Capstone Report - Airbnb New User Bookings

Problem Statement:

Predict in which country a new user will make his or her first booking. The evaluation metric is NDCG (Normalized discounted cumulative gain) at k = 5. In other words, making a maximum of 5 predictions on the country of the first booking at the used id level

Benefits

Airbnb Leadership

Model will allow Airbnb to perform targeted marketing to new users by offering clients competitive listings to drive conversion

Aribnb Clients

With more relevant listings and content, clients will have a better user experience by shortening their time on researching possible destinations. This will help overall client conversion on bookings.

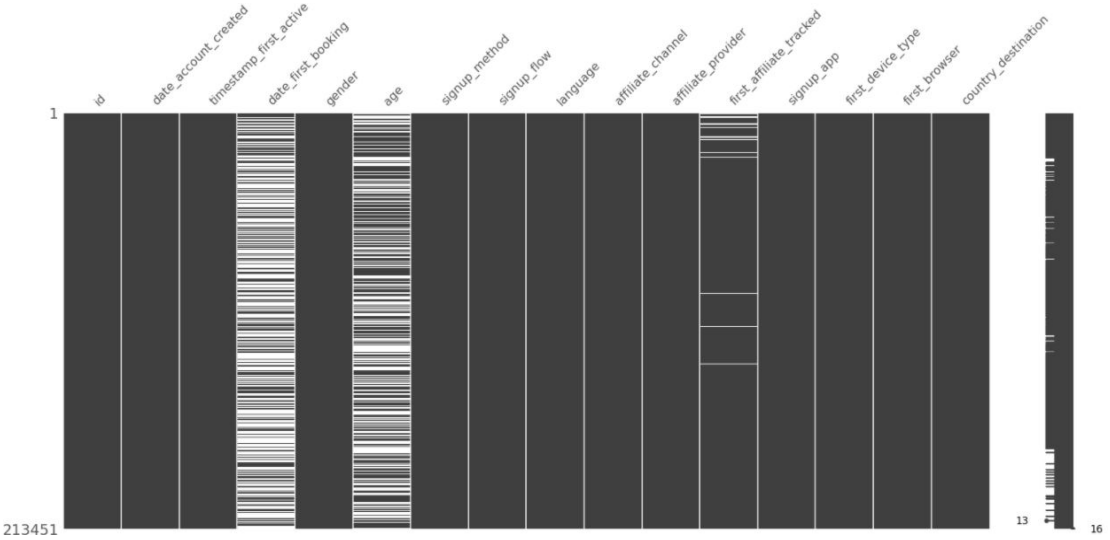
Dataset Description

Obtained from Kaggle, data are in the form of 6 dataset

- train: includes 4 datetime variables, 1 continuous variable, and 10 categorical variables. Target variable is column 'country_destination' Size: (213451, 16)
- test: same as train set, without 'country_destination' column. Size: (62096, 15)
- session: web session log for users, 5 categorical variables, 1 continuous variable. Size: (10567737, 6)
- summary statistics on age, gender and countries
- Data Dictionary

Cleaning and Wrangling

- Missing data.



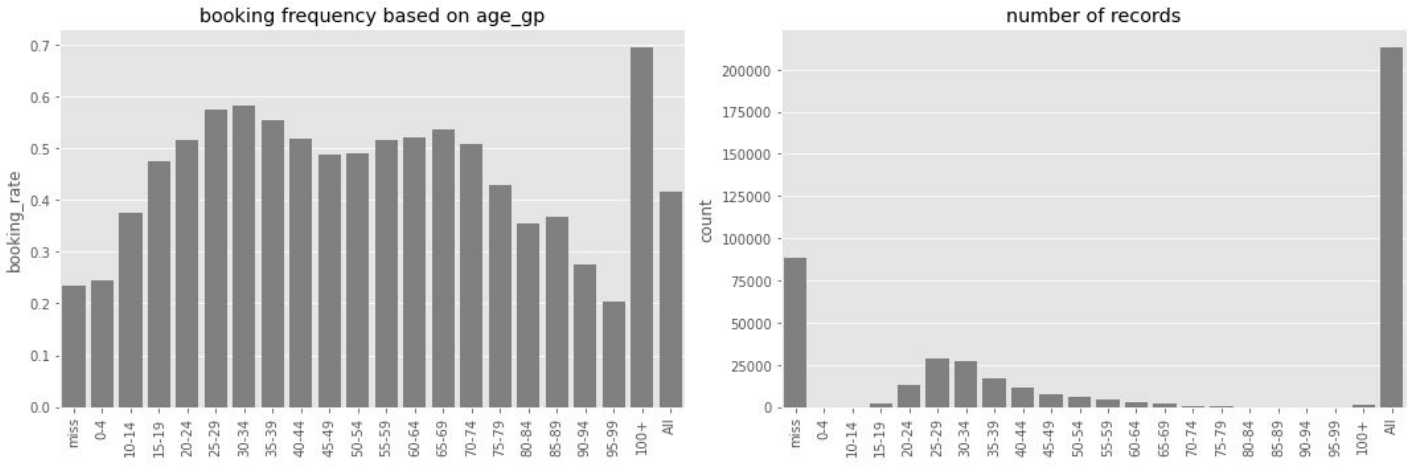
- Cleaning steps taken include:
    - converted all date/time related columns into DateTime data type
    - converted age into categorical variable through binning, assigned “missing” as a group
    - filling missing values under first_affiliate_tracked columns with “miss”

Exploratory Analysis

The goal is to discover features that may influence the likelihood of booking. Also, remove features that have high collinearity.

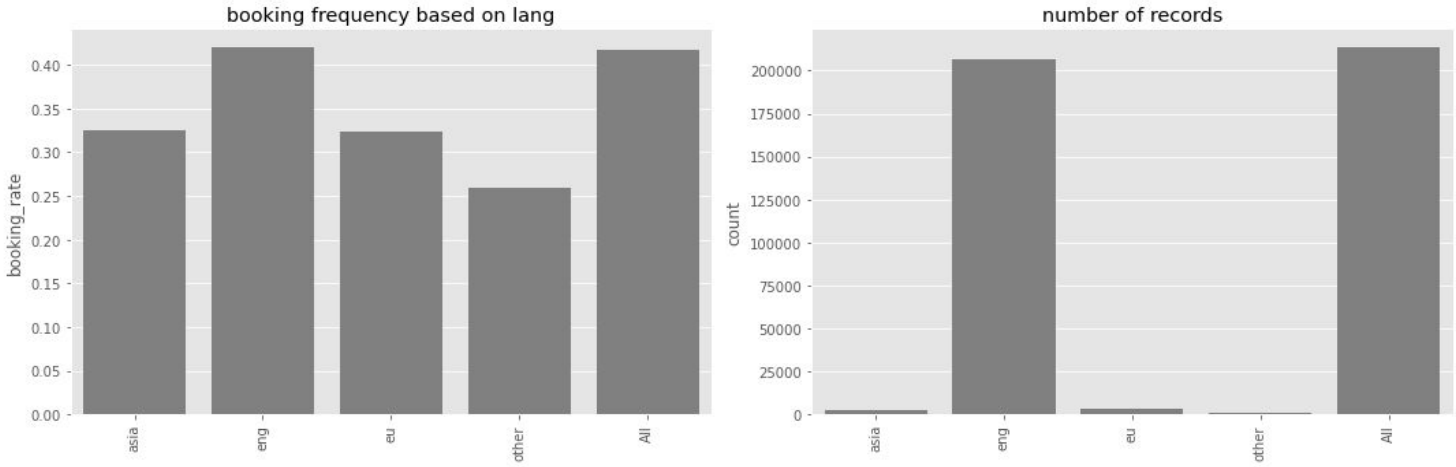
Findings:

Age Group



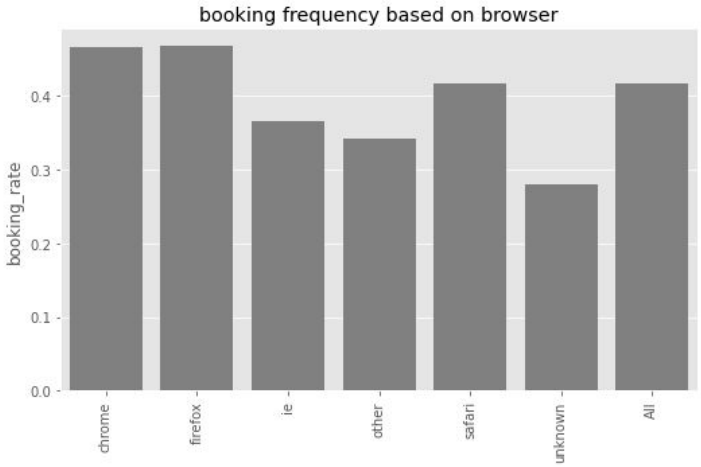
- Age group between 25 to 35 are the is the mainstream user groups.

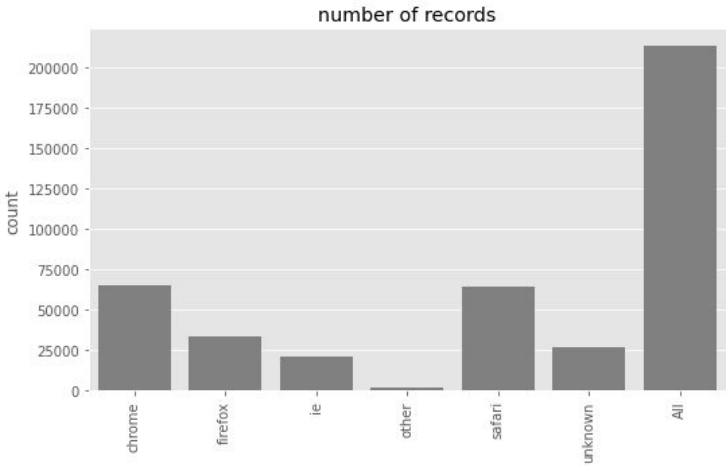
Language



- English is the main language used on the platform. Other languages appear to have lower booking rate

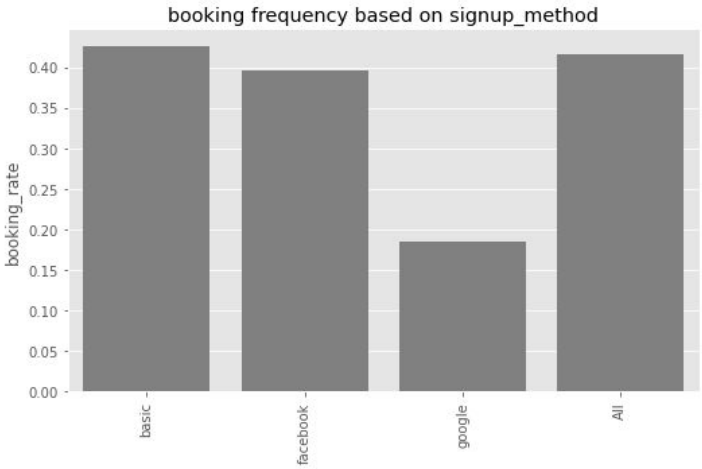
Browser

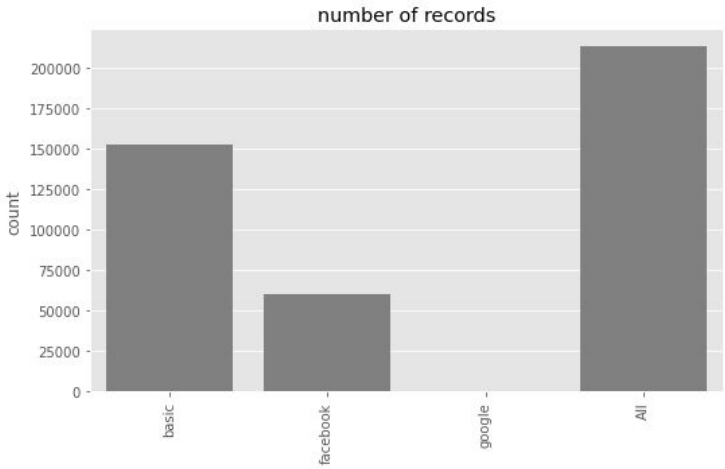




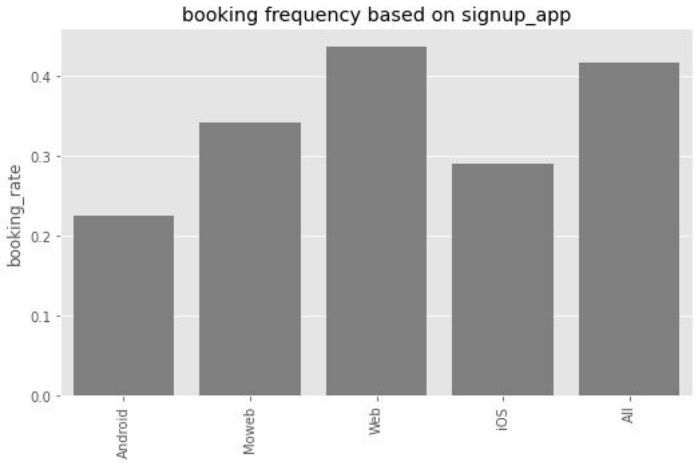
- Chrome and FireFox appear to have a higher booking rate than IE and Safari.

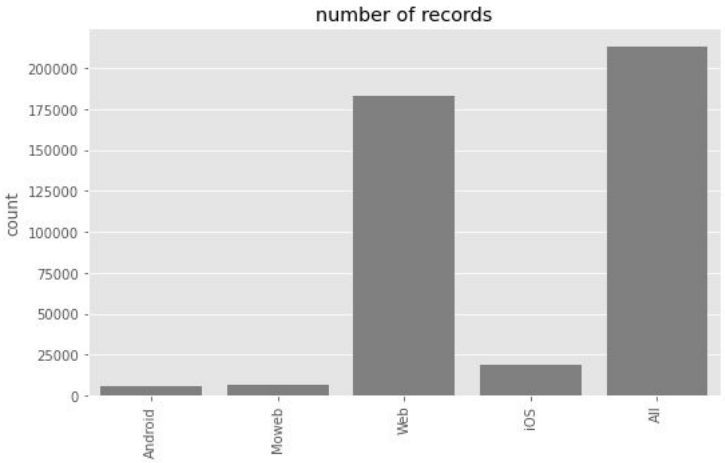
Sign up method, app, and flow



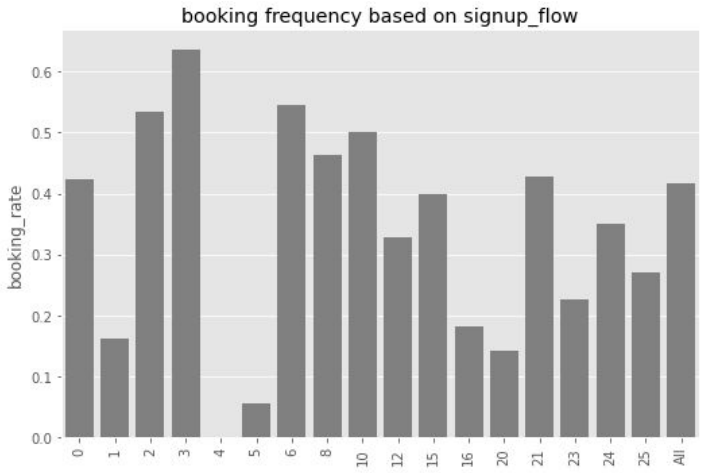


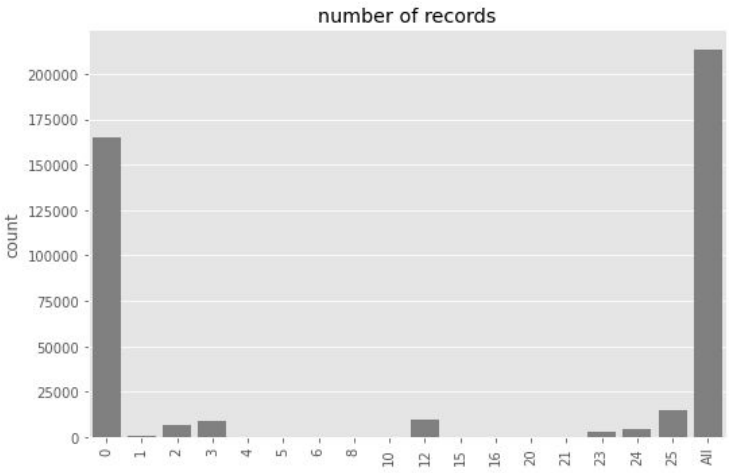
- Google signup method has the lowest booking rate





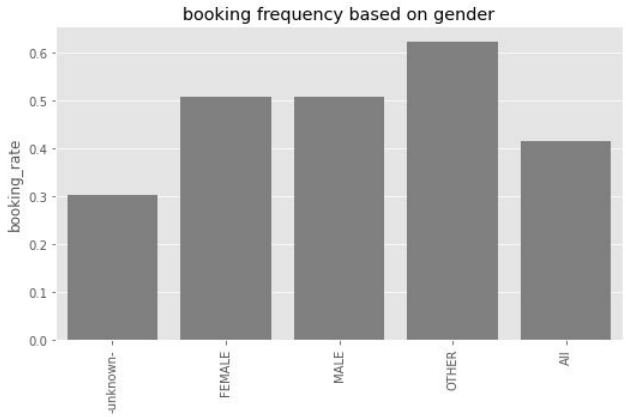
- Web has the highest booking rates, followed by iOS

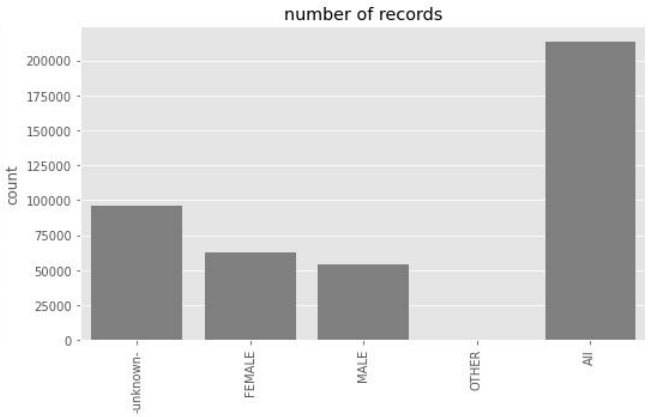


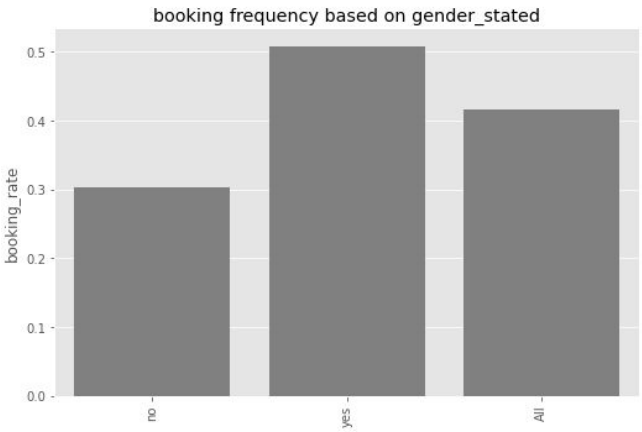


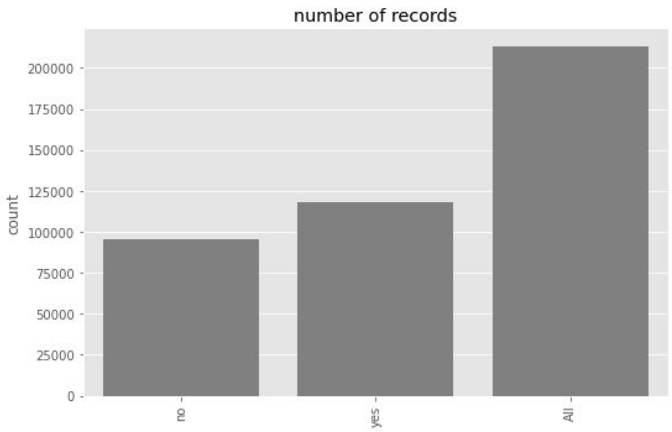
- The Highest booking rate is on page visit 0-3, 12 or above 23. 0 pages users probably already know what they want and that session just for booking. If they browse more than 20 pages, which could mean that they are really serious above making a booking.

Gender









- Booking rate is almost even between male and female
- There is a big difference in the booking rate on whether gender is stated

Statistical Test

All the above features passed chi-square test. Two more features, affiliated provider, and action type under session dataset, also passed chi-square test indicating significance on influencing booking rates

Summary

- Almost all variables being tested had some influence on the booking rate.
- Drop affiliate provider as it is highly predictive by the affiliate channel
- The combination of actions and action details are likely to have predictive power to booking, which we leverage machine learning algorithms to test.
- New features created are age group, language group, and browser category.
- Focus on last action in the user's browsing sequence
- Target variable will need label encoding and label binarizer transformation

Machine Learning

Strategy

- Apply feature selection method to narrow down numbers of binary variables from action details and type. Use LassoCV, followed by Recursive Feature Elimination (RFE) using Logistics Regression and Random Forest as regressors.
- Combine features set as identify from exploratory analysis with the action details binary variables
- Compare base model performance across five types of classification models (Logistics Regression, Extra Tree, Random Forecast, Gradient Boosting, and Lightgbm). Pick top three models for hyperparameters tunning
- Apply deep learning method using Keras to identify potential performance improvement

Part 1: Feature Extraction/Selection

- Applied one-hot encoding to all categorical variables, a total of 359 features are formed
- Applied LassoCV, obtained 183 features that have non-zero coefficients
- Applied RFE using Lightgbm and Xgboost estimators, use voting method to obtain 80 features
- Final dataframe has 124 features

Part 2: Compare model performance and hyperparameter tuning

- Applied 4 fold cross-validation, optimize for ndcg score, identified tree base model (i.e. Random Forecast and Extra trees) have the best performance, followed by Lightbgm
- Applied Bayesian optimization approach to identify the best parameter that lead to the highest ndcg score. Results of the tuned models on the train set as follows:

Models	ndcg score
Extra Tree	0.822
Random Forecast	0.806
Lightgbm	0.824

Part 3: Deep Learning Model

- Using Keras in Tensorflow to develop a one hidden layer neural network model, achieved 0.8251 ndcg score

Model Performance (ndcg on Kaggle)

Model	ndcg score	Time
dummy	0.6609	
extra tree	0.85212	1m 42s
extra tree (tuned)	0.86675	8m 49s
random forest	0.85788	1m 7s
random forest (tuned)	0.85359	1m
lightgbm	0.86849	22s
lightgbm (tuned)	0.86769	41s
deep learning (Keras)	0.86891	44s

Recommendation/Next Steps

Lightgbm (no tunning) performed better than the tree-based models and is recommended as the winning model.

Target variable is skewed toward US (70%). Perhaps analysis could be done at a more micro level within US to make the predictions more precise – need extra data points from Airbnb.

Keras deep learning model had a slight improvement over the lightgbm with a rather simple set up of one hidden layer. This suggests a further improvement in ndcg is possible should clients preferred to investigate further. However, the computation will get expensive.

More features could be generated under the session activities data since this study only considered the last action details. Robust evaluation is needed to determine what features to choose, as the dataset could become complex and result in an overfitting problem.