# Machine Learning Engineer Nanodegree: Capstone proposal

## Arvato Solutions: Customer Targeting

### Author: Mike Griffin

### Date: 21st April 2020

**Domain background**

This is a Udacity capstone project representing a real-life data science problem using data provided by Betrelsmann Arvato Analytics.

Aravto is a global services company that develops and implements solutions for business customers in a diverse range of industries, with a focus on innovations in data analysis and automation.

For this project, Aravto provided services to help a German mail-order sales company understand demographics and improve customer acquisition, in order to design more effective future campaigns. This project will explore unsupervised clustering to identify customer groups as well as classification modelling to predict which individuals will respond to campaigns.

**Problem statement**

The overarching objective is to identify which individuals are most likely to respond to the campaign and become customers of the company. This is a binary classification problem using individual-level attributes to estimate the likelihood of a customer response. Specifically, the project will aim to maximise the ROC score on a held-out test set on Kaggle.

**Datasets and inputs**

Demographics information has been provided for both the general population at large as well as for prior customers of the mail-order company in order to build a model of the customer base of the company.

* Udacity\_AZDIAS\_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
* Udacity\_CUSTOMERS\_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
* Udacity\_MAILOUT\_052018\_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
* Udacity\_MAILOUT\_052018\_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Information is stored at the individual-level. 365 fields are common across the dataset with information on the fields and labels stored in two data dictionaries:

* [DIAS Information Levels - Attributes 2017.xlsx](https://viewqk2nn0jlef.udacity-student-workspaces.com/edit/DIAS%20Information%20Levels%20-%20Attributes%202017.xlsx) – high-level descriptions of attributes
* [DIAS Attributes - Values 2017.xlsx](https://viewqk2nn0jlef.udacity-student-workspaces.com/edit/DIAS%20Attributes%20-%20Values%202017.xlsx) – detailed mappings

All data is used under T&Cs associated with Aravato and should not be

**Solution statement**

The solution will have two related components:

1. **Customer segmentation**– a customer segmentation report which will separate the population in to distinct clusters and explore the variation in attributes by clusters. The approach will consider dimensionality reduction and cluster numbers to explain variance.
2. **Customer acquisition** - a set of models will be trained and compared on the binary classification of whether an individual will respond. Performance will be assessed using train/validation splits.Measure performance on a held-out test set on Kaggle

**Benchmark model**

A simple decision tree will be trained to determine baseline performance with clear interpretability.

**Train and validation**

**Evaluation metrics**

Assessment for each segment of problem

The target dataset contains demographics information for targets of a mailout marketing campaign.

For the supervised prediction problem I will use the Area under the Recover Operator curve (ROC) to align with the Kaggle competition, This metric is bounded between 0 and 1 with 1 representing perfect performance on the test set

**Project design**

I will follow a standard data science pipeline for this problem:

1. Exploratory data analysis – I will explore value distributions and missingness, with visualisations
2. Data processing and feature engineering – PCA will be explored to reduce dimensionality ahead of clustering
3. Clustering analysis – unsupervised clustering methods will be used to determine the appropriate number of customer clusters
4. Supervised modelling – will need to consider class imbalance
5. Tuning and model selection – train, validation and test splits will be used to allow hyperparameter tuning and model selection
6. Kaggle submission – a limited number of Kaggle submissions will be made to avoid overfitting the held-out test set
7. Model interpretation – I will explore measures of feature importance/partial dependence to understand which features matter