Capstone Proposal

Machine Learning Engineer Nanodegree

Nicolina Wenzler, August 2020

Domain Background

Arvato offers financial solutions in different ways, as e.g. payment processing or risk management activities. This is also the field of this capstone project. Arvato aims to use its available datasets to support a new client base. To achieve this, the datasets will be explored to identify relevant attributes and demographic features that can help segment customers of interest.

With the possibility to gather high-dimensional of the customers and understand and explain it, a variety of possibility arise. First, the customer's decisions can become more transparent. This is ambitious, but once decisions can be explained, also future behavior can be modeled and predicted.

A general customer segmentation is a relevant step for companies that already have a broad basis of customers and aim to gain more information about their customers. This information can be used for personalization and targeted address and advertisement for specific groups. For example, a specific new or existing customer could only be addressed with products that are of interest for him, measured by his peer group/customer segment.

In the use case addressed here the knowledge gained from data shall be used to predict, whether a customer will respond to a specific marketing campaign or not. With this, the effectiveness of a campaign can be increased by addressing only promising potential customers.

Customer segmentation is a relevant and interesting task for many companies. So there are many papers published in this area, addressing different challenges from different industries. A general starting point for exploring the data and searching for pattern is described in the following blog post by Riley Predum (Predum, 2020).

Problem Statement

The goal that shall be achieved here is to identify new customers with a high potential to respond to a mailing campaign. With this, the costs for the campaign could be reduced and targeted to those customers with a high probability of resonance.
The problem is that often marketing campaigns are rolled out to a broad variety of new customers. As customers differ in many dimensions and have diversive needs.
The effectiveness of a campaign can be increased by addressing only those customers who are willing to respond in a given format.

Datasets and inputs

The datasets being used to find a solution to the problem statement are the ones provided by Bertelsmann/Arvato in the context of the Udacity Machine Learning Engineer Nanodegree. They are also published in the Kaggle challenge: https://www.kaggle.com/c/udacity-arvato-identify-customers.

There are four datasets:

- `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).

In addition, there are two metadata files for these datasets:

1. DIAS information levels - attributes2017.xlsx
2. DIAS Attributes - Values 2017.xlsx

Solution statement

As a preprocessing step to be able to identify a new customer as potentious or not, the existing customers shall be segmented into meaningful groups. So different attributes will be taken into account and groups will be formed. In the next step, a supervised model will be trained. It will be trained to recognize, which of the identified groups of customers responded to the mail campaign. To divide these two groups of potential customers, different attributes will be examined to find out which are the ones that probably have an impact to the target variable (whether a customer respond to the mail campaign or not).

As a first step in data preprocessing dimensionality reduction is applied, using PCA.
To cluster the customers with the reduced dimensions, an unsupervised clustering method is used. K-means often is the method of choice. K-means is the method-to-go for clustering tasks, as it is easily and fast implemented and shows good results. Iteratively, the data points are assigned to clusters and the centroid of the clusters are computed with the data points. It works iteratively and hence the clustering improves over time. Once the data points do not change clusters anymore the algorithm is called to have *converged* and hence the assignments to the cluster can be seen as final (Bishop, 2006).
The crucial point when using K-means usually is to determine a k, the number of clusters. In this project it will be determined by using the “elbow”-method. Here the algorithm is run until convergence with different numbers of k and the mean distance to the centroid as a function of k is plotted. The k where the graphs shifts is the k to take. It can be validated using crss validation. (Trevino, 2020)
To train the model to detect, whether a group is a target group for a mail campaign, supervised methods are using to train the model. One possibility is to use logistic regression.

Benchmark model

As a benchmark, a linear regression model will be trained on the dataset.

Evaluation metrics

To evaluate the quality of the clustering algorithm, the similarity between two assignment is measured. This can be done using the adjusted Rand index (scikit-learn, 2020).
For the second part, where the model is trained with supervised methods to learn the response rate of the mail campaign for a specific group, precision, accuracy and recall will be taken into account.

Project design

The steps to approach this challenge are:

1. Data understanding: The first step always is to get an idea of what the data looks like and understand the structure.
2. Data cleanup and visualization: Afterwards, data is checked for missing values and miscorded values. Data might be neglected and cleaned. With some visualizations, broad patterns can be identified.
3. Feature engineering: Reduction of the dimensions without losing too much variance. A possible method to do this is PCA.
4. Model selection: With K-means clustering algorithm, data can be segmented into clusters. In the next step, supervised techniques are used to train and evaluate whether a personal shall be addressed or not. Here, logistic regression, decision trees, random forests and gradient boosted classifiers are possible methods.
5. Model tuning: The most promising model from step 4 is taken and by tuning its hyper parameters, the performance will be optimized.
6. Test and predict: The best model will be taken to make prediction on the test data.

Literaturverzeichnis

Bishop, C. M. (2006). *Pattern recognition and machine learning.* Springer.

Predum, R. (15. 08 2020). *Customer Segmentation Analysis with Python*. Von https://towardsdatascience.com/customer-segmentation-analysis-with-python-6afa16a38d9e abgerufen

scikit-learn. (18. 08 2020). *Clustering*. Von https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation abgerufen

Trevino, A. (18. 08 2020). *Introduction to K-means Clustering*. Von https://blogs.oracle.com/datascience/introduction-to-k-means-clustering#:~:text=The%20K%2Dmeans%20clustering%20algorithm%20is%20used%20to%20find%20groups,groups%20in%20complex%20data%20sets. abgerufen

Wiedenbeck, M. &. (2001). Klassifikation mit Clusteranalyse: Grundlegende Techniken hierarchischer und k-means-Verfahren.