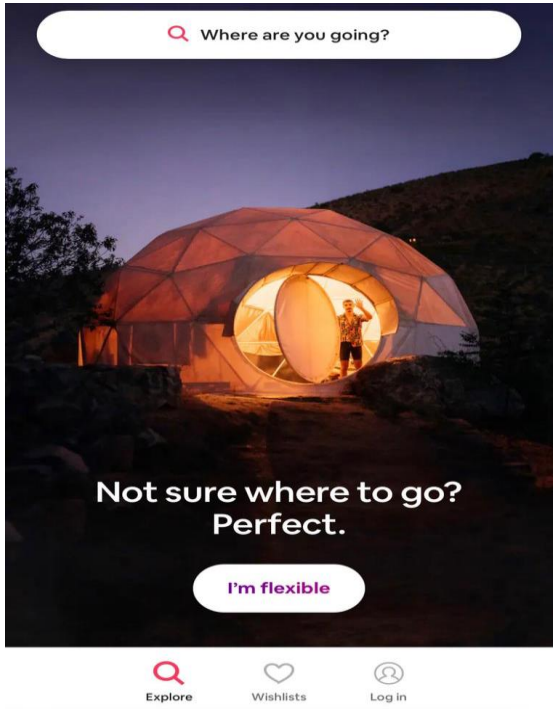
AirBnB
New User Booking

Prepared by:
Mohamed Ahmed Abdelmaguid

Motivation

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.



Agenda

- Objectives
- Data Understanding
- Data Pre-processing
- Challenges
- Data Exploration & Hypothesis testing
- Data Preparation
  - Handling non-realistics data, outliers and missing values.
  - Feature Selection and Scaling
  - Feature Engineering
- Model Selection & Evaluation metric
- Fine-Tune the model
  - Hyperparameters Optimization
  - Feature Importance selections
- Results
- Other Approaches
- Business Recommendations

Objectives

Objectives

Business Challenge :

- Given a list of users along with their **demographics, web session records, and some summary statistics.**
- Can we predict **where a new AirBnB user will make their first booking?**

Project Aim :

- The first objective of this project is to recognize **key factors that will use to know the first destination of new Airbnb Users**.
- The second one is to develop a **to predict which country a new user's first booking destination will be**.

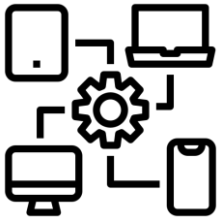
Dataset Understanding

Data understanding

The following data was provided.

- **Demographic:** age, gender, sign-up method, device used, browser used, etc.

- **Web Sessions:** sequences of actions taken on the website, e.g. click, view results, request booking, verify email.a.



- **Users**: From USA only
- **Data Dates:** from 2010 to 2014

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

Challenges

- The data was unreasonable to solve our problems, With the current data, it is possible to predict whether users will **Book** or **Not Book** with some accuracy
- A lot of missing and incorrect data
- The model doesn’t get better with more examples.
- Data Highly skewed

Data Pre-processing

Sessions dataset:

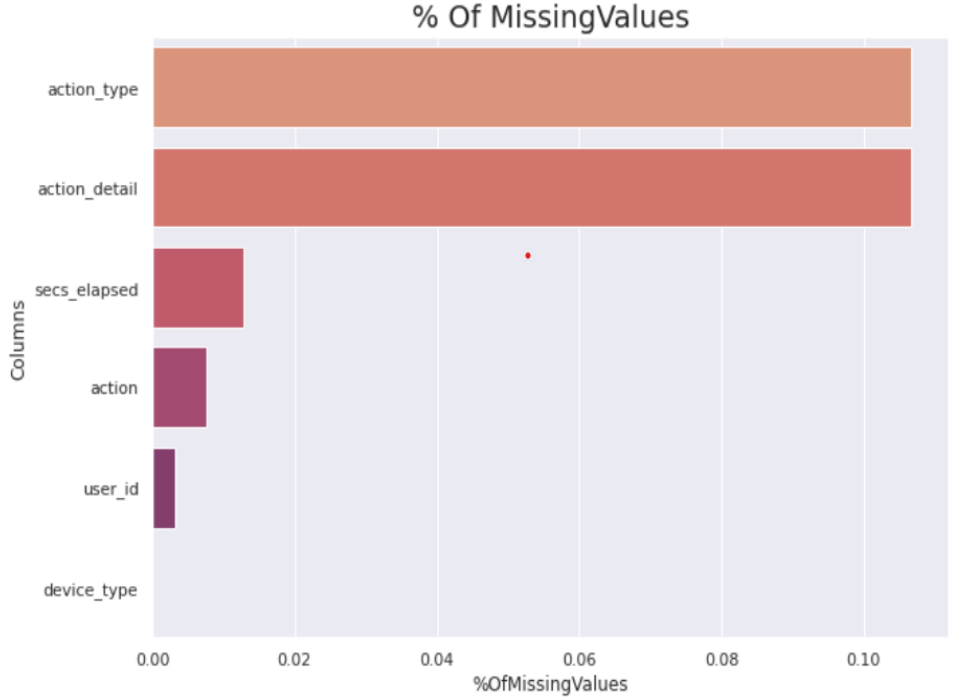
user_id	action	action_type	action_detail	device_type	secs_elapsed
ld5pygrsxi	show	view	p3	iPhone	46.0
p3fo4xkeau	show	view	p3	Android Phone	431.0
mlc9sos6g3	show	NaN	NaN	Windows Desktop	1034.0
2fiobp0qpu	update	submit	update_listing	Mac Desktop	134272.0
z8en38as6c	search	click	view_search_results	iPhone	28557.0
eavj04j4oc	create	submit	signup	iPhone	NaN
z5347wkq3o	lookup	NaN	NaN	Windows Desktop	650.0
hmemv026xu	similar_listings	data	similar_listings	Windows Desktop	632.0
dikpgnlgco	index	view	message_thread	Mac Desktop	1200.0
cujr6q1ccs	ajax_photo_widget_form_iframe	-unknown-	-unknown-	Windows Desktop	6.0
1gcnp2htnj	other_hosting_reviews_first	-unknown-	-unknown-	Mac Desktop	190.0
kjard7fcok	index	data	reservations	iPhone	9092.0

Data Pre-processing -cont

Sessions dataset:

- Dropped all users_id have NaN values
- Replaced '-unknown-' with NaN
- Imputed the missing values in secs_elapsed

	secs_elapsed	secs_elapsed_fillna_with_median	secs_elapsed_fillna_with_mean
count	1.039776e+07	1.053324e+07	1.053324e+07
mean	1.941124e+04	1.917631e+04	1.941124e+04
std	8.890920e+04	8.835953e+04	8.833556e+04
min	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.290000e+02	2.370000e+02	2.370000e+02
50%	1.146000e+03	1.146000e+03	1.188000e+03
75%	8.442000e+03	8.191000e+03	9.225000e+03
max	1.799977e+06	1.799977e+06	1.799977e+06



Data Pre-processing - cont

Users dataset:

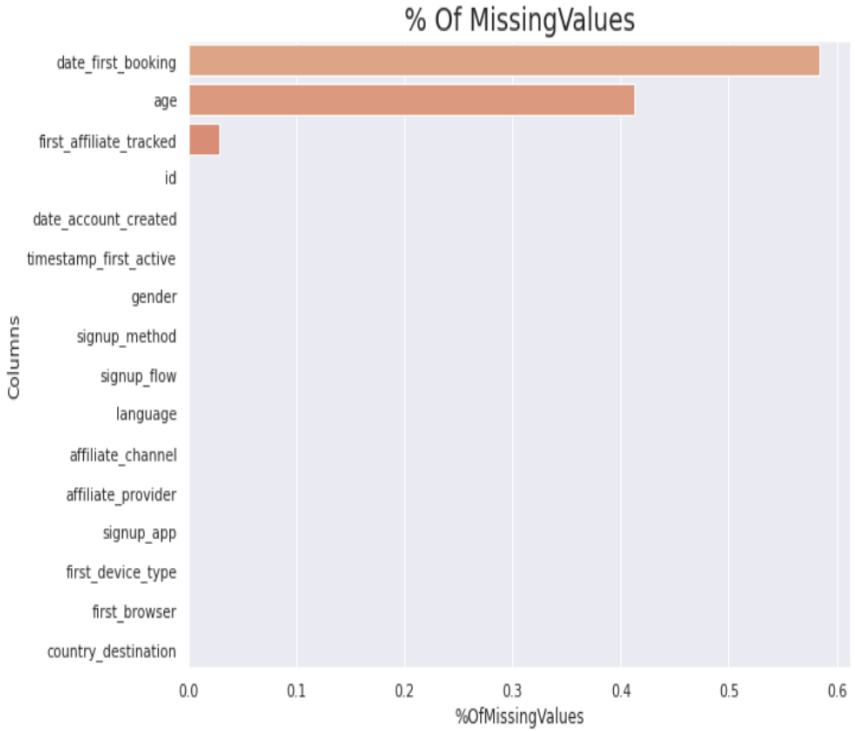
#	Column	Non-Null Count	Dtype
0	id	213451 non-null	object
1	date_account_created	213451 non-null	object
2	timestamp_first_active	213451 non-null	int64
3	date_first_booking	88908 non-null	object
4	gender	213451 non-null	object
5	age	125461 non-null	float64
6	signup_method	213451 non-null	object
7	signup_flow	213451 non-null	int64
8	language	213451 non-null	object
9	affiliate_channel	213451 non-null	object
10	affiliate_provider	213451 non-null	object
11	first_affiliate_tracked	207386 non-null	object
12	signup_app	213451 non-null	object
13	first_device_type	213451 non-null	object
14	first_browser	213451 non-null	object
15	country_destination	213451 non-null	object

id	date_account_created	timestamp_first_active	date_first_booking	gender	age	signup_method	signup_flow	language
1haefu2ypw	2012-01-23	20120123190718	NaN	MALE	40.0	basic	2	en
f08ryn8zdi	2012-09-11	20120911144044	NaN	FEMALE	49.0	facebook	0	en
vtnclw1sqo	2011-07-11	20110711143804	NaN	MALE	43.0	facebook	2	en
8tya25gm0m	2012-08-10	20120810204449	2012-08-12	FEMALE	34.0	facebook	0	en
trmu030wpo	2013-04-02	20130402215932	2013-04-02	MALE	NaN	basic	0	en
q1u4eymb7t	2013-10-17	20131017202107	NaN	MALE	51.0	basic	0	en
j4dqsliq6g	2014-04-24	20140424234623	NaN	FEMALE	73.0	basic	0	en
u9twkzb2m0	2014-04-14	20140414035854	NaN	-unknown-	NaN	basic	0	en
cpj4fzn2be	2012-07-14	20120714060220	NaN	FEMALE	2014.0	basic	0	en
67znibnfz1	2013-12-23	20131223053415	NaN	MALE	25.0	facebook	0	en

Data Pre-processing - cont

Users dataset:

- Dropped data_first_booking
- Relacing ‘-unknown-’ with NaN
- Fixed data type & convert timestamp to date form
- Fixed unrealistic values in the age feature
    - By replacing less than 18 or more than 119 with NaN

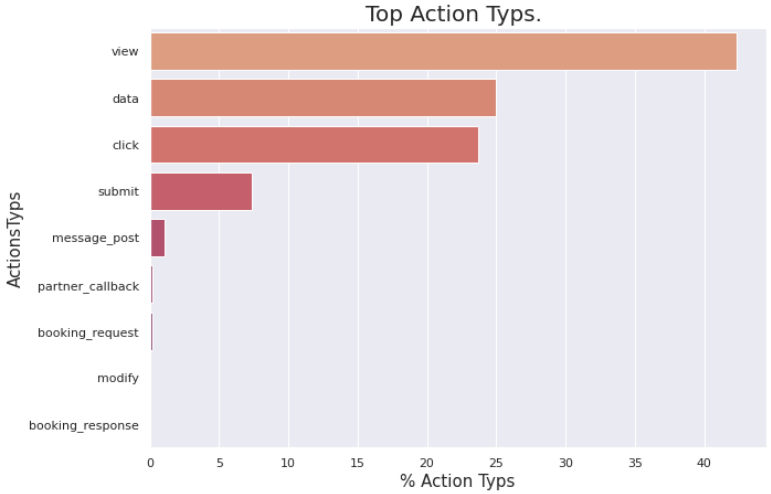


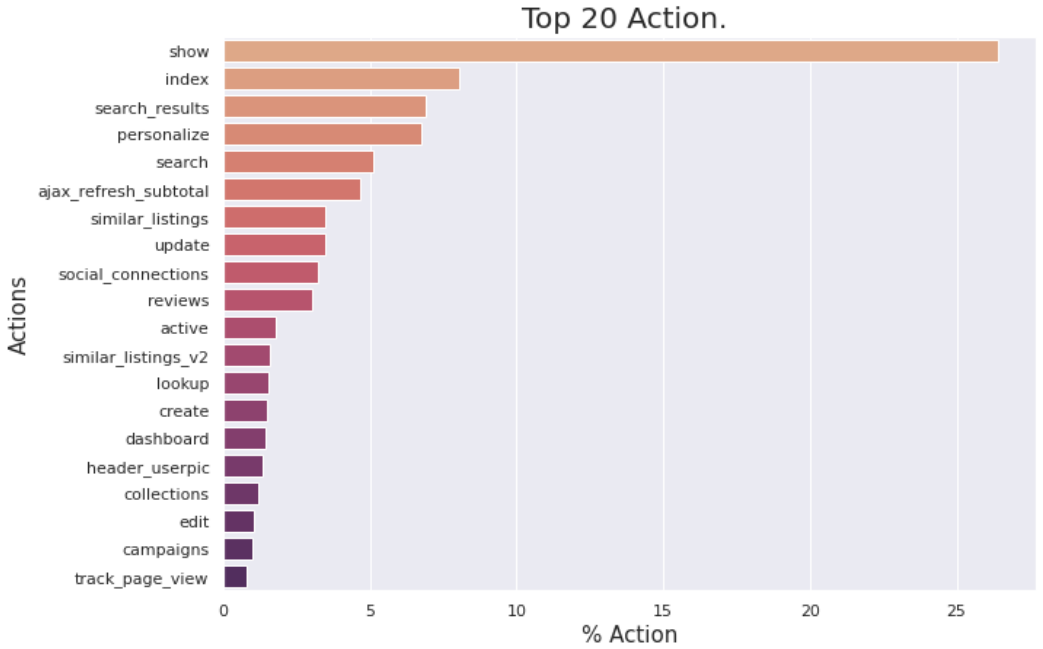
Data Exploration & Hypothesis testing

Data Exploration & Hypothesis testing

Sessions dataset:

- Actions

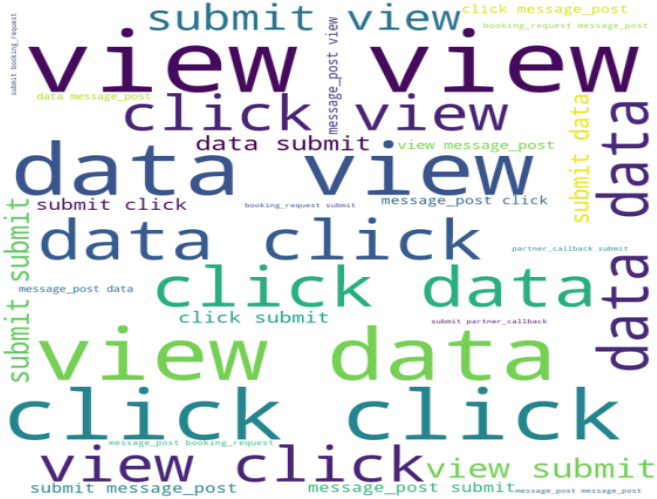




Data Exploration & Hypothesis testing - cont

Sessions dataset:

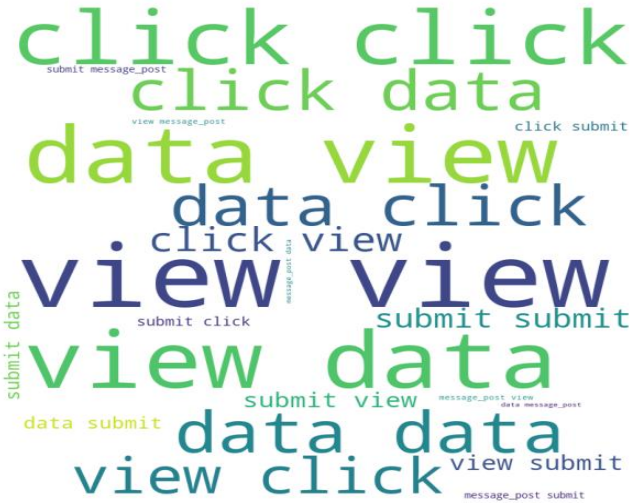
- Actions without NDF

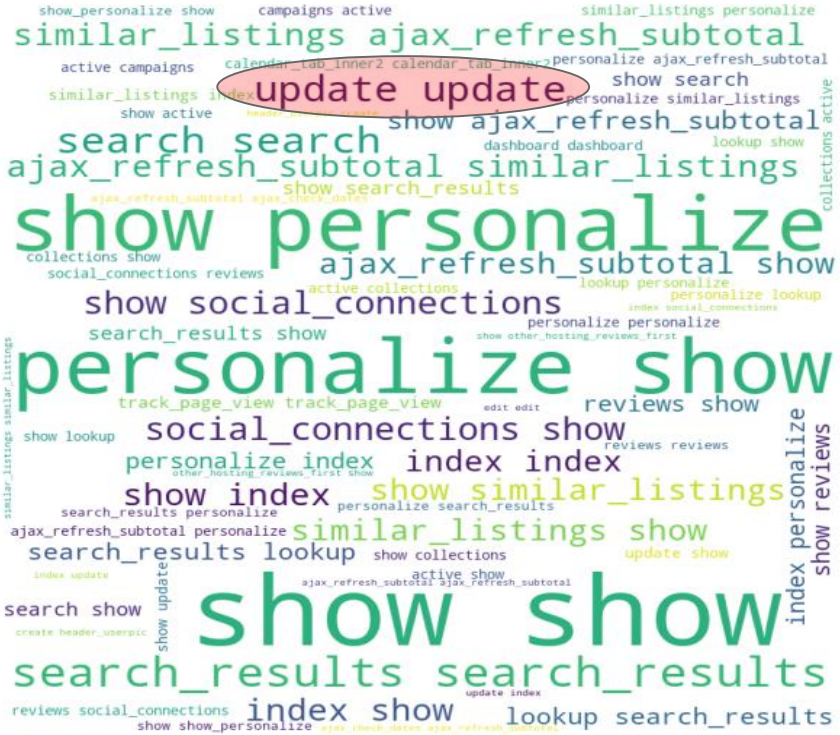


Data Exploration & Hypothesis testing - cont

Sessions dataset:

- Actions for NDF only

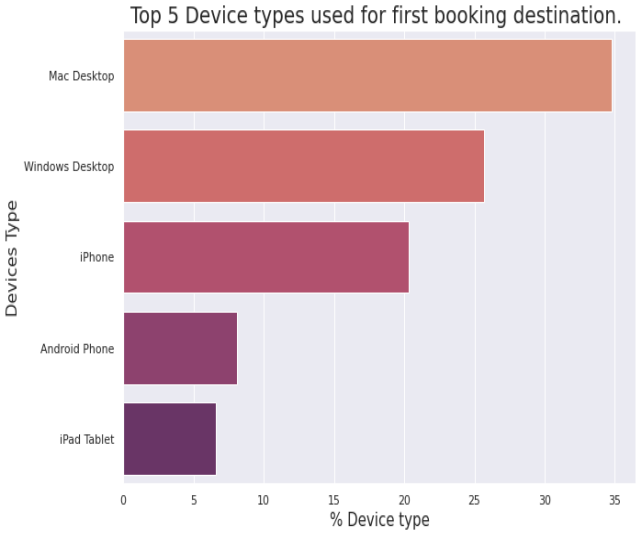




Data Exploration & Hypothesis testing – cont

Sessions dataset:

- Device type



Without NDF (Booked)

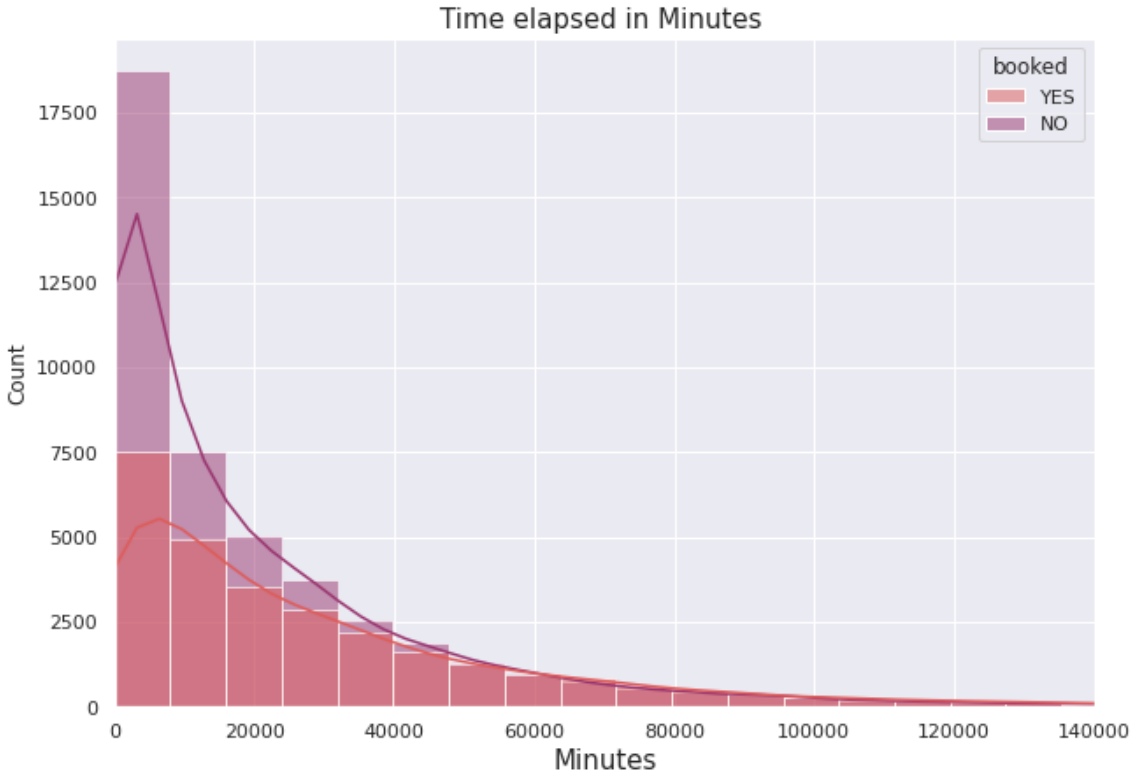
Only NDF(Not booked)

Data Exploration & Hypothesis testing - cont

Sessions dataset:

- Time elapsed

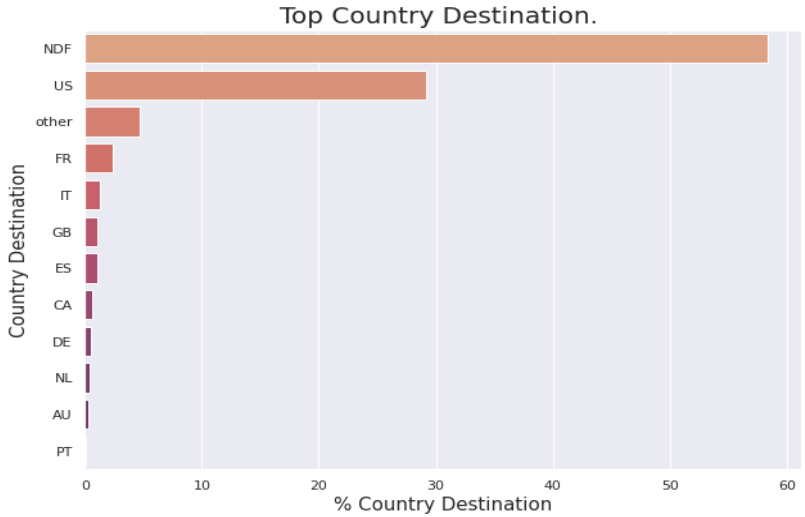
	booked	NO	YES
mins_elapsed	count	45041.000000	28774.000000
	mean	21392.937064	31254.892611
	std	28513.531256	35725.991493
	min	0.000000	0.000000
	25%	2972.316667	7485.395833
	50%	11463.933333	19817.425000
	75%	28924.350000	42229.241667
	max	637022.716667	523221.533333

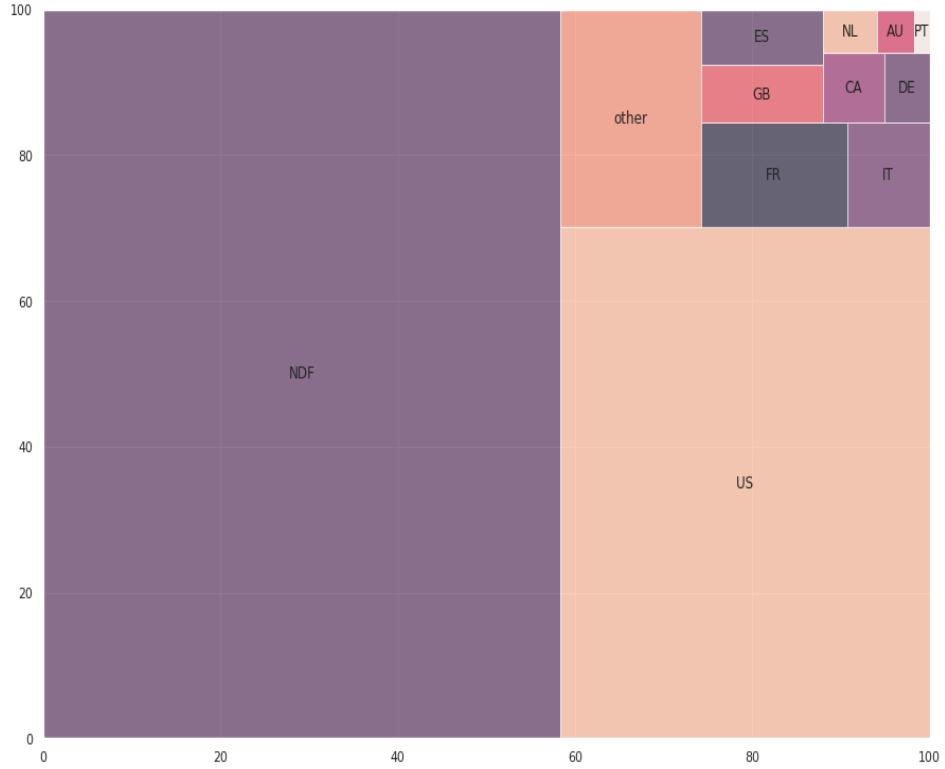


Data Exploration & Hypothesis testing - cont

Users dataset:

- Country Destination



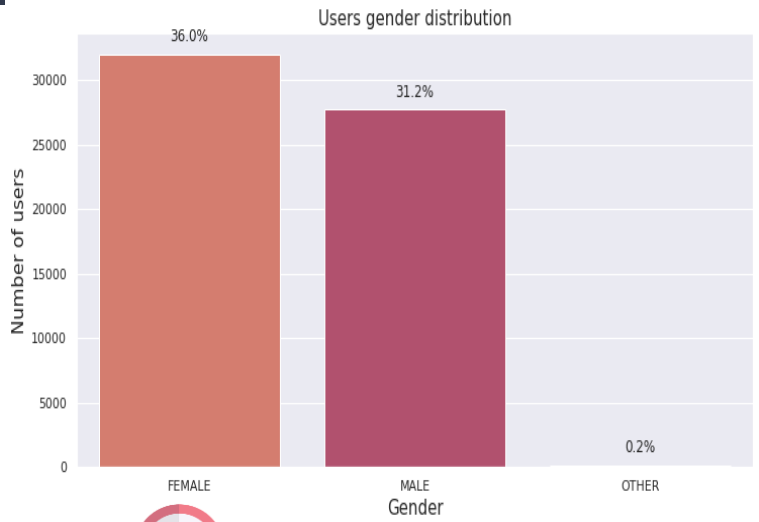


Data Exploration & Hypothesis testing – cont

Users dataset:

- Gender who booked

country_destination	AU	CA	DE	ES	FR	GB	IT	NL	PT	US
gender										
FEMALE	195	419	325	800	1796	817	1004	230	70	21118
MALE	174	445	386	623	1246	646	644	261	63	18396



H0: No relationship between country preference and the gender of the user.

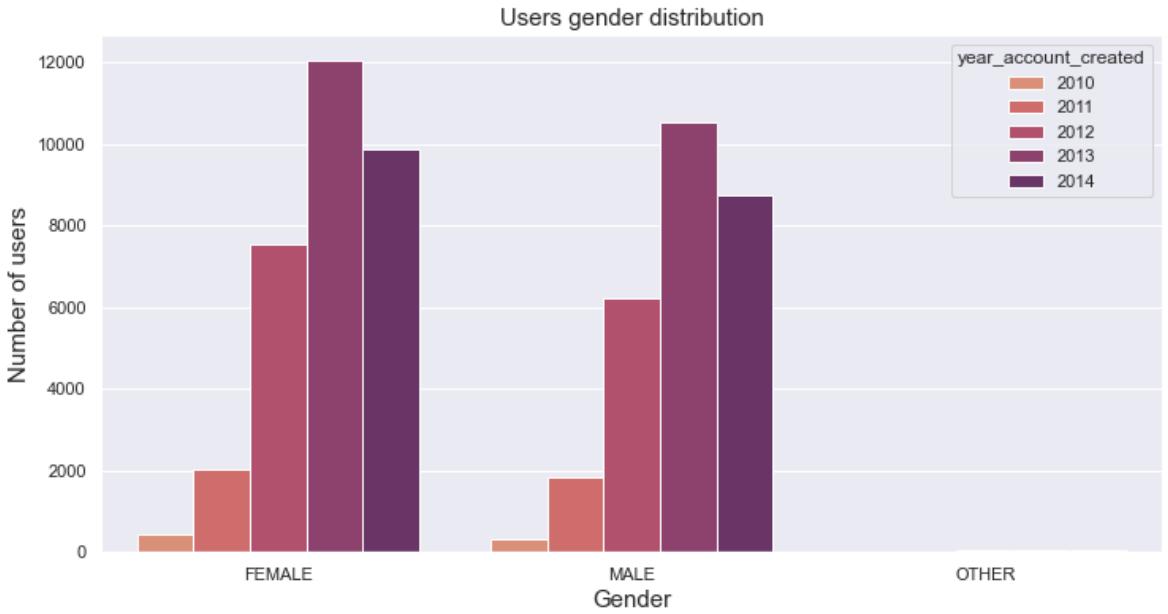
H1: There is a relationship between country preference and the gender of the user.

With Significance level 0.05

Data Exploration & Hypothesis testing - cont

Users dataset:

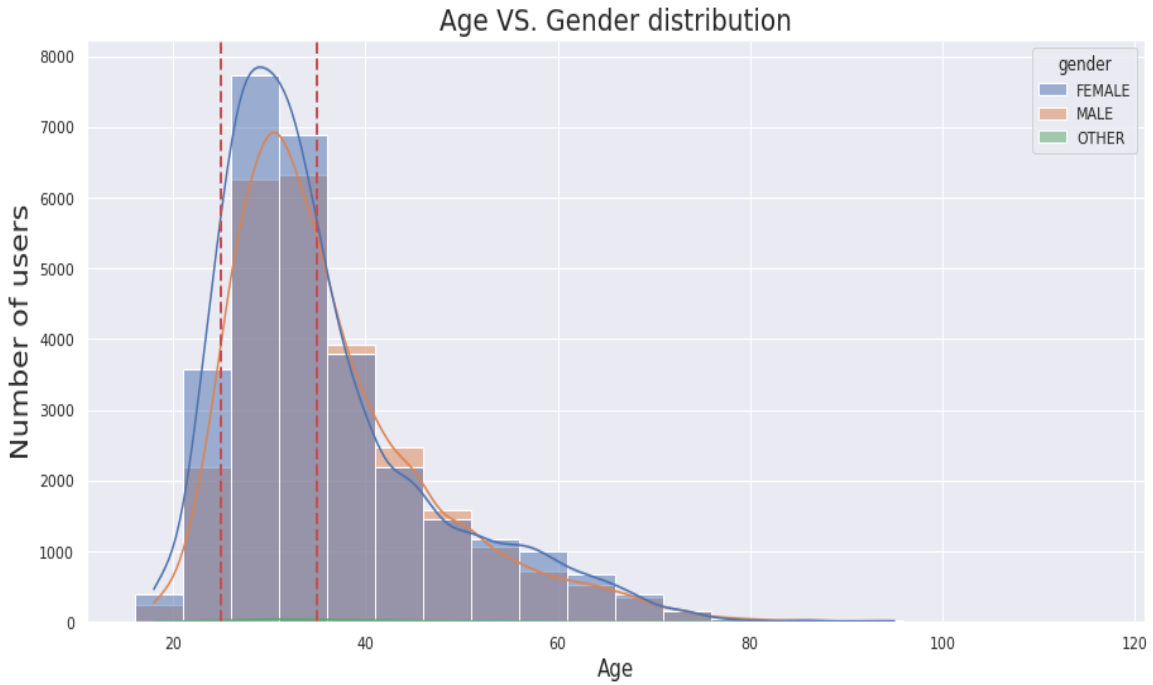
- Gender who booked



Data Exploration & Hypothesis testing - cont

Users dataset:

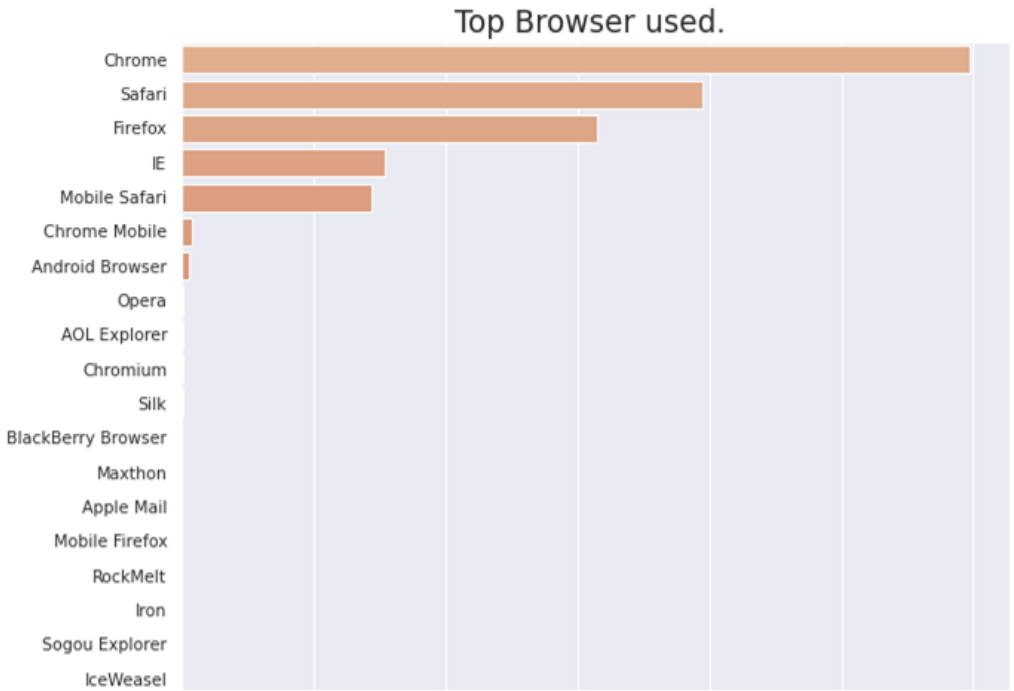
- Age



Data Exploration & Hypothesis testing - cont

Users dataset:

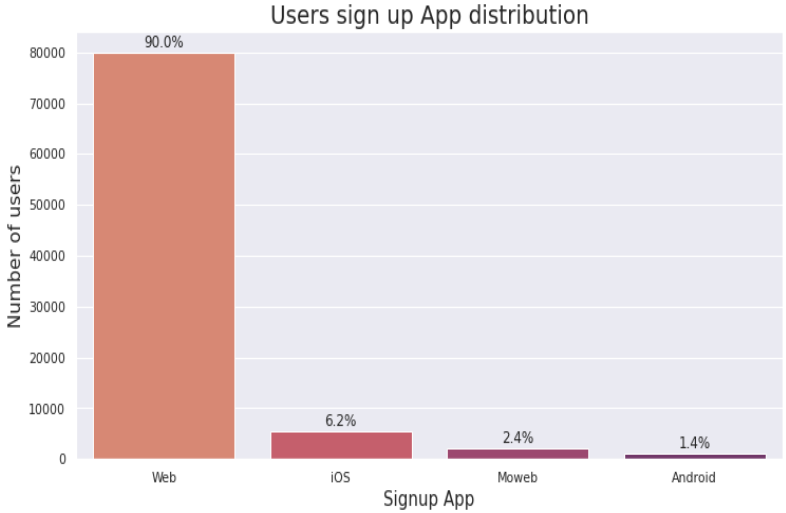
- Browser used

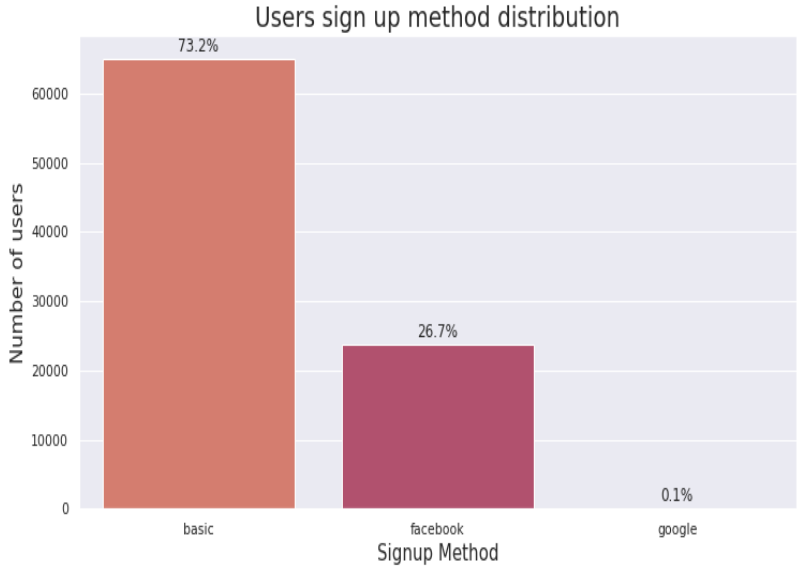


Data Exploration & Hypothesis testing - cont

Users dataset:

- Sign up

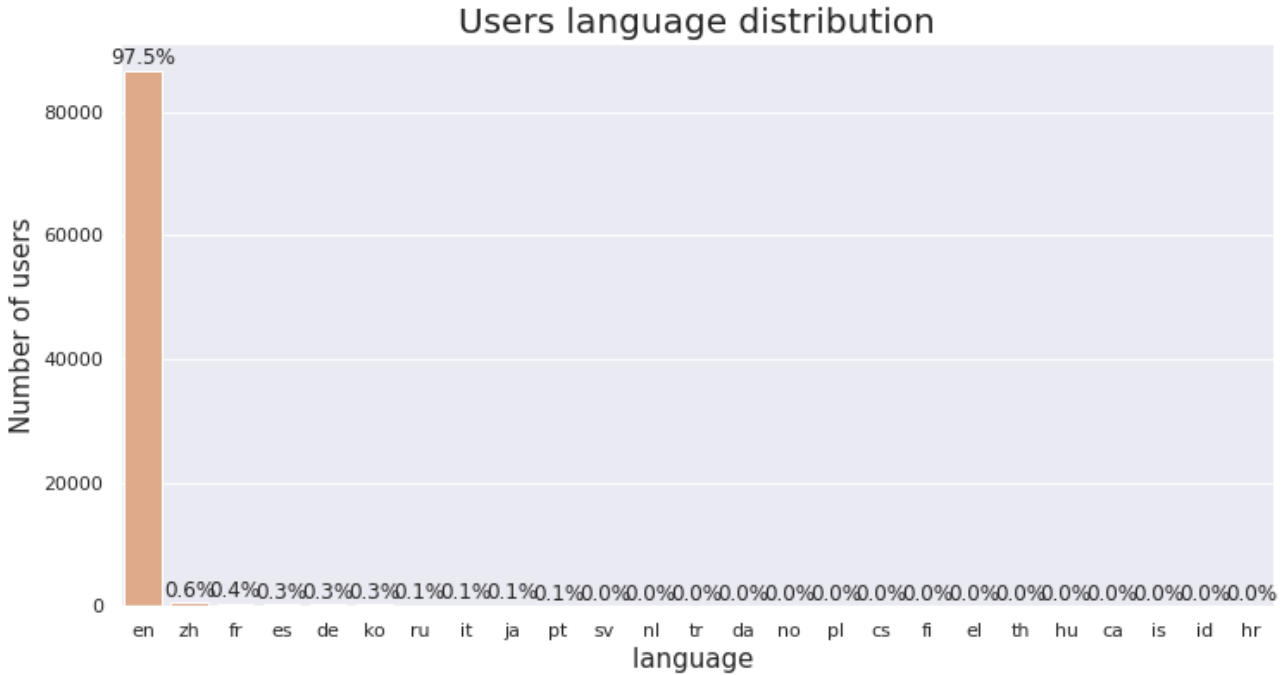




Data Exploration & Hypothesis testing - cont

Users dataset:

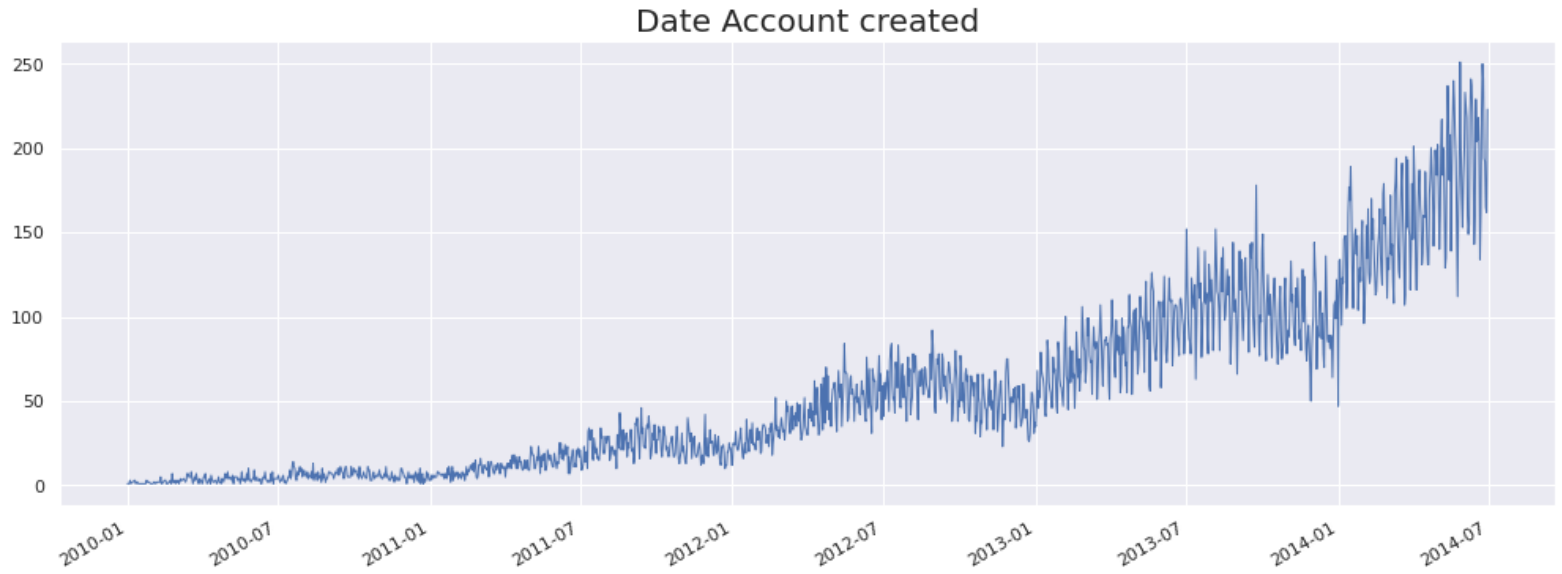
- Language



Data Exploration & Hypothesis testing - cont

Users dataset:

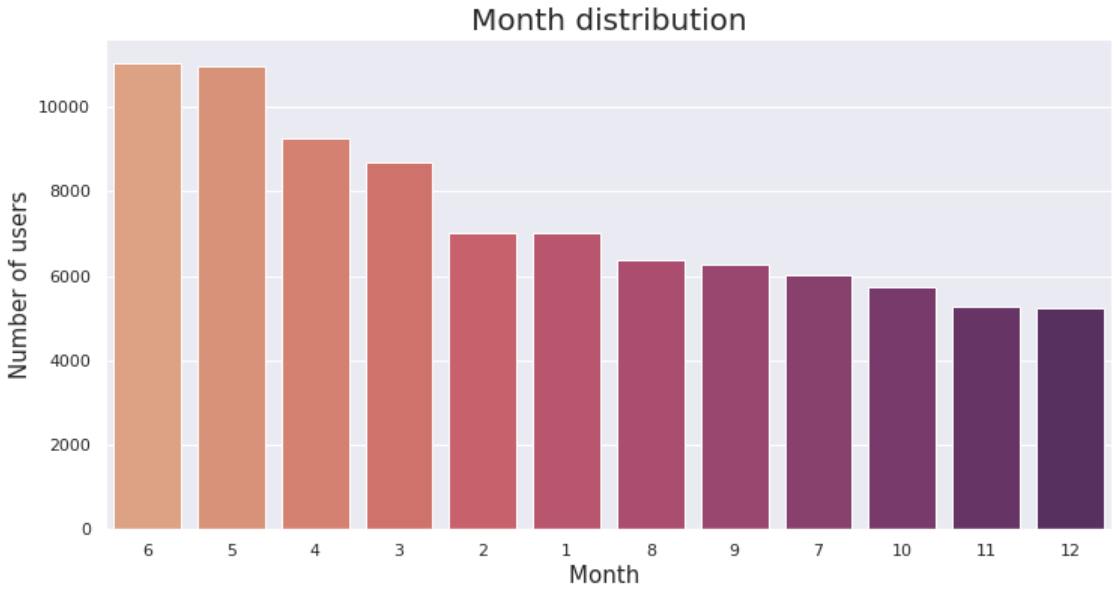
- Trend of users booking



Data Exploration & Hypothesis testing - cont

Users dataset:

- Trend of users booking monthly



Data preparation

Data preparation

Sessions dataset:

- Filling NaN with Unknown
- Calculated the sum and count for time elapsed for each action type, action.

user_id	action	count
023iyk9l	Unknown	1
	ajax_refresh_subtotal	2
	callback	0
	confirm_email	1
	dashboard	4

action_type	id	Unknown_Count	booking_request_Count	booking_response_Count	click_Count	data_Count	message_post_Count
0	00023iyk9l	3.0	1.0	0.0	4.0	9.0	1.0
1	0010k6l0om	20.0	0.0	0.0	16.0	9.0	0.0
2	001wyh0pz8	11.0	0.0	0.0	66.0	2.0	0.0
3	0028jgx1x1	1.0	0.0	0.0	9.0	5.0	0.0
4	002qnbzfs5	260.0	1.0	0.0	140.0	140.0	16.0

- Merged sessions with user train and test dataset by user_id

Data preparation - cont

Train_Sessions_merge dataset:

- Extract year and month and day of the week , hour from **Date features**

- Continuous variables such as age were binned into groups, to handle missing values in age.
  - 4 categories (Young, middle , old , other).

- Reduced for each variable in such a way so as to increase disparity
  - Language : convert to only English or Foreign.
  - Affiliate : consider only direct and google and anything else Other
  - Browsers : consider only major browsers and the rest make it Others
  - One hot encoding, label encoding for the target

Now our dataset ready to **feed the machine learning model**

Model Selection

Model Selection & Evaluation metric

- **Data was trained on three tree-based classification models to find out which best fit and know our baseline:**
  - **XGBoost classifier**
  - **Decision Tree**
  - **Random Forest**

- **Evaluation Metric : NDCG** (Normalized discounted cumulative gain)

Top 5

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

Where

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

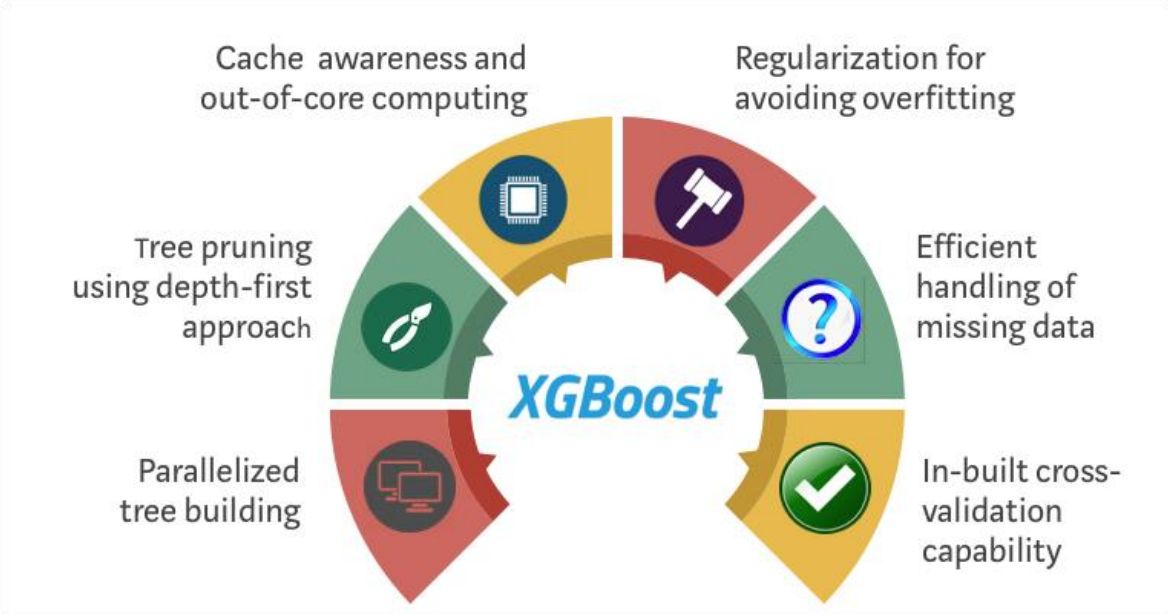
Model Selection & Evaluation metric

- Results

Model	Train Score	Test Score
Decision Tree	99%	65%
Random Forest	99%	84%
XGBoost	91%	85%

Model Selection & Evaluation metric

- Why XGBoost



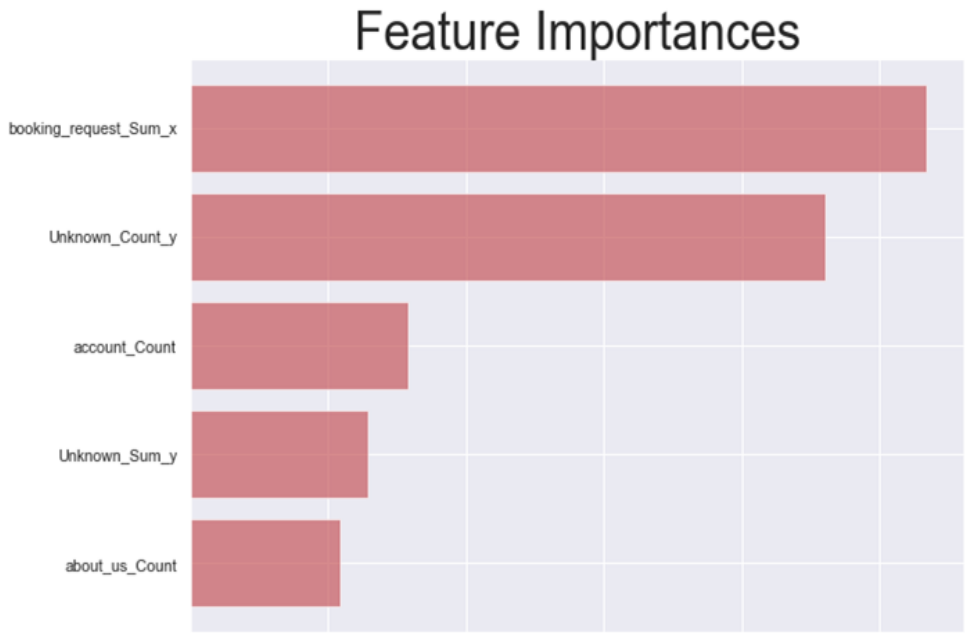
Fine-Tune the Model

Fine-tune the model

Hyperparameter	values	Best
'Max_depth'	sp_randint(3, 20)	9
'Learning_rate'	0.001, 0.01, 0.1, 0.2]	0.001
'Gamma'	[0, 0.25, 0.3,0.35,0.45,0.5,0.6,0.8,1.0]	0.8
'Reg_lambda'	[0, 0.25, 0.3,0.35,0.45,0.5,0.6,0.8,1.0]	1.0
'Subsample'	[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	0.7
'Min_child_weight'	[0.25,0.5, 1.0, 3.0, 5.0, 7.0]	0.5
'N_estimators'	[100,200,500,1000,2000]	500
'Colsample_bytree'	[0.1,0.3,0.5,1]	1

Fine-tune the model - cont

- **Top 5 Features**



Results

- **Result before selecting important features - (476 features)**

submit.csv an hour ago by Mohamed Ahmed Abdelmaguid Mohamed add submission details	0.88279	0.87654	☐

- **Result after selecting important features - (44 features)**

Submission and Description	Private Score	Public Score	Use for Final Score
submit.csv a few seconds ago by Mohamed Ahmed Abdelmaguid Mohamed add submission details	0.88413	0.87920	☐

Results - cont

- **Kaggle benchmarks the highest scores achieved on the dataset [Link]**

P.O.C	1st kaggle competitor who submitted notebooks	Our proposed model
Scores	87%	88.5%
Number of features	435	44

Other Approaches

- We can make Hierarchical classification

  To reduce the problem of imbalanced data

- Using deep learning model to capture more hidden patterns

Business Recommendations

Business Recommendations

- For new user-facing, a problem called **cold start**, to solve it we can make the user enter their **preference.**
- Due to most of the USA users tend to book an Airbnb within the country, it will be better to recommend a host near to him/her.
- Improve Android app User Experience, because most of those who book from an app use iOS devices
- Suggest to the user to install browser cookies, which are small text files that store data and can identify a user when they visit different pages on a browser for more improved service.

THANK YOU!

vodafone