Capstone Project

Saverio Pulizzi

Udacity - Machine Learning Engineer Nanodegree

July 4th, 2021

Customer Segmentation Report for Arvato Financial Services

Domain Background

Arvato Financial Services is a company operating in the mail-order sales business in Germany and is a subsidiary of Bertelsmann.

The company wants to grow their customer base by better targeting clusters of the general population with their marketing campaigns.

Demographics data of the general population and of prior customers of the business will be used in order to identify those individuals who are more likely to respond to the marketing campaign and to become customers of the mail-order company.

At the beginning of the project, unsupervised learning methods will be used to analyze attributes of established customers and the general population in order to create customer segments.

Finally, the previous analysis will be used against a target dataset with attributes from targets of a mail order campaign in order to build a machine learning model that predicts whether or not each individual will respond to the marketing campaign.

Problem Statement

The problem that will be worked on this project is:

> *"Can Arvato Financial Services predict individuals who are more likely to convert to become new customers?"*

The problem that this project is trying to solve is whether machine learning and in particular unsupervised learning and supervised learning could be used to predict those individuals from a general population dataset that are more likely to convert and to become new customers.

Datasets and Inputs

The data has been provided by Bertelsmann Arvato Analytics, and consists of demographics data of the general population of Germany, for prior customers of the company and of individuals targeted on a marketing campaign.

In particular, there are four datasets:

- 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.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file.
The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether or not each recipient became a customer of the company. For the "TRAIN" subset, this column has been retained, but in the "TEST" subset it has been removed.

Solution Statement

In order to predict individuals who are more likely to convert to become new customers, the project will involve using the following techniques:

- Unsupervised learning to identify segments of the general population that are very similar to historical customers of the company. In particular, Principal Component Analysis will be used for dimensionality reduction, followed by K-means clustering to obtain the needed clusters.
- Supervised learning to build a model that will be able to predict the probability that a targeted individual will convert to become a new customer. This is a classification task and different models will be tried from simple (e.g. Linear Classifier) to more sophisticated (e.g. Random Forest Classifier)

Benchmark Model

A Linear Regression Model could be used as a benchmark model with outcomes "1" for likely to convert and "0" for not likely to convert.

Evaluation Metrics

Once the model will be trained, it will be used to make predictions on the campaign data from the Kaggle Competition and the score on the leader-board will be used as the evaluation metric.

The evaluation metric for this project 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).

The line plotted on these axes depicts the performance of an algorithm as we sweep across the entire output value range. We start by accepting no individuals as customers (thus giving a 0.0 TPR and FPR) then gradually increase the threshold for accepting customers until all individuals are accepted (thus giving a 1.0 TPR and FPR).

The AUC, or area under the curve, summarizes the performance of the model. If a model does not discriminate between classes at all, its curve should be approximately a diagonal line from (0, 0) to (1, 1), earning a score of 0.5.

A model that identifies most of the customers first, before starting to make errors, will see its curve start with a steep upward slope towards the upper-left corner before making a shallow slope towards the upper-right.

The maximum score possible is 1.0, if all customers are perfectly captured by the model first.

Project Design

Following is a summary of the theoretical workflow used to reach a solution:

1. Download or otherwise retrieve the data
2. Cleaning and data exploration
   a. explore the data
   b. cleaning (e.g. handling null and empty values)
   c. visualizations
   d. correlation studies
3. Data preparation and transformation
   a. Feature engineering (PCA)
4. Upload the processed data to S3
5. Model development and training
   a. Develop a model
   b. Train a model
6. Model validation and evaluation
   a. Hyperparameters tuning
   b. Select the best performing model based on the test results
7. Deploy the trained model to perform classification on the Kaggle competition test data

Reference

- Arvato. In Wikipedia. Retrieved from: https://en.wikipedia.org/wiki/Arvato#cite_note-3
- Customer Segmentation (Online Definition). In SearchCustomerExperience. Retrieved from: https://searchcustomerexperience.techtarget.com/definition/customer-segmentation
- Udacity+Arvato: Identify Customer Segments. In Kaggle. Retrieved from: https://www.kaggle.com/c/udacity-arvato-identify-customers

- Receiver Operating Characteristic. In Wikipedia. Retrieved from: https://en.wikipedia.org/wiki/Receiver_operating_characteristic