# Machine Learning Engineer Nanodegree

Capstone Project Proposal

Customer Segmentation for Arvato Financial Solutions

**Zhaoyun Ma**

**01/16/2021**

## 1 Domain Background

Arvato is a global services company headquartered in Gütersloh, Germany. Its services include customer support, information technology, logistics, and finance. Bertelsmann Arvato are providing consultation for a mail-order company to better understand their customer demographics compared to the general population in German and identify the customer segments for future marketing campaigns. **Customer segmentation** aims to divide **customers** into clusters based on common characteristics so that companies can market to each cluster effectively and appropriately. Customer segmentation is very helpful for companies to understand their customers and the targeted marketing can save companies significant cost compared to marketing towards the whole population.

This project is one of the proposed capstone projects in Udacity Machine Learning Nanodegree program. This project aims to help a mail-order company to optimize the customer targeted marketing campaign as well as predict the probability of customer conversion for better targeting. The first part of this project focuses on customer segmentation in order to identify the clusters of population that best describe the core customer base of the company compared to the general population. The main task of second part is to identify a supervised learning model that can predict the likelihood of customer conversions with a reasonable accuracy.

## 2 Problem statement

- Identify the customer segments that can describe the core customer base of the company compared to the general population for future marketing campaign
- Identify a supervised machine learning model that can predict the customer conversion probability based on the output of a customer segmentation (Will the customer respond to the tailed campaign?)

## 3 Solution statement

With respect to the problem statement:

- In order to identify the customer segments that can describe the core customer base of the company compared to the general population, principal component analysis (**PCA**) should be first conducted to extract the important features and reduce the data dimensionality. With that, unsupervised learning technique, **K-Means** clustering in this project, will then be performed to identify the customer segments. The customer segments will then be compared to the general population to find the distinguish components and features.
- In order to identify a supervised machine learning model that can predict the customer conversion probability. A benchmark model **XGBClassifier** and evaluation metric **AUROC** (Area under Receiver Operating Curve) are selected, and then a few supervised learning classifiers will be tested and compared to the benchmark model. The best model will then

be further improved using **Bayesian Optimization** [2] to quickly narrow down the parameters, and finally **GridSearchCV** will be used to find the best model.

## 4 Project data

The data used is provided by Udacity partners at Bertelsmann Arvato Analytics, and represents a real-life data science task. Four demographic data files are provided with some general information, including:

- Udacity_AZDIAS_052018.csv (general population of German, 891 211 rows by 366 columns)
- Udacity_CUSTOMERS_052018.csv (customers of a mail-order company, 191 652 rows by 369 columns)
- Udacity_MAILOUT_052018_TRAIN.csv (targets of a marketing campaign, 42 982 rows by 367 columns)
- Udacity_MAILOUT_052018_TEST.csv (targets of a marketing campaign, 42 833 rows by 367 columns)

The first two data files (azdias and customers) are used for the customer segmentation and figure out the similarity and difference between the customer and general population. The mailout_train data file is used to build up a supervised machine learning model and then the mailout_test data file is then used to test the model accuracy in a Kaggle competition.

In addition, two excel spreadsheets are provided with detailed documentation of the features in the demographic data files, which are:

- DIAS Information Levels — Attributes 2017.xlsx
- DIAS Attributes — Values 2017.xlsx

## 5 Evaluation metrics

In this project, since the **mailout** data is highly imbalanced with more than 98% of the data response is negative and less than 2% of the data response is positive, accuracy is not appropriate for the model evaluation. Since our goal is to predict the customer conversion probability, AUROC(Area under Receiver Operating Curve) is used as the evaluation metric with considering the contributions of both true positive rate and true negative rate. ROC plots the true positive rate (TPR, the proportion of actual customers that are labeled as 1) against the false positive rate (FPR, the proportion of non-customers labeled as 1) as shown in Fig 1.

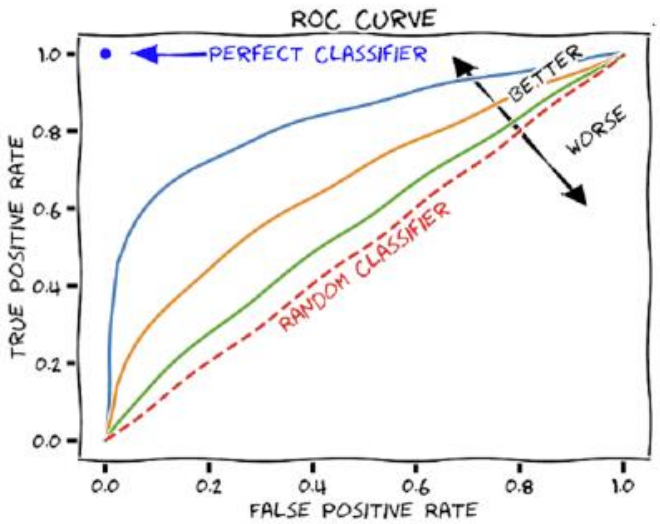

Fig 1 ROC curve [1]

## 6 Benchmark Model

The benchmark model I selected is XGBoost XGBClassifier base model. XGBoost can provide a high-performance implementation of gradient boosted decision trees, and it becomes many Kaggle competition winning model. I will train this base model on 70% of the train data and record the AUROC score on the verification data (30% of the train data). I will then compare the future results with this bench mark and further improve the model. Considering the high dimensionality of the data, I would try to achieve 0.8 AUROC score at the end (top score on Kaggle leader board is 0.84).

## 7 Project design and outline

The project is designed following 3 main parts shown below:

1. **Data exploration and cleaning**: the provided dataset will be explored and cleaned through building up an **ETL pipeline** aided with appropriate data visualization.
    - Unknown data will be identified and replaced with NaN,
    - Rows and columns with a lot NaNs will be removed based on a selected threshold,
    - The rest NaN will be appropriately imputed with the mode of the data,
    - High dimensional categorical features will be removed.
2. **Customer segmentation report**:
    - Principle component analysis (**PCA)** will be first conducted on the cleaned dataset to analyze the principal component and feature importance, **selected features** that can explain most of the variance will then be kept and used for **K-Means** clustering.
    - **K-Means** clustering will be conducted, and k will be selected based on **Elbow method** [2]. Based on the cluster results, the important components and features will be interpreted and insights will be delivered.

3. **Customer conversion probability prediction**: the **mailout** dataset is used for customer conversion probability prediction through supervised learning modeling. A few classifiers are tested with the base model and compared with the benchmark model (**XGBClassifier**) the best one will be selected for further model improvement.
    - **Bayesian Optimization** [3] will be used to fast find the ideal hyperparameters for classification modeling process.
    - **GridSearchCV** will then be used for fine tuning to further improve the model performance.

## References


[1] https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg
[2] https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/
[3] https://github.com/fmfn/BayesianOptimization