AirBnB New User Bookings

Study to predict the destination country for a new user booking

Springboard

Capstone Project

Made By
Kshitiz Khatri

Overview

- Introduction
- Exploratory Data Analysis (EDA)
- Modeling
- Ideas for further research

Introduction

- Aim – To predict the destination country for a new user registered with the client
- Client – AirBnB
- Data – Kaggle AirBnB recruitment challenge
  - ✓ Countries
  - ✓ Age gender buckets
  - ✓ Test and Train users
  - ✓ Sessions

Data

- Countries
  - ✓ US – USA
  - ✓ CA – Canada
  - ✓ NDF – No Destination Found

12 such destination countries

- Age gender buckets
  - ✓ Age bucket
  - ✓ country destination

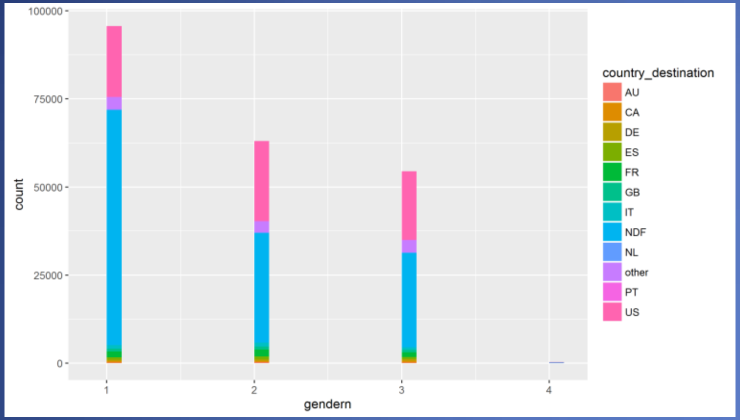
Data

- Age gender buckets
  - ✓ Gender
  - ✓ Population in thousands
  - ✓ Year
- Sessions
  - ✓ Action – lookup, show, personalize etc.
  - ✓ Action_type – data, view, click etc.
  - ✓ Device Type & time_elapsed in secs

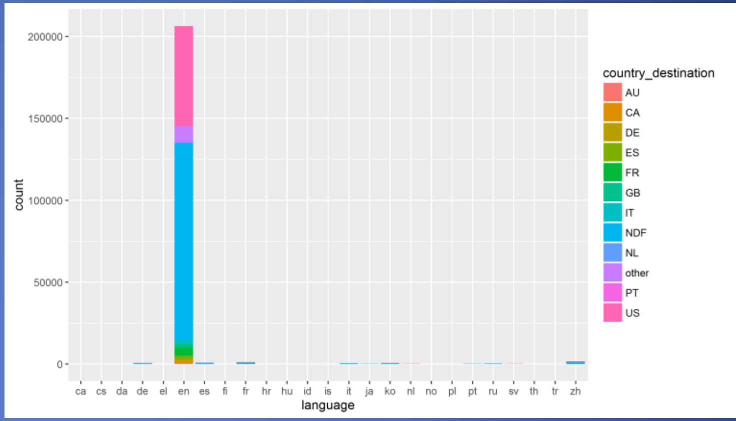
Data

- Training & Test set
  - ✓ date of account created & timestamp first active
  - ✓ signup method, signup flow
  - ✓ gender, age
  - ✓ language, affiliate channel, affiliate provider
  - ✓ first affiliate tracked, first device type
  - ✓ first browser, **country destination**

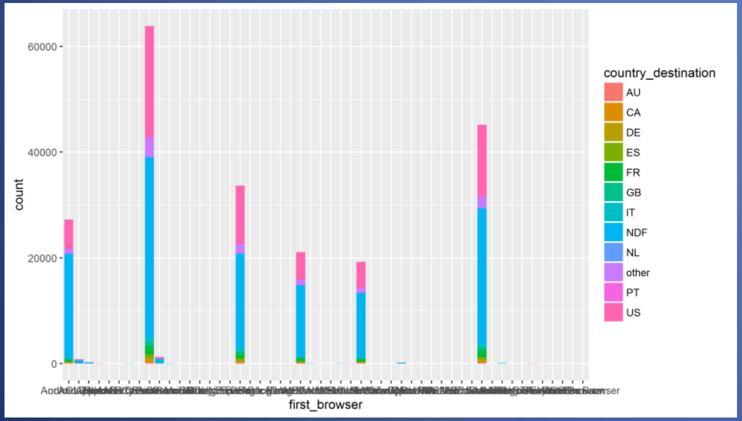
Exploratory Data Analysis



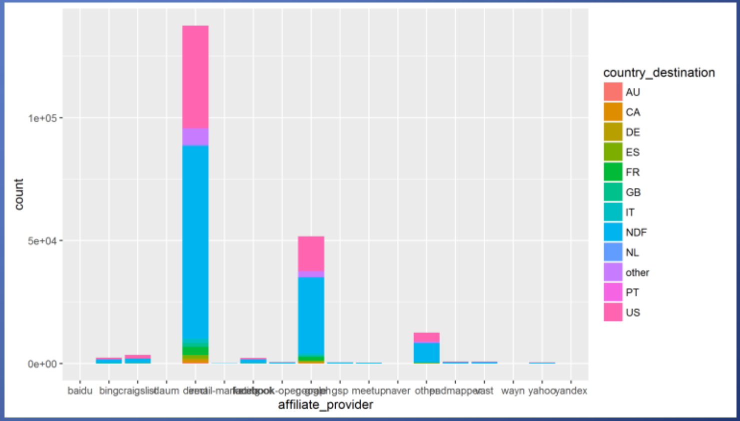
Gender



Language

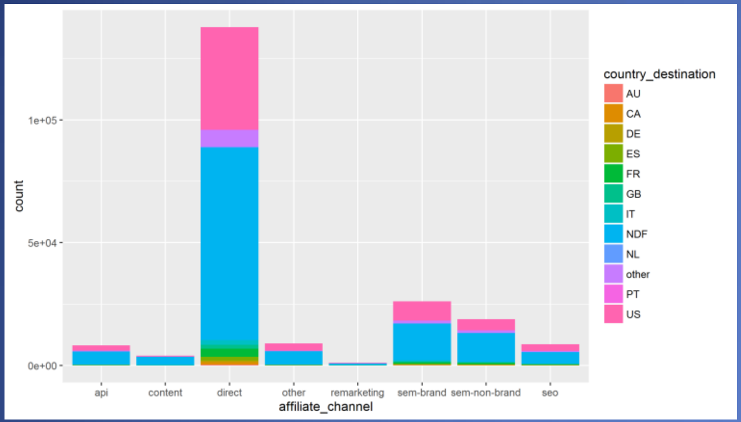


Browser

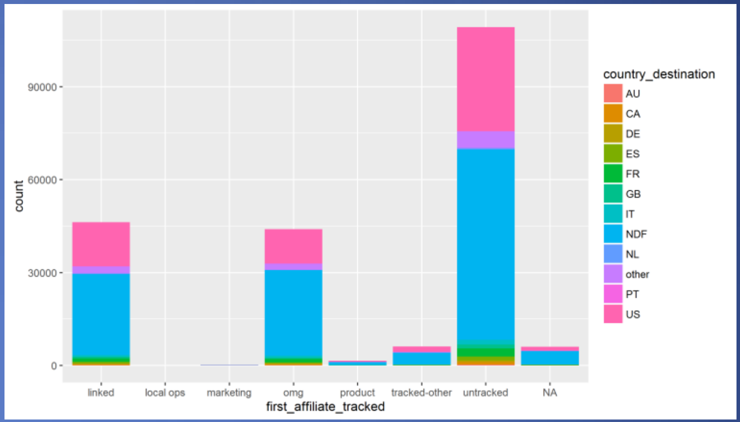


Affiliate provider

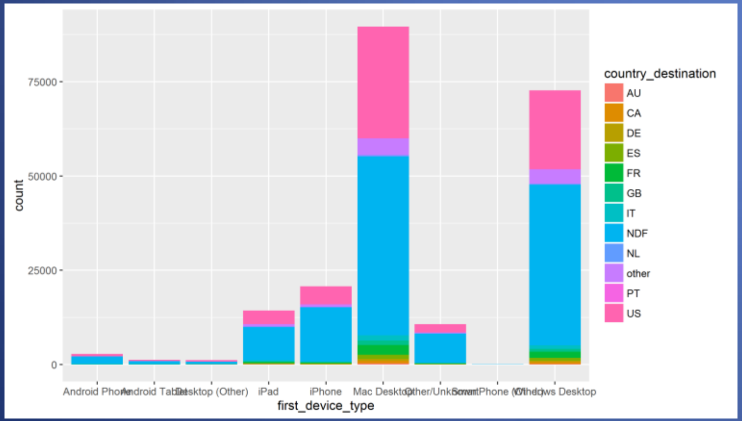
Exploratory Data Analysis



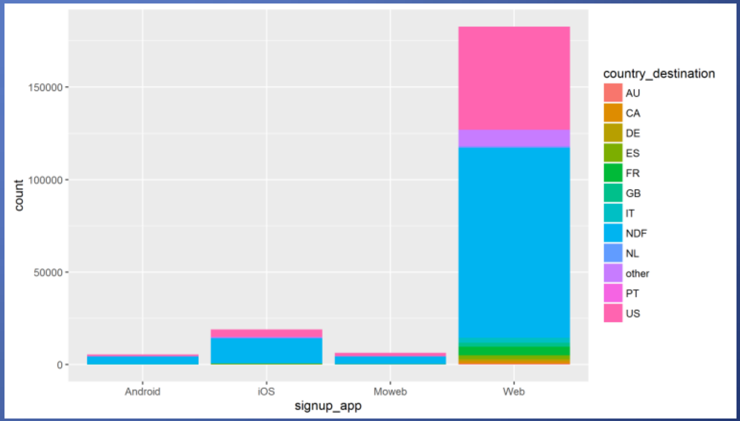
Affiliate channel



First Affiliate tracked

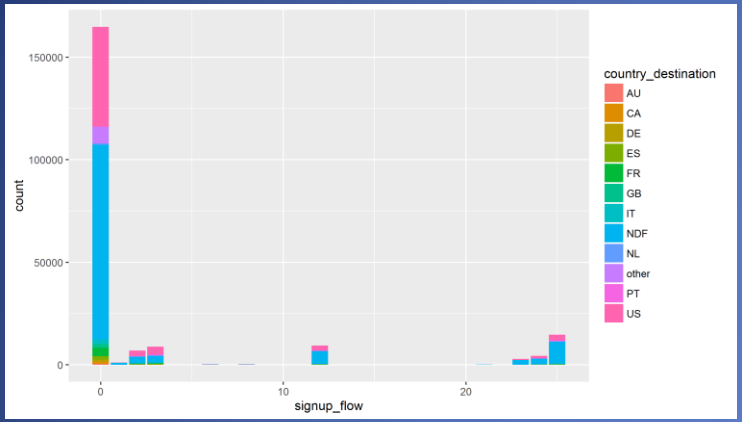


First Device Type

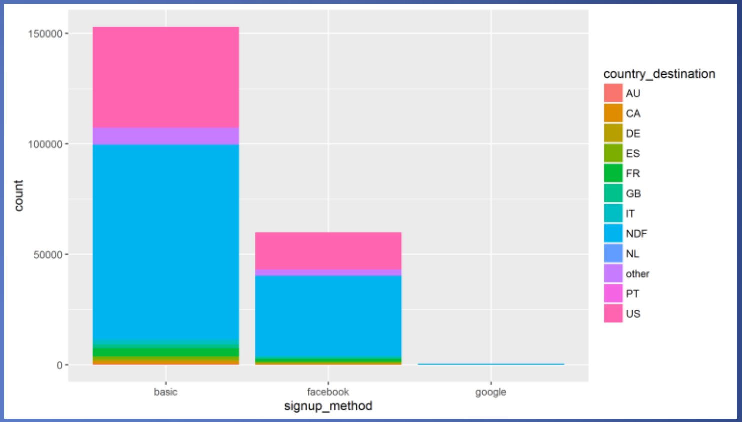


Signup app

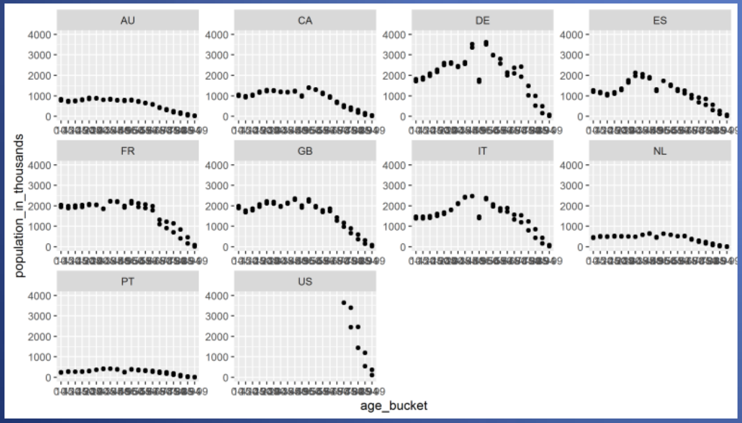
Exploratory Data Analysis



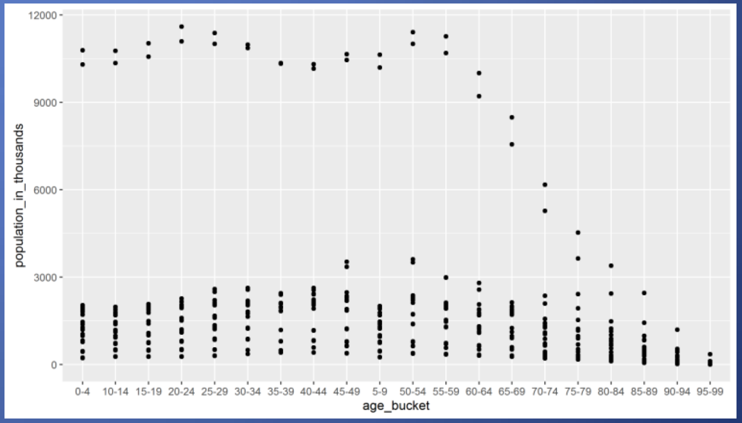
Signup flow



Signup method



Age bkt vs Population, country wise



Age bkt vs Population

Analysis

- Modeling Techniques Used:
  - ✓ Random Forest
  - ✓ xgboost decision trees
- Feature Engineering:
  - ✓ Categorical Variables to Numeric
  - ✓ Dates to individual days, months and year
  - ✓ Predicted binary variable for each destination country

Analysis

- Parameter Tuning: Random Forest

Model	Independent Variable	No. of Trees	Node Size	Accuracy (as per Kaggle)
model1	Age, gender, signup language, affiliate provider, first browser	400	25	66.66 %
model2	Age, gender, affiliate provider, first browser	400	25	62.73 %
model3	Age, gender, signup language, signup flow, affiliate provider, first browser	400	25	66.29 %
model4	Age, gender, signup flow, affiliate provider, first browser	400	25	61.89 %
model5	Age, gender, signup language, affiliate channel, first browser	400	25	67.86 %
model6	Age, gender, signup language, first affiliate tracked, first browser	400	25	67.9 %

Analysis

- Parameter Tuning: xgboost model

Parameters explored for xgboost model:

- ✓ max.depth
- ✓ eta
- ✓ nthread
- ✓ booster
- ✓ nrounds

Best accuracy achieved : 65.10 %

Analysis

- General Challenges while modeling
  - ✓ High Class Imbalance in response variable
  - ✓ Multiple minority class in response variable
  - ✓ High number of independent variables demanded complex model
- Practical challenge
  - ✓ Better system requirements to efficiently process the data if working with simple models like Random forests

Analysis

- Best Model Comparison

Model	Accuracy	Shortcoming
Random Forest	67.9%	Highly skewed results due to high class imbalance
Xgboost classifier	65.10%	Lower accuracy, but captures the minority class as well

Ideas for further research

- Hyper parameter optimization
- Limiting the classes by eliminating the least frequent
- Further exploration of the data
- Devising new features using predictors
- Ensembling

Kaggle Submissions

- Top 2 Kaggle Submissions

Model	Accuracy
Random Forest	68.39%
Random Forest	67.90%