Capstone Proposal

Desared Osmanllari

March 25, 2020

1 Domain Background

In the financial world, most of the services are becoming more and more data-oriented. Arvato Financial Solutions offers all financial services related to payments and cash flow - from risk assessment to the emergence of receivables through invoicing and settlement[1]. Arvato is analysing various datasets and attributes to calculate credit scores, detecting fraud by analysing the payment behaviour, and finding the proper customers in a given population. These problems can be solved by applying machine learning.

In terms of business, the goal of a data-driven solution is to increase the efficiency. More specifically, acquiring new clients can be a difficult problem. In the past, without a data-driven solution, a mail-order company would contact the whole population. This brings difficulties in terms of cost and management. On the other hand, finding the proper clients by introducing a machine learning model can reduce the cost, improve the speed of delivery and the service accuracy.

2 Problem Statement

In this report, supervised and unsupervised learning techniques will be proposed to analyze demographics data of customers of a mail-order company in Germany against demographics information of the existing clients. Therefore, the research question that arises is:

What are the potential customers and how can the company acquire them efficiently?

[1] https://finance.arvato.com/en/

The main purpose of this project is to identify the right people, who can be potential customers. It will increase the efficiency of the customer acquisition process, by targeting the proper clients. On the other hand, without a data driven approach, the company would advertise their product to all people in Germany, which can be both unefficient and costly.

3 Datasets and Inputs

The data is provided by Bertelsmann Arvato Analytics, and it includes the general population dataset, the customer segment dataset, the mailout campaign dataset and a test dataset.

Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns). Some of the feature values are missing. The data types are categorical, binary, and numerical.
Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns). Compared to *Udacity_AZDIAS_052018.csv*, 3 more features are added in this dataset: CustomerGroup, OnlinePurchase, and ProductGroup.
Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns). Compared to *Udacity_AZDIAS_052018.csv*, the information about the reaction to the mailout campaign is added.
Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns). Same features as in *Udacity_AZDIAS_052018.csv*.

The features are described by two excel sheets:
DIAS Information Levels.xlsx: Attributes name, the description, the values and the meaning of the features.
DIAS Attributes.xlsx: More detailed information on the features.

4 Solution Statement

First, attributes of existing clients will be analysed and matched against the attributes of people living in Germany. Consequently, we can figure out which people in Germany are most likely new customers for the company. In this step, dimensionality reduction and clustering will be used to find latent features and group the customers based on their characteristics.

Then, using the training data set, a machine learning model will be trained to predict the potential new customers for the company. Since the model will be supervised, different evaluation metrics can be used to measure its performance in the validation set.

5 Benchmark Model

Some benchmark models in the same problem can be extracted from the Kaggle competition: Identify Customer Segments. XGBoost model is widely used in this competition. Some of the techniques used in one of the benchmark projects are: data normalization using Standard Scaler, dimensionality reduction using PCA, and clustering using KMeans.

The best score achieved in the competition is 0.81. By combining several techniques, especially on unsupervised learning, this score can be further improved. Clustering is a hard problem to evaluate, but there are many new approaches which can surpass K-Means results.

Secondly, using ensemble models instead of pure supervised learning algorithms can improve the performance significantly. Ensemble methods combine several machine learning techniques into one predictive model in order to decrease the variance and the bias. They also tend to have a better accuracy and higher consistency, which results in less overfitting. There are two main approaches in ensemble learning: bagging and boosting. In this project, the focus will be in boosting since it tends to teach the ensemble model.

6 Evaluation Metrics

Depending on the distribution of data, different evaluation techniques can be applied. By a first observation, the number of people responding to a campaign is pretty low. Therefore, the dataset is highly imbalanced. Accuracy, recall, precision, and F1-score can be measured in a classification problem. Since the classes are unbalanced, the F1-score will be selected as the best metric to evaluate the model.

Another approach for model evaluation can be the AUC/ROC curve. The curve is plotted by combining the specificity and sensitivity in a graph. The area under the curve shows how good the model is.

Additionally, plotting the learning curves on both training and validation set can help on predefining if the model has overfit or if it is biased. For example, a high score in the training set, associated with a low score in the validation set, can be a sign of overffiting.

7 Project Design

This project can be designed in four main sections.

1. Data Preprocessing: Data preparation, cleaning, and transformation occurs in this section. This is a very important step which helps the analysis.
2. Customer Segmentation: In this section, unsupervised learning techniques will be used to perform customer segmentation. Particularly, the general population and customer segment data sets will be analysed.

Two main steps occur in this section: dimensionality reduction and clustering. My first approach is to use PCA (principal component analysis) for dimensionality reduction, and K-Means algorithm for clustering. However, other techniques like hierarchical clustering or DBSCAN [1] will be used for population segmentation.

Determining the most optimal number of clusters is one of the most frequent problems in data clustering. Even though, several techniques can be applied to overcome this problem. The elbow method or silhouette coefficients [2] can be used to identify the most optimal number of clusters for the K-Means algorithm. Dendrograms [3] can be used to perform the most optimal split in case of hierarchical clustering. If DBSCAN is selected, there are two main parameters to be predefined: *epsilon* and *minSamples*. The KNN (k nearest neighbors) distance plot can help on identifying the most optimal value for *epsilon*. There is no theoretical basis how to select *minSamples* (minimum number of samples), but it usually depends proportionally on the data volume.
3. Supervised Learning Model: Several supervised machine learning models will be built by using the marketing campaign response dataset. These models will be later used to predict the potential future customers.

Different classifiers and ensemble methods will be tested and the best model will be selected. The ensemble methods can be adapted for both classification and regression problems. They combine several estimates from different models in order to reduce the variance of a single estimate.

Next, using Grid Search, model optimisation will be performed. Grid search helps on testing various combination of hyperparametrs using cross validation. The best set of hyperparameters will be selected after testing them on the validation set.

4. Kaggle Competition: The model will be applied to the test data. Finally, the results will be submitted to the Kaggle competition.

References

[1] Kantardzic, Mehmed. *Data mining: concepts, models, methods, and algorithms.*

[2] Celebi, M Emre and Aydin, Kemal *Unsupervised Learning Algorithms.*

[3] Phipps, JB. *Dendrogram topology.*