Customer Segmentation Proposal for Arvato Financial Services

Ellen Zhang (Yarong.zhang@gmail.com)

Domain background

This project is the capstone project for Machine Learning Nano-degree of Udacity[1].

In this project, we will analyse demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population.

Then we will use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then, we'll apply what we learned from former step to a third dataset with demographics information for targets of a marketing campaign for the company, and use a model to predict which individuals are most likely to convert into becoming customers for the company.

The data for this project is provided by Bertelsmann Arvato Analytics, and represents a real-life data science project.

Problem statement

1. Customer Segmentation Report

Access the AZDIAS and CUSTOMERS dataset by using unsupervised learning methods to analyse attributes of established customers and the general population in order to create customer segments.

2. Supervised Learning Model

Use the extracted features from the part 1 to build a machine learning model that predicts whether each individual will respond to the campaign, based on the MAILOUT training dataset.

3. Kaggle Competition

Use the model built in the second part to make predictions on the campaign data (MAILOUT test dataset) as part of a Kaggle Competition.

Datasets and inputs

There are four data files associated with this project[2]:

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).

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).

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighbourhood.

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 our final predictions will be assessed in the Kaggle competition.

Otherwise, all of the remaining columns are the same between the four data files.

In addition to the above data, there are two additional meta-data:

5. DIAS Information Levels — Attributes 2017.xlsx: a top-level list of attributes and descriptions, organized by informational category.

6. DIAS Attributes — Values 2017.xlsx: a detailed mapping of data values for each feature in alphabetical order.

Solution statement

1. Data pre-processing is the first and essential step, mainly to deal with the missing/unknown value in the dataset and re-encode the features. I will abide by the workflow of data analysis learned from DAND to pre-processing those datasets.

2. Customer Segmentation Report

In this part, first of all, I will use the most common dimensions reduction method is PCA (Principal Component Analysis) to try to reduce the dimensions of the dataset. And second of all, I will use the K-means clustering method to perform unsupervised learning to separate the general population into a few groups.

3. Supervised Learning Model

I am going to use ensemble methods. Use a few different Regressor like Light Gradient Boosting Regressor, XGBoost Regressor, Support Vector Regressor, etc. And then stack them up and optimize them, remove the weakest model. Blend them together to get the best performance.

Benchmark model

In this project [kaggle](https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard) competition public leader board, the best score(AUC) is 0.80819. The article[3] described the champion model.

The benchmark model's key parameters as below:

1. Drop columns with missing value higher than 20%
2. Keep 260 principal components

3. The cumulative variance is more than 95%
4. 12 clusters
5. Regressor: LGBMRegressor
6. 5 folds validation

I will implement the benchmark model first and then make some improvements.

Evaluation metrics

In the segmentation part, explained variance ratio is be used in the PCA process.

$$\frac{\sum_n s_n^2}{\sum s^2}$$

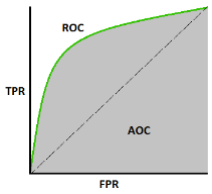
[image source] (https://github.com/udacity/ML_SageMaker_Studies/blob/master/Population_Segmentation/Pop_Segmentation_Exercise.ipynb)

Explained variance accounts for the ability to describe the whole feature variance, the more the explained variance, the more import of the component.

In the supervised model prediction parts, mean squared log error and AUC are used as main metric.

$$L(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{N}(log(y_i+1) - log(\hat{y}_i+1))^2$$

[image source] (https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/mean-squared-logarithmic-error)



[image source] (https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5)

Project design

1. Exploratory Data Analysis (EDA):

To get some initial insight from the data. Data cleaning, data wrangling, data visualization will be done.

2. Dimensionality reduction with PCA:

It is very hard for K-means to figure out which features are most important in a higher dimension. So before clustering this data, PCA will be hired to reduce the number of features within a dataset. Try to retain the "principal components".

3. Clustering data with k-means:

Use the unsupervised clustering algorithm, K-means to segment customers using their PCA attributes. Will use 'Elbow Graph' to guide me to find a "good" K.

4. Supervised modelling:

Define and train a binary logistic classifier to effectively separate two classes of MAILOUT data.

Reference:

1. Udacity Machine Learning Nano-degree: https://www.udacity.com/course/machine-learning-engineer-nanodegree--nd009t
2. Bertelsmann/Arvato Project Workspace: Arvato Project Workbook.ipynb
3. Shihao Wen: Investigating Customer Segmentation for Arvato Financial Services https://medium.com/@shihaowen/investigating-customer-segmentation-for-arvato-financial-services-52ebcfc8501