

# Machine Learning Engineer Nanodegree

## Capstone Project Proposal

### Airbnb New User Booking Dataset

Saumitra Rawat

Jan 3rd, 2019

### Domain Background

Airbnb, which is an online marketplace where people list, discover, and book accommodations around the world. It has collected various datapoints about users. This data about the usage 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. Airbnb had posted this on Kaggle as a Recruitment Challenge. Using user data effectively can help organizations increase metrics such as sales, user experience, customer retention and customer satisfaction. Machine Learning techniques can help organizations attain useful predictions using these data. The motivation for pursuing this project is to understand how to work on real world datasets and challenges that companies like Airbnb consider to be important and valuable for their companies and learn to provide similar value for organizations that I work with in the future.

### Problem Statement

By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

Using the data from Airbnb New User Bookings (<https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings>) dataset, the challenge is to predict the destination of choice for the users' first booking. This data includes demographics of users and their session data. The model will utilize these demographics and session data to make models that can predict the destinations.

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

## Datasets and Inputs

The dataset is composed of 5 CSV files. It has been obtained from a Kaggle Competition provided by Airbnb. [\[link\]\(https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data\)](https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data).

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` and has 213451 rows. 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 and has 62096 rows. We have a good amount of data to work with to produce meaningful models.

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`)

## Solution Statement

The solution will largely utilize the fact that similarities in user demographics is likely to be correlated to the choices made by the users on the platform. This will be helpful for us to test supervised learning models to predict the behaviour of new users. I will use the first 15 columns of the `users` data as input to these models and the `country_destination` as the target.

I will then test various models such as SVM, Decision Trees, Random Forest etc. 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.

## Benchmark Model

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.

Along with that, the goal will be to place our final model in the top 20% of the leaderboard on Kaggle using the evaluation metric describe below

## Evaluation 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:

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

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

## Project Design

The project will be composed of the following steps:

- *Data Exploration*: Visualizing the dataset, detect outliers, remove null values, cleaning the dataset, check relevance of every column to the target column, cluster the dataset using unsupervised techniques to see if we can engineer new features, splitting training dataset into training and testing sets etc.
- *Training*: Consider multiple supervised ML models and select the best one, use techniques such as cross validation, and optimizing using GridSearchCV for hyperparameter optimization.
- *Testing and Optimizing*: Optimizing the model offline, using the trained models to test on Kaggle and improve the rank on kaggle.