arvato
BERTELSMANN
Financial Solutions

Udacity Machine Learning Engineering Nanodegree

Capstone Project Proposal

Arvato

Daniela Aragaki

Table of contents

Domain background

Arvato is a company that focuses on development and implementation of innovative solutions across the globe. Their main expertise are Supply Chain solutions, Financial services and IT services, where automation and data/ analytics plays an important role on innovation and development of news solutions.

The company is owned by Bertelsmann, a media, services and education company operating in more than 50 countries.

Being innovative has been crucial for many companies, and Arvato Financial Services can optimize the financial performance of their clients using data-driven, cutting-edge solutions. The use of machine learning can help companies discover hidden patterns and understand more their customer behavior, allowing them to make better decisions about their products and services.

Problem statement

In this project we will dive into a real-life data science problem. Our goal is to help Arvato acquire new customers in Germany, so this problem can be split into the following ones:

- How can we describe Arvato's customer base?
- How can we identify which targeted individuals are more likely to become a customer?

To describe Arvato's customer base, we will have to compare their current customer database against information of the general population. The second question can be answered by finding patterns from customers and later checking them against the targeted individuals.

Both problems can be solved using machine learning techniques.

Datasets and inputs

The datasets used in this project were provided by Udacity. They comprise of demographic data for the whole german population, demographic data from Arvato's customer database, demographic data from a marketing campaign with and without the target response whether they became a customer or not. We also used information from two files containing metadata. The list below details the datasets used in the analysis.

Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (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 neighborhood.

Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns). The file contains three extra columns

(CUSTOMER_GROUP, ONLINE_PURCHASE, and PRODUCT_GROUP), which provide broad information about the customers depicted in it.

Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns). It includes the RESPONSE target, indicating if it became a customer or not.

Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

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

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

Solution statement

In order to solve the problem, we will break it in three parts.

First, we will have to get to know the data, explore it, identify missing data and treat it. It is difficult to state how the missing data will be handled. It depends on the % of missing data for a variable, total amount of variables with high % of missing data as well as their importance.

The second step is to segment the data. To be able to use an unsupervised learning technique, the data will need to be processed a bit more, this means we will have to encode some variables and standardize them. Some algorithms require this, otherwise the results might be misleading. Applying feature selection and/or PCA should be considered, as the dataset contains 366 features. The number of clusters should not be arbitrary. There are a few different methods which can be applied to determine the optimal number of clusters, and in this project we will explore only one, the elbow method. Once we have determined the optimal number of clusters, then we can train and predict which cluster each individual belongs to.

In the last part we will apply a supervised learning to the MAILOUT dataset and predict whether an individual is likely to become a customer or not. Additionally we will test our predictions in a Kaggle competition.

Benchmark model

The benchmark model will be a logistic regression, since it is a simple model easy to be trained. Besides that, we will use a decision tree classifier. This is a good strategy to compare performances and decide which model is more suitable for the problem.

Evaluation metrics

Different evaluation metrics will be used in this project, as different approaches will be applied during the development.

Customer segmentation: being an unsupervised learning technique, we will apply K-Means with different numbers of clusters and verify the relation between the number of clusters and inertia. Using the elbow method, we can then select the proper number of clusters.

Customer acquisition: we have a classification problem, so we should be looking at accuracy, how the confusion matrix is, and AUC ROC. Depending if the target variable is quite unbalanced or not, the relevance of one of those metrics might change.

Project design

Here is a brief explanation of the steps of the project:

Data Cleaning and Visualization

In this section we aim to explore the data we have, how is the quality, identify missing values, outliers and define and apply strategies to handle them.

Feature Engineering

In this section we aim to prepare the dataset for modelling, creating other features, selecting the features that are relevant, encoding categorical variables, standardizing the data, and applying PCA if needed.

Modelling

This section will contain modelling with different algorithms.

An unsupervised learning algorithm in the customer segmentation problem, where we will evaluate a different number of clusters.

Supervised learning algorithms for the customer acquisition problem, using a few different algorithms for classification, such as logistic regression and random forest, from which we will decide which algorithm is more suitable for the problem.

Hyper parameter tuning

Having selected an algorithm and seen its performance, we will further work on it to improve it, tuning its parameters.

Prediction

In the last section we will predict the target variable in the test dataset provided, and upload the predictions to the respective Kaggle competition.

References

[1] Arvato-Bertelsmann, "Arvato". Available at https://www.bertelsmann.com/divisions/arvato/#st-1