PREDICTIVE MODELING OF NEW USER BOOKINGS

Nick Kessler

Problem Background

- Airbnb users provide numerous data points about themselves when signing up
- This information is collected both actively and passively
  - Active: Age, Gender, Language Preference, Location
  - Passive: Browser, OS, Device Type, Device Language
- Can this information be used to predict where a user will choose to book their first trip using Airbnb's service?

Data

Data Information

- Data provided by Airbnb as part of Kaggle competition
- Two datasets:
  - User data (demographic and signup device information)
    - User data includes information for users who signed up between 2010-2014
  - Session Data (web session logs)
    - Only available for users who signed up after 1-Jan-2014
- Data is mostly categorical

User Data

Column	Type	Description
id	String	User identifier.
date_account_created	Date	Date of account creation.
timestamp_first_active	String	Timestamp of first activity. This can occur before account creation as users can browse the site before signing up. Stored as a string value in the dataset.
date_first_booking	Date	Date of first booking. This value is null if the user never booked a trip.
gender	Categorical	User supplied gender. Male, Female, Other or Unknown.
age	Numeric	User supplied age value.
signup_method	Categorical	User signup method. Indicates whether the user created a direct Airbnb account, or signed up with their Facebook or Google account.
language	Categorical	User language preference
affiliate_channel	Categorical	Type of paid marketing, if applicable.
affiliate_provider	Categorical	Where marketing came from. E.g. Google, craigslist, etc.
first_affiliate_tracked	Categorical	First marketing the user interacted with prior to sign up.
signup_app	Categorical	The app the user used to sign up. E.g. Web, iOS, Android.
first_device_type	Categorical	The device type used to sign up. E.g. Windows PC, Mac, iPhone.
first_browser	Categorical	Web browser used to sign up, if applicable
country_destination	Categorical	The destination chosen by the user for their first booking. This is the target variable in this analysis.

Session Data

Column	Type	Description
user_id	String	User identifier. Corresponds to id column in User dataset.
action	Categorical	Description of the action
action_type	Categorical	Action category
action_detail	Categorical	Detailed action description
device_type	Categorical	Type of device (e.g. Windows Desktop, Mac, iPhone)
secs_elapsed	Numeric	Number of seconds elapsed in session.

Target Categories

Country Labels	Full Country Name
AU	Australia
CA	Canada
DE	Germany
ES	Spain
FR	France
GB	United Kingdom
IT	Italy
NL	Netherlands
PT	Portugal
US	United States
NDF	No Destination Found (i.e. None)
Other	Other

- Twelve possible target values

Data Cleaning & Preparation

User Data

- Data was largely pre-processed and clean
- Timestamp values converted from string to datetime datatypes
- Missing values for *Gender* were substituted with ‘UNKNOWN’
- Certain *Age* values were entered as birthyear, and were adjusted to age by subtracting birth year value from current year
- Missing *Age* values substituted with 0
- Additional derived features engineered from provided data.

Data Cleaning & Preparation

User Data – Derived Features

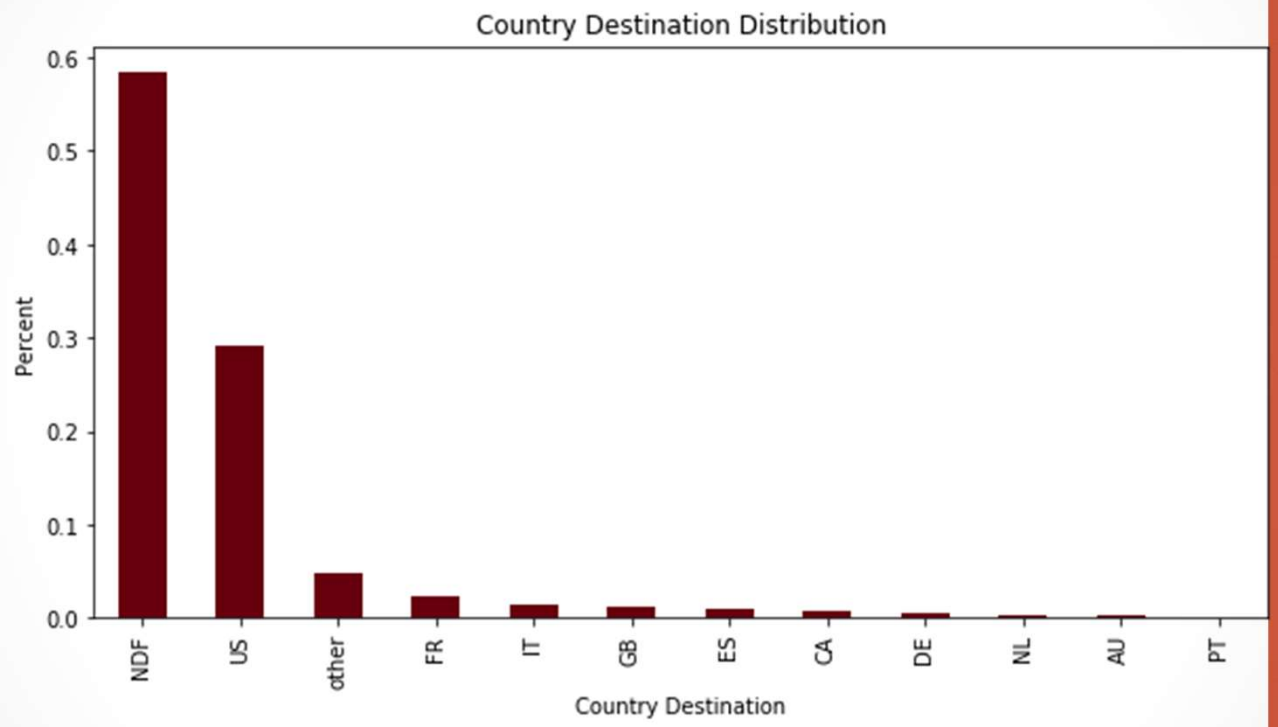
Column	Type	Description
age_group	Categorical	Derived from Age value. Six age-range group.
days_to_book	Numeric	Difference in days between account creation and date of first booking (if applicable).
device_class	Categorical	Derived from Device Type, simplifies devices into four classes: PC, Tablet, Phone, or Other
affiliate_provider_type	Categorical	Derived from Affiliate Provider, groups providers into four classes: Search, Social, Direct, and Other

Data Cleaning & Preparation

Session Data

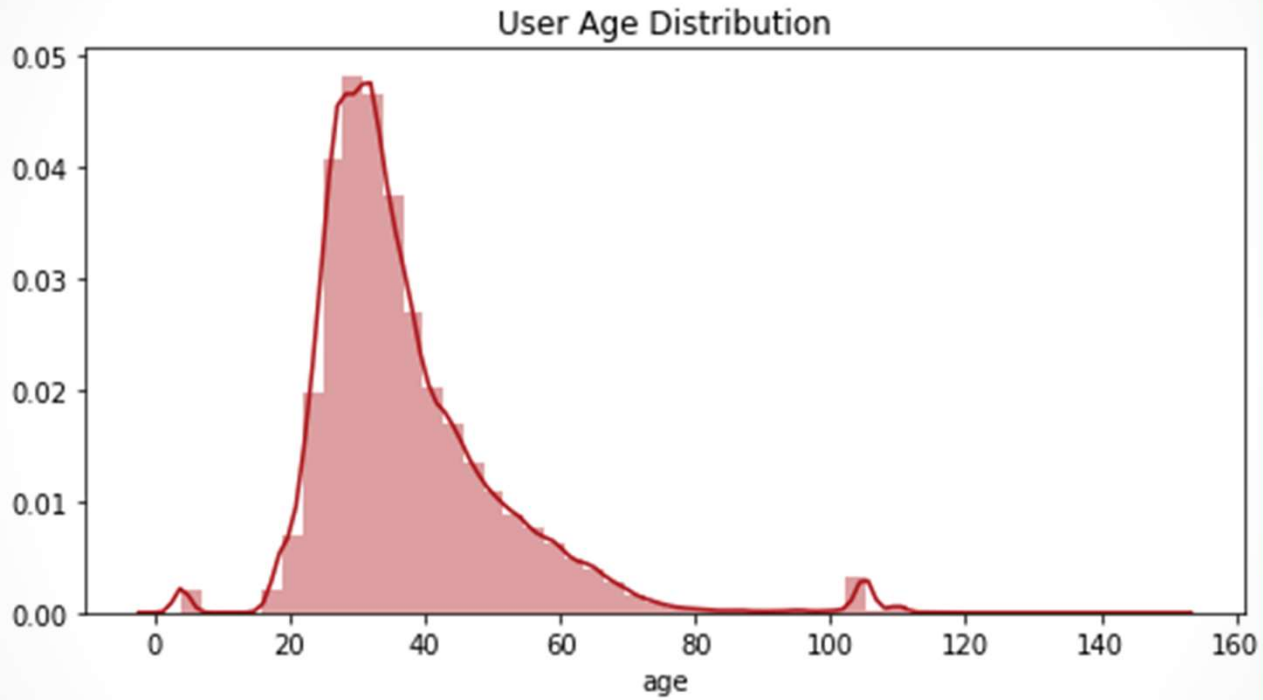
- Data contains web logs showing each action grouped by category.
- To join with user data, session data needed to be reduced to a single row for each user ID
- Counts of each action type were compiled for each user ID, creating a dataset that contains the total number of actions in each category initiated by the user.
- Seconds Elapsed was summed for each user ID to create a total value.

Exploratory Data Analysis



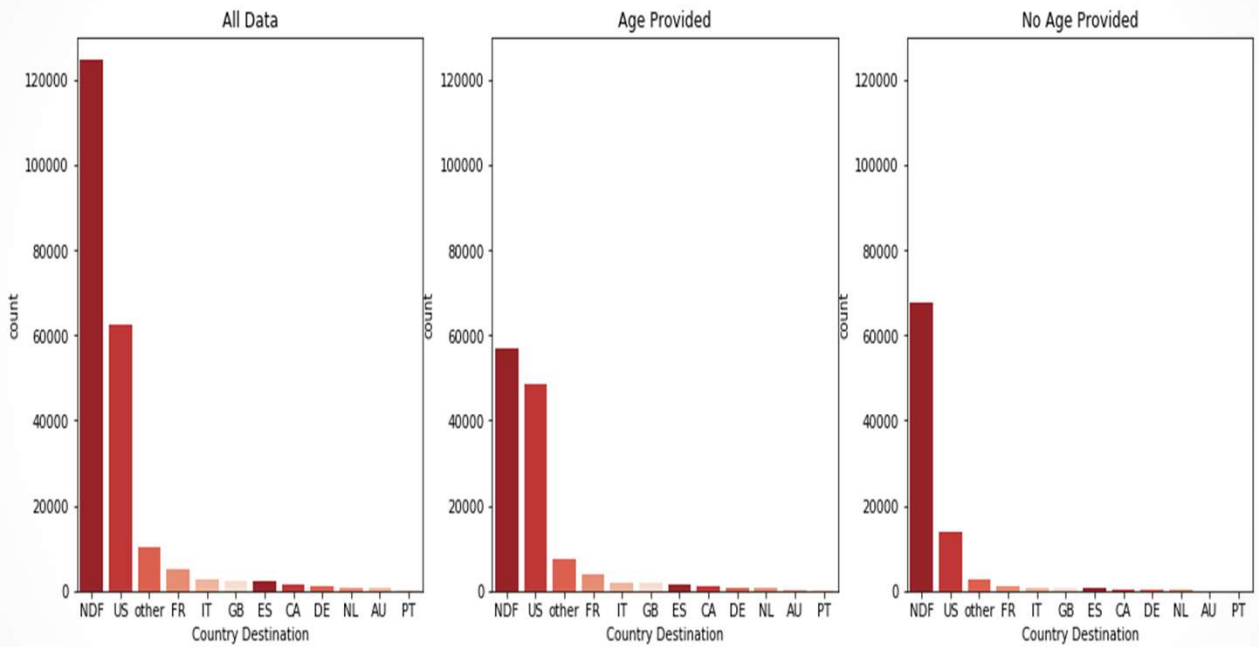
Distribution of Destinations

- Majority of users who signed up did not book a trip
- US is most popular destination
- Other is more popular than all non-US top destinations
- NDF, US, and other comprise 92% of the total data.



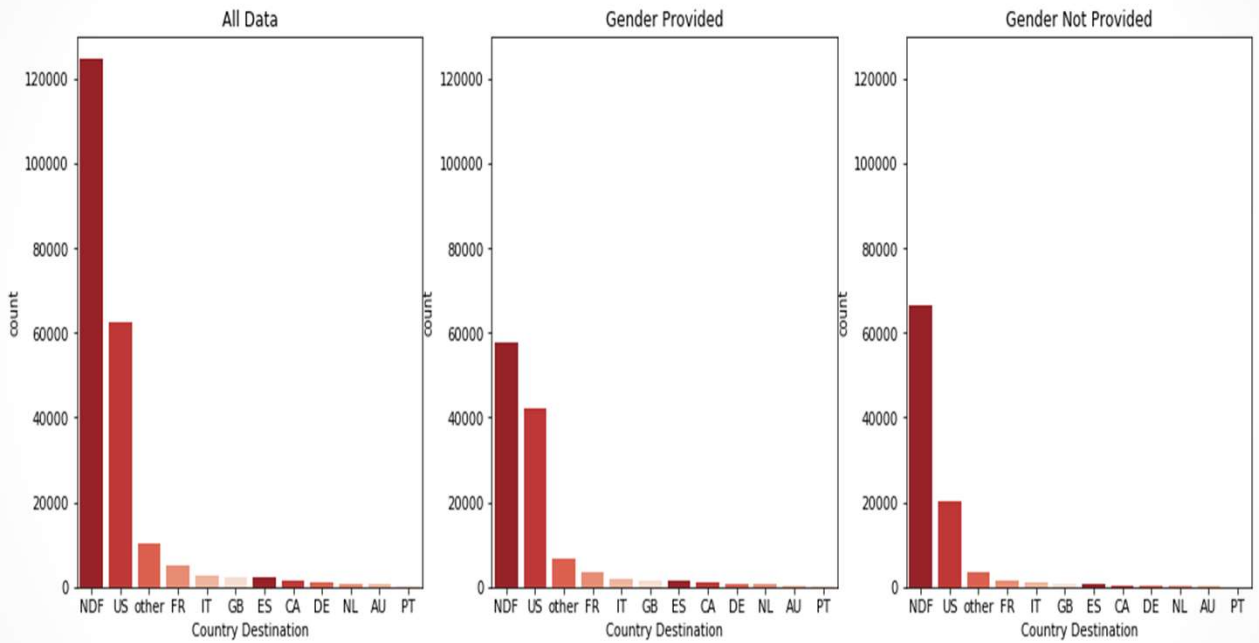
Distribution of Ages

- Age distribution shows most users clustered in 20-60 range
- Density Plot reveals small number of extreme outliers which are likely errant entries



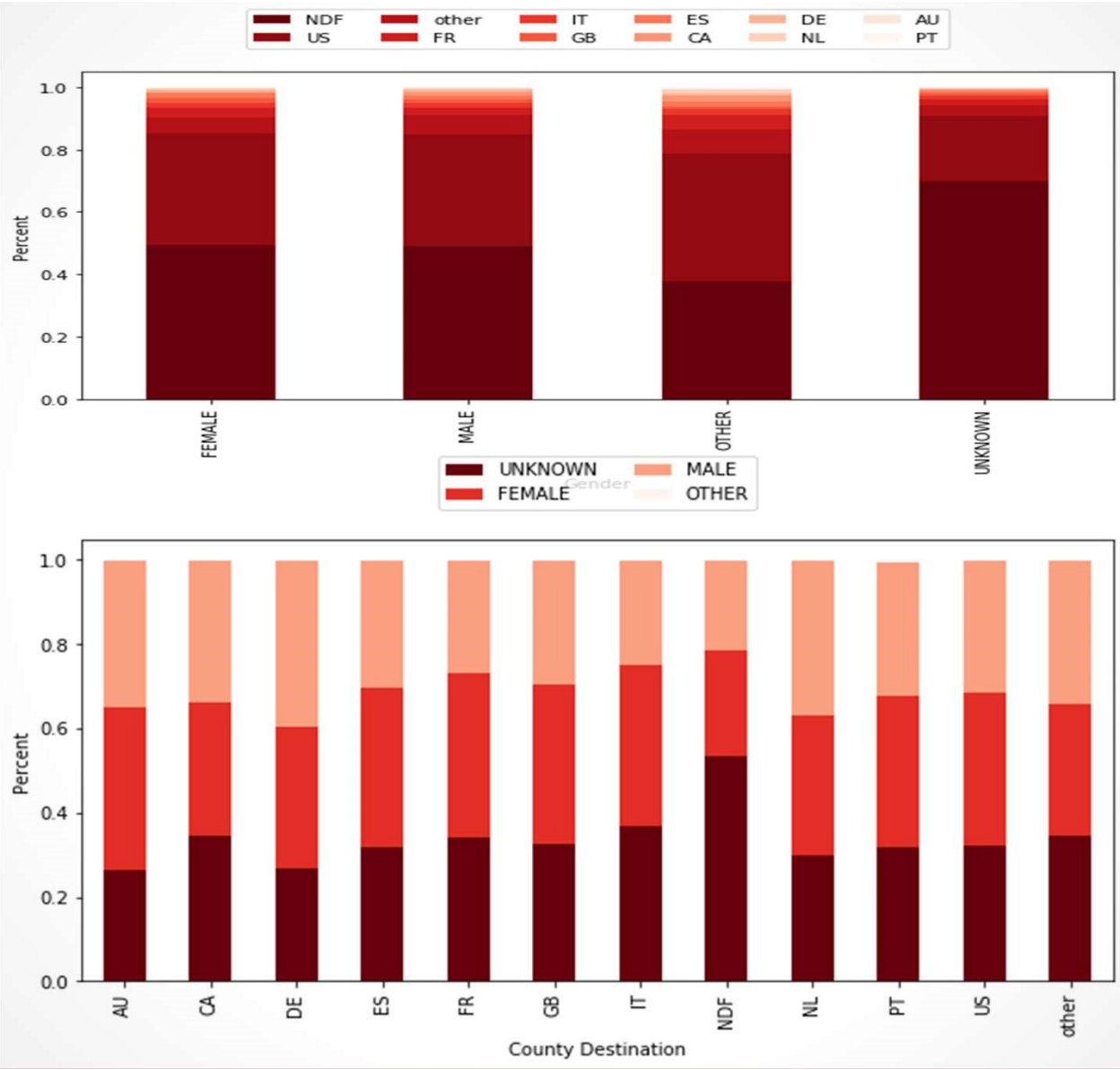
Bookings by Age Provided

- Age is an optional user-provided data point
- Users who supplied an age were significantly more likely to book a trip than those that did not (Destination of NDF indicates no trip booked)



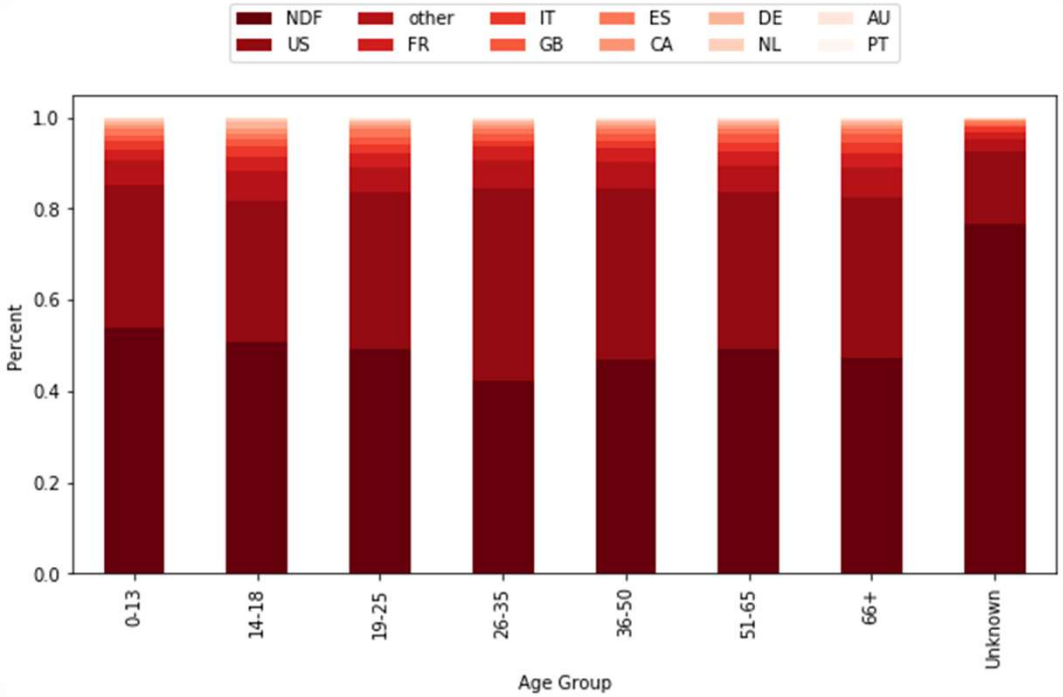
Bookings by Gender Provided

- Gender, like Age, is an optional value when signing up
- Users who supplied an gender were significantly more likely to book a trip than those that did not.
- Conclusion: Users who supplied more information about themselves during signup were far more likely to book.



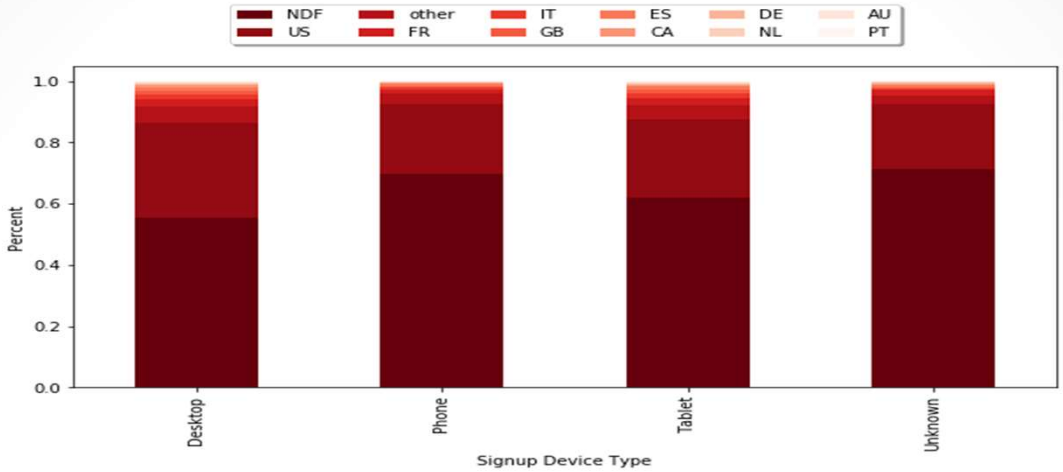
Destination By Gender

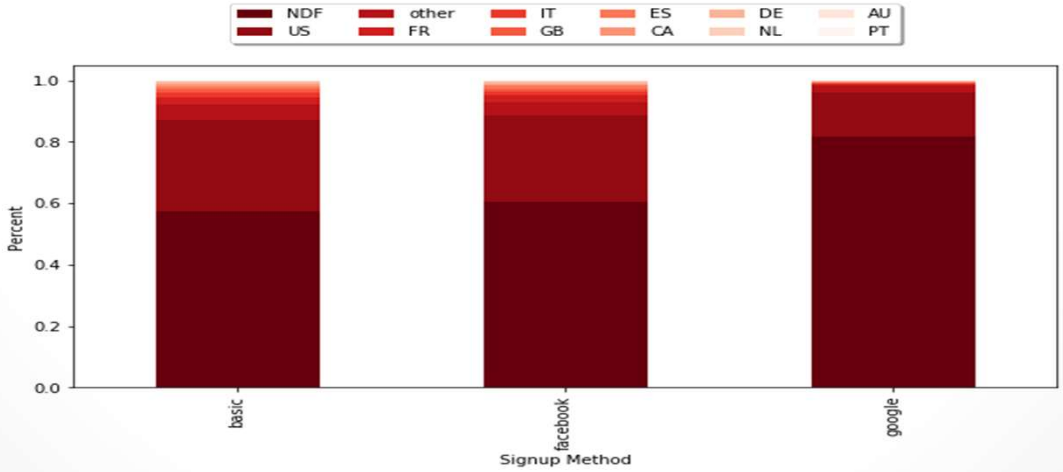
- Male and Female have very similar preferences for destination countries
- Other slightly more likely to book, but represent very small portion of the data
- Unsupplied gender associated with NDF



Destination By Age Group

- Decision to book or not is similar across age groups, though we see that users in the 26-35 age group were somewhat more likely than others to book a trip after signing up.
- Unsupplied Age associated with NDF





Destination By Signup Method

- Users who signed up using a PC or Tablet were significantly more likely to book a trip than phone signups.
- Users who signed up with their Google account were significantly less likely to book than those who signed in with Facebook or a direct Airbnb account

EDA - Conclusions

- **More data supplied → more likely to book**
  - Users who provided Gender or Age information during signup were significantly more likely to book.
  - Indicates users more serious about booking were more likely to supply this information.
  - Encouraging users to supply demographic info like Age and Gender during signup may lead to increased bookings following signup
- **Men and women have similar preferences across all age groups.**
  - Some minor differences. Men slightly more likely than women to book to 'other' destinations.
- **PC & Tablet Users → more likely to book**
  - Users who signed up on PCs or Tablets were far more likely to book a trip than phone users.
  - Phone users might be more inclined to browse and are less serious about booking.
  - Mobile web designers may want to find ways to encourage phone users to book.
    - Make mobile signup and booking simpler
    - Display listings in a way that encourages booking

Classification Analysis

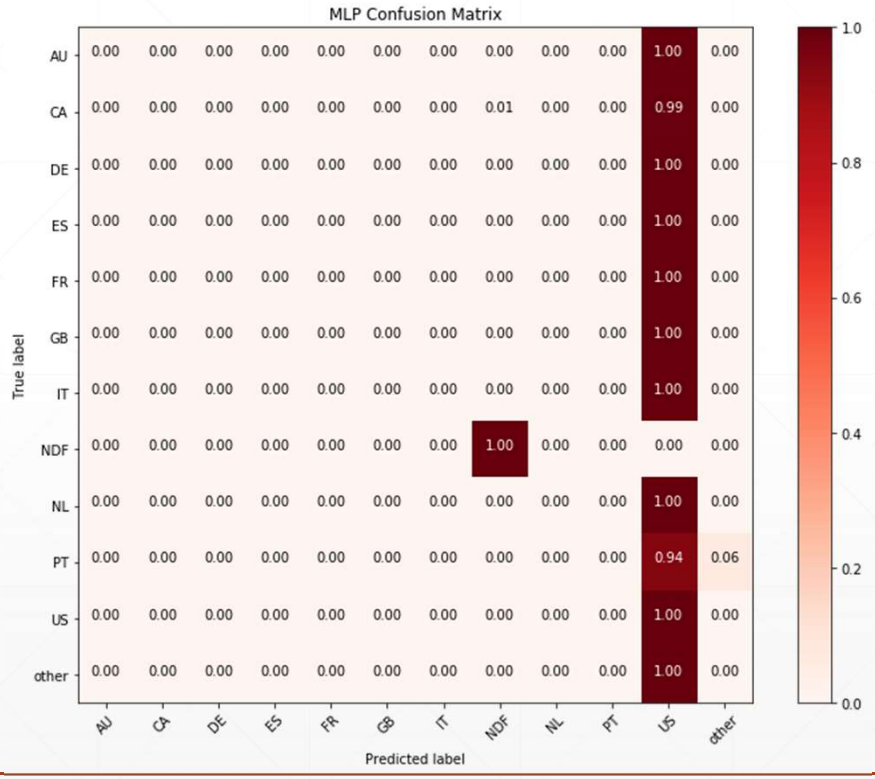
Twelve Category Classification

Results Summary

Classifier	Score	Avg. Precision	Avg. Recall
Decision Trees	0.8116	0.81	0.81
SVM	0.8808	0.80	0.83
Random Forest	0.8494	0.81	0.83
Gradient Booster	0.8808	0.80	0.83
Extra Trees	0.8345	0.81	0.82
Gaussian Naïve Bayes	0.8107	0.81	0.81
MLP Classifier	0.8808	0.80	0.88

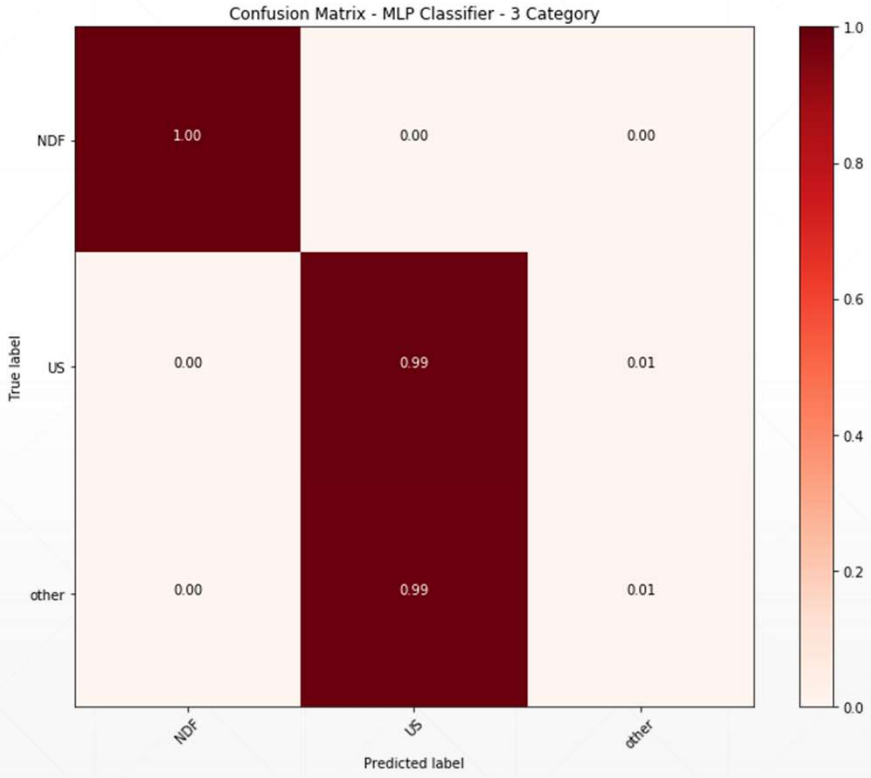
- All classification models attempted achieved reasonably high accuracy
- However, this is mostly due to the fact the data is largely concentrated in just 2 categories

Twelve Category Classification – MLP Classifier



- MLP Classifier achieved accuracy of > 88%
- But this was achieved by predicting US as the destination for any target other than NDF
- Observations between US and other destinations not distinct enough to classify accurately, or not enough data in less popular categories.

Three Category Classification



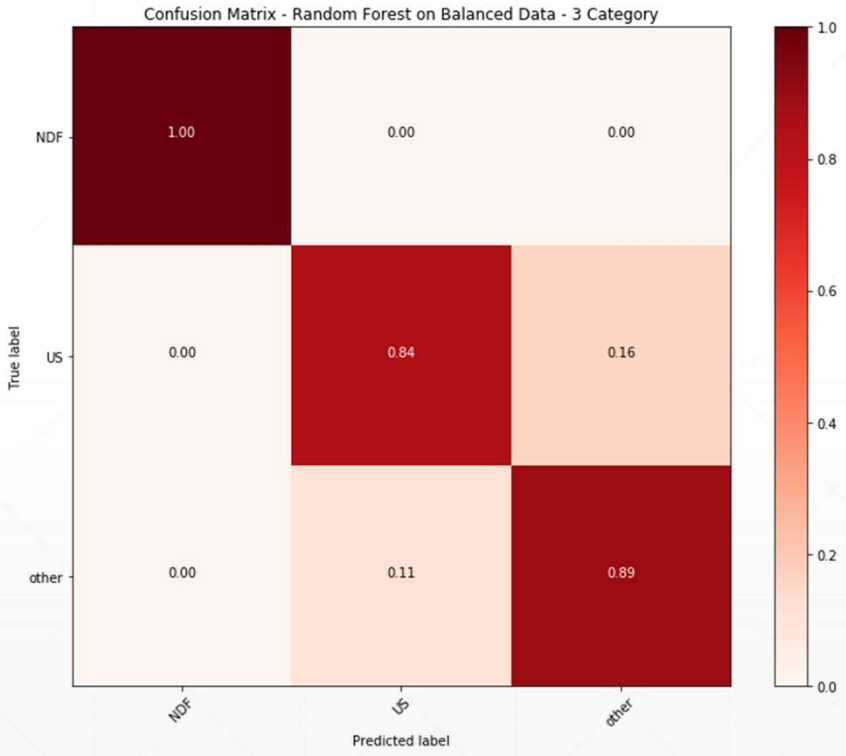
- Because 92% of the data was contained in three categories (NDF, US, other), reformatted the problem to a three-category classification.
- Country destinations other than US and NDF were labeled ‘other’
- Similar result to 12-category classification
  - All destinations that were not NDF were labeled ‘US’

Twelve Category Classification – Balanced Data

Classifier	Score	Avg. Precision	Avg. Recall
MLP	0.3696	0.34	0.34
Random Forest	0.8904	0.90	0.89

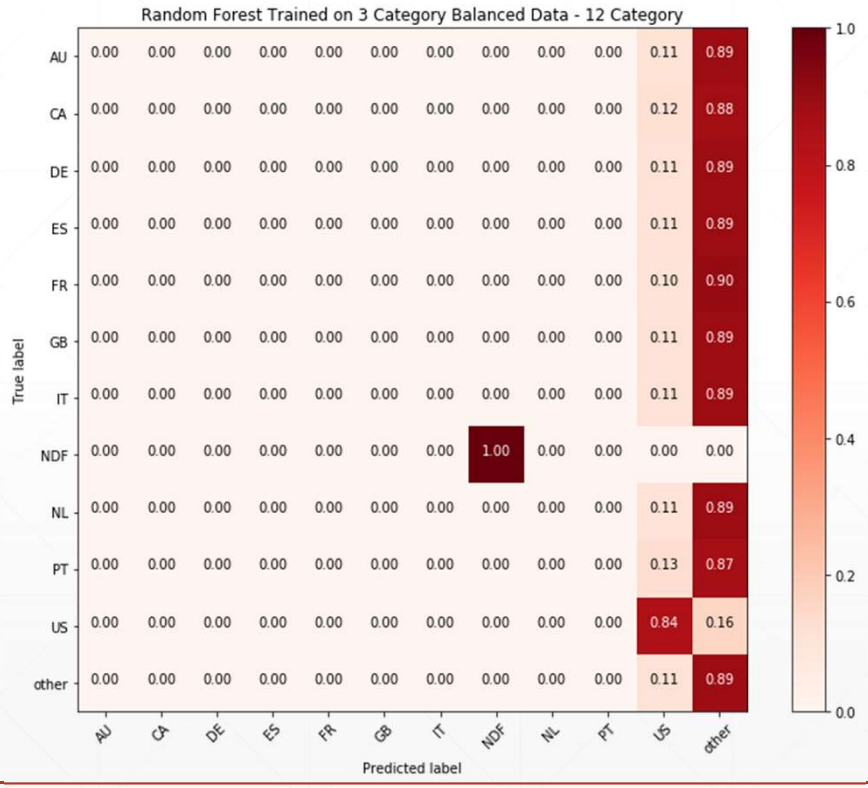
- NDF, US, and other comprise 92% of the data
- Less popular destinations like Netherlands are only 0.35%
- To supply model with additional data for less chosen destinations, created balanced dataset with equal number of observations for each destination
- Bootstrapped data (sampling with replacement) to create dataset with an even number of observations for each category.
- Random Forest showed improvement with this method, while MLP had large downgrade in accuracy.

Three Category Classification – Balanced Data



- Performing three-category classification on balanced data, model was able to produce much more accurate predictions.

Twelve Category Classification – Balanced Data



- When performing 12-category predictions on model trained with 3-category data, very accurate in identifying US and Other destinations.

Classification Analysis - Conclusions

- **Skewed Data → Skewed Results**
  - Models were able to attain high accuracy, but this was done by predicting only ‘US’ as the destination for observations that were not NDF
- **Available features not enough to accurately distinguish destinations**
  - Model can easily tell users who booked a trip vs. those who didn’t, but can’t accurately distinguish individual countries.
  - Additional features may increase model accuracy.
- **More Data → More Accurate Models**
  - Additional observations in less-popular categories may improve prediction accuracy
  - This can be observed by the improved accuracy in the models when using balanced data set with even number of observations in each category.