# DNB Data Scientist for Enterprise Nanodegree Program

## Capstone Proposal

John Gunnar Blostrupsen
November 21st, 2018

## Proposal

### Domain Background

Worldwide, a lot of non-profit organizations depend on donations from contributors and sponsors, in order for them to be able to carry on with their work. It is of great interest for them to be able to target those potential contributors that are most likely to make donations, both in order to be cost-effective, and to avoid potential irrelevant communications.

Traditional marketing techniques like using demographic segmentations, can be challenged by approaches using machine learning techniques that enable us to predict an individual's income, using features from public available data, such as census data.

### Problem Statement

Migraine is one of the most common and disabling neurological disorders in the world. Just looking at UK as an example, there are 190,000 migraine attacks every day, costing the society billions of euros every year in reduced work efficiency, absent from work days, costly treatments and personal discomfort. https://www.migrainetrust.org/. In this capstone project I would like to investigate on how machine learning techniques may help such organizations to be able to target potential contributors more effectively, using features from publically accessible census data. In doing so, I hope to indirectly help such organizations to carry on with their very important work.

## Datasets and Inputs

The data I want to use as an example in this project is a dataset from the UCI Machine Learning Repository, containing multivariate data from the US 1994 census database, in order to predict whether individuals have an annual income of at least $50,000. This dataset should be appropriate with regards to exemplify how non-profit organizations may enable themselves to operate more efficiently when targeting potential sponsors.

## Solution Statement

I will aim at finding a solution that enables non-profit organizations like The Migraine Trust to target their most potential sponsors, and show the importance of how they may narrow down their focus on certain features in the data that are more important than others. To do this, I will carry out an analysis testing several supervised learning models, and evaluate the outputs by how accurate they are, to recommend an appropriate model for future use.

## Benchmark Model

As a benchmark model, I will use Decision Trees – as I am aware of that it is a quite common technique used in customer targeting in several organizations. Comparing the accuracy score and F-score between Ensemble Methods (AdaBoost), and Support Vector Machines (SVM) against the more traditional Decision Trees will be helpful when communicating the potential opportunities for the organization.

## Evaluation Metrics

Accuracy score and F-score will be used to evaluate which model that works best in this context.

## Project Design

First, the data at hand will be gathered, explored and assessed. Then, the data will be prepared so that machine learning techniques may be applied to it. This preprocessing implies cleaning, formatting and restructuring the data. Furthermore, any skewness in the data may demand some transformations. Also, scaling numerical features is good practice to perform. Then, the data will be split into test- and train-sets, before applying appropriate models. Further, I will create a pipeline for training the models and making predictions. Finally, I will optimize the model that works best, in order to make conclusions on a recommendable way for The Migraine Trust (and other non-profit organizations) to target their most potential sponsors. The findings will be summarized in a report.