Udacity + Arvato Financial Solutions: Identify Customers from a Mailout Campaign

Capstone project

Olivier Gobron

July 8, 2020

1 Introduction

This competition [1] is connected to one of Udacity's capstone project options for the Machine Learning Engineer Nanodegree program, in connection with Arvato Financial Solutions, a Bertelsmann subsidiary.Arvato Financial Solutions is a company that provides professional B2B financial services to other companies - allowing them to leave their credit management to a professional, so they can focus on what matters most for their business.

2 Domain background

Since the first mail sent in 1971, and the massive development of internet in the 90s, the use of mails has been developped to marketing. Indeed, the companies now use them to communicate with their customers. The purposes are numerousous. It could be to keep their customer aware about their last items or offers, or teaching about their brand and keep their clients attracted by their company or products between purchases, or in particular, it can be used to attract new customers amoung their clients. In this case, it is possible to send a mail to a massive amount of people and hopefully target amoung them the potential new customers. However, in order to highlight the values of a brand and increase the impact of these emails instead of filling the spam folder of a lot of mailboxes, it is important to select carefully the recipients and target them efficiently as potential new customers. Doing so, the companies can expect an increase in their customer acquisition process. This selection can be made thanks to past experience or intuition, but we generally observe better results using machine learning techniques and data as back up for these decisions.

In the project, a mail-order sales company in Germany is interested in identifying segments of the general population to target with their marketing in order to grow. Demographics information has been provided for both the general population at large as well as for prior customers of the mail-order company in order to build a model of the customer base of the company. The target dataset contains demographics information for targets of a mailout marketing campaign. The objective is to identify which individuals are most likely to respond to the campaign and become customers of the mail-order company.

3 Problem statement

We have to predict which people to target with an email marketing process to make them become new customers. I will tackle the problem in 2 steps :

- Customer Segmentation
  I will use unsupervised learning methods to identify the group with the maximum of customers from the global popullation.

- Binary classification
  Once I have identified this group, I will develop a binary classifier that predict with a maximum accuracy which persons is a member of this group.

4 Datasets and inputs

As part of the project, half of the mailout data has been provided with included response column. For the competition, the remaining half of the mailout data has had its response column withheld; the competition will be scored based on the predictions on that half of the data.

5 Solution statement

The solution will be the predictions using the binary classifier performed on a test dataset. Here are the steps that I will follow :

1. I will visualise the data in order to have a better understanding of what I have to deal with, cleaning the data if required.

2. I will train a feature reduction on the data and keep the features so that the variance is above a certain threshold in order not to loose too much information. The purpose of this step is to maximize the efficiency of the next step.

3. I will train an unsupervised model cluster the popullation in different groups and see in which group(s) we have the more customers.

4. I will label the data so that we identify the persons that are in the group(s) of the customers.

5. I will then developp a binary classifier that tells if a person is amoung the same group(s) as the customers.

6 Benchmark model

I will train 3 different models to solve this problem :

1. For the feature reduction, I will use a Principal Companent Analysis model.
2. For the popullation segmentation, I will use a K-Nearest-Neighboor model.
3. For the binary classification, I will use test several of them and compare their performances. Maybe XGBoost, or Random Forest or another one.

7 Evaluation metrics

The evaluation metric for this competition is AUC for the ROC curve, relative to the detection of customers from the mail campaign. A ROC, or receiver operating characteristic, is a graphic used to plot the true positive rate (TPR, proportion of actual customers that are labeled as so) against the false positive rate (FPR, proportion of non-customers labeled as customers).

8 Project design

I workflow that I have in mind follow the following steps :

1. Data exploration
   Shape of the data, missing values, Data visualization
2. Data preparation
   Removing missing value or impute them, format the data for training
3. Features reduction
   Training of the model and evaluation using the variance of the data)
4. Popullation segmentation
   Training of the model and characterization of the customer group
5. Binary classification
   Traning of several models
6. Evaluation of the classifier
   Evaluation using the AUC or ROC curve on each classifier trained and selection of the best one

References

[1] "Apply machine learning techniques to predict customers using data provided by arvato financial solutions." https://www.kaggle.com/c/udacity-arvato-identify-customers/overview, 2019.