Capstone Proposal

Customer Segmentation and Acquisition Optimization for Arvato Financial Solutions

Shilton Jonatan Salindeho

Domain Background

Arvato is a global services company headquartered in Gütersloh, Germany which offers customer support, information technology, logistics, and finance. Today, Arvato is one of eight divisions of Bertelsmann, the media, services and education group [1]. In this project, we will be conducting a project which will analyze data provided by Arvato and represents a real-life data science task handled by Arvato.

The project simulates a real-world problem where one of Arvato's client, a mail-order company, who is interested to acquire new clients more efficiently.

Techniques employed in order to tackle this problem are customer segmentation in order to identify parts of the population that best describe the core customer base, combining two datasets of general population and the company's customer base; as well as predictive analytics in order to predict which individuals are most likely to convert into becoming the company's customer from a third dataset with demographic information.

Problem Statement

The problem statement of this project is as follows:

"How can the client (the mail-order company) convert new customers more Efficiently, provided overall German demographics data, customer base data, and potential new customers data?"

More specifically, we are trying to predict whether someone is likely to become a customer of our client provided his/her demographic data. This is possible to quantify and measure since from segmenting our initial data (overall demographics and existing customer base) we will be able to construct a model that will be able to calculate the probabilities based on how similar the potential new customers are to the existing segments/clusters.

Datasets and Inputs

The project makes use of four datasets:

- 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)

Solution Statement

Our final goal is to enable the client company to firstly gain further knowledge about their current customer base, in order to target new customers among the German population, by making use of the information gained in the first stage to be able to predict with confidence who will become a new customer for them.

Firstly, some exploratory data analysis will be done on all dataset in order to gain more information about the data. Afterwards we will utilize unsupervised learning in order to carve out customer segments from the customer base; this will also include PCA (Principal Component Analysis) which will be used to apply Dimensionality Reduction on the data since it might work to our advantage to work with less variables, in addition to clustering algorithm such as K-Means Clustering which will be used to obtain 'clusters' of customers.

Then, in the second stage of the project where we will need to predict which individuals are most likely to convert, we will shift from unsupervised to supervised learning which will consist of different possible supervised algorithms, such as Decision Tree Classifier and Random Forest Classifier. Do note that we have not decided on one confirmed algorithm since one classifier might give us better accuracy than others, depending on the behavior of the data.

Benchmark Model

Our benchmark model for the second stage of the project (supervised learning) will be the Logistic Regression classifier since this is the standard model with 1 as "converted into new customer" score and 0 as the "not converted into new customer" score.

Evaluation Metrics

The initial findings in the population dataset and the customer dataset shows us that we will be dealing with imbalanced data, since the population dataset is roughly 4 times larger than the customer dataset. Due to this, accuracy will not be a good evaluation metric.

Thus, the possible remaining choices for our evaluation metrics are: F1 score, precision, recall, and area under the receiver operating curve (ROC) otherwise known as AUC. We will be using the AUC to evaluate performance of different models because it is one of the best options for the imbalanced data. AUC can be interpreted as the probability that the model ranks a random positive example more highly than a random negative example. In addition, the Kaggle Competition [2] also uses AUC as the evaluation metric, thus by also using AUC we will be able to compare our results to the Kaggle leaderboard.

Project Design

The summary of the proposed workflow for this project is as follows:

Data Exploration and Cleaning
- Exploratory Data Analysis
- Data Cleaning
- Data Visualization

Feature Engineering
- Principal Component Analysis (PCA)

Model Selection
- Selection of Unsupervised Learning Algorithm for Stage 1 (e.g. K-Means)
- Selection of Supervised Learning Algorithm for Stage 2 (e.g. Decision Tree, Random Forest)

Model Training and Tuning
- Model Training
- Hyperparameter Tuning

Model Testing
- Model Testing with Evaluation Metrics (AUC)

References

[1] Arvato. In *Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Arvato

[2] Udacity+Arvato: Identify Customer Segments. In *Kaggle*. Retrieved from: https://www.kaggle.com/c/udacity-arvato-identify-customers