# Machine Learning Engineer Nanodegree

## Capstone Proposal

## Customer Segmentation – Arvato Financial Solutions

### **Introduction**

Arvato Financial Solutions provides a professional B2B (Business to Business) financial services. Few services offered by Arvato are diverse segments, payment processing etc.

In this capstone project, a dataset is provided by Arvato is used to support mail-order companies selling various products to acquire new clients.

### **Problem Statement**

The problem statement for this project is “How to acquire new clients to sell client’s products”.

The project is sub-divided into following tasks:

1. Unsupervised learning approach – In this approach, part/region of population are to be selected who would most likely can be approached as new customers using the given demographic data.
2. Supervised learning approach – In this approach, from the above selected region of population which customer segments can be targeted as potential customers for advertising campaign,
3. Kaggle competition, predict if customer would likely convert as company customers or not.

### **Datasets**

There are 4 datasets available:

* 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)

### **Evaluation Metrics**

1. To determine which population region to be considered as potential high customer base, I can check if the proportion of customers in that particular region is greater than the general population proportion. And these selected population regions will become the data set for the next problem.
2. After the population region/area to target is separated. In this section, my algorithm should classify if the customer is likely to become a customer or not.

The final decision on which evaluation metrics to use highly depend on the information obtained through explanatory data analysis.

### **Project Design**

1. Data Cleanup
2. Data Visualization
3. Feature Engineering
4. Model Selection
5. Model Tuning
6. Test and Prediction

**Solution Statement**

1. **For the first part of the problem, which is Unsupervised learning approach, I was planning to use K-means Clustering. As it has higher rate of success in such problems in general.**
2. **For the second part of the problem, which is the Supervised learning approach. Any supervised algorithm can be implemented, e.g. - SVM, logistic regression, k-nearest neighbor algorithm, neural networks or decision trees. I have not yet been able to narrow to which algorithm would produce best result.**