Project Report: Airbnb New User Bookings

**Introduction**

New users on Airbnb are able to book accommodation in thousands of cities and hundreds of countries. Using machine learning, we can predict which country a new user is most likely to book their first travel experience. If Airbnb is able to accurately predict where a new user will make their first booking, the company can reduce average time to first booking, market opportunities more effectively, and improve demand forecasting.

This project uses machine learning algorithms to predict the top 5 countries where a user will make their first booking on Airbnb.  The project is a 12-class classification problem and uses Normalized Discounted Cumulative Gain (NDCG) as its evaluation metric.

**Dataset Description and Data Cleaning**

The data is provided by Airbnb as part of a [Kaggle competition](https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings), and can be accessed [here](https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data).  The dataset is composed of user data for 213,451 unique Airbnb new users, each with 15 feature variables that can be used as input to the Machine Learning models. Data includes datetime values (e.g. date account created and timestamp first active), categorical values (e.g. gender, signup method, language, affiliate channel, affiliate provider, signup app, first device type, first browser) and numeric values (age).

The following steps were taken to clean and prepare the data for machine learning:

1. Convert nonstandardized dates from string format into datetime format using [pd.to\_datetime](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.to_datetime.html)
2. Remove unreasonable age data. Airbnb users must be at least 18 years old and are not likely to be over 100 years old. Age above or below these thresholds are converted to NaNs
3. Age data is categorized into 5 buckets, each bucket containing a similar number of users. Splits are based on [quintiles](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.quantile.html). NaN age data are placed into their own bucket for unknowns
4. Given that some machine learning algorithms [require numeric data as input](https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/), non-numeric values are converted into numeric values through [one-hot encoding](https://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html)
5. Convert target labels into numeric form through a [LabelEncoder](http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html)

**Exploratory Data Analysis**

Jupyter notebook for Exploratory Data Analysis and Initial Findings can be accessed [here](https://github.com/jon-lo/airbnb_new_user_bookings/blob/master/airbnb-exploratory-analysis.ipynb).

**Machine Learning Model**

This model employs a Gradient Boosting Classifier and is able to predict the top 5 countries a user is most likely to make his/her first booking. The model generates a list of users and the top 5 countries for each user (‘submission.csv’). When evaluating the predicted countries to the actual countries, an NDCG score of **0.865** is achieved.