# Machine Learning Engineer Nanodegree

## *Capstone Proposal*

Sten Ruben Strandheim
August 7th, 2019

# Proposal

## Domain Background

Domain: Media publishing

Background: Intermediates (Google, Facebook, and so forth) try to profit first-party data published by newspapers. In order for the publishers to compete back, they need to customize the way they present their content and ads.

Why: By identifying a reader's demography, a web publisher can expose more customized content to the visitor, such as premium content or ads.

My dataset is from a Norwegian newspaper, made by my colleague. I will try to identify Pensioners in this dataset, because they have a specific consumption pattern.

By doing this, the publisher can
1) Leverage on its first-party content in competition with other intermediates.
2) Sell targeting of ads, in order to increase the revenue potential.

## Problem Statement

Problem: Identify pensioners at a level that justify spent time.

As to illustrate the level of difficulty, calculate:
- general % of pensioners in the dataset
- specific % of pensioners that visited each of the magazines

Example on an academic paper where machine learning was applied to predict age/demographics: "How well can machine learning predict demographics of social media users?"
https://arxiv.org/ftp/arxiv/papers/1702/1702.01807.pdf
Quote from the Conclusion: Other demographic traits such as, age and race/ethnicity are more challenging to predict.

## Datasets and Inputs

Dataset: Contains data from two sources:

- Collected over a timeframe from the publisher's cookies, each visitor gets its own ID from the cookie
- Demographic data is collected through surveys among visitors, conducted by the same cookies providing the same visitor ID.

Data was then (inner-)joined, aggregated and prepped in various age segments. Pensioners were here stated to be aged 60 and above.

Aggregate: Number of times the customer visited a given kind of section in the various magazines/newspapers owned by the publisher, over a timeframe. In this dataset, the sections has been anonymized for it to be used in this project. These sections may be sports, culture, celebrities, etc

The visitor's choice of sections may help indicate the age, as the reader-ID from the cookies doesn't disclose demography in daily use.

Number of rows: 4214
Number of features (magazine sections): 12.
Target: Pensioner is marked as 1, else 0

## Solution Statement

I will suggest a classification model that is capable of predicting readers of age above 60. This may be more challenging as the dataset will be unbalanced, as this age group is less active on the web than younger groups.

## Benchmark Model

Among all the unique readers that returned the demographic survey, there are 13% pensioners. This is the benchmark I will align my results to.

## Evaluation Metrics

Verify results from each algorithm with:

- F_beta scoring, providing an mix of Precision and Recall:
  - Precision
    - P = TP/(TP+FP)
    - Good content customization for the readers. Argument for higher ad prices.
  - Recall
    - R = TP/(TP+FN)
    - Optimal distribution of the publisher's content and ads
  - F_beta:
    - F_beta = (1+beta^2)*P*R/(R+P*beta^2)
    - Beta=0.5 provides an accepted mix with emphasis on precision, wich I think will suite an imbalanced dataset
- Show result with ROC curve and AUC
- Show cumulative lift to assess business value
- Evaluate % of pensioners in predicted d ataset.

## Project Design

Present correlation matrix

Prepare dataset before classifying:

- Normalize all features
- Fix highly skewed features with logarithmic transformation

Illustrate the level of imbalance in the dataset:

- general % of pensioners in the dataset
- specific % of pensioners that visited each of the sections

Split into training and test datasets.
Stratification while splitting: Keep % pensioners in both datasets.

Address possible unbalanced dataset:

- Remove outliers from training dataset. Try with use of DBscan clustering
- Upsample training dataset with SMOTE or ADASYN

Illustrate remaining variance in the training dataset with PCA and correlation matrix.

Predict age with different basic algorithms, with tuning of hyperparameters:

- XGboost classifier
- GaussianNB
- Logistic regression classifier
- RandomForest classifier

Some algorithms may be prone to overfitting with small datasets. In order to illustrate this, learning curves will be displayed. A known algorithm that may easily overfit, is the Light GBM classifier