Customer Segmentation for Arvato Financial Services
Capstone Project Proposal

Weijia Chen

June 20, 2021

Contents

1 Domain Background

Customer segmentation is essential for big companies to identify and classify existing and potential customers. It allows companies to better tailor their marketing effort to a wide range of audience subsets. For example, Millenials and Generation Z can have different spending habits and use different source to find merchants. It could be more effective to use social media to target the younger generation than baby boomers. By identifying the right customer segment for specific product and service, companies can boost sales, create targeted promotion campaigns and better manage client relationships.

A company can segment customers using various factors, such as age, gender, employment status, location and etc. The more factors used, the more detailed and more accurate customer segmentation can be. Traditionally customer segmentation are done through surveys and telephone interviews. Now as people go shopping online, collecting customers' shopping information online helps characterize a customer's image and classify them into smaller groups.

Nowadays, with advanced data analytics tools and more data available collected through various resources, it is possible to target customer groups more accurately with machine learning techniques.

2 Problem Statement

Arvato is a global service company whose services include customer support, information technology, logistics and finance[1]. In this project, I am tasked to analyse demographics data for customers of a Arvato's client, figure out the common characteristics of these customers and make recommendations for a marketing campaign to determine the potential customers from another population group.

In the first part, analysis needs to be done on the existing customer data. To better characterize the customers and quantify their features, it is required to compare it against the demographics of the general population in Germany, in order to construct a segmentation of the customers. In the second part, another two dataset of customers with same number of features is provided. It is required to train a supervised machine learning model on the training set and use the model on the test set to make predictions on whether each person is likely to become a customer and therefore should be covered by company's marketing campaign.

Finally, after a model is created, it is required to generate the prediction result on the test data to participate in a Kaggle competition.

3 Datasets and Inputs

There are four dataset provided in 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).

[1] Wikipedia: https://en.wikipedia.org/wiki/Arvato

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 information from the first two files can be used to figure out how customers ("CUSTOMERS") are similar to or differ from the general population at large ("AZDIAS"). The third and fourth file are files to be used for supervised learning: use the analysis result to make predictions on these files ("MAILOUT") and predict which recipients are most likely to become a customer for the mail-order company.

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; it is against that withheld column that your final predictions will be assessed in the Kaggle competition. Otherwise, all of the remaining columns are the same between the three data files.

It is important to note that due to the nature of marketing campaigns, most targeted population do not respond to mail-out advertisement or become customers. Therefore, the data in the "TRAIN" set is highly imbalanced in terms of the value in the "RESPONSE" column. Actually only around 1% of the total recipients became customers. The imbalance of target classes need to be taken extra care of when doing data analysis and making predictions.

Other than that, two more spreadsheets are provided on information about the columns depicted in the files: `DIAS Information Levels - Attributes 2017.xlsx` is a top-level list of attributes and descriptions, organized by informational category.
`DIAS Attributes - Values 2017.xlsx` is a detailed mapping of data values for each feature in alphabetical order.

4 Solution Statement

In the first part of the project, I will perform customer segmentation using unsupervised machine learning techniques. By comparing the demographics of the company's existing customers and the general population, I will find out the factors that discriminate the existing customers from others using methods such as principal component analysis (PCA). I also plan to conduct cluster analysis and inspect if such methods are effective to segment customers from the general population. I will complete with a list of features that have the most discriminating power and interpret them with real-world meanings. After the first part, I can characterize qualitatively and quantitatively the customer image of the mail-order sales company.

In the second part, I will first use the selected features to filter out the most relevant information, and then train different models on the data. I will compare and select the best model based on different metrics. I will then make predictions on a new population group on the test set using the best model and figure out whether each potential customer should be included in the new campaign. Models that I plan to use include tree-based methods such as random forest and boosting.

Finally, after finding the best model and generate the test results, I will upload it to participate in the Kaggle competition.

5 Benchmark Model

For the second part of the problem, a K-nearest neighbor model with k=1 can be used as a benchmark. Essentially, this means labelling the person with the same class (customer or not) the same as its closest neighbor from the training set. Intuitive as it sounds, deliberation is needed to calculate distance between neighbors, considering different types of features (categorical, ordinal and continuous) in the data set.

6 Evaluation Metrics

The evaluation metrics used in this project is the AUC-ROC metric. The ROC curve, or receiver operating characteristic curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied[2]. It is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. AUC, or area under curve provides an aggregate measure of performance across all possible classification thresholds. 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. One way of interpreting AUC is as the probability that the model ranks a random target person more highly than a random non-target person. ROC analysis provides a tool to select possibly optimal models for the given customer prediction task.

7 Project Design

The project includes the following parts:

1. **Data Exploration And Data Cleaning**: In this part, I will conduct an overview of what the data look like: what features are there; which ones are categorical, ordinal or continuous; what meanings do they represent; what does the distribution of the data look like, etc. After that, I will clean the data by removing noisy points and outliers. An initial screening of features to discard meaningless ones can also be done in this part.
2. **Unsupervised Learning for Customer Segmentation**: Following the initial data exploration, I will conduct cluster analysis and principal component analysis on the customer data and the general population data. At the end of this part, a few features with the most discriminating power will be selected. A weighted sum of these features is expected to draw a line between the customers and the majority of the general population.
3. **Supervised Learning for Customer Identification**: In the third part of this project, I will train supervised learning models and build a binary classifier to determine whether a person is likely to become a potential customer. The models will be trained on the given training set and tested on the given test set. I may also need to split out a validation set from the training data for validation purpose during training.
4. **Upload Testing Result for Kaggle Competition**: Finally, after finding the best model from part three, I will generate test result from the test data and upload them to participate in the online Kaggle competition.

[2] Wikipedia: https://en.wikipedia.org/wiki/Receiver_operating_characteristic