# Machine Learning Engineer Nanodegree

## Capstone Proposal

Vladimir Baranov December 17th, 2017

## Proposal

### Domain Background

Instead of waking to overlooked "Do not disturb" signs, Airbnb travelers find themselves rising with the birds in a whimsical treehouse, having their morning coffee on the deck of a houseboat, or cooking a shared regional breakfast with their hosts.

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. 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.

The aim of this project is to predict in which country a new user will make his/her first booking. This project is a Kaggle competition that was run by Airbnb from Nov 2015 to Feb 2016 [1].

### Problem Statement

In this project we are going to predict which country will be the first destionation of a new user given a dataset containig information of a user, his activities on airbnb web site, summary statistics of a user and destination countries.

### Datasets and Inputs

In this project, we are given a list of users along with their demographics, web session records, and some summary statistics. All the users in this dataset are from the USA. There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. 'NDF' is different from 'other' because 'other' means there was a booking, but is to a country not included in the list, while 'NDF' means there wasn't a booking.

The training and test sets are split by dates. In the test set, we will predict all the new users with first activities after 7/1/2014. In the sessions dataset, the data only dates back to 1/1/2014, while the users dataset dates back to 2010.

The datasets were obtain from kaggle [1].

### Solution Statement

To tackle the problem described in Section 2, we will use Supervised Machine Learning. We will find up to five most probable solutions to this problem. We believe that the most suitable method to solve this problem is Logistic Regression, because it is computationally inexpensive baseline algorithms in most frameworks, which is widely used in applications [2]. But most inportantly, this method returns well calibrated probability predictions by default as it directly optimizes log-loss [3]. In case this this method will give poor results, the other option is to use the other methods such as SVM or Adaboost. However, these classifiers should be calibrated manually [3].

### Benchmark Model

In this model we will use all-NDF benchmark. In other words, we will fill the fields to be predicted with 'NDF'. This submission yields 0.68411 NDCG (for details see below) score in the Kaggle leaderboard. Needless to say, any reasonable submission should surpass that score.

### Evaluation Metrics

The evaluation metric for this competition is $NDCG_k$ (Normalized discounted cumulative gain) where k=5. NDCG is calculated as:

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

$$NDCG_k = \frac{DCG_k}{IDCG_k},$$

where $rel_i$ is the relevance of the result at position $i$ and $IDCG_k$ is the maximum possible (ideal) $DCG$ for a given set of queries. All $NDCG$ calculations are relative values on the interval 0.0 to 1.0.

When we predicting the countries, the ground truth country is marked with relevance = 1, while the rest have relevance = 0. 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 $NDCG = \frac{2^0-1}{log_2(1+1)} + \frac{2^1-1}{log_2(2+1)} = \frac{1}{1.58496} = 0.6309$

### Project Design

- Programming language : Python 2.7

- Libraries : Scikit-learn, Pandas
- Workflow:
    1. Data Wrangling.
       The first step is to load the given datasets into Pandas dataframes and clean them. This is followed by various wrangling methods to make data suitable for further analysis and predictions.
    2. Data Analysis and Visualisation.
       The second step is to statistically analyse and create visualisations for the cleaned data and try to come up with a few preliminary hypothesis. The graphs, charts and other visualisations will also be a very important part in creating our story for the project.
    3. Machine Learning
       The next step is to build a predictive model using ML methods. The model will be incrementally improved by testing it against the test users dataframe created from one of the Airbnb datasets.
    4. Data Story, Results and Conclusion
       The final step is to report the results of the analysis performed and the accuracy of the predictive model. This step will involve creating a story around the initial problem, the problems it aims at solving and the insights gained from the data. This will be followed by explaining the intuitions involved in building the ML model, the incremental improvements involved, the accuracy and future prospects of improvement.

### References

[1] https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings.

[2] P. Harrington, Machine learning in action. Manning publication co. (2012).

[3] http://scikit-learn.org/stable/modules/calibration.html.