Airbnb
New User Bookings Project

Introduction

Research Objective

Airbnb's business covers 34,000+ cities across 190+ countries. Being able to accurately identify where new users are heading to is important as it allows Airbnb to recommend rentals that suit customers' needs and as a result, decrease the average time to first booking and improve the site's booking rate overall.

Research Question

Predict which country new users will book their first trip based on users' demographic data, web session records, and some summary statistics of different countries.

Data

The datasets for this project are available on Kaggle. They are:

Training & Testing data:

- Includes information related to Airbnb accounts such as when the user signed up/ made his/her first booking, sign-up flow, language preference, etc.

Country data:

- Includes geographic information of different countries

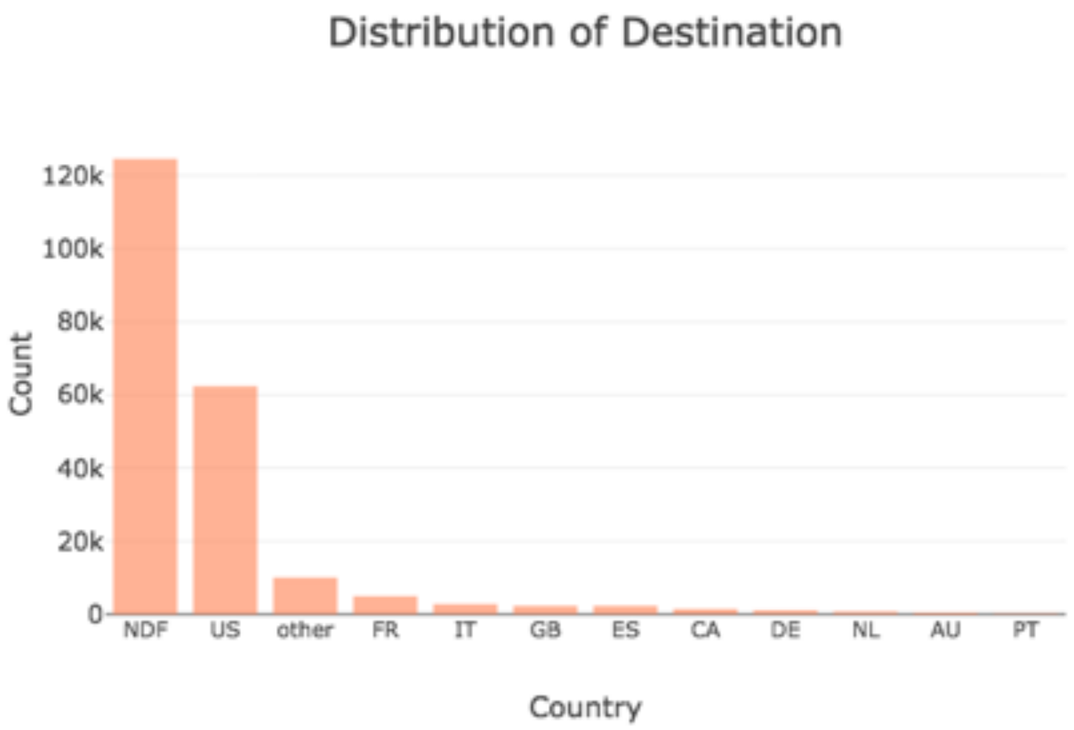
Session data:

- Users' web session log

Age Gender data:

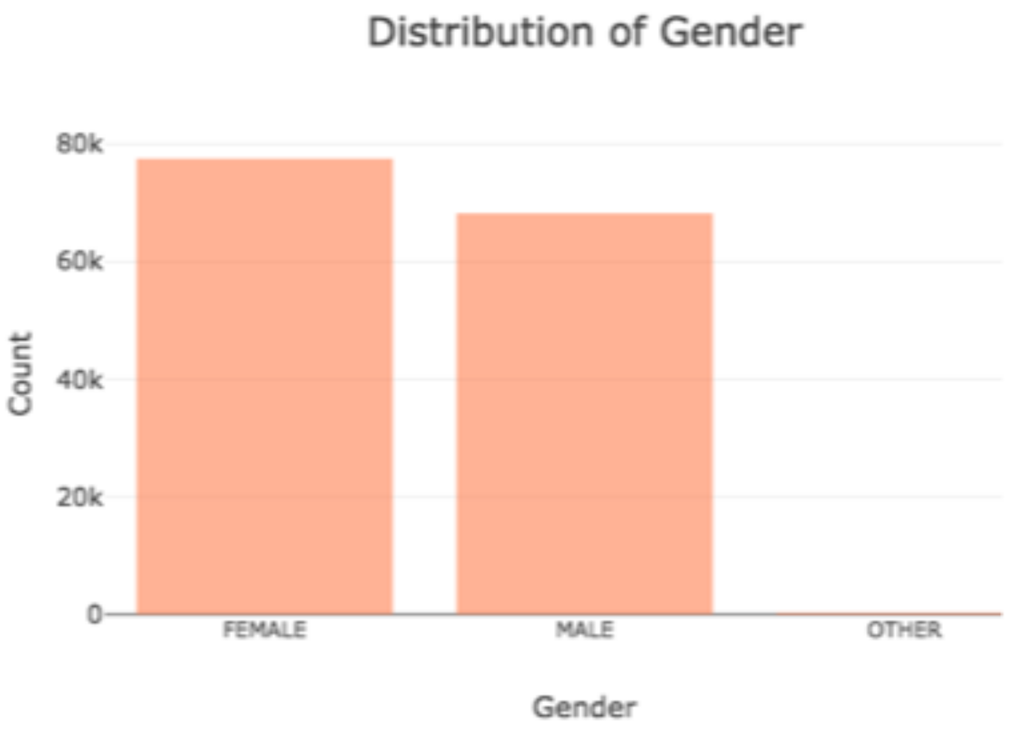
- Includes different countries' age/ gender splits

Exploratory Analysis



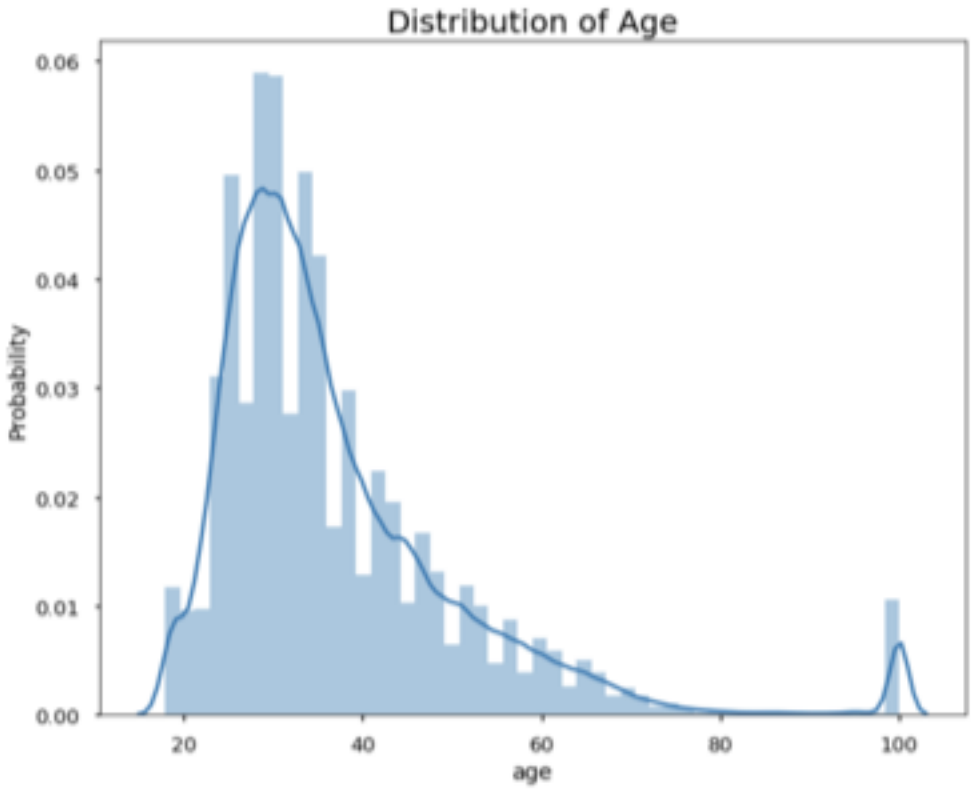
- Most users did not end up booking a trip; for those who ended up booking, most of them booked a trip to the U.S.

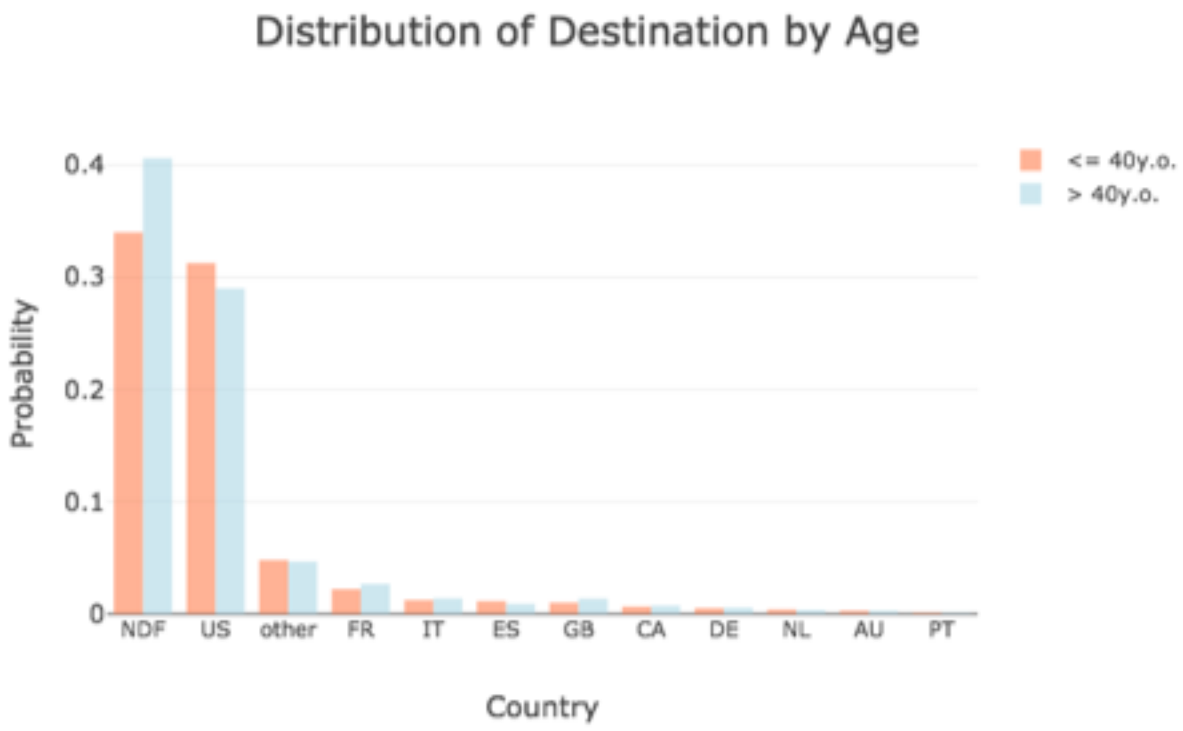
Exploratory Analysis



- We have slightly more female users than male. However, male and female users did not seem to show different preference when picking their first destination.

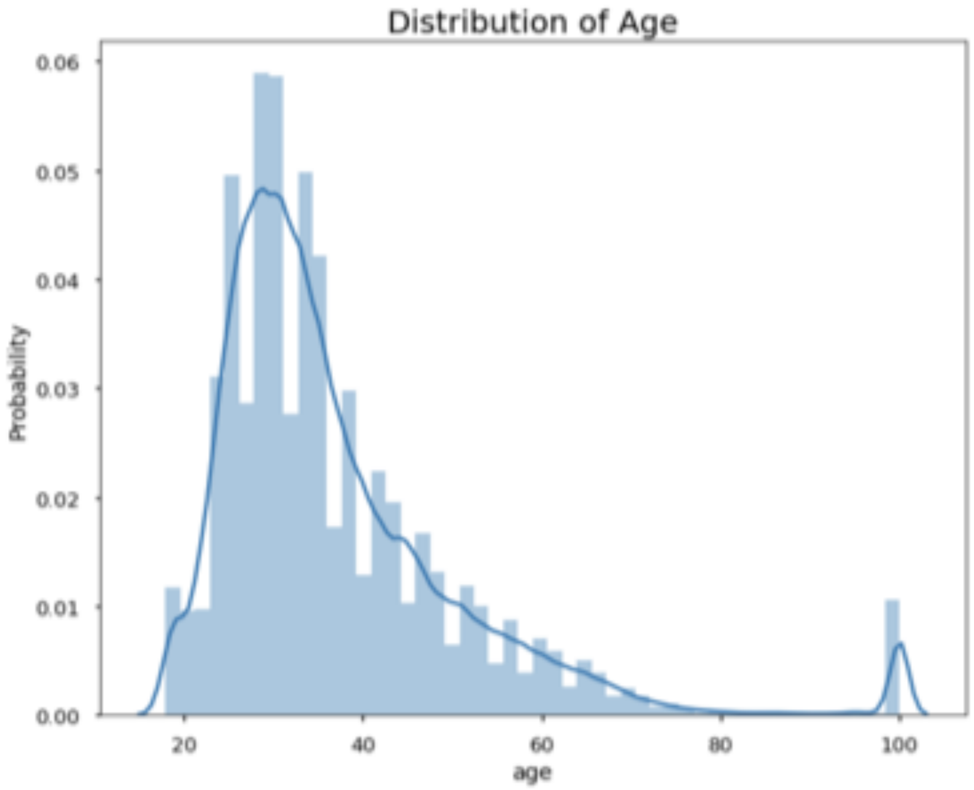
Exploratory Analysis

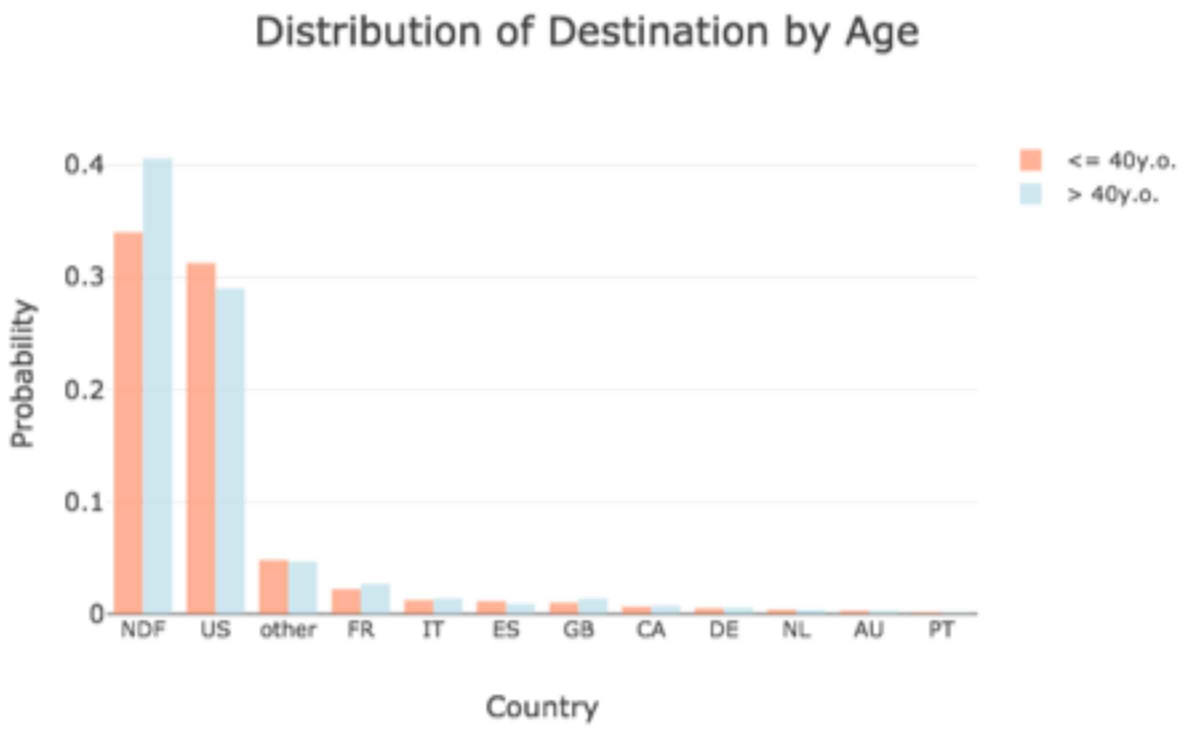




- The majority of Airbnb users are under 40. Younger users (<= 40 years old) had a higher probability of booking a trip (i.e. lower NDF) than those who are > 40 years old.
- A one-tail z test is conducted to test for this hypothesis and the p-value of the test is 7.20e-68. Since the p-value is very close to 0, we reject the null hypothesis and conclude that younger users had a higher probability of booking a trip.

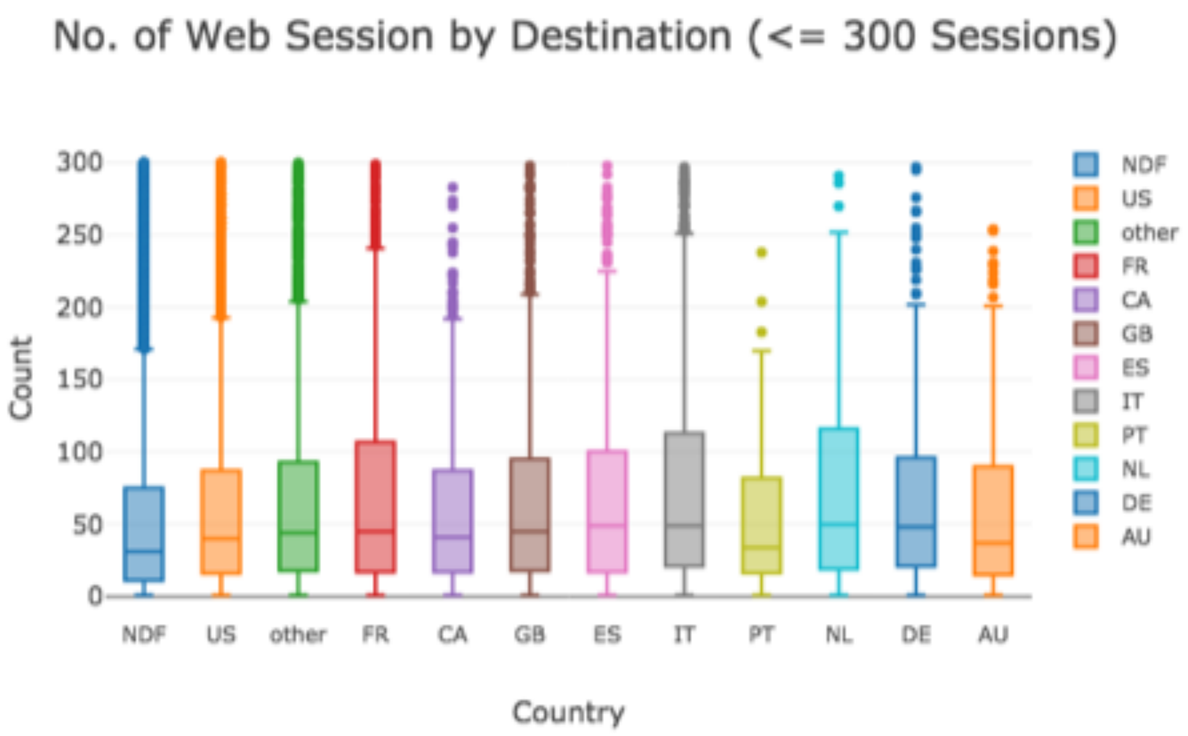
Exploratory Analysis





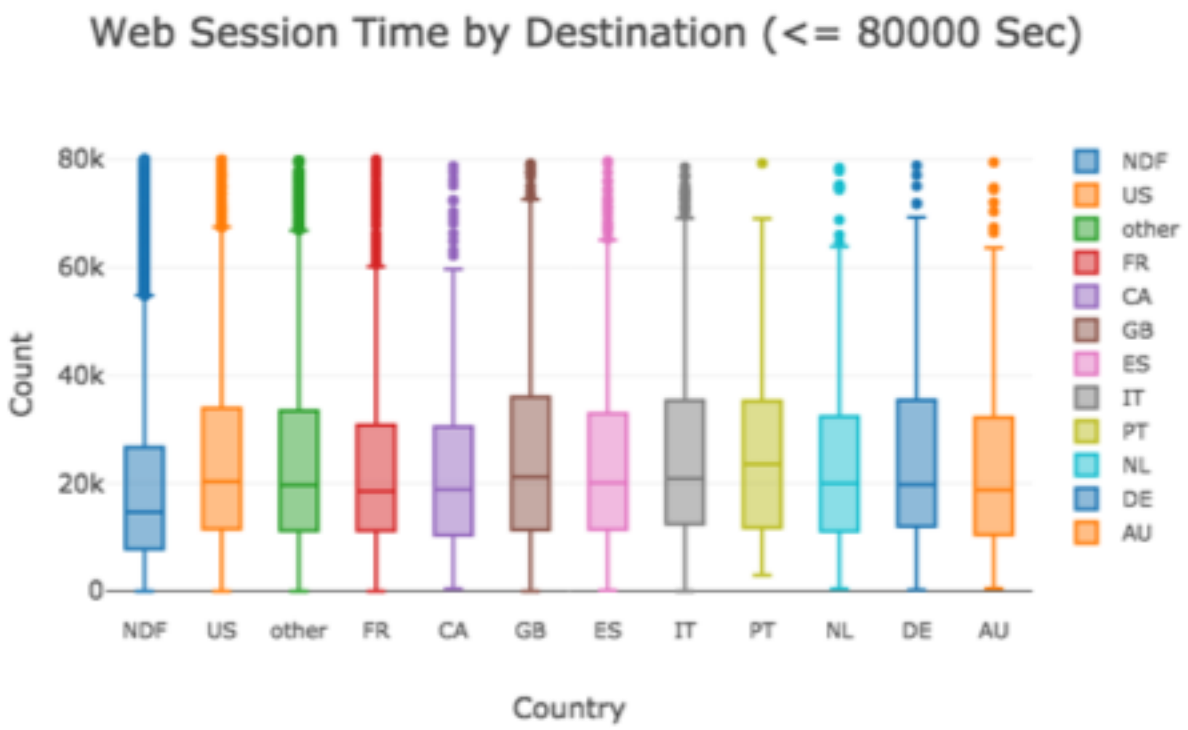
- Younger users were also more likely to pick U.S. as the destination of their first trip.
- Again, a one-tail z test is conducted and the p-value is 3.45e-66. We reject the null hypothesis because the p-value is very close to 0 and conclude that younger users had a higher probability of booking a trip to the U.S.

Exploratory Analysis



- Users who didn't book a trip seemed to visit Airbnb's website/ app less frequently.
- The p-value of our one-tail z test is 5.12e-104. Since the p-value is very close to 0, we reject the null hypothesis and conclude that those who booked a trip visited Airbnb's website/ app more frequently.

Exploratory Analysis



- Regarding the length of web sessions, those who didn't book a trip had shorter sessions.
- The p-value of our one-tail z test is 4.68e-34 and we conclude that those who booked a trip had longer web sessions .

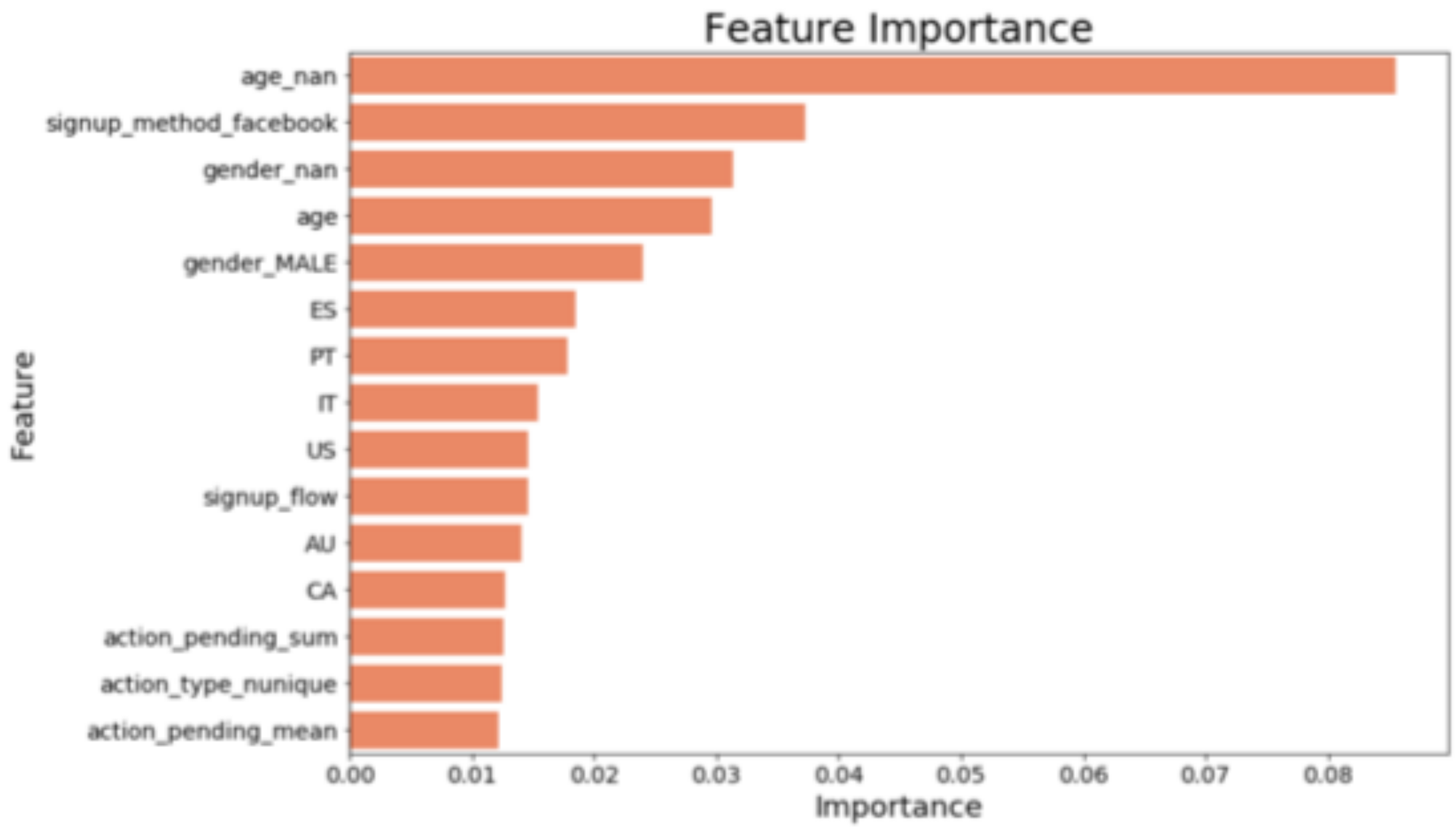
Machine Learning

- Multi-class classification
  - Outcome variable: Users' first destinations
  - Features: Users' demographic data, web session records, and some summary statistics of different countries
- The top 5 countries with the highest predicted probability are chosen as the predicted output for each user
- Predictions based on the test data are submitted to Kaggle and are evaluated using NDCG (Normalized discounted cumulative gain).

Model I – Random Forest

- Reduce overfitting by introducing randomness through bagging with different combinations of features
- Grid search with K-Fold cross validation (k=3) to determine the best parameters
- Best parameters:
  - max_features = ‘auto’
  - min_samples_split = 0.0005
- Best NDCG = 0.87521

Model I – Random Forest

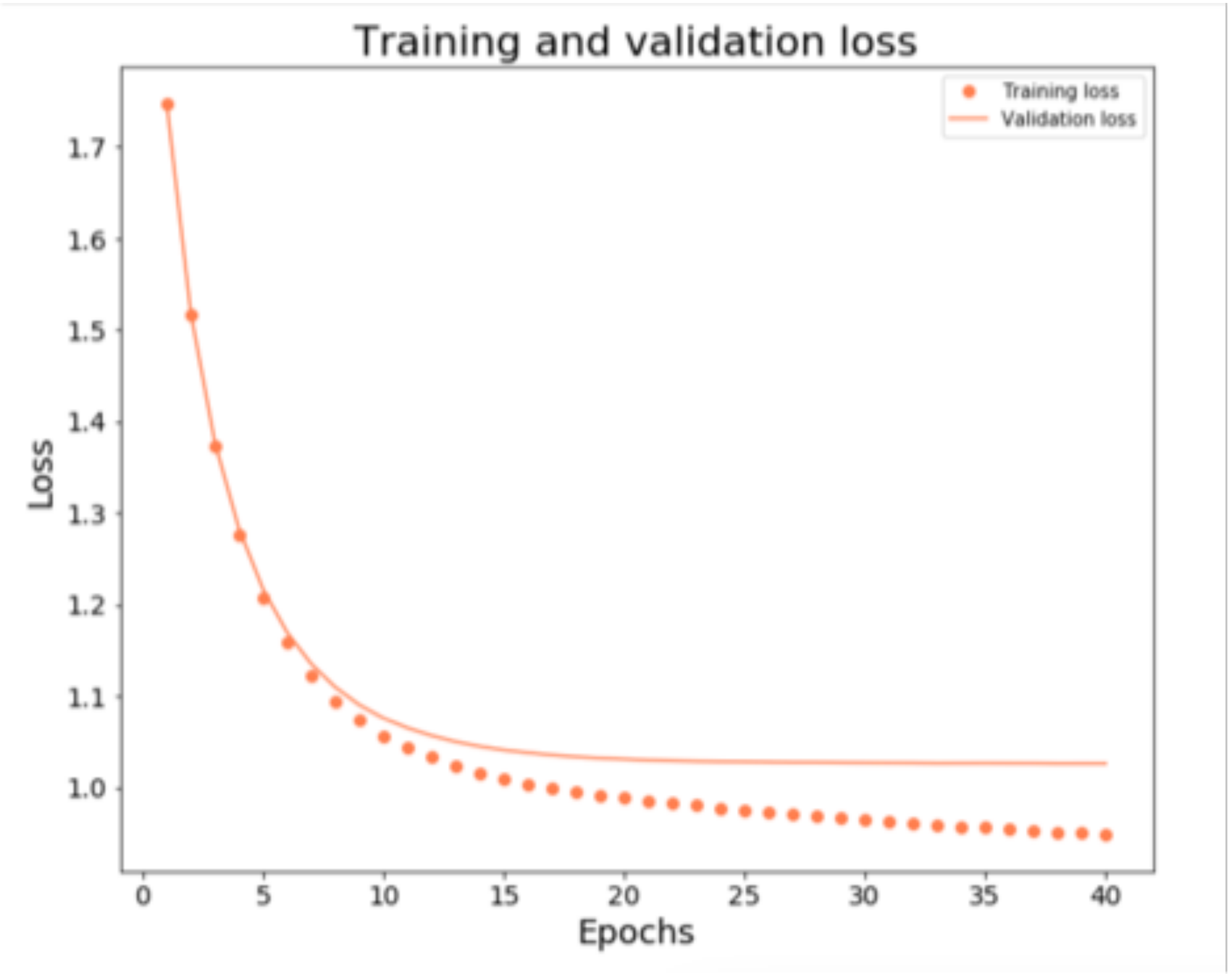


- Most important feature: whether the user leaves ‘age’ blank (which makes sense because those who did not intend to book a trip would probably leave ‘age’ blank)
- Other important features: whether the user signs up through Facebook, gender and age

Model II – XGBoost

- Outperforms other algorithms and is usually one of the winning solutions in Kaggle competitions
- XGBoost with number of rounds = [10, 20, 40] are run.
- Early stopping: training and validation loss is monitored, and the training process will stop if the validation loss has not improved for 5 rounds to avoid overfitting.

Model II – XGBoost



- Best model: number of round = 40
- Best NDCG: 0.87837 (although validation performance doesn't seem to improve much after 15 rounds)
- Early stopping didn't kick in because the validation loss was still decreasing slightly

Model III – Feedforward Neural Networks

- Neural Networks are flexible - by varying the number of layers and nodes, neural networks can fit different data with different complexity
- Features are normalized before feeding into the Neural Networks
- Feedforward Neural Networks with varying depths and number of nodes are tried
- Early stopping is implemented

Model III – Feedforward Neural Networks

- Batch size = 50
- No. of epoch = 20 although the model hits early stopping after 10 epochs
- Best model: 4 layers with number of nodes = [250, 125, 125, 12]
- Best NDCG = 0.87414

Model IV – Ensemble Model (Soft Voting)

- Weighted mean of the predicted probabilities from the best Random Forest, XGBoost and Neural Network model
- The top 5 countries with the highest weighted mean are chosen as the predicted output for each user
- Different combinations of weights are tried
- Best model: Weight of [Random Forest, XGBoost, Neural Networks] = [1, 2, 1]
- Best NDCG = 0.8780.

Final Model

Model	NDCG
Random Forest	0.87521
XGBoost	0.87837
Neural Networks	0.87414
Ensemble	0.87800

- My final model is XGBoost Model with 40 rounds of training with a NDCG score of 0.87837. It ranks 325 out of 1462 on Kaggle Leaderboard.

Future Research Recommendation

- More hyperparameter tuning:
  - Try to tune more hyperparameters to optimize my models if given more time and resources

- Try more classification algorithms:
  - Try more algorithms such as AdaBoost, CATBoost, etc.

Thank you!