Machine Learning Engineer Nanodegree

Capstone Proposal

Qiyue Ren

October 14th, 2021

Domain Background

Arvato is a globally active services company that designs and delivers award-winning customer service and back-office processing services for customers. It help its clients achieve strategic objectives and delight its customers. [1]

In this project, Arvato is helping a mail-order company with efficient customer acquisition. As a typical customer segmentation problem, it will identify the similarities of different groups and build up connections with clients in the groups. The customer segmentation analysis will enable marketing teams to identify customers with high efficiency and accuracy so as to maximize the value of customers.

We will analyze the attributes of existing clients and employ machine learning models to help identify most likely new clients based on certain characteristics.

Problem Statement

As stated by the Arvato manager, the main goal from a business perspective is: given the attributes and demographic information of existing clients, how can the mail-order company acquire new German clients more efficiently?

In specific, we need to solve the following problems:

- Identify what demographic features are corresponding to customers of the mail-order company;
- Predict possible customers based on demographic information.

In order to solve these problems, first of all, we will use unsupervised learning models to help analyze the attributes and characteristics of the established customer to create customer segmentation. Secondly, we will use the previous analysis to build supervised learning models to predict whether a person will be a potential customer.

Datasets and Inputs

The data are provided by Arvato company:

- 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).

Solution Statement

As mentioned in the problem statement, we want to understand the relations between demographics features and existent customers and predict potential customers. In order to achieve this goal, the problem is divided into three steps:

- Data analysis: The data will be explored and cleaned. If there are any problems, such as missing values and data scale, we will fix them and prepared them for further analysis.
- Customer segmentation: We will try to employ the Principal Component Analysis (PCA) Algorithm for dimension reduction. In order to segment the whole population to different parts, we will use K-Mean Algorithm to obtain clusters and assign data into groups based on selected features.
- Customer prediction: Based on the previous segmentation, we will try several models and choose the most accurate one for the prediction: logistic regression, decision tree, XGB classifier, Random Forest classifier.

Benchmark Model

The benchmark model would be the logistic regression model with 1 stands for a potential new customer and 0 stands for not a potential new customer. It will establish initial results upon which further improvements can be made to assess the relative improvements.

Evaluation Metrics

The project will be measured by Kaggle competition which would give an evaluation based on AUC for the ROC curve [2]. To be specific, 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 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

A theoretical workflow for approaching the solution:

- Data exploration, standardization and visualization:
  - ➢ Clean the datasets: Identify the missing data, outliers, etc.;
  - ➢ Visualize the overall distribution, the relationships between different features;
  - ➢ Pre-process data, such as standardizing data to speed up training process;
- Unsupervised analysis:
  - ➢ Identify principal features by PCA algorithm;
  - ➢ Cluster data based on K-Mean algorithm.
- Supervised analysis:
  - ➢ Make a prediction based on benchmark model;
  - ➢ Try to build a prediction model with several algorithms (decision tree, XGB classifier, Random Forest classifier) or a combination of those algorithms. And select the algorithm with the best accuracy.
- Evaluation by Kaggle:
  - ➢ Test the model with Kaggle evaluation methods and better the built model.

Reference

[1] https://www.arvato.co.uk/

[2] https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/evaluation