or age group with the booking of destination.
4. Imputing missing values is a good idea when missing values are less than 10% of the data set in general. Imputing large number of missing values can lead to an erroneous model.
5. The number of people not booking any destination is highest in the account holders reported in the training data set. This is followed by the bookings done in US.
6. Bookings done for other countries are very less as compared to US followed by Spain and Italy.

Modeling:

1. **Modeling techniques used:**

- Random Forest – Given that I have a categorical variable, as my response variable, I could have used multinomial logit or decision tree.
  I tried using multinomial logit (mlogit), but my system didn't support it for the number of independent variables I wanted to use. I used the same number of variables in Random Forest, and it was working fine for a limited number of trees.
- Xgboost – This is another package which supports heavier computations and it is faster than other techniques given the amount of calculations it does.

2. **Feature Engineering:**

The unique feature engineering I have used is the binary predictors for each destination country along with the other variables converted to sparse matrix. These features have been made up using xgboost library as no other model could handle so many independent variables at the same time. I haven't devised any new feature for other models, but I have converted all the categorical variables to the numeric values for making the calculation faster, except for the response variable.

3. **Parameter Tuning & Confusion Matrix:**

- Below is the table for the different parameters used for a **Random Forest** model along with the accuracy of each model.

Model	**Independent Var.**	**No. of trees**	**Node size**	**Accuracy (as per Kaggle)**
model1	Age, gender, signup language, affiliate provider, first browser	400	Default	66.668%
model2	Age, gender, affiliate provider, first browser	400	25	62.73%
model3	Age, gender, signup language, signup flow, affiliate provider, first browser	400	25	66.293%
model4	Age, gender, signup flow, affiliate provider, first browser	400	25	61.89%
model5	Age, gender, signup language, affiliate channel, first browser	400	25	67.860%

model6	Age, gender, signup language, first affiliate tracked, first browser	400	25	67.898%

The last model has given the highest accuracy. It can be observed from accuracy numbers that the parameter tuning hasn't resulted in significant increase in the accuracy.

- All the xgboost models had an accuracy around **65%-66%**, for any type of parameter tuning done manually.
  The better way to go about could be hyper parameter optimization, but I haven't tried that yet.

- **Confusion Matrix classwise for Random Forest Model:**

Here I will present the confusion matrices for the validation set.

Before that I will present the full table for the actual vs the predicted output:

	AU	**CA**	**DE**	**ES**	**FR**	**GB**	**IT**	**NDF**	**NL**	**other**	**PT**	**US**
AU	0	0	0	0	0	0	0	119	0	0	0	37
CA	0	0	0	0	0	0	0	312	0	0	0	99
DE	0	0	0	0	0	0	0	236	0	0	0	71
ES	0	0	0	0	0	0	0	526	0	0	0	129
FR	0	0	0	0	0	0	0	1158	0	1	0	293
GB	0	0	0	0	0	0	0	532	0	2	0	143
IT	0	0	0	0	0	0	0	660	0	0	0	163
NDF	0	0	0	0	0	0	0	32468	0	10	0	3225
NL	0	0	0	0	0	0	0	167	0	0	0	55
Other	0	0	0	0	0	0	0	2282	0	20	0	642
PT	0	0	0	0	0	0	0	50	0	0	0	14
US	0	0	0	0	0	0	0	13362	0	2	0	4668

Confusion Matrix destination country wise for **Random Forest** model:

- AU

	PREDICTED		
ACTUAL		AU	Not AU
	AU	0	156
	Not AU	0	61290

- CA

	PREDICTED		
ACTUAL		CA	Not CA
	CA	0	411
	Not CA	0	61035

- DE

	PREDICTED		
ACTUAL		DE	Not DE
	DE	0	307
	Not DE	0	61139

- ES

	PREDICTED		
ACTUAL		ES	Not ES
	ES	0	655
	Not ES	0	60791

- FR

	PREDICTED		
ACTUAL		FR	Not FR
	FR	0	1451
	Not FR	0	59995

- GB

	PREDICTED		
ACTUAL		GB	Not GB
	GB	0	675
	Not GB	0	60771

- IT

	PREDICTED		
ACTUAL		IT	Not IT
	IT	0	823
	Not IT	0	60623

- NDF

	PREDICTED		
ACTUAL		NDF	Not NDF
	NDF	32468	3235
	Not NDF	19404	25718

- NL

	PREDICTED		
ACTUAL		NL	Not NL
	NL	0	222
	Not NL	0	61224

- Others

	PREDICTED		
ACTUAL		Others	Not Others
	Others	20	2924
	Not Others	5	58487

- PT

	PREDICTED		
ACTUAL		PT	Not PT
	PT	0	64
	Not PT	0	61382

- US

	PREDICTED		
ACTUAL		US	Not US
	US	4668	13364
	Not US	4871	38543

On the previous page I have created confusion matrix. Below are the figures for accuracy, recall and precision based on the analysis done on last two pages.

Destination Country	Accuracy	Precision	Recall
AU	0.997	NaN	0
CA	0.993	NaN	0
DE	0.995	NaN	0
ES	0.989	NaN	0
FR	0.976	NaN	0
GB	0.989	NaN	0
IT	0.986	NaN	0
NDF	0.719	0.625	0.909
NL	0.996	NaN	0
Other	0.952	0.800	0.006
PT	0.998	NaN	0
US	0.703	0.489	0.258

Overall Accuracy of the Model - $\frac{\text{sum (diagonal elements)}}{\text{sum of validation set (sum of rows)}}$

Hence, Overall accuracy is - $\frac{37156}{61446} = 0.604$

- **Confusion Matrix classwise for xgboost method**

	AU	CA	DE	ES	FR	GB	IT	NDF	NL	other	PT	US
AU	0	0	0	0	1	0	0	75	0	0	0	86
CA	0	0	0	0	0	0	0	198	0	1	0	229
DE	0	0	0	0	1	0	0	178	0	2	0	137
ES	0	0	0	1	1	0	0	342	0	1	0	330
FR	0	0	0	0	0	0	2	763	0	3	0	739
GB	0	0	0	0	0	0	0	350	0	0	0	347
IT	0	0	0	0	2	0	1	447	0	1	0	400
NDF	0	0	0	0	2	1	0	32268	0	5	0	5087
NL	0	0	0	0	0	0	0	116	0	0	0	113
other	0	0	0	0	2	0	1	1569	0	12	0	1444
PT	0	0	0	0	0	0	0	38	0	0	0	27
US	0	0	0	0	5	0	0	9292	0	10	0	9406

Confusion Matrix destination country wise for **xgboost** model:

- AU

	PREDICTED		
ACTUAL		AU	Not AU
	AU	0	162
	Not AU	0	63874

- CA

	PREDICTED		
ACTUAL		CA	Not CA
	CA	0	428
	Not CA	0	63608

- DE

	PREDICTED		
ACTUAL		DE	Not DE
	DE	0	318
	Not DE	0	63718

- ES

	PREDICTED		
ACTUAL		ES	Not ES
	ES	1	674
	Not ES	0	63361

- FR

	PREDICTED		
ACTUAL		FR	Not FR
	FR	0	1507
	Not FR	14	62515

- GB

	PREDICTED		
ACTUAL		GB	Not GB
	GB	0	697
	Not GB	1	63338

- IT

	PREDICTED		
ACTUAL		IT	Not IT
	IT	1	850
	Not IT	3	63182

- NDF

	PREDICTED		
ACTUAL		NDF	Not NDF
	NDF	32268	5095
	Not NDF	13368	13305

- NL

	PREDICTED		
ACTUAL		NL	Not NL
	NL	0	229
	Not NL	0	63807

- Others

	PREDICTED		
ACTUAL		Others	Not Others
	Others	12	3016
	Not Others	23	60985

- PT

	PREDICTED		
ACTUAL		PT	Not PT
	PT	0	65
	Not PT	0	63971

- US

	PREDICTED		
ACTUAL		US	Not US
	US	9406	9307
	Not US	8939	36384

On the previous page I have created confusion matrix. Below are the figures for accuracy, recall and precision based on the analysis done on last two pages for the xgboost model .

Destination Country	Accuracy	Precision	Recall
AU	0.997	NaN	0
CA	0.993	NaN	0
DE	0.995	NaN	0
ES	0.989	1	0.001
FR	0.976	NaN	0
GB	0.989	0	0
IT	0.986	0.25	0.001
NDF	0.711	0.707	0.863
NL	0.996	NaN	0
Other	0.952	0.342	0.004
PT	0.998	NaN	0
US	0.715	0.512	0.502

Overall Accuracy of the Model - $\frac{\text{sum (diagonal elements)}}{\text{sum of validation set (sum of rows)}}$

Hence, Overall accuracy is - $\frac{41688}{64036}$ = 0.6510 or 65.10 %.

Ideas for further research

I have following ideas to improve the accuracy of the model:

1. To devise new feature which can be substituted instead of the age variable in the model, such that I do not have to impute such a large number of missing values.

2. I can try to find out better and faster models which might support my system, so as to run higher number of iterations so as to improve the accuracy of the model.

3. The information gathered through the EDA can be used to devise or modify the current variables, to improve the model accuracy. For e.g. I will try to limit the classes for all the categorical variables as it was evident from the results that there were very few categories in all the variables which were dominating the count. This might lead to a better model.

4. Due to time constraint, I couldn't go back to the sessions dataset, which seems to have quite a lot of information about particular users. It might give me better understanding of user behavior.

5. While analyzing the data, I didn't merge the test and the training set, as I assumed that training data set has enough data to analyze all the variables due to time constraint. I will join these data sets and try to analyze if I can derive any more insights out of it.

Recommendations to the Client:

Based on my findings and observations of the study, I will give following recommendations to the client:

1. Based on the model, the destination country for users can be predicted with 67.8% accuracy, hence, based on these results, predicted users can be targeted for higher conversion rates.
2. Direct marketing channel is working wonders for customer acquisition along with google and facebook as affiliate providers, so company may want to think about the redistribution of funds to the affiliate providers, based on the performance, for e.g. cutting off the funds from yandex and wayn and focusing on the ones which are working for it.