UDACITY

Proposal Report

Machine Learning Engineer Nanodegree

Julio Guijarro Hernández

06-05-2020

Índice

1. 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 and data/analytics.

Arvato is inside of Bertelsmann, that is a company founded in 1835 dedicated mainly to the communication sectors and also present in the service and education sectors.

On the other hand, the problem to be analyzed in this project, is population segmentation, that is a very common project in the marketing and sales sector. This project pretends to be able to identify in a more detailed way possible clients to which to direct their campaigns, by detecting possible patterns or information present in the data that define the consumer and in that way manage to optimize their sales and increase the benefits.

2. Problem Statement

The problem in this project, as mentioned above, is population segmentation.
With the segmentation of the population, Arvato aims to classify in an optimal way the population of which it has information in order to be able to consider it as a target population or not for its sales and marketing campaigns, in this specific case, for its mail order campaign.

This problem can be solved in several steps to obtain the final result of knowing who should be included in the company's sales campaign. The first of these phases would be to obtain a summary of the characteristics that define each person of whom we have information, and of this available information which is really useful for the project and its objective.

After this you can proceed to group the users by groups with common characteristics that enrich the information you have about the customer, to finally decide who to send the campaign to.

3. Datasets and Inputs

The information to be used in this project is that provided by Udacity and Arvato for the project.There are four data files associated 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 files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood. Two more metadata files are provided:

- DIAS Information Levels – Attributes 2017.xlsx: top lervel list of attributes and descriptions.
- DIAS Attributes – Values 2017.xlsx: detailed mapping of data.

4. Solution statement

In order to solve the end-to-end problem proposed by Arvato and Udacity, I will develop a series of steps that could be divided into two large sections.

The first of these steps will be the pre-processing of the data that Arvato and Udacity have provided for the project. The goal of this phase is to understand the data and know at a high level what the data says about the consumers it describes, using data cleaning and data mining techniques, as well as Unsupervised Learning techniques such as PCA and clustering algorithms such as Kmeans, in order to detect patterns in the data that will allow us to group the population that is most likely to buy in Arvato's sales campaign.

5. Benchmark Models

As first references for modeling the data, we could use the models already present in the Kaggle competition that are present in this project, which use XGBoost models and have achieved up to 80 % of correct classifications. A very important point as the model in this project will be the type of data that feeds this supervised learning model, which will be tested with information only from training campaigns, or mixed with information from the population provided to us.

6. Evaluation Metrics

In this Project a classification is going to be elaborated, so in classification models one of the most used and recommended metrics is the AUC, which will be the one I will use in this project.

Also due to the low ratio of positive cases in the data, you will see the Precision Recall AUC metric that is recommended for very small ratio classification cases.

7. Project Design

This Project will consist of two main parts as mentioned above.
In the first part the following tasks will be developed:

- Understanding the data and first observations: the quality of the data and its main characteristics will be analysed.
- Data clean-up: cleaning the data to eliminate those features that are not of interest to us or that are not properly filled in, as well as homogenizing the characteristics.

- Development of PCAs: development of a PCA model to try to simplify the data describing the data.
- Clustering development: development of a Kmeans model to group consumers by the main characteristics detected by the model and to observe whether within these groups, consumers are more abundant in some than in others.

In the second part, a supervised model will be developed to identify customers more likely to go to Arvato's sales campaigns. From these models, the main metrics that define them will be obtained and the best one will be selected to be sent to the Kaggle competition.

8. References

https://es.wikipedia.org/wiki/
https://www.bertelsmann.com/divisions/arvato/st-1
https://medium.com/@wengsengh/customer-segmentation-with-pca-6ddf9681ce1d