Predicting User Next Action Through Site Activities

Data Scientist:

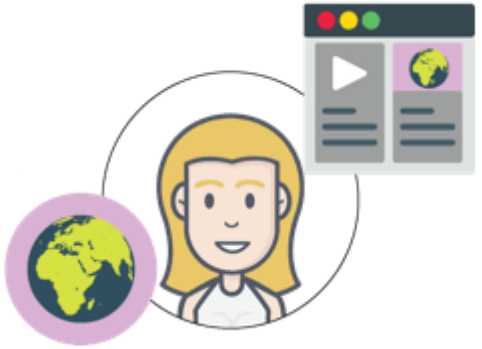
Paul Yap

Dataset:

Kaggle - Airbnb New User Bookings

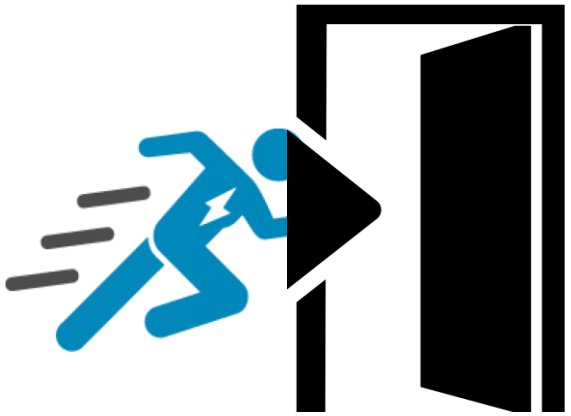
Business Problem:

predicting where a new user will book for their first travel



Enrich Engagement

Through more personalized content delivery



Improve Conversion Time

Decrease average time to first booking

Better Supply Management

By better forecasting demand

Greater Challenge:

turning internet activities into opportunities

16

Unique Features

360

Unique Actions

56,232,142

Hours Spent On Site

Dataset

Dataset

Train Set

213,451

Unique Rows

Test Set

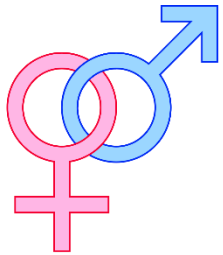
62,096

Unique Rows

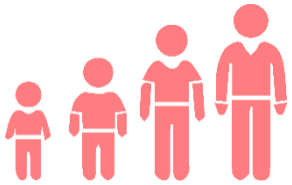
Dataset

three main types of information

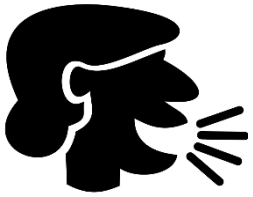
Sociodemographics



Gender



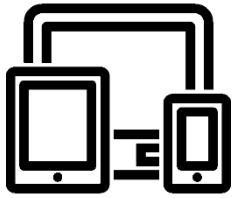
Age



Language

Surfing Preferences

Signup Method



Devices

Affiliate Channels

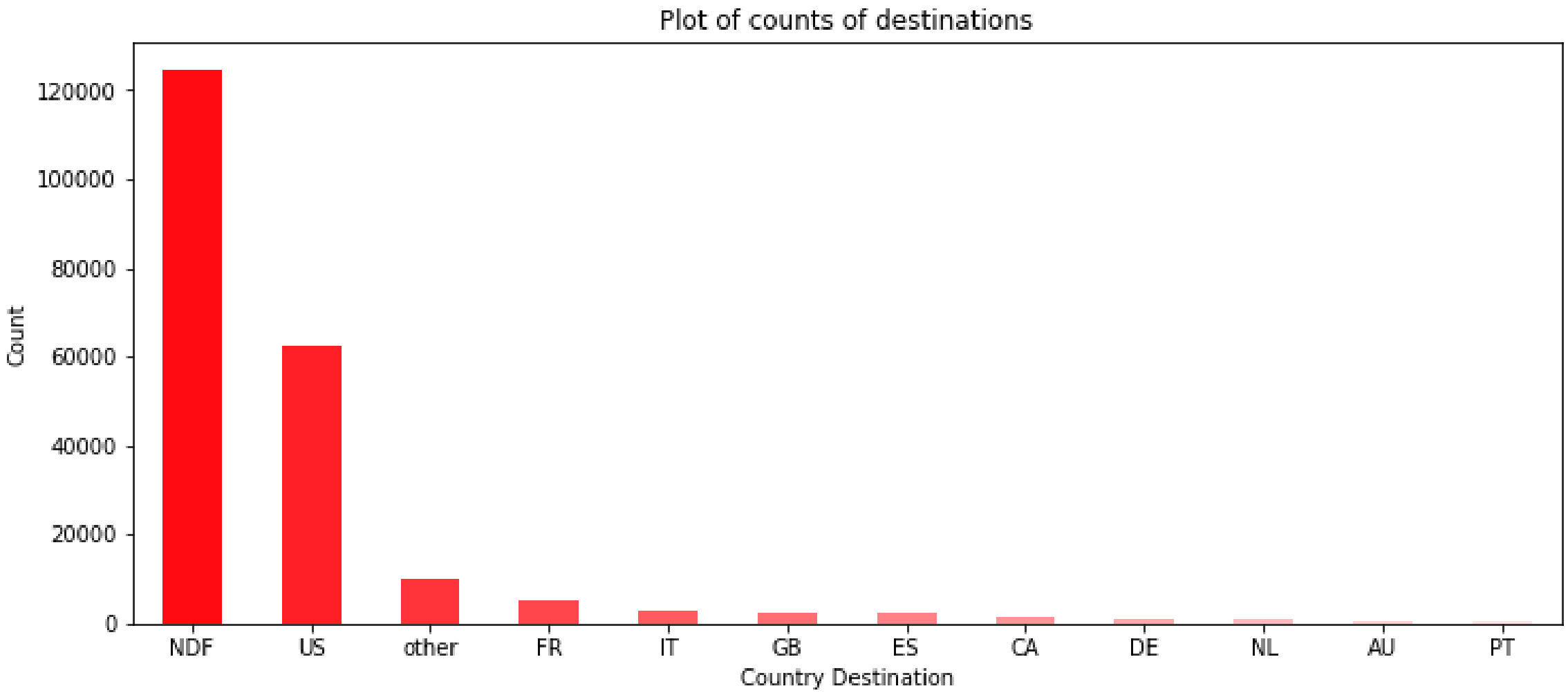
Session Logs

10.1 Million Logs

EDA

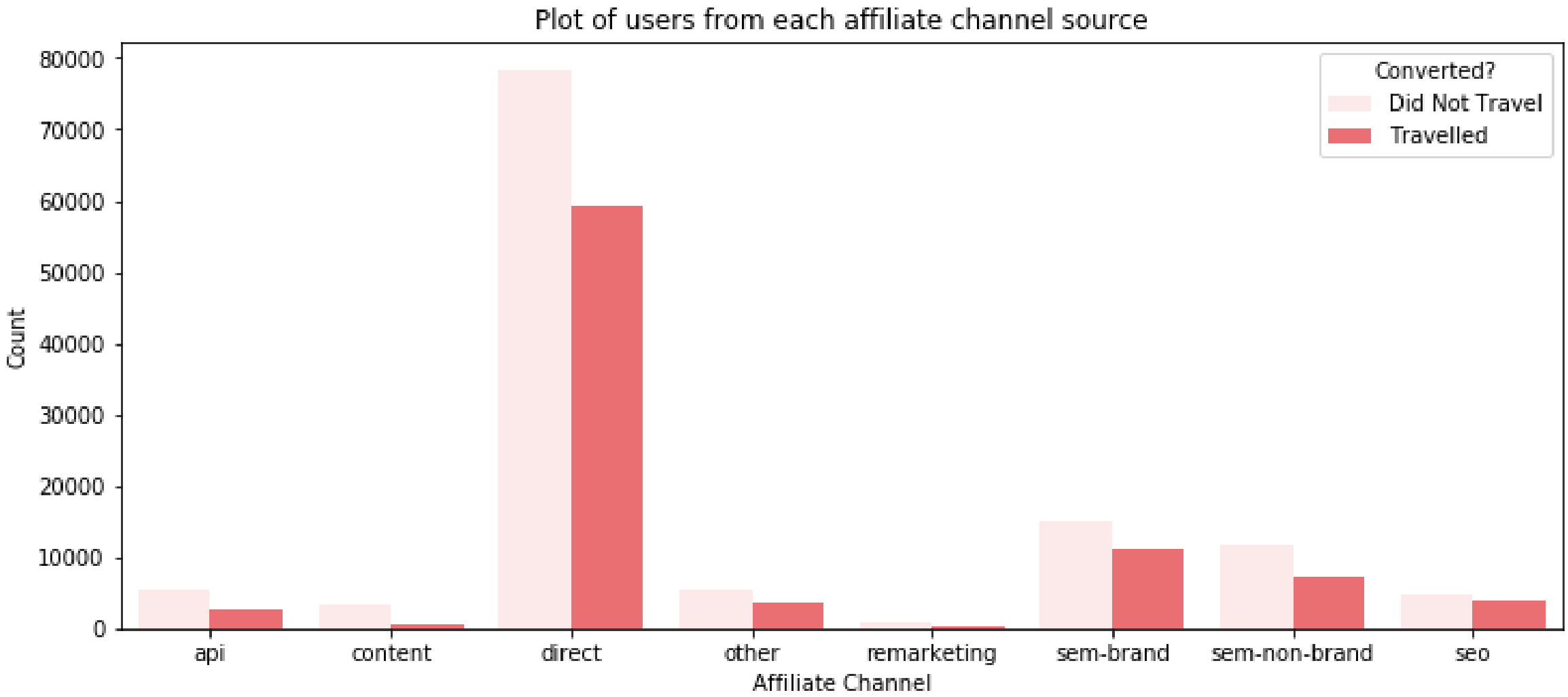
EDA

major class imbalance



EDA

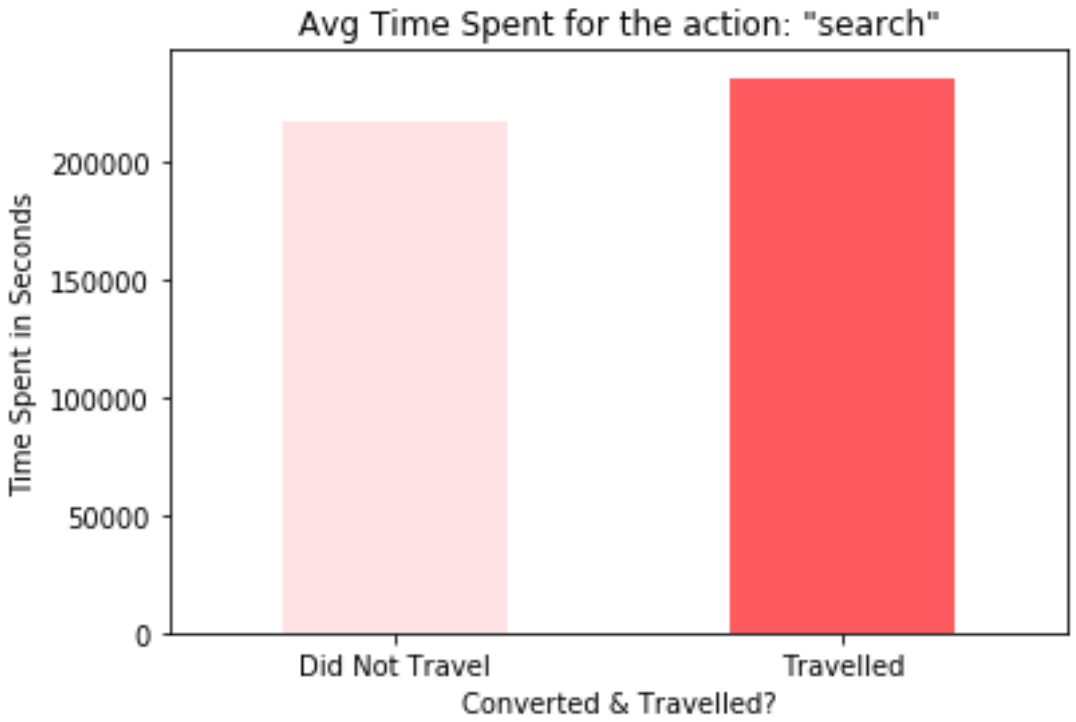
not much difference between affiliate channels…



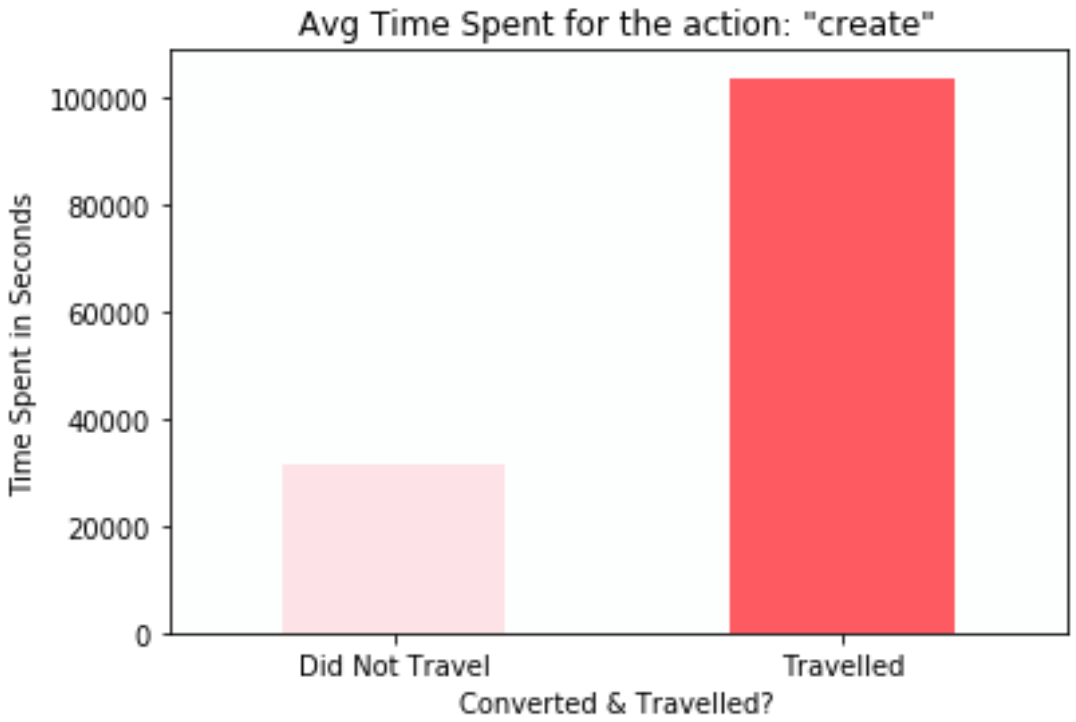
EDA

finding distinction between classes by looking into site activities

No Distinction

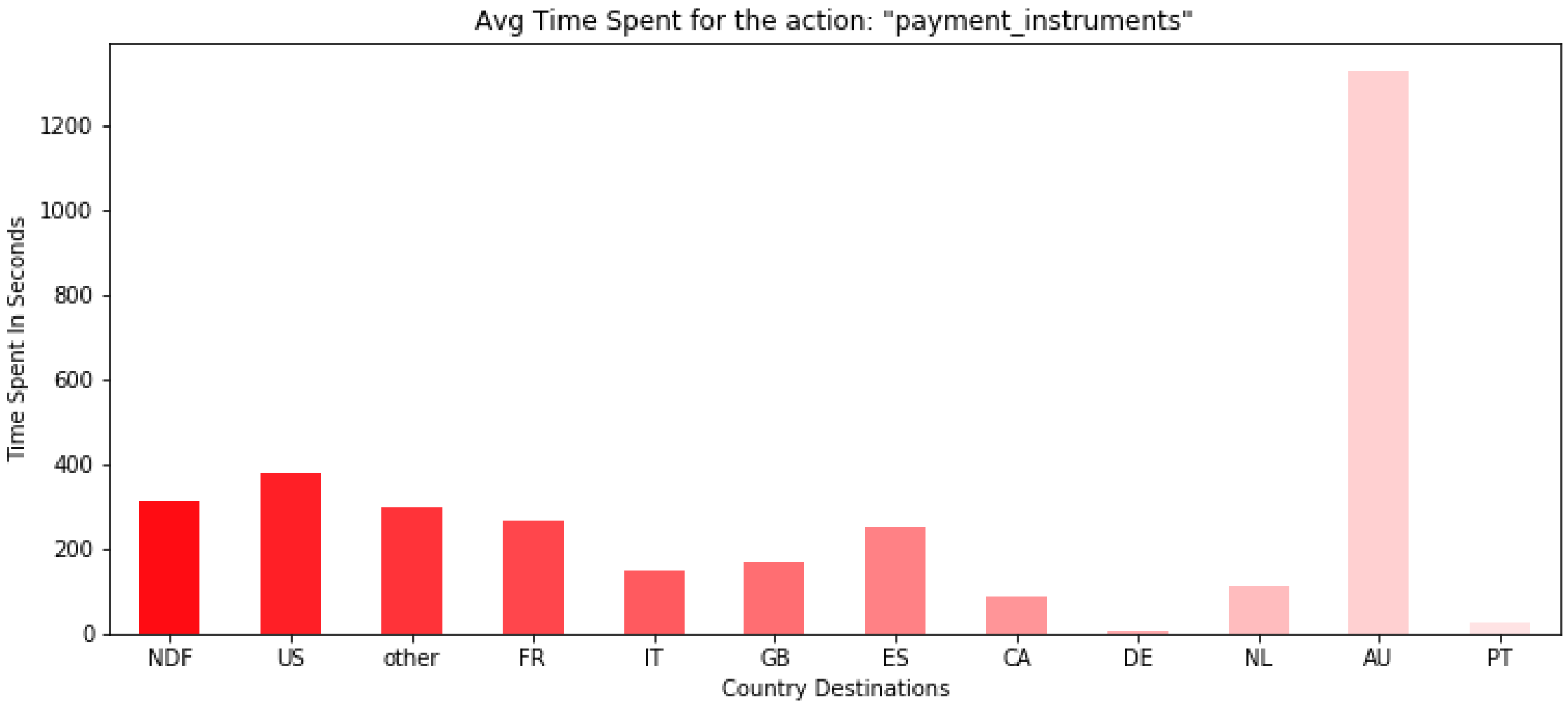


Noticeable Distinction



EDA

finding distinction between classes by looking into site activities cont…

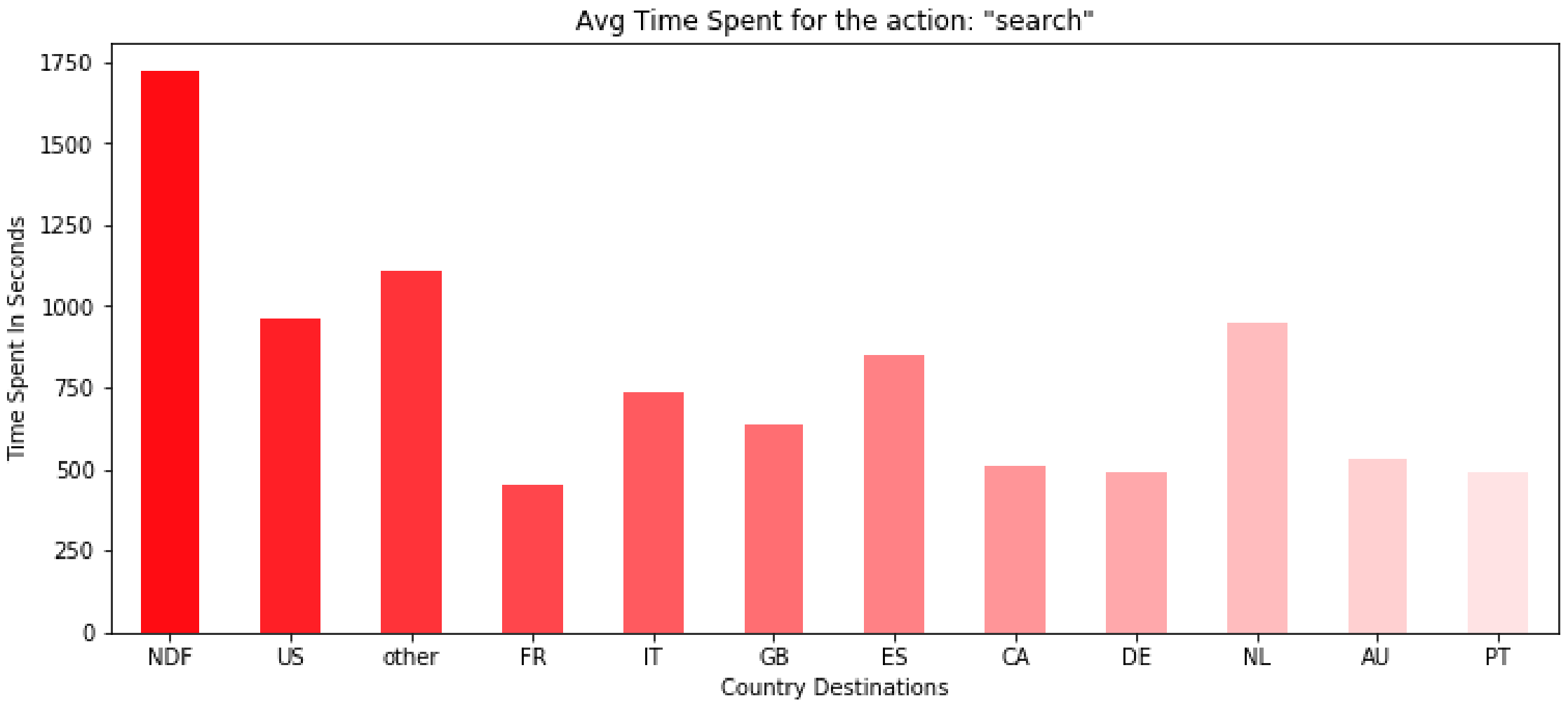


EDA

finding distinction between classes by looking into site activities cont…

EDA

finding distinction between classes by looking into site activities cont…



Modelling

Modelling

evaluation metric;
NDCG (Normalized discounted cumulative gain)

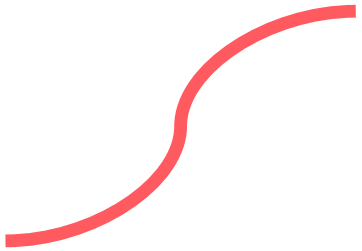
$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)},$$

$$nDCG_k = \frac{DCG_k}{IDCG_k},$$

- Allow to make up to 5 guesses
- The further the answer is away from the true value, the lower the score for that entry

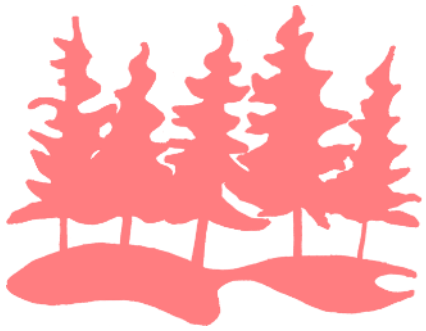
Modelling

summary – not all models are equal



Logistic Regression – **0.446**

Neural Network – **0.843**



Random Forest – **0.849**

XGBoost – **0.872**

Stacking Results

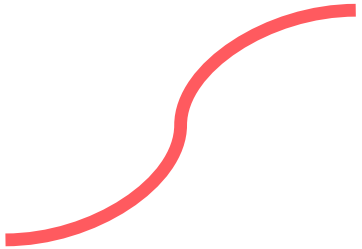
Random Forest – **0.835**

XGBoost – **0.858**

Neural Network– **0.858**

Modelling

logistic regression



Logistic Regression – **0.446**

Why this model?

- Observe how well a simple model matches up

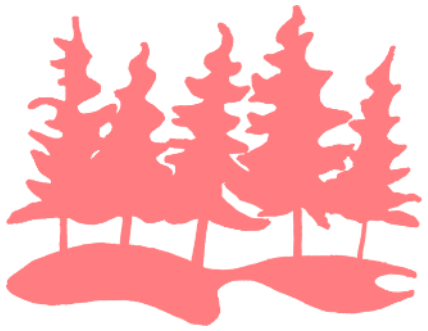
Findings

- Relationship between features seems quite non-linear
- Model is too simplistic to pick it up
- Challenge when dealing with a multiclass problem with severe class imbalance

Modelling

summary – not all models are equal

XGBoost – **0.872**



Random Forest – **0.849**

Why this model?

- Learned from log reg to pick a model capable of handling non-linear features

Findings

- Tree-based models seems to work best, though interpretability is lacking
- More robust and able to generate non-zero probabilities for minority classes

Modelling

stacking – too much noise

Stacking Results

Random Forest – **0.835**

XGBoost – **0.858**

Neural Network– **0.858**

Why this model?

- Try to piggy back on superior results to further boost it

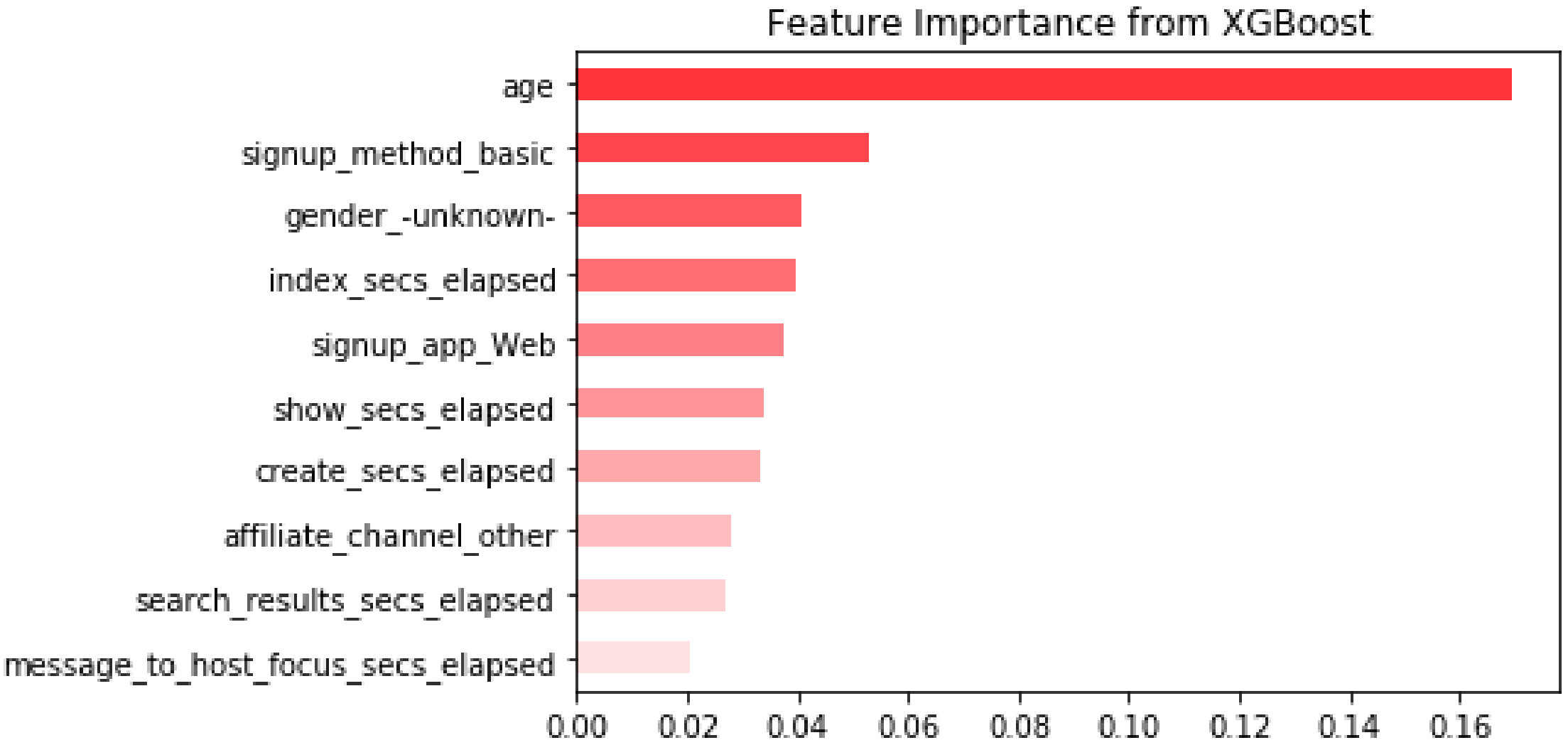
Findings

- Weights of input models are equal, creating more noise for the superior model instead
- More superior model such as XGBoost gets dampened

Findings

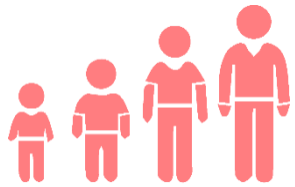
Findings

top 10 feature importance



Findings

Age



- People at different stages of their lives have different travel destination goals

Signup Method

- Signups through the web app using a desktop have higher chance of converting

Missing Value

- Can generally ignore targeting IDs with any missing information

Findings

putting it all together

Age: **26**
Gender: **Male**
Signup Method: **Basic**
Signup Medium: **Web**
Time Spent: **“index” 57,033 secs**
Time Spent: **“update” 89,270 secs**

Pred Destinations	
US	50.0%
FR	2.5%
IT	2.0%

Targeting

BY AIRBNB / MAY 4 2018
COMMUNITY, DESTINATIONS, NEWS

Airbnb Unveils Top 10 Most Hospitable Cities in the U.S.

Airbnb anchors a customer service center in Wasquehal (Hauts-de-France) in partnership with Acticall Site!

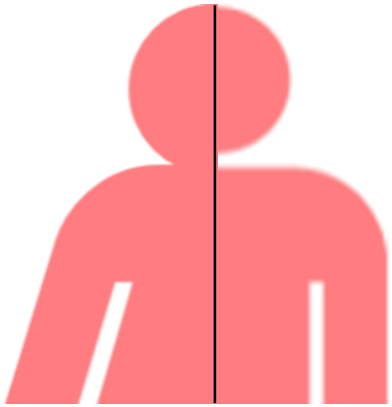
For the first time in its history, Airbnb announces the establishment, in Wasquehal, near Lille in the Hauts-de-France region, of its first customer service centre dedicated to the Airbnb Community in France.

Airbnb in this idyllic Italian town for 3 months for free

Catherine Clifford | 10:00 AM ET Sat, 19 Jan 2019

Findings

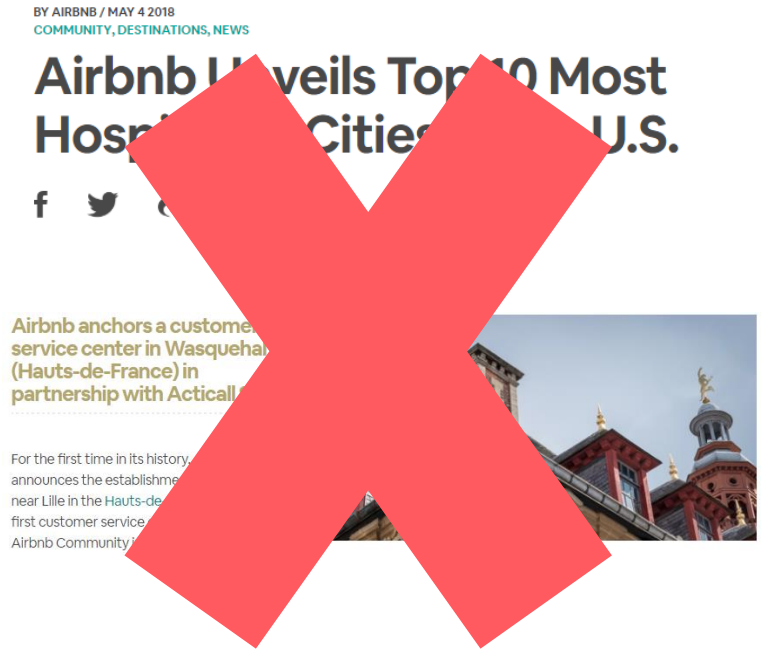
putting it all together



Age: **37**
Gender: **-unknown-**
Signup Method: **Basic**
Signup Medium: **Web**
Time Spent: **“create” 26,418 secs**
Time Spent: **“update” 10,452 secs**

Pred Destinations	
No Travel	89.5%
US	5.9%
other	1.8%

Targeting

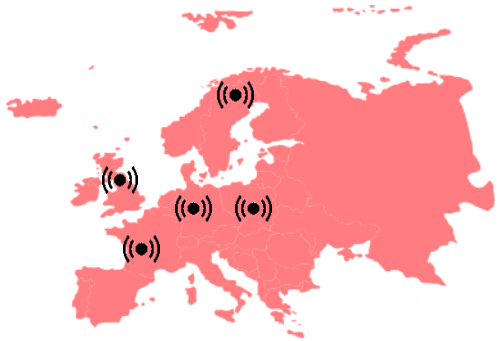


Limitations

Limitations

Lacks Interpretability

It is difficult to further dissect the model to gain more insights



Specificity Limit

The business still has to decide with of the 5 outputs to target market for

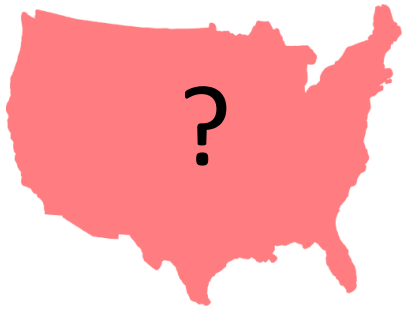
Limited Feature Information

Some of the information given have to be inferred. Which increased uncertainty

Limitations

possible information to improve model

Search Terms



Device Location

More Observations on Minority Classes

Thank You!

Data Scientist:

Paul Yap