Capstone Proposal
Customer Segmentation Report for Arvato Financial Services

Course: Machine Learning Engineer Nanodegree
Author: JRA Supuli Jayaweera
Date: 10th July 2020

Domain Background

Arvato Financial Services, a mail order company in Germany is interested in analyzing the demographics data of its customers against the general population to identify which segments in the general population could be potential targets for their marketing plan. For this task, Arvato has provided demographics data of the general population at large as well as the data of the previous customers. In addition to the modelling and identifying this customer base, this project will also identify individuals who will respond favourably to the marketing campaign and become an Arvato customer.

Problem Statement

The problem statement of the project can be divided into four main sections.

Section1: In this section, data preprocessing is done including data cleaning, reformatting data types and filling the missing data. Further, this section will decide which features to keep or drop and which features needs regrouping.

Section2: Identify the demographics features that describe and differentiate the company's existing customer base from the general population. This will be useful to explain which part of the general population is more likely to become the company's customers base.

Section3: In this section, develop a model to identify which individuals will become an actual customer for this company after the marketing campaign. For this task, supervised classification models like random forest, decision trees, SVM and boosting models are considered. Once a successful model is developed, used it on a test dataset to predict the outcome of a selected individual being a customer.

Section4: In the final section, the predictions made on the test dataset will be submitted to the Kaggle competition to evaluate the performance.

Datasets and Inputs

This project provides four main datasets with identical demographics features.

1) ***Udacity_AZDIAS_052018.csv:*** Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
2) ***Udacity_CUSTOMERS_052018.csv****:* Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns). The extra 3 columns

are for CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP', which provide broad information about the customers.

3) ***Udacity_MAILOUT_052018_TRAIN.csv:*** Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
4) ***Udacity_MAILOUT_052018_TEST.csv:*** Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

In addition to the above four data files, two additional files were provided with details description of the attributes.

i) ***DIAS Information Levels - Attributes 2017.xlsx:*** A top-level list of attributes and descriptions, organized by informational category.
ii) ***DIAS Attributes - Values 2017.xlsx:*** A detailed mapping of data values for each feature in alphabetical order.

Solution Statement

To achieve the goals of the project the following solution statement is proposed.

1) The data preprocessing and cleaning will be done using exploratory data analysis and the use of support documentation provided on attribute description.
2) After a dimensionality reduction of the features (using PCA), unsupervised learning techniques like k-means clustering will be used to perform customer segmentation. This will be used to identify which demographic features best describe the core customer base for the company.
3) Then, supervised classification algorithms will be used to develop a model to identify which individuals will be an actual customer after the marketing campaign. Decision trees, SVM, random forest, and boosting methods (AdaBoost and XGBoost) are selected as the initial classification algorithms. The model selection and tuning will be done against the training data provided. The model is used again on thr test dataset to get predictions. The model evaluation is based on the area under the ROC curve which will be described in detail in this proposal later.

Benchmark Model

Due to the nature of the data provided, a random forest (RF) model is selected as the benchmark model as it has high accuracy for classification among current algorithms and runs efficiently on large databases [1]. Also, RF can give estimates on what variables are important in the classification which can be effective to explain the predictions.

Evaluation Metrics

Based on the Kaggle competition guidelines, the evaluation metric for this project is the area under the receiver operating characteristic (ROC) curve relative to the detection of customers from the mail campaign [2]. A ROC is a graph used to plot the true positive rate (TPR, the proportion of actual customers that are labelled as so) against the false positive rate (FPR, the proportion of non-customers labelled as customers).

Project Design

1) Understand the Data structures and additional files.
2) Data Preprocessing – Data cleaning, filling the missing data, formatting or regrouping selected features.
3) Dimensionality reduction (using PCA) followed by unsupervised learning model for customer segmentation. For this k-means clustering method will be used.
4) Apply Supervised learning algorithms on the training data to tune a model to identify customers for the company successfully. Algorithms like decision trees, boosting methods and SVM will be used. Finally, compare the performance of the selected model with the benchmark model.
5) Apply the selected model on the test dataset to get predictions.
6) Submit the predictions to the Kaggle competition.

References

[1] Breiman L, Cutler A. Random Forests. [accessed 2020 Jul 11]
https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[2] Udacity+Arvato: Identify Customer Segments. [accessed 2020 Jul 11]
https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/evaluation