Report for Arvato Financial Solutions Proposal

1- Domain background

I ll work with a recommendation system, looking for the similar behavior and characteristics, helping companies to find the clients, for this we ll analyse the current client database. I decided for this project because it's a very interesting and big company. Also I am already good with time series so ll be good to work with a different approach. I need to read much more abou it i know but the first step is to understand the data and segment, i think this 2 articles ll help me alot :
AI meets marketing segmentation models | by Jan Teichmann | Towards Data Science
(PDF) Marketing Segmentation Through Machine Learning Models: An Approach Based on Customer Relationship Management and Customer Profitability Accounting (researchgate.net)

2- Problem statement

Take decisions based on data instead of gut feel, i ll need to prove with metrics and math that is the best way to find the clients and showing with supervised learning the result of a training model.

3- Datasets and inputs

I have 4 different dataframes to analyse:

- **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).**

It's a very different project for me because we have 2 different datasets just to understand the data and the problem, and another 2 to train and test.

We have a lot of columns on dataframes, so first i ll clean up the nans or empty values, doing the data wrangler, after that i ll analyse data and create so feature selection or extraction, like pca we use in the people segmentation project.

4- Solution statement

I create a pipeline with a Data wrangler, feature extraction and unsupervised algorithms like knn, after that I'll find the best supervised learning algorithm for the classification.

5- Benchmark model -

i ll use for benchmark logistic regression, with no hyperparameters tuning.

6- Evaluation metrics

Well here the most important thing ll be the precision because i need to find the false positives, who can be clients but are not. so i'll use the metric i saw on amazon sagemaker, precision_at_target_recall, for binary classification.

7- Project design

I'll develop a solution inside the sagemaker to work with this big data, and I'll use the metrics and benchmark to evaluate my model, testing a lot of different models, watching to don't have underfitting or overfitting.