




# Final Project - Airbnb Kaggle



Matt Cano

October 17, 2016



# Problem Statement & Hypothesis

---

- **Problem:** Airbnb wants to know where a person is going to book before they do it.
- **Why?** If they know before, they can serve geo specific marketing material to increase the likelihood of booking

## Hypotheses:

- Non-mobile more likely to be international
- Booked faster => US booking (need more time for international booking)

# Dataset Description

## Dimensions:

213,451 rows x 16 col

## Target:

country\_destination  
(multi-class classification)

**Obtained:** Kaggle/Airbnb

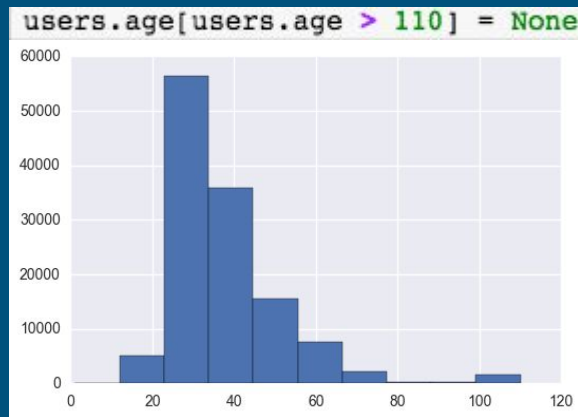
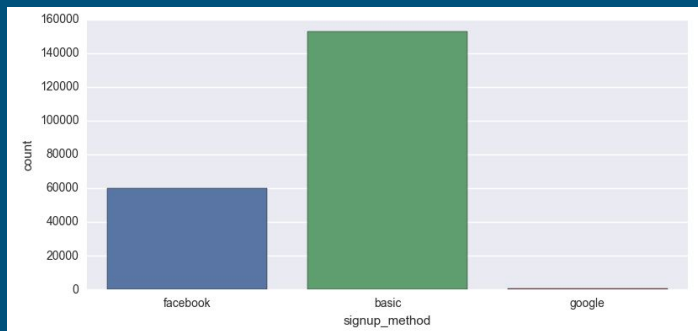
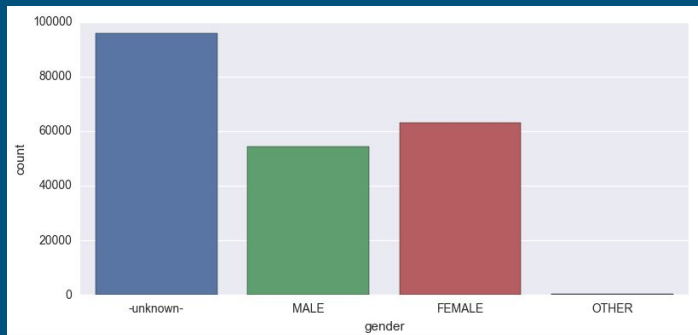
Name	Type	Description
id	text	- unique identifier
date_account_created	text	- full account created
timestamp_first_active	text	- first active on site
date_first_booking	text	- when first booked
gender	text	{fem, male, other, un}
age	float	(cleaned)
signup_method	text	{basic, facebook...}
signup_flow	int	(noisy)
language	text	(97% english)
affiliate_channel	text	{direct, sem, seo...}
affiliate_provider	text	{direct, google, craigslist...}
first_affiliate_tracked	text	{untracked, linked, omg...}
signup_app	text	{web, moweb, ios...}
first_device_type	text	{mac, windows, iphone...}
first_browser	text	{chrome, safari...}
country_destination	text	{NDF, US, other, FR...}

# Preprocessing Steps - Data Cleaning

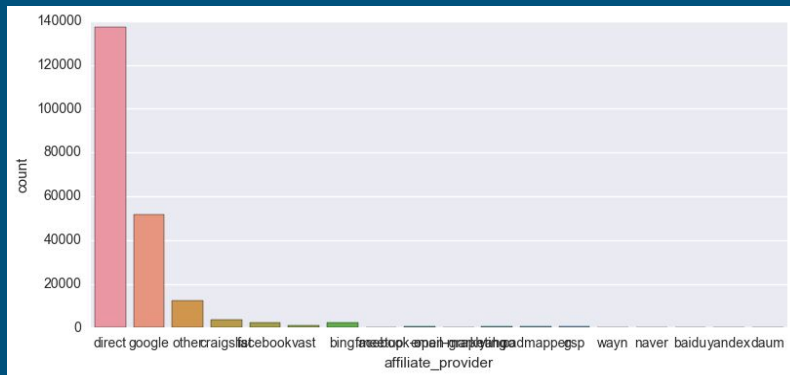
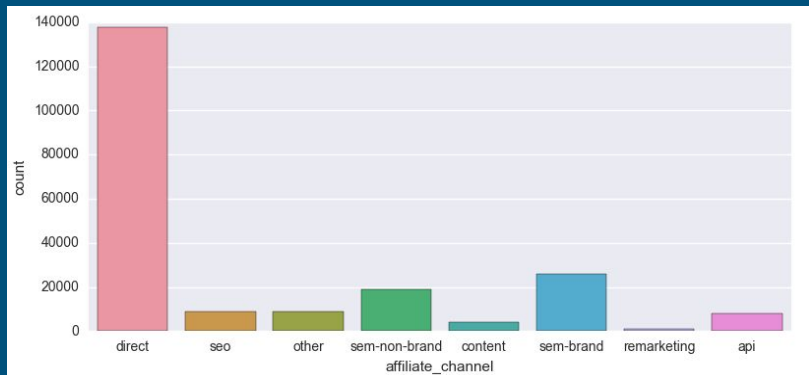
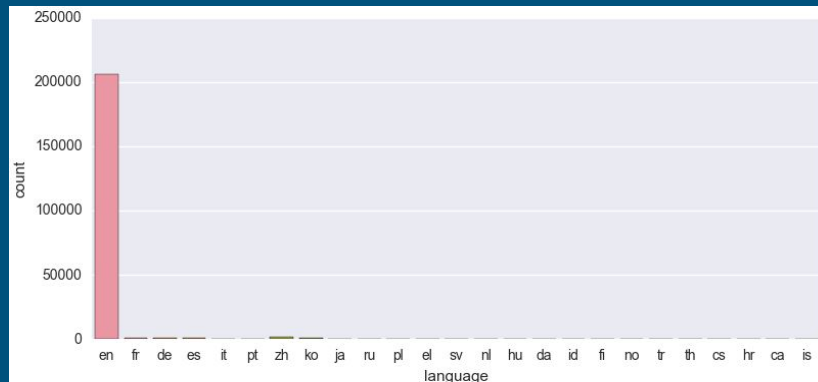
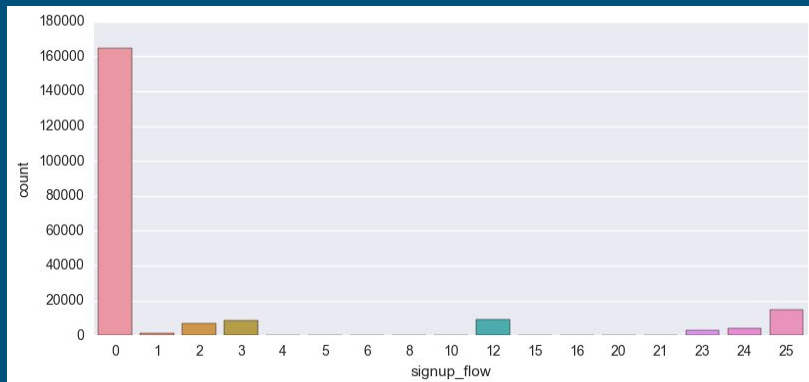
---

1. To DateTime:
  - a. account\_created
  - b. first\_active
  - c. First\_booking (strange format)
2. Age:
  - a. 800 values > 110 (stored as year, 1952) => NULL
  - b. 88k null values
3. Gender:
  - a. female, male, other or unknown

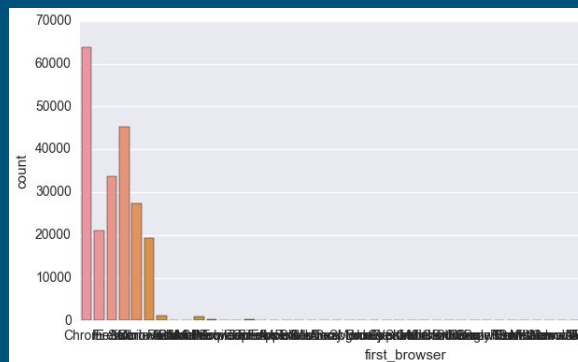
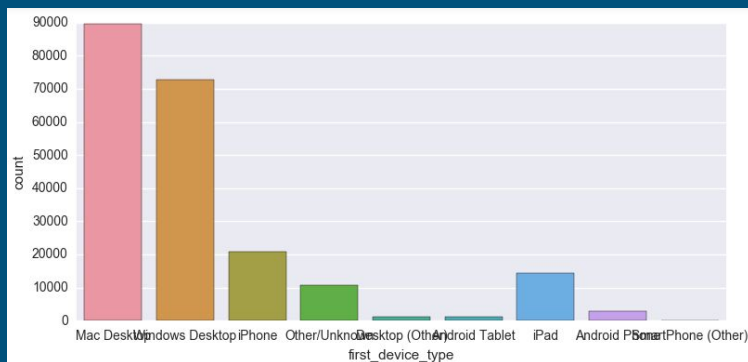
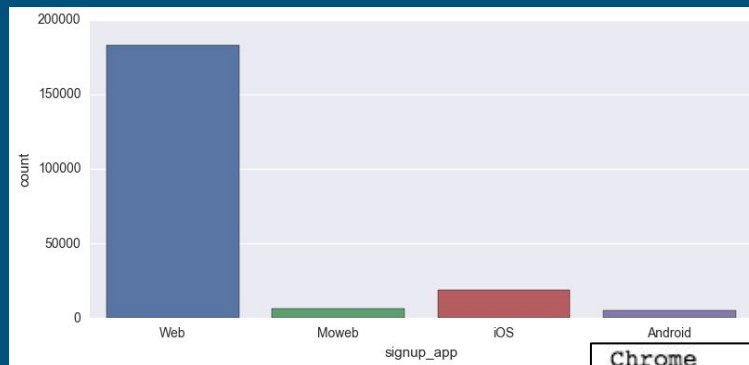
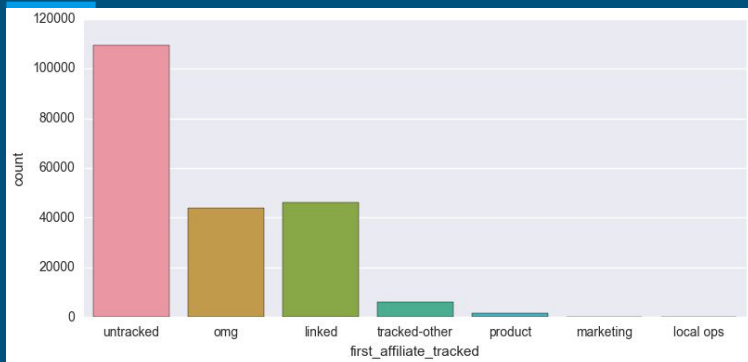
# Exploratory Data Analysis Insights



# Exploratory Data Analysis Insights

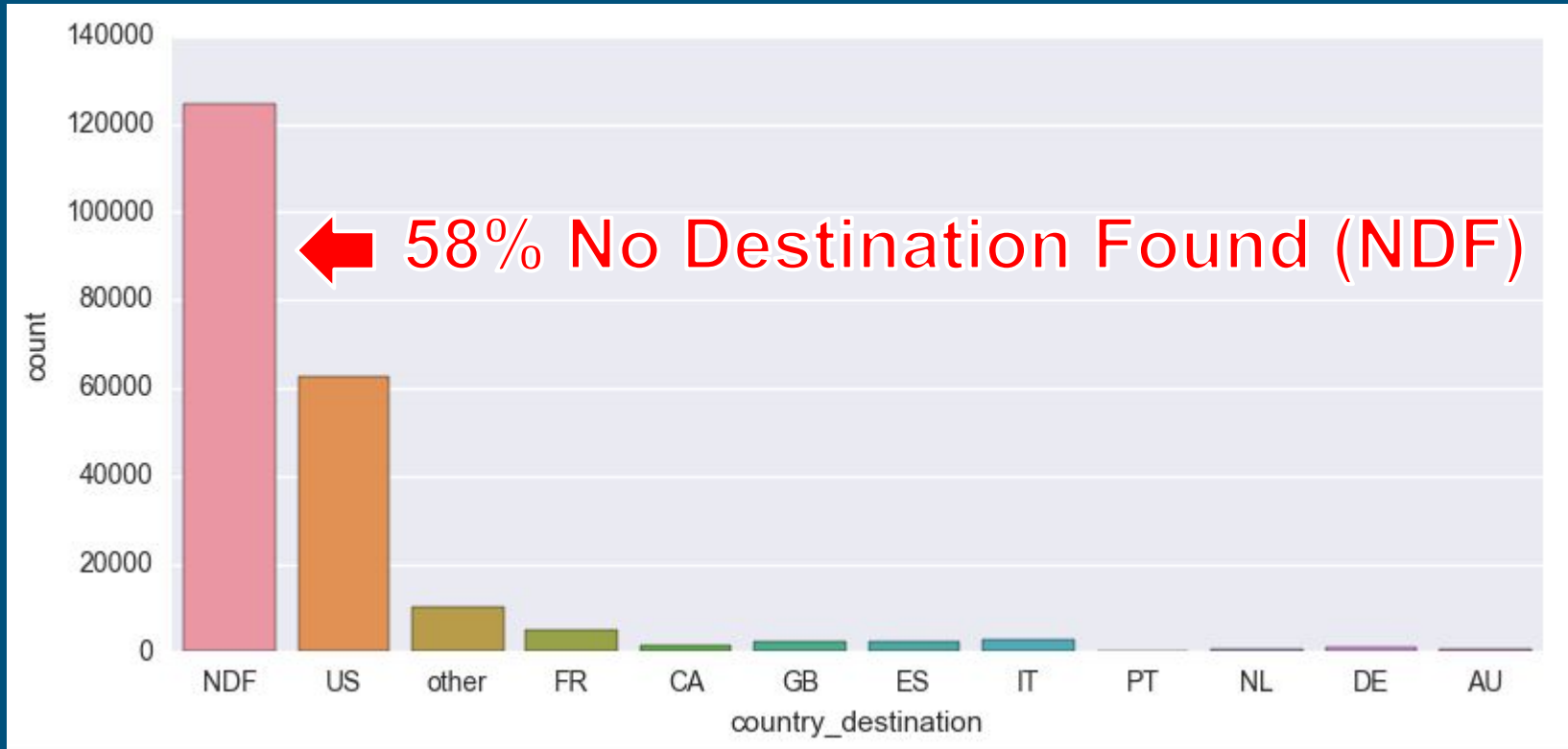


# Exploratory Data Analysis Insights



Chrome	63845
Safari	45169
Firefox	33655
-unknown-	27266
IE	21068
Mobile Safari	19274
Chrome Mobile	1270
Android Browser	851
AOL Explorer	245
Opera	188
Silk	124
Chromium	73
BlackBerry Browser	53
Maxthon	46
IE Mobile	36
Apple Mail	36
Opera Mobile	20

# Exploratory Data Analysis Insights





# Feature Engineering

---

1. Time between activities as a measure of how engaged people are

- Diff\_created\_active
- Diff\_created\_booked

2. Account created by day of week, month or season to understand temporal differences

- Account Created: month, day of week, season

3. Age by decade buckets to better group (since age and decisions are not linear)

- Age\_decade

4. Dummy variables (since mostly categorical data)

# Model - Predicting Country of Booking


Goal: Predict which country a user will book

Model Type: Logistic Regression, KNN, Decision Tree, Random Forest

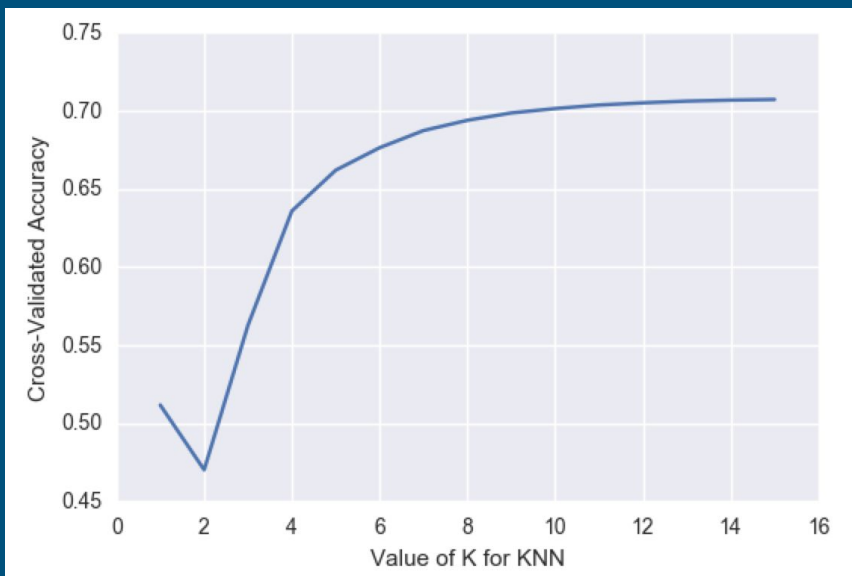
## Model #1: Logistic Regression

- CV Accuracy: 70.9%

	precision	recall	f1-score	support
AU	0.00	0.00	0.00	123
CA	0.00	0.00	0.00	269
DE	0.00	0.00	0.00	212
ES	0.00	0.00	0.00	458
FR	0.00	0.00	0.00	907
GB	0.00	0.00	0.00	469
IT	0.00	0.00	0.00	487
NL	0.00	0.00	0.00	157
PT	0.00	0.00	0.00	50
US	0.70	1.00	0.82	11874
other	0.00	0.00	0.00	1942
avg / total	0.49	0.70	0.58	16948

 Guesses "US" every time

# Model #2 - KNN



K = 15

- CV Accuracy: 70.7%

	precision	recall	f1-score	support
AU	0.00	0.00	0.00	123
CA	0.00	0.00	0.00	269
DE	0.00	0.00	0.00	212
ES	0.00	0.00	0.00	458
FR	0.00	0.00	0.00	907
GB	0.00	0.00	0.00	469
IT	0.00	0.00	0.00	487
NL	0.00	0.00	0.00	157
PT	0.00	0.00	0.00	50
US	0.70	1.00	0.82	11874
other	0.20	0.00	0.01	1942
avg / total	0.51	0.70	0.58	16948

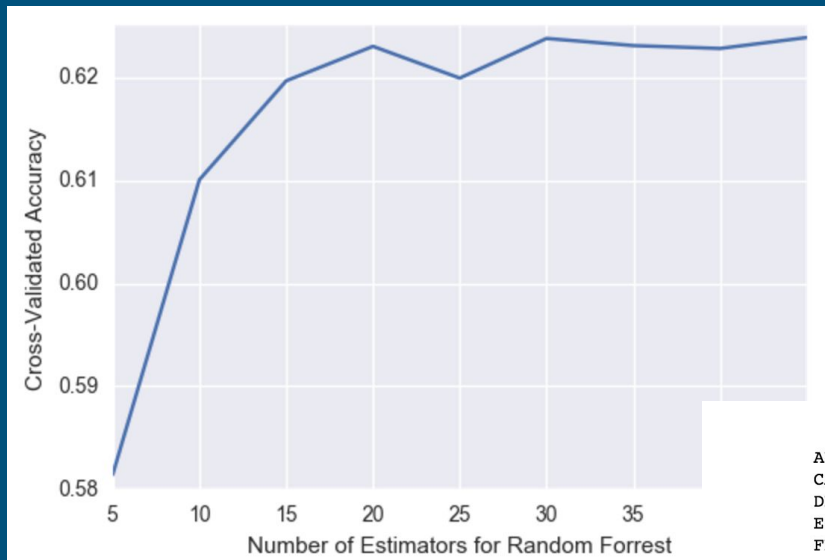
# Model #3 - Decision Tree

- CV Accuracy: 46.7%

	precision	recall	f1-score	support
AU	0.01	0.01	0.01	123
CA	0.02	0.02	0.02	269
DE	0.01	0.01	0.01	212
ES	0.03	0.03	0.03	458
FR	0.05	0.06	0.05	907
GB	0.03	0.03	0.03	469
IT	0.04	0.04	0.04	487
NL	0.00	0.00	0.00	157
PT	0.00	0.00	0.00	50
US	0.70	0.68	0.69	11874
other	0.12	0.12	0.12	1942
avg / total	0.51	0.49	0.50	16948

	Features	Importance Score
0	diff_created_active	0.362496
64	browser_Chrome	0.026461
11	age_3.0	0.025676
47	dow_2	0.023432
97	browser_Safari	0.022681
48	dow_3	0.022236
73	browser_Firefox	0.021759

# Model #4 - Random Forest



- N\_estimators = 30
- CV Accuracy: 60.0%

	Features	Importance Score
0	diff_created_active	0.362496
64	browser_Chrome	0.026461
11	age_3.0	0.025676
47	dow_2	0.023432
97	browser_Safari	0.022681
48	dow_3	0.022236
73	browser_Firefox	0.021759

	precision	recall	f1-score	support
AU	0.00	0.00	0.00	123
CA	0.01	0.00	0.00	269
DE	0.03	0.01	0.02	212
ES	0.03	0.02	0.02	458
FR	0.06	0.03	0.04	907
GB	0.01	0.01	0.01	469
IT	0.04	0.02	0.03	487
NL	0.02	0.01	0.02	157
PT	0.05	0.02	0.03	50
US	0.70	0.85	0.77	11874
other	0.13	0.06	0.08	1942
avg / total	0.51	0.61	0.55	16948

# Success & Challenges

---

## Successes:

- Multiple models
- Feature engineering
- Comfort with user data
- Comfort with time series data

## Challenges:

- NDF (no destination found) not taken into consideration => too many leaks
- Sessions data
- Actual business insights

# Next Steps

---

## Business Insights:

- Optimize for visitors to quickly create an account

## Further Analysis:

- Resample so NDF and US are not majority of selections
- Include NDF as possible predicted outcome (prevent leaks)
- Session data (# of interactions...)
- Understand what leads to booking at all

# Conclusion & Key Learnings

---

- Model quickly to get baseline
  - Creating the models is actually pretty easy
- Put a lot of time aside for multiple iterations of FEATURE ENGINEERING
- More time for EDA
  - Correlations between columns
  - Identify potential leaks
- Start modeling with fewer columns

Questions?