# proposal

June 15, 2017

## 1 Machine Learning Engineer Nanodegree

### 1.1 Capstone Project Proposal

#### 1.1.1 Airbnb New User Booking Dataset

Rohan Verma
June 15th, 2017

#### 1.1.2 Project Overview

In this project, I plan to use Machine Learning Techniques to predict in which country a new user will make their first booking using the Airbnb New User Bookings. This project will involve data cleaning, data exploration using visualizations, and testing various algorithms for classification for the same.

#### 1.1.3 Problem Statement

Airbnb, which is a online marketplace where people list, discover, and book accomodations around the world. It has collected various datapoints about users. This data about the patterns of its present user base can be utilized to predict patterns about its future users to provide them with customized suggestions to serve Airbnb's customers better.

Using this data, the challenge is to predict the destination of choice for the users' first booking.

#### 1.1.4 Datasets and Inputs

The dataset is composed of 5 CSV files. It has been obtained from a Kaggle Competition provided by Airbnb. [link]

The most important file is the `train_users` file which has 16 columns containing user id, dates of account creation, first booking date, gender, age, signup method, signup app, destination etc along with the target variable `country_destination`. The `test_users` is similar to the previous file discussed but does not have our target variable and we have to use these to predict the destination.

The other three files contain web session logs (`sessions.csv`) for the users, summary statistics of destination countries (`countries`) and summary statistics of about the users age group, gender, etc. (`age_gender_bkts.csv`)

### 1.1.5 Solution Statement

The solution will largely utilize similarites in user behaviour considering people having similar demographics are likely to perform similar actions. This will be helpful for us to test supervised learning models to predict the behavious of new users. I will then test various models we have learned in this course along with techniques such Grid-SearchCV to optimize and other models such as XGBoost which are used effectively in competitive environments such Kaggle.

### 1.1.6 Metrics

Since this is a Kaggle Challenge, we already have an evaluation metric, that is the NDCG (Normalized Discounted Cumulative Gain)

For each new user, we are to make a maximum of 5 predictions on the country of the first booking. The ground truth country is marked with relevance = 1, while the rest have relevance = 0.

$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{\log_2(i+1)}$

$nDCG_k = \frac{DCG_k}{IDCG_k}$

where $rel_i$ is the relevance of the result at position $i$ and $k = 5$.

For example, if for a particular user the destination is FR, then the predictions become:

[ FR ] gives a $NDCG = \frac{2^1-1}{log_2(1+1)} = 1.0$

[ US, FR ] gives a $DCG = \frac{2^0-1}{log_2(1+1)} + \frac{2^1-1}{log_2(2+1)} = \frac{1}{1.58496} = 0.6309$

To determine a baseline benchmark, we will find the metric value obtained by predicting the 5 most common outcomes [NDF, US, OTHER, FR, IT] against the train and test datasets.

### 1.1.7 Project Design

The project will be composed of the following steps:

- Data Exploration: Visualizing the dataset to gain insight into patterns and decide upon relevant features and improving the quality of the dataset.
- Training: Train models by considering different supervised models using techniques such as cross validation and optimizing using GridSearchCV.
- Testing and Optimizing: Using the trained models to test on Kaggle.