Customer Segmentation Report for Arvato Financial Solutions

Capstone Proposal

Vajo Lukic, September 2021

Domain background

Arvato is an international company providing services on a global scale from multiple different domains: finance, customer support, logistics and information technology. Its history goes back all the way to 1835 and it was introduced under a current name in 1999. Through mergers and acquisitions the company has grown to more than 65 000 employees worldwide and has a total turnover close to 4 billion euros.

One of many business areas where Arvato provides its services is customer acquisition. Arvato helps its customers to acquire new customers effectively and with a low cost. Arvato achieves this by providing very accurate marketing campaigns.

Machine learning has been successfully applied in the marketing domain. There are many research papers from that field and one of them is: ‘’*Customer churn prediction in telekom using machine learning in big data platform*” by Abdelrahim, Jafar and Kadan, published in the “Journal of Big Data” on 20th of March 2019.

Problem statement

The problem statement we are going to solve is to find out: “How can a mail-order company which is selling organic food - acquire new clients more efficiently?” When creating a marketing campaign, we have to decide who will be recipients of marketing material. Every time when a recipient does not respond to an offer, that increases the cost of the campaign and reduces the expected profit. For that reason, we want to narrow down the list of choices and to send offers only to those recipients who are the most likely to respond positively. But how can we know which contacts are potentially the best prospects? Well, we can use our past experiences and learn from them, or our gut feeling. Or we can also try to use the data if we have it, which will be more explained in the following sections of this proposal.

Since our goal is to try to predict whether a person is a potential customer or not, this is a pure classification problem. Inputs to our classification model will be relevant features of existing customers, and output will be a prediction - a clear “potential customer” yes/no answer, for a given list of individuals.

Datasets and inputs

We are going to use four data files with this project which were provided by Udacity and Arvato:

- 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).

By looking at the data in “azdias” (population) dataset and a customer dataset, we can see that we will be dealing with imbalanced data, since a population data set is roughly 4 times larger than the customer data set.

Solution statement

For the first part of the project, customer segmentation report, the idea is to use attributes and demographic data from existing customers of the mail order company. From this data we can obtain insights about which groups of people are most likely to buy such products. Then we can use those insights to find potential new customers that share the same traits and attributes as existing customers. In this step we will analyze, process and prepare the data for the second step.

For the second part of the project, building a supervised learning model, we will use pre-processed data from the previous step, and train different supervised classification models, trying to find the one that performs the best on this data. We will test the model performances on a validation dataset, select the best performing model, tune its parameters and try to get the maximum performance out of it.

For the third part, Kaggle competition, we will use data processing methods which we have developed in the first step, and the best performing model from the second step. Then we will apply them to the provided test data and see how the model performs.

Benchmark model

To be able to properly evaluate different models, we will have to establish some baseline to compare to. For that, we will use a Logistic regression model as a benchmark. The reason for that is that this model is one of the simplest classification models, and is easy and quick to train.

Evaluation metrics

Customer segmentation report: for this part we will use the PCA method to reduce the number of features in the dataset, so that we can apply a clustering method and segment the data. What we can measure in this step is the proportion of explained variance by principal components. Then we can choose only those components that explain the most of variance.

In order to find the optimal number of clusters with a clustering algorithm, we will use an elbow method.

Supervised learning model: here we will be dealing with a classical binary classification methodology. As mentioned, we will be dealing with imbalanced datasets. Possible choices for the evaluation metrics in that case are: F1 score, precision, recall and area under the receiver operating curve.

We will be using the area under the receiver operating curve to evaluate performance of different models because it is one of the best options for the imbalanced data.

Kaggle competition: the leaderboard of the competition is indicating the use of the area under receiver operating characteristic curve as a score.

Project design

Project will be consisted of following parts:

- **Data exploration and visualization**: we need to perform statistical analysis of the data to learn about its distributions, potential missing data etc. Both numerical and visual analysis are needed to be able to find and recognize patterns within the data.
- **Data cleaning and feature engineering**: any inconsistencies in the data values and data types need to be removed so that we can construct valuable features. Here are some of techniques that will be used for data preprocessing:
    - Since we're dealing with a classification problem, we have to convert any categorical values to numerical values.
    - We have to see if any values are labeled as "missing" or "unknown" and replace them with real null values
    - We have to replace any null values with real values, choosing the right approach (median value, most frequent value etc.) depending on the type of data
    - We'll be using the K-means clustering algorithm which is sensitive for skewed data, and requires variables with the same average values and the same variance. For that reason we have to log transform the data, scale it and standardize it.
- **Feature selection**: by applying unsupervised learning techniques we will reduce the number of features used for modelling, by choosing the most valuable features and to remove any redundant features. Here we will use techniques for dimensionality reduction like PCA (principal component analysis).

- **Model creation, selection and tuning**: we will divide data into training and validation data sets. Then we will try different types of supervised learning algorithms and evaluate how they perform on training data. We will choose the most appropriate benchmark metrics based on the shape of the features. Then we will choose the best performing models, tune them and evaluate them on a validation data set. Models we are going to try are: Logistic regression, XGBoost, Gaussian Naive Bayes, k-nearest neighbor, AdaBoost, Random Forest and Gradient boosting. For the parameter tuning we will use a GridSearchCV which will consider combinations of different given parameters. Robustness of the model will be ensured through the use of cross validation.
- **Model evaluation and prediction**: final step will be to evaluate the selected best performing model on the test data set and produce predictions for that data.

References

https://finance.arvato.com/en/

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics

https://scikit-learn.org/stable/model_selection.html