## Domain Background

A mail-order sales company offers the service of buying things without the need of going to the shop in person, that is, tele-buying (tele is a prefix that stands for ‘in the distance’). The remote methods can be: postal mailing, telephone call, and online via web browser or mobile app. This last methos is known as online shopping or e-comerce.

The mail-order business is very old. The first case related to this business is in 1498, in Venice, where a catalogue of books was printed. Later, in 1667 in England, a seed catalogue was published. Once in England, the spread to the Brithis America took place in 1744. However, the first modern mail-order company was set up in 1856, and by 1880 he had more tan 100.000 customers not only in the United Kingdom, but also in the British Colonies. The big shock to the inductry came up with the invention of the internet, where a company’s website became the more usual way to order.

It is important for the company to target the most prone clients so that it can increment the sales in the best optimized way, and it is imoprtant for the client to recieve the offers that best fits its needs, that is, customer intelligence. Relevant academic researches in the field are: “US eCommerce Forecast: 2013 to 2018”, Forrester Research and “What is Customer Intelligence”, CRM Today. I’m interested in this field because I was one in an internship at a bank in the department of Marketing, doing cross-selling and up-selling propensión models.

## Problem Statement

The problem we want to solve is: find the most prone citizens to become new customers of the mail-order company.

## Datasets and Inputs

The data that we will use has been provided by Bertelsmann Arvato Analytics, and represents a real-life data science task.

There are four data files associeted with this project:

- 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).

Each row of the demographics represents a single person, but also includes information outside of individuals, including information about their household, building and neighborhood.

## Solution Statement

To solve this problem, wi first apply unsupervised learning to know what features best describe our target clients. For this, we first reduce the dimensión of the dataset with principal component analysis and then use k-means clustering algorithm. Finally, we will train a fully connected neural network in this features to predict a grade of how much a citizen is prone to become a new customer.

## Benchmark Model

According to refrence number 4, previous benchmark studies suggest that there really exists a difference in performance between white box models and black box models. We will use the logistic regression as the benchmark model for the supervised classification task.

## Evaluation Metrics

The performance of the solution model presented will be evaluted with the área under the curve (AUC) metric in the test dataset. This is a good metric because the dataset is imbalanced (most of the citizens don't answer marketing campaign). Other metrics like precision, recall and F1 score are also viable measures.

## Project Design

First of all, a preprocesing of the data is mandatory, deleting records with not avaliable (NA) information. The columns which are highly linerly correlated will be deleted also. Categorical variables will be transformed into n-1 dummy variables where n is the number of classes in the variable. Before standarazing the data a brief descriptive summary of the variables will be done (mean, mode and cuantiles). Some ilustrative graphs will be plot before standarazing. After this step, k-means algorithm will be use to achieve unsupervised learning. Finally, a fully connected layer neural network will be set up for modeling.

## References

1. https://en.wikipedia.org/wiki/Mail_order
2. https://en.wikipedia.org/wiki/E-commerce#References
3. https://en.wikipedia.org/wiki/Customer_intelligence
4. https://towardsdatascience.com/benchmarking-simple-machine-learning-models-with-feature-extraction-against-modern-black-box-80af734b31cc