Data Mining and Predictive Analytics (BUDT758T)

Project Title: Airbnb New User Bookings Prediction

Team Members: Kushagra Sinha

Sagar Khanwalkar

Shambhavi Kumar

Suvrodeep Ghosh

Yasho Vardhan

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

	Typed Name	**Signature**
Contact Author	Kushagra Sinha	KS
	Sagar Khanwalkar	SK
	Shambhavi Kumar	SK
	Suvrodeep Ghosh	SG
	Yasho Vardhan	YV

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY

Airbnb is one of the most popular online marketplace for finding accommodation. By connecting local hosts with travelers, the company's innovative platform created an entirely new supply of real estate rental, and anyone can now easily access the accommodation market. Since its onset in 2008, Airbnb has expanded to more than 34,000 countries in 191 countries, serving more than 60 million users.

Our project involves predicting in which country a new user will make his or her first booking, out of 12 possible outcomes for the destination country: United States, France, Canada, Great Britain, Spain, Italy, Portugal, Netherlands, Germany, Australia, Other, or No Destination Found (NDF). Our objective is to accurately predict where a new user will book their first travel experience so that Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

II. DATA DESCRIPTION

The source for our data is Kaggle. (Link: https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data).

Airbnb provided 6 different files for this challenge:

1. Age_gender_bkts: Summary statistics of users' age group, gender, country of destination.
2. Sample_submission: Format for submitting prediction.
3. Sessions: Web log for users.
4. Test_users: The test set for users.
5. Train_users_2: The train set for users.

We have primarily used train_users_2 for building our prediction model. One challenge we faced with test_users was that the final destination of the respective user was not there, so we split the training data 70:30 for training and testing. The dataset has 213,451 records (n) of users from the year 2010 to 2014. It has 16 variables (k). Following are the descriptions to the variables in the dataset:

Name	Description	Format	Type
id	user id	String	Categorical
date_account_created	the date of account creation	Date	Numerical
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	Timestamp	Numerical
date_first_booking	date of first booking	Date	Numerical
gender	gender of user	String	Categorical
age	age of user	Number	Numerical
signup_method	through website, Facebook or Google	String	Categorical
signup_flow	the page a user came to signup up from	Number	Numerical
language	international language preference	String	Categorical
affiliate_channel	what kind of paid marketing	String	Categorical
affiliate_provider	where the marketing is e.g. google, craigslist, other	String	Categorical
first_affiliate_tracked	the first marketing the user interacted with before the signing up	String	Categorical

signup_app	signup through Web or Mobile	String	Categorical
first_device_type	device first used to access website	String	Categorical
first_browser	browser first used to access website	String	Categorical
country_destination	this is the target variable to predict	String	Categorical

Sample Data:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	id	date_account_created	timestamp_first_active	date_first_booking	gender	age	signup_method	signup_flow	language	affiliate_channel	affiliate_provider	first_affiliate_tracked	signup_app	first_device_type	first_browser	country_destination
2	gxn3p5htnn	6/28/2010	20090319043255		-unknown-		facebook	0	en	direct	direct	untracked	Web	Mac Desktop	Chrome	NDF
3	820tgsjxq7	5/25/2011	20090523174809		MALE	38	facebook	0	en	seo	google	untracked	Web	Mac Desktop	Chrome	NDF
4	4ft3gnwmtx	9/28/2010	20090609231247	8/2/2010	FEMALE	56	basic	3	en	direct	direct	untracked	Web	Windows Desktop	IE	US
5	bjjt8pjhuk	12/5/2011	20091031060129	9/8/2012	FEMALE	42	facebook	0	en	direct	direct	untracked	Web	Mac Desktop	Firefox	other
6	87mebub9p4	9/14/2010	20091208061105	2/18/2010	-unknown-	41	basic	0	en	direct	direct	untracked	Web	Mac Desktop	Chrome	US
7	osr2jwljor	1/1/2010	20100101215619	1/2/2010	-unknown-		basic	0	en	other	other	omg	Web	Mac Desktop	Chrome	US
8	lsw9q7uk0j	1/2/2010	20100102012558	1/5/2010	FEMALE	46	basic	0	en	other	craigslist	untracked	Web	Mac Desktop	Safari	US
9	0d01nltbrs	1/3/2010	20100103191905	1/13/2010	FEMALE	47	basic	0	en	direct	direct	omg	Web	Mac Desktop	Safari	US
10	a1vcnhxeij	1/4/2010	20100104004211	7/29/2010	FEMALE	50	basic	0	en	other	craigslist	untracked	Web	Mac Desktop	Safari	US
11	6uh8zyj2gn	1/4/2010	20100104023758	1/4/2010	-unknown-	46	basic	0	en	other	craigslist	omg	Web	Mac Desktop	Firefox	US
12	yuuqmid2rp	1/4/2010	20100104194251	1/6/2010	FEMALE	36	basic	0	en	other	craigslist	untracked	Web	Mac Desktop	Firefox	US
13	om1ss59ys8	1/5/2010	20100105051812		FEMALE	47	basic	0	en	other	craigslist	untracked	Web	iPhone	-unknown-	NDF
14	k6np330cm1	1/5/2010	20100105060859	1/18/2010	-unknown-		basic	0	en	direct	direct		Web	Other/Unknown	-unknown-	FR
15	dy3rgx56cu	1/5/2010	20100105083259		FEMALE	37	basic	0	en	other	craigslist	linked	Web	Mac Desktop	Firefox	NDF
16	ju3h98ch3w	1/7/2010	20100107055820		FEMALE	36	basic	0	en	other	craigslist	untracked	Web	iPhone	Mobile Safari	NDF
17	v4d5rl22px	1/7/2010	20100107204555	1/8/2010	FEMALE	33	basic	0	en	direct	direct	untracked	Web	Windows Desktop	Chrome	CA
18	2dwbwkx056	1/7/2010	20100107215125		-unknown-		basic	0	en	other	craigslist		Web	Other/Unknown	-unknown-	NDF
19	frhre329au	1/7/2010	20100107224625	1/9/2010	-unknown-	31	basic	0	en	other	craigslist		Web	Other/Unknown	-unknown-	US
20	cxlg85pg1r	1/8/2010	20100108015641		-unknown-		basic	0	en	seo	facebook		Web	Other/Unknown	-unknown-	NDF
21	gdka1q5ktd	1/10/2010	20100110010817	1/10/2010	FEMALE	29	basic	0	en	direct	direct	untracked	Web	Mac Desktop	Chrome	FR
22	qdubonn3uk	1/10/2010	20100110152120	1/18/2010	-unknown-		basic	0	en	direct	direct		Web	Other/Unknown	-unknown-	US
23	qsibmuz9sx	1/10/2010	20100110220941	1/11/2010	MALE	30	basic	0	en	direct	direct	linked	Web	Mac Desktop	Chrome	US
24	80f7dwscrn	1/11/2010	20100111031438	1/11/2010	-unknown-	40	basic	0	en	seo	google	untracked	Web	iPhone	-unknown-	US
25	jha93x042q	1/11/2010	20100111224015		-unknown-		basic	0	en	other	craigslist	untracked	Web	Mac Desktop	Safari	NDF

Why the data are of interest:

The data are of interest because we can use it to understand visitor behavior that lead to a conversion for a specific country. Once there is clear understanding on what channels are most effective, Airbnb can identify which users to target for country-specific advertisements and offers depending on their usage stats. This will help optimize website performance by decreasing the average time to first booking, and better forecasting of demand.

III. RESEARCH QUESTIONS

The primary focus of our analysis is to identify how the variables in the data affect the final destination of a user. Questions that we investigated using the data were:

- What are the most important features for predicting the destination of a new user?
- Is there any correlation between demographics and the destination country?
- How does the sign up method affect the destination of a new user?
- How does the first device type affect the destination of a new user?

IV. METHODOLOGY

We used Tableau to perform exploratory analysis on the data. We examined created several visualizations and examined them to understand visitor behavior. This also helped us establish our data cleaning strategy too.

Data Cleaning

To clean the data we followed the below steps. It is worth mentioning that to tackle missing values we went through a process of replacement of missing values by category averages. We identified categories as the sign-up method since

1. Replace missing age by categorical averages considering category by sign-up method
2. Extract **date_first_active** from **timestamp_first_active** column
3. Replaced missing **date_first_booking** by adding mean time difference between **date_account_created** and **date_first_active**
4. Calculated **time_to_first_book** by calculating time difference between **date_first_booking** and **date_first_active**

After cleaning the data, we ran the following models for training and prediction:

1. Multinomial Logistic Regression: As a general rule of thumb, we started with Logistic Regression on multiclass predictors. Multinomial Logistic Regression fits many independent logistic regression models through a neural network and has less dependency on collinearity of variables. We got an accuracy of 0.6048.

2. Random Forest: We used Random Forest because overcomes the overfitting problem encountered by decision trees. It can handle thousands of input variables without variable deletion and is comparatively fast and scalable. Random Forest proved to be a significant improvement over Multinomial Logistic Regression resulting in an accuracy of 0.8752.

3. XGBoost: By far one of the most accurate prediction model. Achieves better

computation time than most ensemble methods. Every predictor variable must be in numerical format though. This was the best model which gave us an accuracy of 0.8752.

V. RESULTS AND FINDINGS

By running XGBoost, we found out that ‘time taken to first bookings’ is the most important feature.

Feature	Importance
timeto_first_book	96.75
age	1.01
affiliate_channel	0.32
first_browser	0.28
first_device_type	0.27
gender	0.26
signup_flow	0.26
affiliate_provider	0.24
language	0.23
first_affiliate_tracked	0.20

The respective accuracies that we got are as follows:

Model	Accuracy
Multinomial Logistic Regression	0.6048
Random Forests	0.8752
XGBoost	0.8752

VI. CONCLUSION

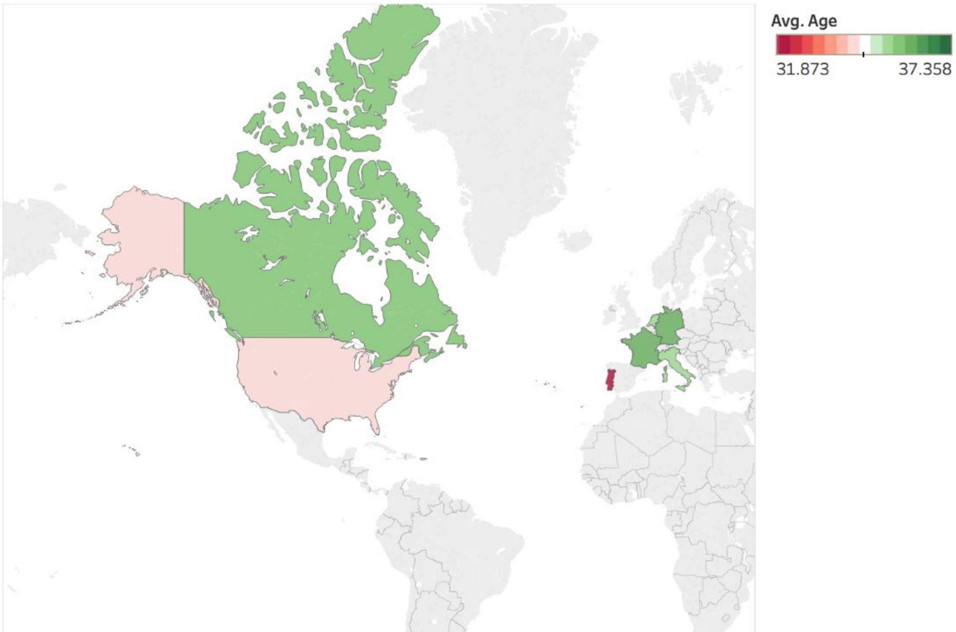
By XGBoost, we examine that ‘time taken to first bookings’ is an extremely important feature. Each destination country has a varying average time taken to the first booking. On average, an American who decides to visit Australia takes a much longer time for their first booking whereas an American visiting Spain makes their first booking in less than twenty days. Also, demographics such as age and gender are crucial to identifying customer segment. This would aid in targeted marketing. Average age for an American visiting the UK is much higher than the average age for an American using AirBnB for the first time to visit say, Portugal.

Basis our analysis, we have a few additional recommendations for Airbnb.
Airbnb needs to recognize their key affiliate providers through which they garner users. They should identify top performing affiliate providers such as Facebook over Google.More data on demographics should be gathered to facilitate customer segmentation and targeted marketing
Also, users who decline to enter age and gender express low interest and tend to browse rather than book soon.

VII. APPENDIX

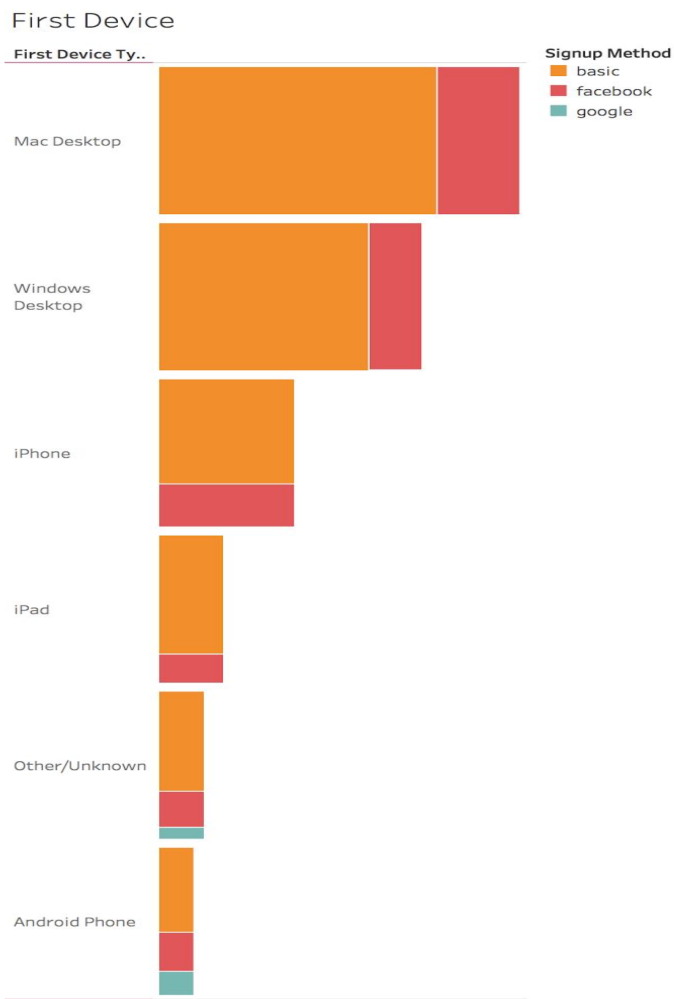
Appendix 1: Average age by destination

Avg Age/Dest



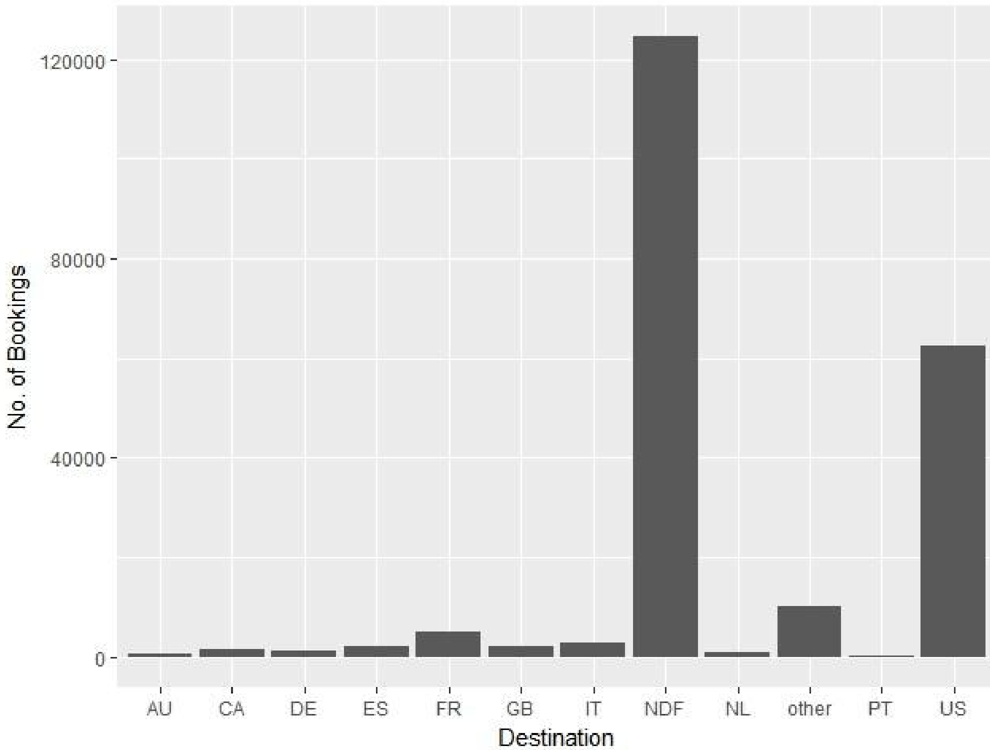
Map based on Longitude (generated) and Latitude (generated). Color shows average of Age. Details are shown for Country Destination. The data is filtered on Gender, which keeps FEMALE and MALE.

Appendix 2: First device type

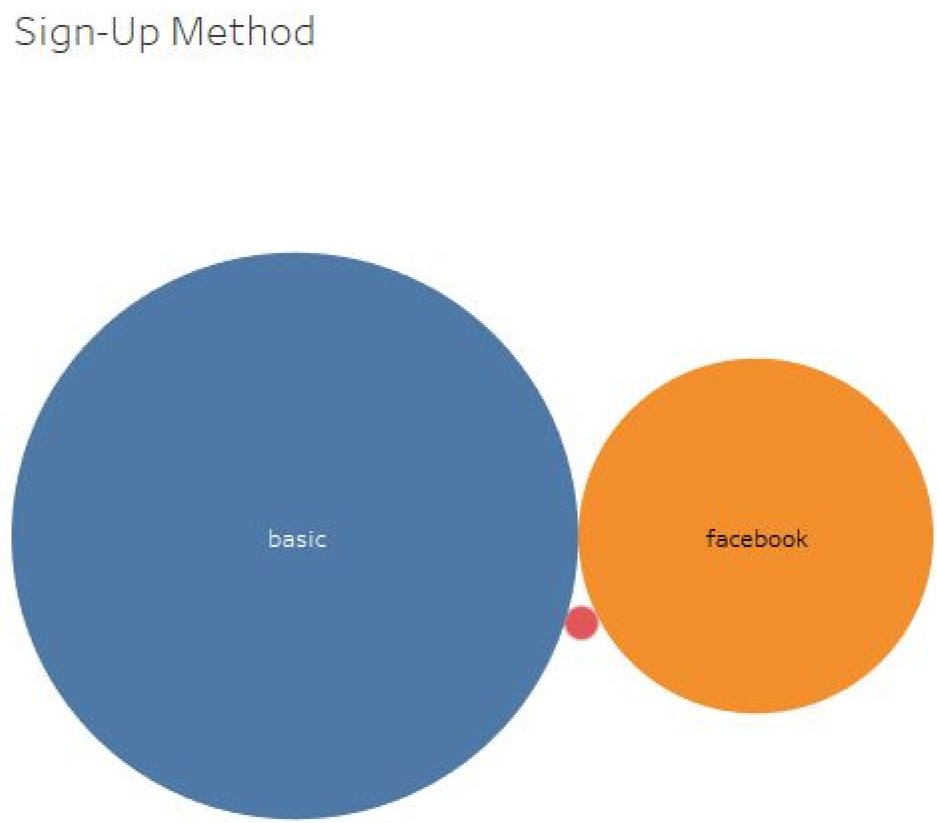


Signup Method (color) and count of Id (size) broken down by First Device Type. The view is filtered on Signup Method and First Device Type. The Signup Method filter keeps basic, facebook and google. The First Device Type filter excludes Android Tablet, Desktop (Other) and SmartPhone (Other).

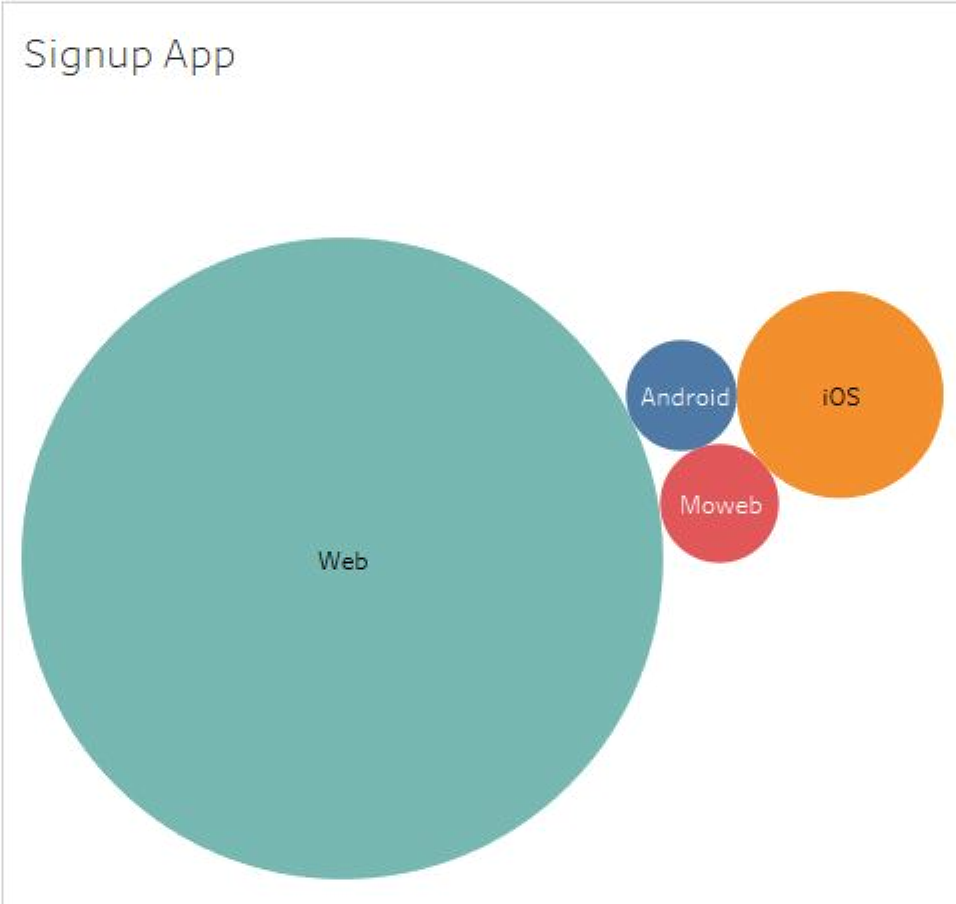
Appendix 3: Number of bookings by destination



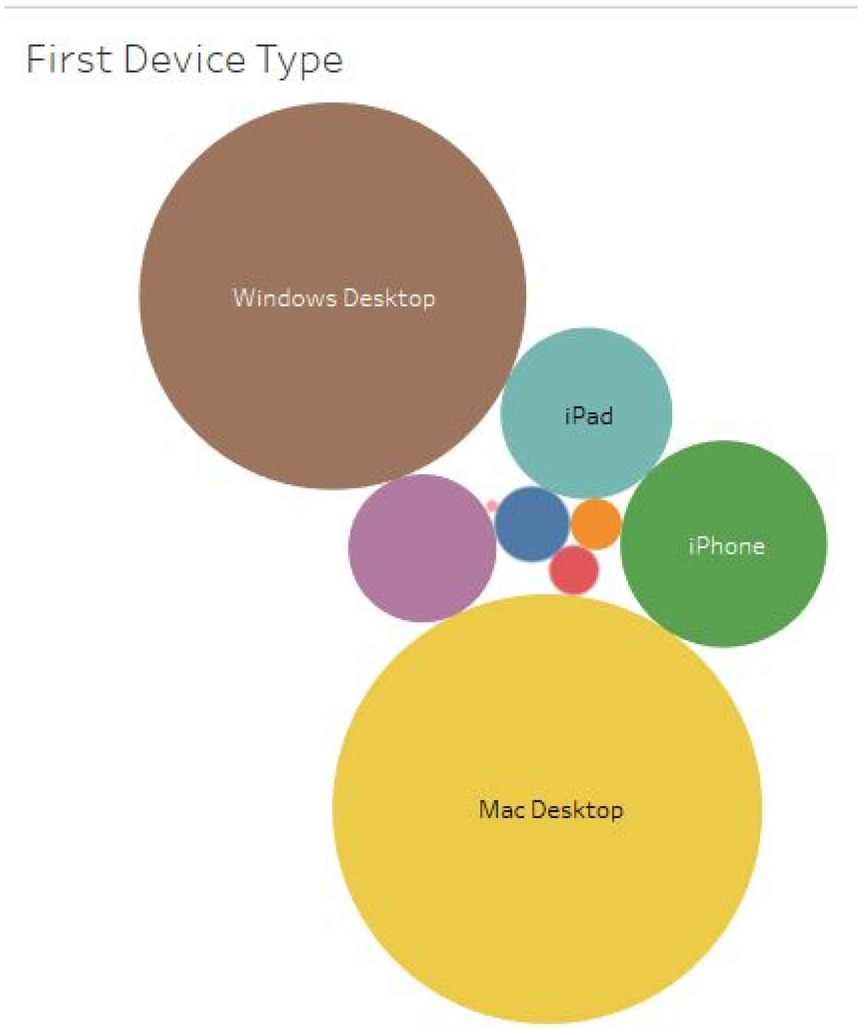
Appendix 4: Sign up method used by visitors



Appendix 5: Sign up App used by visitors



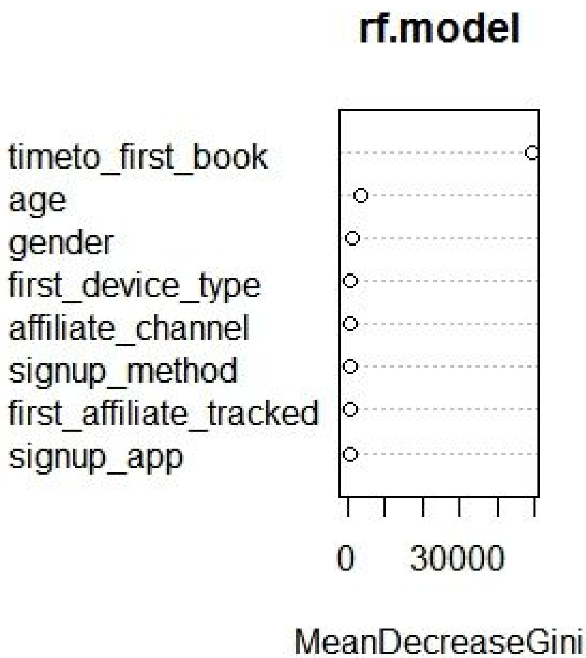
Appendix 6: First device type used by visitors



Appendix 7: Features listed by Importance (XGBoost)

Feature	Importance
timeto_first_book	96.75
age	1.01
affiliate_channel	0.32
first_browser	0.28
first_device_type	0.27
gender	0.26
signup_flow	0.26
affiliate_provider	0.24
language	0.23
first_affiliate_tracked	0.20

Appendix 8: Features listed by Importance (Random Forest)



Appendix 9: Results of Multinomial Regression

<u>Multinomial Logistic Regression:</u>

```
Overall Statistics

               Accuracy : 0.6048
                 95% CI : (0.6023, 0.6073)
    No Information Rate : 0.5835
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1684
 Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6	Class: 7
Sensitivity	0.00000	0.000000	0.000000	0.00000	2.843e-04	0.00000	0.00000	0.8779
Specificity	1.00000	1.000000	1.000000	1.00000	1.000e+00	1.00000	1.00000	0.3130
Pos Pred Value	NaN	NaN	NaN	NaN	1.000e+00	NaN	NaN	0.6416
Neg Pred Value	0.99747	0.993308	0.995028	0.98946	9.765e-01	0.98911	0.98672	0.6466
Prevalence	0.00253	0.006692	0.004972	0.01054	2.354e-02	0.01089	0.01328	0.5835
Detection Rate	0.00000	0.000000	0.000000	0.00000	6.692e-06	0.00000	0.00000	0.5122
Detection Prevalence	0.00000	0.000000	0.000000	0.00000	6.692e-06	0.00000	0.00000	0.7984
Balanced Accuracy	0.50000	0.500000	0.500000	0.50000	5.001e-01	0.50000	0.50000	0.5954

	Class: 8	Class: 9	Class: 10	Class: 11
Sensitivity	0.000000	4.246e-04	0.000000	0.31669
Specificity	1.000000	1.000e+00	1.000000	0.84594
Pos Pred Value	NaN	7.500e-01	NaN	0.45908
Neg Pred Value	0.996426	9.527e-01	0.998983	0.74991
Prevalence	0.003574	4.729e-02	0.001017	0.29222
Detection Rate	0.000000	2.008e-05	0.000000	0.09254
Detection Prevalence	0.000000	2.677e-05	0.000000	0.20158
Balanced Accuracy	0.500000	5.002e-01	0.500000	0.58132

Appendix 10: Results of Random Forest

<u>Random Forest:</u>

Overall Statistics

```
               Accuracy : 0.8752
                 95% CI : (0.8726, 0.8777)
    No Information Rate : 0.5835
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7678
 Mcnemar's Test P-Value : NA
```

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6	Class: 7
Sensitivity	0.000000	0.000000	0.000e+00	0.00000	0.00000	0.000e+00	0.00000	1.0000
Specificity	1.000000	1.000000	1.000e+00	1.00000	1.00000	1.000e+00	1.00000	0.9981
Pos Pred Value	NaN	NaN	0.000e+00	NaN	NaN	0.000e+00	NaN	0.9986
Neg Pred Value	0.997486	0.993316	9.950e-01	0.98947	0.97648	9.891e-01	0.98672	1.0000
Prevalence	0.002514	0.006684	4.966e-03	0.01053	0.02352	1.089e-02	0.01328	0.5835
Detection Rate	0.000000	0.000000	0.000e+00	0.00000	0.00000	0.000e+00	0.00000	0.5835
Detection Prevalence	0.000000	0.000000	1.562e-05	0.00000	0.00000	1.562e-05	0.00000	0.5843
Balanced Accuracy	0.500000	0.500000	5.000e-01	0.50000	0.50000	5.000e-01	0.50000	0.9990

	Class: 8	Class: 9	Class: 10	Class: 11
Sensitivity	0.000000	0.000e+00	0.000000	0.9981
Specificity	1.000000	1.000e+00	1.000000	0.8249
Pos Pred Value	NaN	0.000e+00	NaN	0.7018
Neg Pred Value	0.996439	9.527e-01	0.998985	0.9990
Prevalence	0.003561	4.729e-02	0.001015	0.2922
Detection Rate	0.000000	0.000e+00	0.000000	0.2917
Detection Prevalence	0.000000	4.685e-05	0.000000	0.4156
Balanced Accuracy	0.500000	5.000e-01	0.500000	0.9115

Appendix 11: Results of XGBoost

XGBoost:

```
Overall Statistics

               Accuracy : 0.8752
                 95% CI : (0.8726, 0.8777)
    No Information Rate : 0.5835
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7678
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 0 Class: 1  Class: 2 Class: 3  Class: 4  Class: 5  Class: 6
Sensitivity          0.000000 0.000000 0.000e+00  0.00000 0.000e+00 0.000e+00 0.000e+00
Specificity          1.000000 1.000000 1.000e+00  1.00000 1.000e+00 1.000e+00 1.000e+00
Pos Pred Value            NaN      NaN 0.000e+00      NaN 0.000e+00 0.000e+00 0.000e+00
Neg Pred Value       0.997486 0.993316 9.950e-01  0.98947 9.765e-01 9.891e-01 9.867e-01
Prevalence           0.002514 0.006684 4.966e-03  0.01053 2.352e-02 1.089e-02 1.328e-02
Detection Rate       0.000000 0.000000 0.000e+00  0.00000 0.000e+00 0.000e+00 0.000e+00
Detection Prevalence 0.000000 0.000000 1.562e-05  0.00000 4.685e-05 1.562e-05 3.124e-05
Balanced Accuracy    0.500000 0.500000 5.000e-01  0.50000 5.000e-01 5.000e-01 5.000e-01
                     Class: 7 Class: 8  Class: 9 Class: 10 Class: 11
Sensitivity            1.0000 0.000000 9.908e-04  0.000000    0.9978
Specificity            0.9982 1.000000 9.999e-01  1.000000    0.8250
Pos Pred Value         0.9987      NaN 2.727e-01       NaN    0.7019
Neg Pred Value         1.0000 0.996439 9.527e-01  0.998985    0.9989
Prevalence             0.5835 0.003561 4.729e-02  0.001015    0.2922
Detection Rate         0.5835 0.000000 4.685e-05  0.000000    0.2916
Detection Prevalence   0.5843 0.000000 1.718e-04  0.000000    0.4154
Balanced Accuracy      0.9991 0.500000 5.004e-01  0.500000    0.9114
```

VIII. REFERENCES

- NYC Data Science Academy Blog - https://blog.nycdatascience.com/student-works/capstone/predicting-new-users-first-travel-destination-airbnb-capstone-project/
- Airbnb- Predicting New User Bookings - http://rstudio-pubs-static.s3.amazonaws.com/197502_9bf4cf621a824e3093abc48d5a04e6de.html
- Data Science: A Kaggle Walkthrough – Understanding the Data - http://brettromero.com/wordpress/data-science-a-kaggle-walkthrough-understanding-the-

data/