Machine Leaning Capstone Proposal

Topic: Customer Segmentation – Arvato Financial Solutions

¶

Author: Li Tang

Date: 10/18/2020

I. Domain Background

Arvato is a global service company, providing financial services, information technology services, and supply chain management solutions for business customers. Arvato's customers come from a variety of industries from insurance companies, e-commerce, energy providers, IT, and Internet providers.

A major buiness of Arvato is data analytics, helping its customers get valuable insights from data to gain insights and facilitate making business decisions.

For this project, Arvato helps a Mail-order company selling organic products in Germany to understand its customer segments in order to identify next possible customers. The existing customer data and the demographic population data in Germany will be combined to study customer segments. Afterwards, a model will be built to predict a future potential customer.

II. Problem Statement

Main statement: identifying a potential new customer of Mail-order.

Part 1: customer segment pattern will be modeled with a unsupervised machine leaning algorithm and therefore the demograhic feature/pattern of the existing Mail-order customers could be discovered.

Part 2: a potential new customer could be predicted with a supervised machine learning algorithm based on his/her demographic data.

III. Dataset and Inputs

Four data files associated with 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).

Note: AXDIAS file is general demographic data and CUSTOMERS file has 3 additional columns with their customer information to Mail-order company.

Two meta-data files have been provided for attribute information:

- DIAS Attributes - Values 2017.xlsx: data values for each attribute/feature in an alphabetical order
- DIAS Information Levels Attributes 2017.xlsx: overview of a list of attributes

Note: TRAIN and TEST files will be mainly used for supervised machine leaning algorithms.

IV. Solution Statement

Part 1:

- Data exploration and data cleaning (missing data)
- Data preprocessing (categorical data re-encoding and data scaling)
- Feature engineering (e.g., PCA)
- Customer segmentation with unsupervised machine leaning algorithm (e.g., K-means clusting)

Part 2:

- Data exploration and data cleaning (missing data)
- Data preprocessing (categorical data re-encoding and data scaling)
- Feature engineering (PCA)
- A machine leaning model with supervised algorithm (e.g., Logistic Regression, Decision Tree, Random Forest, XG Boost)
- Model hyperparameter tuning
- Potential new custmer prediction with the above supervised machine leaning model

V. Benchmark Model

Logistic regression model will be used as a benchmark model for this project, as it is relatively easy and simple.

VI. Evaluation Matrics

Part 1:

- The "explained variance ratio" of each feature will be used to select the minimum number of dimensions for featur engineering using PCA.
- The "squared error" will be used to evaluate K-mean clustering, a unsupervised machine leaning algorithm.

Part 2:

- "accuracy", "confusion matrix", and "area under the receiver operating curve" will be used to evaluate the supervised algorithms.

VII. Project Outline Design

1. Data Exploration, Cleaning, and Visualization
2. Data Preprocessing (categorical data re-encoding and data scaling)
3. Feature Engineering
4. Model Building
5. Medel Hyperparameter Tuning
6. Prediction