Machine Learning Engineer Nanodegree Capstone Report

Customer Segmentation – Arvato Financial Solutions

Li Tang

November 05, 2020

Contents

I. Problem Definition

1. Problem 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 including insurance companies, e-commerce, energy providers, IT, and Internet providers.

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

2. Dataset

There are 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: AZDIAS file is general demographic data and CUSTOMERS file has 3 additional columns with their customer information to Mail-order company. These files will be mainly used for unsupervised machine learning algorithms.

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

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

3. Problem Statement

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

Part 1: customer segment pattern will be modeled with an unsupervised machine leaning algorithm and therefore the demographic 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.

4. Evaluation Metrics

Part 1: Customer segmentation using unsupervised learning

The dimensionality reduction technique PCA was used to reduce the number of feature dimensions. The explained variance ratio of each feature could be the reference in selecting the number of dimensions for the later use. The minimum number of dimensions explaining as much variation as possible in the dataset was chosen. Also, in case of segmenting the customers into different clusters, an unsupervised learning using K-Means clustering was performed. Also, in this case the number of clusters is selected on the squared error i.e. the distance between all the clusters with the help of an elbow graph.

To be specific:

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

Part 2: Customer identification using supervised learning

The second task is to predict a new customer for the Mail-order company. The provided training data will be split into training, validation, and test datasets. The supervised machine learning model will be trained on the training dataset and the validation dataset will be used to prevent overfitting. After model training, the model will be evaluated on the test dataset. The evaluation metrics for classification can be used in this step.

In this study, "Accuracy" and "Confusion matrix" will be used to evaluate the supervised algorithms.

II. Problem Analysis

1. Exploratory Data Analysis and Preprocessing

1.1 Fixing mixed type columns

Columns 18 and 19 (i.e., 'CAMEO_DEUG_2015' and 'CAMEO_INTL_2015') contains mixed features and some mis-recorded values. Therefore, the feature type was changed to one common type and the mis-recorded values – 'X', 'XX' were replaced with NaN values.

1.2 Fixing "unknown" values

All the "unknown" values were replaced with NaN values.

1.3 Addressing non-existent values in 'LP_*' columns

These 'LP_*' columns give the information about a person's family status and financial status they are in.

- These columns with '0' do not correspond to any category specified in the Attribute information data. These columns with '0's have been converted to NaN values
- The 'LP_LEBENSPHASE_FEIN' and 'LP_LEBENSPHASE_GROB' have too much granular information. The FEIN data consist of fine information about life information and finance information. This information has been divided to represent finance information as one feature and life information as one feature and was saved into the same two columns.
- The columns 'LP_FAMILIE_FEIN' and 'LP_STATUS_FEIN' have been dropped because they contain duplicate information with the corresponding '_GROB' columns

1.4 Features Re-encoding

The below features have been re-encoded:

- EINGEFUGT_AM: This column represents the date on which the person has joined or the date the entry was made. This column has been converted to datetime column and only year has been extracted as a feature.
- ANREDE_KZ: This represents the Gender, which was encoded with values 1,2 for male and female, is re-encoded to contain 0-male and 1-female
- CAMEO_INTL_2015: This column contains information about the status of a person according to international standards. This column has been divided into two different columns to consist information about International Family status and International Wealth status.
- WOHNLAGE: This column also has mis-recorded values, which were replaced with NaNs
- LNR: This column corresponds to an ID given to each person and this feature has been neglected during the analysis process.

1.5 Missing values

- Columns: the percentage of “NaN” values in each column was analyzed. The columns which had missing values in customers data also have missing data in the general population data. The distribution of the missing data per column is similar between these two. Therefore, the columns that have more than 30 percentage missing values were dropped from both customers data and general population data. A total of 11 columns have been dropped in this step.

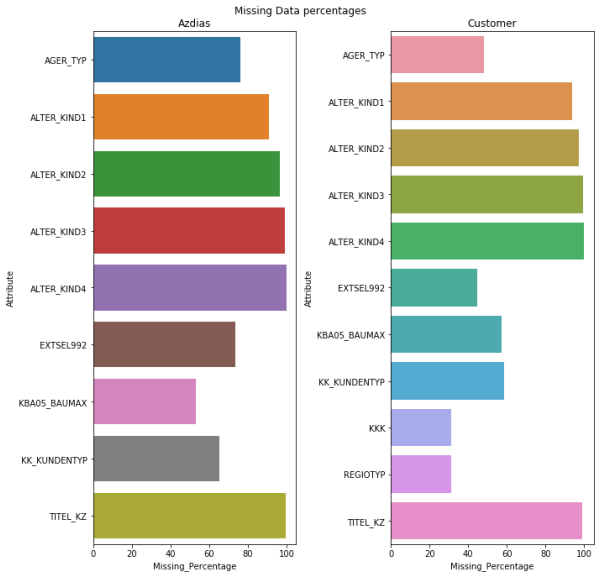

Figure 1. Columns with more than 30% missing values

- Rows: the number of “NaN” values per row was analyzed afterwards. Any row with more than 50 missing features was dropped with a total of 1,53,933 rows from general population data (8,91,211 original rows in total). Similarly, a total of 57,406 rows were dropped from customers data (1,91,652 original rows in total).

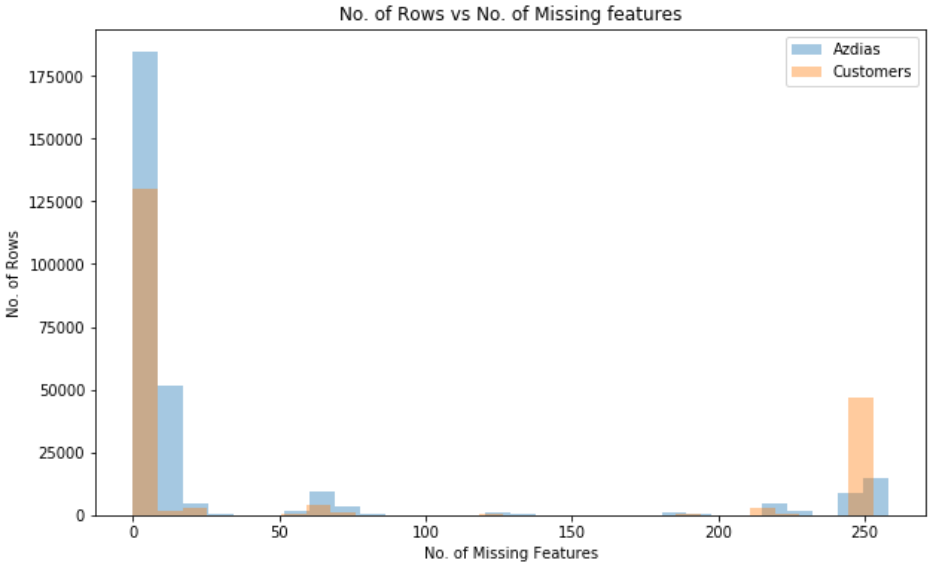

Figure 2. Missing values distribution before removing rows with > 50 missing values

1.6 Imputing missing values

After removing the columns and rows which “NaN” values, the dataset still has some remaining “NaN” values in range of 1-50, as shown in Figure 3. These “NaN” values have been replaced with the most frequently occurred observation in each of the corresponding feature.

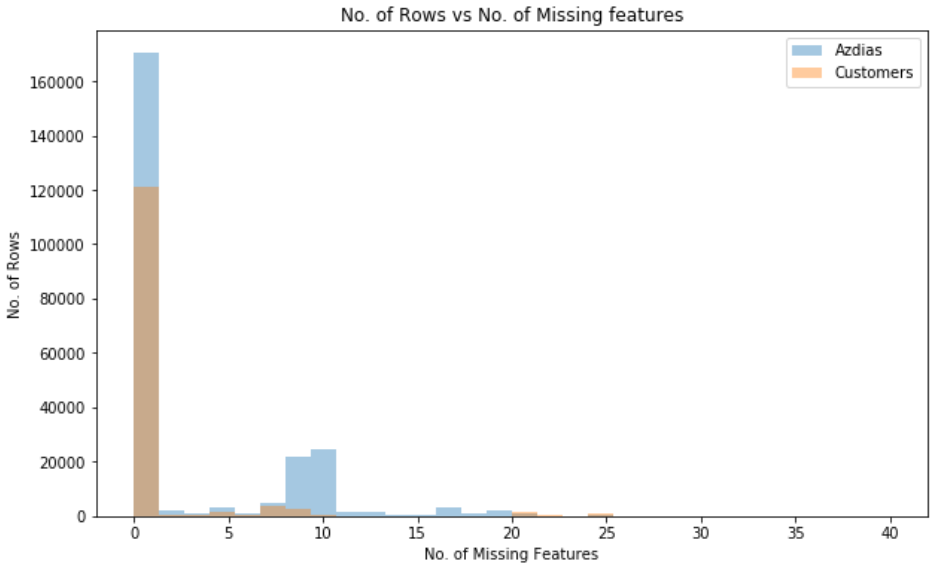

Figure 3. Missing values distribution before removing rows with 1- 50 missing values

1.7 Feature scaling

A standard scaler was used to standardize all the features to the same range. It is an important step to eliminate feature dominance when applying dimensionality reduction.

III. Machine Learning Process

1. Customer Segmentation

1.1 Dimensionality reduction using PCA

PCA was performed on the given data to reduce the number of dimensions. There were 352 features after the data cleaning and feature engineering step, it is important to understand which features will be able to explain the variance in the dataset.

Based on the calculation shown in the Notebook ("unsupervisedML_PCA_K-Means.ipynb"), although we have 352 features almost 75% of the variance in the data can be explained with the help of 98 components of PCA. With this step we will be able to reduce the number of features from 352 to 98.

1.2 PCA component analysis

These 98 components can be further explained by looking at the feature weights the PCA algorithm has given to the original features. For example, the component '0' explanation is shown in Figure 4.

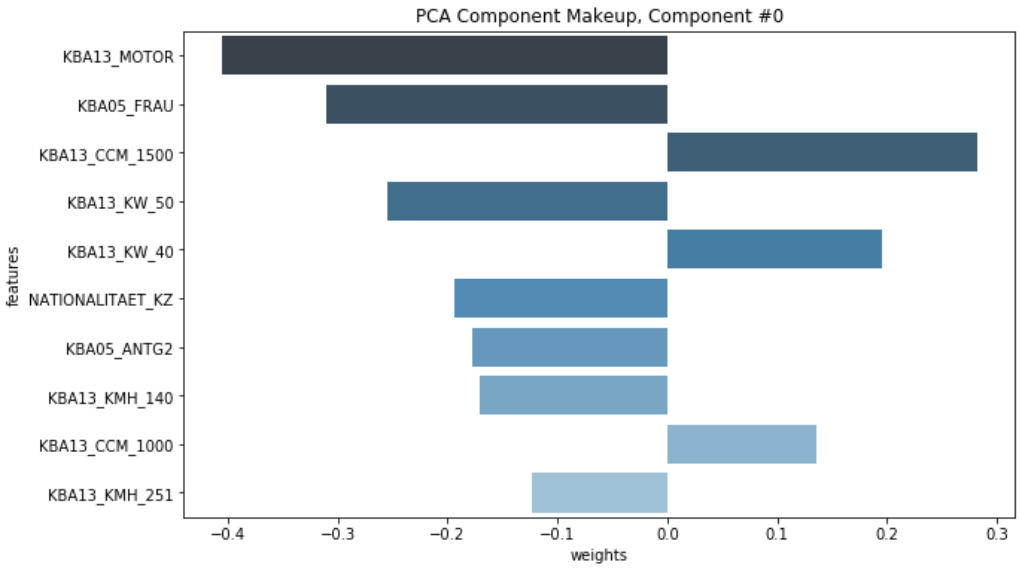

Figure 4. PCA component 0

The PCA component 0 has a high positive weight to KBA13_CCM_1500, followed by KBA13_KW_40 and then KBA13_CCM_1000. On the other hand, component 0 has a high negative weight to KBA13_MOTOR, followed by KBA05_FRAU and KBA13_KW_50.

1.3 Clustering using K-Means

After the feature reduction using PCA, the following step is to cluster the general population and customers into different segments. K-Means clustering algorithm has been chosen for this task due to its simplicity and its ability to measure the distance between two observations to assign a cluster. This algorithm will help us in separating the general population with the help of the reduced features into a specified number of clusters. And use this cluster information to understand the similarities in the general population and customer data.

The number of clusters is a critical hyperparameter for the clustering algorithms. The basic idea behind the clustering algorithms is to select the number of clusters to minimize the intra-cluster variation, with the points in one cluster are as close as possible to each other. In this study, an elbow graph was used to decide the number of clusters for the K-Means algorithm. The elbow graph plots the Sum of Squared distances in each cluster for the specified list of number of clusters.

This plot helps in understanding how the number of clusters affect the intra-cluster distances. The optimal number of clusters can be the number where the sum of squares of distances starts to plateau. The number of clusters in this case is chosen to be ‘6’, since the sum of squares of distances stops decreasing at a higher rate at this point as shown in Figure 5.

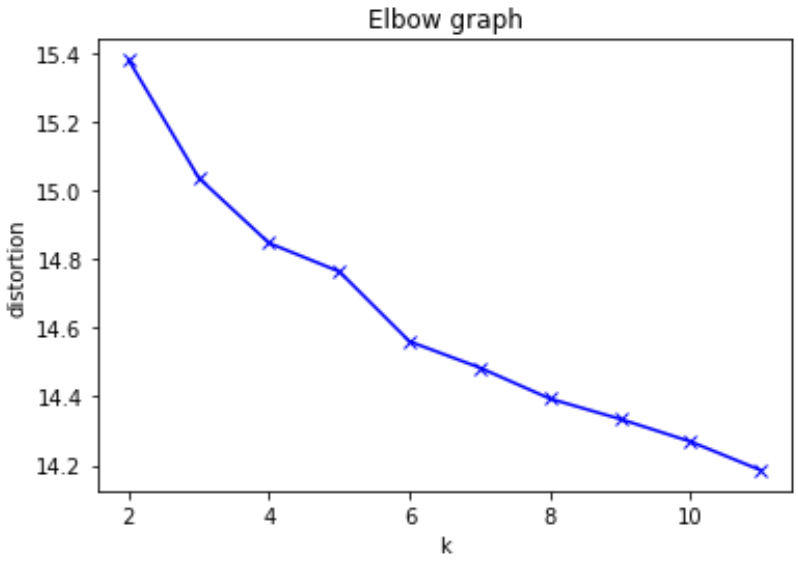

Figure 5: K-Means clustering elbow graph

1.4 Cluster analysis

The general population and the customer population have been clustered into segments with K-means. Figure 6 represents the proportions of population coming into each cluster. The cluster distributions of the general population and the customers are relatively similar, meaning that the general population and customers have been similarly clustered into 6 segments. The top three populations seem to be coming from the clusters '0', '1', and '3'.

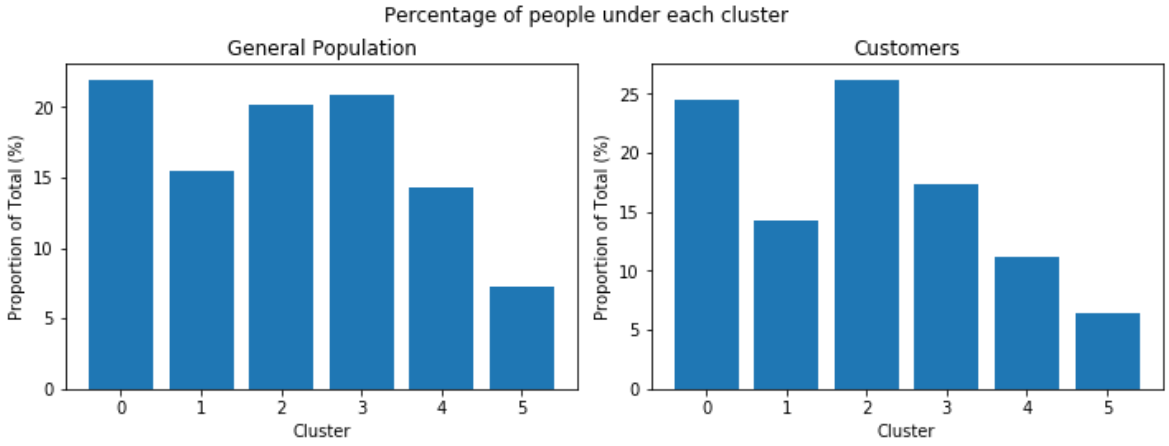

Figure 6: Cluster proportions

We can further confirm this by taking the ratio of proportions of customers segments and general population segments as shown in Figure 7. As seen in Figure 7, if the ratio of proportions is greater than 1 that means this cluster has a greater number of customers in the existing population and has a potential to have more future customers. Whereas if the ratio is less than 1 that means these clusters have the least possibility to have future customers.

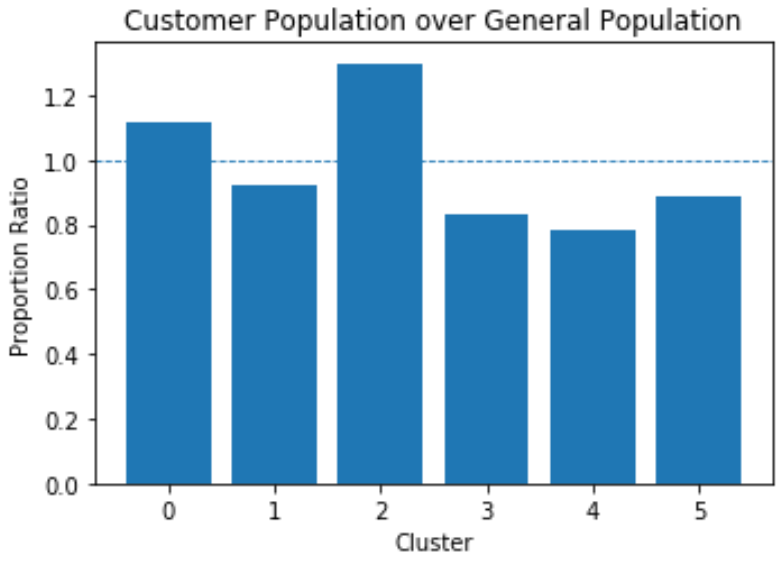

Figure 7: Cluster proportion ratio

2. Customer Identification

A supervised learning algorithm was used to predict whether a person will be a potential customer based on the demographic data. The file 'Udacity_MAILOUT_052018_TRAIN.csv' was provided with the same features as the general population and customers demographic data. An extra column 'RESPONSE' has been provided with this data. The response column indicates whether this person was a customer or not. This data has been cleaned by following similar cleaning and processing steps that were followed for general population and customer data.

The train dataset from the cleaned version of "Udacity_MAILOUT_052018_TRAIN.csv" was split into three parts: training, validation, and test. Training/validation datasets were used for supervised machine learning model training, while test dataset was used for the evaluation of the trained model using the designated metrics: "Accuracy" and "Confusion matrix" in this case.

2.1 XGBoost algorithm supervised model training

XGBoost is a well-known supervised machine learning algorithm based on gradient boosted decision tree for its superior speed and performance. Recently, it has been dominating machine learning for structural data.

Therefore, this algorithm, supplied by AWS SageMaker, was selected in this study.

IV. Result Interpretation

1. Trained XGBoost model evaluation

The “Accuracy" and "Confusion matrix" were used for evaluating the supervised model with the split test data. The detailed calculation can be referred to “supervisedML_XGBoost.ipynb” notebook.

Of the 4296 response, we've correctly predicted 4246 of them (true positives). And we incorrectly predicted 1 that would respond but then ended up not doing so (false positives). There are only 1 that ended up respond, that we predicted would not (false negatives).

predictions	0	1
actual		
0	4246	1
1	49	0

2. Prediction on Test dataset

The final prediction was performed on the test data which was provided in the file ‘Udacity_MAILOUT_052018_TEST.csv’. The same cleaning and pre-processing steps were performed to clean the data. This data was scaled with the scaler which was fit on the training data.

After prediction, the results were re-organized to have the “LNR” as the ID index and saved into a csv file.

V. Conclusions and Future Steps

For the first part, with the data pre-processing/feature engineering steps and unsupervised machine learning algorithm K-Means, potential customers falling in certain clusters can be successfully identified for Mail-out company.

For the second part, a good prediction using our supervised trained model based on XGBoost algorithm has been achieved.

However, there is still scope for improvement including:

- To better understand features and select more relevant ones
- To perform hyperparameters tuning of XGBoost model