Machine Learning Engineer Nanodegree

Capstone Proposal

Clement Palfroy
October 7th, 2021

Domain Background

Arvato is an internationally active services company that develops and implements innovative solutions for business customers from around the world. These include SCM solutions, financial services and IT services, which are continuously developed with a focus on innovations in automation as well as data and analytics.

One of Arvato missions is to help customers extract valuable insights from massive datasets. Those insights are then used to make informed and mathematically based decisions. For many years now, marketing managers searched for ways to optimize their ads targeting. When large datasets are available, Machine Learning represents an ideal tool to identify hidden patterns in customer behaviour (Badea, 2014; Kim & Ahn, 2009).

In this project, Arvato is helping a Mail-order company to understand its customers segments in order to identify with high accuracy, their next customers. General demographic data will be analysed to better understand customer characteristics. From there, a ML model will be built and trained to predict the likelihood of an individual to become a customer.

Problem Statement

The goal of this project is to determine the potential of an individual (given its demographic characteristics) to become client of a specific company.

How are the company clients different than the rest of the population?

The demographic data of the general population will be studied with the help of unsupervised learning algorithms. Segments will be defined. Then, it will be applied to the customer data to discover what features are specific to the company customer.

How can we use the previous findings to better predict the likelihood of a person to become a customer?

Several learning algorithms will be built and train using marketing results from previous campaign.

Datasets and Inputs

Four dataset will be used in this project:

- Udacity_AZDIAS_052018: The demographic data for the general population of Germany. 891211 persons (rows) X 366 features (columns).

- Udacity_CUSTOMERS_052018: The demographic data for the company customers. 191652 persons (rows) X 369 features (columns).

- Udacity_MAILOUT_052018_TRAIN: The demographic data for individuals who were targeted by the marketing campaign, and if they became client or not (1/0). 42982 persons (rows) X 367 features + label (columns).

- Udacity_MAILOUT_052018_TEST: The demographic data for other individuals who were also targeted by the marketing campaign, but without the information if they became client or not. 42833 persons (rows) X 366 features (columns).

Each row of the demographics files represents a single person, and includes information like their household, building, and neighbourhood.

Despite many rows for each dataset (especially for the general population), approximately 11% of the data is missing. Moreover, for some columns, some of the data has been marked unknown/missing with a special number (-1, 0…). The missing values will have to be deleted strategically as to not lose too much information.

Additionally, the training data is highly unbalanced, 98% of the training examples are labelled “0”. To address this problem, specific metrics will be selected.

Those datasets are provided by Arvato in the context of the Machine Learning ND from Udacity. The terms and conditions for the data usage will be included in the project repository.

Solution Statement

1) Data cleaning/preparation

First the dataset will be explored and cleaned. Missing values will be deleted, categorical features will be transformed to numerical format, and all the data will be normalized (so that no features have higher weight than others).

2) Dimensionality reduction

As 366 features can be quite heavy and difficult to apprehend, a PCA will be used to reduce the number of parameters while keeping most of the variance. Precisely how much variance will be decided when exploring the data. The PCA transformer will be built with the general population dataset only.

3) Segmentation

A K-means algorithm will be used to split the general population into clusters. Once trained, it will be used to predict on which of those clusters the customer population tends to fall into.

In the second part of this project, the task will be to build an accurate predictor of an individual potential to become customer.

1) Data cleaning/preparation

Just like in step 1, the train data will be cleaned, and its dimensionality reduced.

2) Model creation and training

Two types of models will be investigated. An XGBoost Classifier, and a classic ANN. For each model, different hyperparameters combinations will be tested (model tuning).

Benchmark Model

A simple linear-regression model could be used as benchmark.

Evaluation Metrics

The training dataset will be divided into a train and a validation set. As the task at hand is a binary classification, the selected metrics will be *Precision and Recall* or *Area under Receiver Operating Curve (AUROC)*. Note that those metrics will be calculated for the validation set exclusively.

Project Design

Here is a summary of the main steps of this project, note that the entire work will be performed within a SageMaker notebook instance (ml.m4.xlarge), and all notebooks will use the “pythorch_p3.6” Kernel.

1) Data cleaning

Using the metadata files, a thorough cleaning and transformation of the data will be performed. An analysis on missing values per feature will be performed to decide which one to keep. A visualization analysis will also be performed to spotlight any relevant patterns.

2) Feature engineering

To facilitate data analyses and modelling, a dimensionality reduction technique will be performed (specifically, a PCA) using the skicit-learn library.

3) Modelling
First, we will identify customer segments using unsupervised learning algorithm. At this stage, the K-Means algo is considered.

Secondly, two supervised learning models will be investigated. An XGBoost Classifier, and a personalized Pytorch ANN. The SageMaker hyperparameter tuning functionality will be used to find the best model architecture.

4) Evaluation
The predictions of the best performing model will be evaluated on the appropriate Kaggle page (udacity-arvato-identify-customers).

Badea, L. M. (2014). Predicting Consumer Behavior with Artificial Neural Networks. *Procedia Economics and Finance, 15*, 238-246. https://doi.org/https://doi.org/10.1016/S2212-5671(14)00492-4

Kim, J., & Ahn, H. (2009). A New Perspective for Neural Networks: Application to a Marketing Management Problem. *J. Inf. Sci. Eng.*, *25*, 1605-1616.