Capstone Proposal

Machine Learning Engineer Nanodegree

Nicolina Wenzler, August 2020

Domain Background

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 could a specific new or existing customer 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.

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 divers 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:

1. Udacity_AZDIAS_052018.csv
2. Udacity_CUSTOMERS_052018.csv
3. Udacity_MAILOUT_052018_TRAIN.csv
4. Udacity_MAILOUT_052018_TEST.csv

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 shall learn, 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. Afterwards the customers will be clustered.
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 benchmark, the suggestions is Gradient Boosting Classifier of a similar dataset which has a performance of about 80%

Evaluation metrics

For the first part of the solution, where unsupervised methods are used, the amount of explained variance can be used to evaluate the usefulness of PCA. Here, the explained variance shall remain high while the dimension of the data decreases.

For the second part, where the model is trained with supervised methods to learn the repondance 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.