Capstone Proposal

Customer Segmentation Report for Arvato Financial Solutions

Project domain background

Arvato is a global services company headquartered in Gütersloh, Germany. Its services include customer support, information technology, logistics, and finance. The history of Arvato goes back to the printing and industry services division of Bertelsmann; the current name was introduced in 1999. Today, Arvato is one of eight divisions of Bertelsmann, the media, services and education group. In 2016, Arvato had about 68,463 employees and an overall turnover of 3.84 billion euros.

In this project, Arvato helps a mail order company to understand the demographics of the population in Germany and match them with their customer data, in order to estimate whether a person is a potential customer or not, based on their demographics.

Problem statement

In summary, the problem is: How can you use demographic data to estimate whether or not a person may be a potential customer?

Datasets and inputs

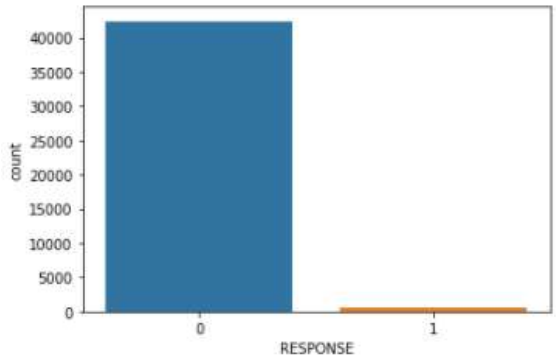
There are four data files associated with this project, all of which are provided as Machine Learning Engineer Nanodegree Program curriculum material:

- 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).

Additionally, 2 metadata files have been provided to give attribute information:

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

It is important to note that in the MAILOUT_TRAIN dataset (composed of 42962 records), the target is highly unbalanced (42430 with value 0 and 52 with value 1). This is important as it effects the sampling and metric choice.



Solution statement

First, a pre-processing and Univariate Analysis stage will be performed on the AZDIAS dataset to clean and explore the demographics data.

Next, a Principal Component Analysis (PCA) will be performed to detect the features that best explain the dataset (we will try to keep between 10 and 15 features), an optimal value of K will be selected, and then a Kmeans estimator will be fitted. With this trained estimator, clusters will be obtained for the AZDIAS and CUSTOMERS dataset, then calculate a ratio between customers and general population in each cluster, and finally, we will analyze feature weights in the clusters with the highest percentage of customers.

In the last part of the project, we will use the MAILOUT_TRAIN dataset, select the features with a less restrictive PCA, split data in train and validation, set a benchmark, test different classification estimators, select the estimators that give the best results for hyperparameter tunning, and finally, we select the best model.

Benchmark model

To obtain a first starting point, we will fit a simple unbalanced estimator with unscaled data, for example, a Logistic Regression estimator.

Evaluation metrics

In the customer segmentation problem, we will use the weight in each cluster of the main features to assess which features are most influential on our customers.

In the second problem, we will use the Accuracy, Confusion Matrix and Area under the Receiver Operating Curve (AUROC)

Project Design

To explain the project design in more detail, please find attached the file project_design.pdf