Machine Learning Engineer Nanodegree

Capstone proposal

Thorsten László Schmid
October 24, 2020

Proposal

Domain Background

Many companies in the B2C sector are utilizing Targeted Advertising to boost their business. 'Targeted advertising is a form of advertising, including online advertising, that is directed towards an audience with certain traits, based on the product or person the advertiser is promoting.'[1] In other words, the objective is to bring together the product with the people likely to buy it.

Arvato, a German services company (more details see [2]) is dealing with this issue in one of its projects as well. The client is a mail-order company selling organic products. Their objective is to acquire new clients more efficiently by reaching out to the people identified as becoming most likely new customers. Therefore, in a second step Targeted Advertising can take place.

Problem Statement

The task is to figure out which people are most likely to become new customers (see [3]). Keeping in mind that we have a high-dimensional problem the solution consists of two parts:

1. Get a feeling on which attributes describe a typical customer of the company. We will generate a Customer Segmentation report for this purpose. The dataset of the company's clients will be matched against a bigger set of Germany's inhabitants.
   Problems of that kind can be typically solved by applying unsupervised learning techniques. We could use a combination of Dimensionality reduction (e.g. by PCA) and a Clustering Algorithm (e.g. KMeans).

2. Build a model that can predict whether a person is likely to become a customer or not. That means we have a binary classification problem. The Data science toolbox contains many items to solve supervised learning tasks of this kind. For example, Logistic Regression, Tree ensemble approaches like Random Forest or Deep Neural Nets.

Datasets and Inputs

We are dealing with demographic data having many features. 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. As the term 'demographic' might not suggest the data also provides information about things like spending behavior and habits in consumption etc.
The primary Data consists of four files:

Filename	#Rows	#Cols	Description

Filename	#Rows	#Cols	Description
Udacity_AZDIAS_052018.csv	891221	366	Demographics data for the general population of Germany
Udacity_CUSTOMERS_052018.csv	191652	369	Demographics data for customers of a mail-order company
Udacity_MAILOUT_052018_TRAIN.csv	42962	367	Demographics data for individuals who were targets of a marketing campaign
Udacity_MAILOUT_052018_TEST.csv	42833	366	Demographics data for individuals who were targets of a marketing campaign

In addition two EXCEL-files '*DIAS Information Levels - Attributes 2017.xlsx*' and '*DIAS Attributes - Values 2017.xlsx*' are included that can be used as legend for the columns, their meaning and the categorical values associated to them.

Solution Statement

The Segmentation task will be solved in two steps.

1. Perform Dimensionality reduction.
   We reduce dimensionality by applying the PCA Algorithm in combination with the Elbow Method.

2. Segmentation.
   After Dimensionality reduction we will use a clustering algorithm to build segments. The standard algorithm is KMeans. Nevertheless, it could be worth looking at others like DBSCAN (see [4] Chapter 3).

A suitable Model will solve the Prediction Task. E.g.:

- Logistic Regression
- XGBoost
- Deep Neural Net

In addition, Model Improvement techniques like Cross Validation, Regularization or tuning of Hyperparameters will be applied depending on the model and its scoring.

Benchmark Model

The Benchmark Model will be the RandomForestClassifier from the Scikit-Learn Package as it is some Kind of intermediate approach regarding the score, at least due to my experience.
Furthermore, there is a Kaggle competition we will participate in.

Evaluation Metrics

Segmentation.
Well we have an unsupervised learning task. So it is hard to give a metric as we don't know what is the truth regarding such tasks. At least we can use the explained Variance to help us select essential components in PCA.

Classification.
As mentioned before we have a Binary Classification problem. We can use the following performance metrics for binary Classification (see [4] Chapter 5).

- Accuracy
- Precision
- Recall
- F1 Score
- Receiver Operating Characteristics curves especially the area under the Curve (AUROC)

We will make heavy use of Accuracy, Precision, Recall and F1 Score. As Kaggle requires AUROC we will compute it at least for the most sucessful model.

Project Desing

The following steps make up the Project Design:

1. Data analysis and pre-processing.
    - Treatment of missing values
    - Identifying and handling outliers
    - Encoding of categories
    - Scaling
2. Apply PCA
    - Use explained Variance
3. Apply Clustering
    - Determine best k for KMeans
4. Setup and train the Prediction Models
    - Consider Hyperparameter tuning and Cross Validation
5. Evaluate the Models
    - Make Predictions
    - Calculate evaluation metrics
6. Choose the best Model
7. Submit the result to Kaggle.

[1]: Wikipedia, "Targeted advertising", Wikipedia. [Online]. Available https://en.wikipedia.org/wiki/Targeted_advertising [Accessed October 24 2020]
[2]: Wikipedia, "Arvato", Wikipedia. [Online]. Available https://en.wikipedia.org/wiki/Arvato [Accessed October 24 2020]
[3]: Youtube, "Arvato Final Project", Udacity. [Online]. Available https://youtu.be/qBR6A0IQXEE [Accessed October 24 2020]
[4]: Müller and Guido (2016), "Introduction to Machine Learning with Python". 1st Edition. O'Reilly