Capstone Proposal

Customer Segmentation
Optimizing the Customer Acquisition Process
Arvato Financial Solutions

Author: Rodrigo P Pacheco de Toledo
Date of Proposal: 8 March 2021

Machine Learning Engineer Nanodegree
Udacity - School of Artificial Intelligence

Customer Segmentation – Arvato Financial Solutions

Capstone Proposal

Rodrigo Polverari Pacheco de Toledo

Domain Background

Arvato is a German global services company service that offers services including customer support, information technology, logistics and finance **[1].** As part of their services, Arvato helps, for example, client companies with questions around client profiles and acquisition. Also, Arvato is wholly owned by Bertelsmann a media, services and education company.

In this project, we will be undertaking a machine learning real-life project that data scientists at Arvato would typically be working on.

More specifically, the underlying business matter we will be focusing on here is the following: a client mail-order company seeks our help/services to acquire new clients more efficiently. To accomplish our project, we will employ customer segmentation: this refers to the practice of dividing a customer base into groups of individuals, depending on well-defined specific features, such as age, gender, interests, spending habits, etc. **[2].**

Problem Statement

The problem I'll be working on this project is:

"From demographic data of a person, how can a mail order company acquire new customers in an efficient way?"

Basically, given the demographics data of a person, what can we do to predict, with the highest accuracy, whether this person will be a new customer to the mail-order company?

Out of all these people with their associated demographics information, can we predict with confidence how many of them could be future customers with high probabilities of becoming customers?

Proposed Solution

The problem can be quantified in the following:

1. Number of current/established customerclusters (customer segmentation unsupervised problem)
2. Probability of being a new customer to the company (supervised problem).

To work on this project, I'll be using Supervised and Unsupervised techniques (as explained above), as follow:

- Using unsupervised learning methods on the data of established customers and the general population's demographics data aiming to create customer segments.

- Using supervised learning methods on a third dataset, we can train a model to predict the probability of a person becoming a new customer (above a certain threshold, the model will assign the person to be a highly probable new customer) and use this model for future predictions.

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).

There´s also 2 metadata files that have been provided to give attribute information:

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

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

Solution Statement

Arvato Financial Solutions' goal is to enable their client company to gain insight from their current established customer base in order to better target the German population atlarge, by predicting in advance and with sufficient accuracy who would become a future customer.

First Step: EDA – Exploratory Data Analysis of the datasets

Second Step: Use of unsupervised learning techniques to identify customer segments

i. Use of PCA to reduce dimensionality

ii. Use of K-Means Clustering to get meaningful clusters of customers

Final Step: Use of supervised learning techniques to predict future potential clients from the German Population dataset, based on insight gained by customer segment.

Benchmark Model

For the final step of this project, where we will apply supervised machine learning techniques to our binary classification problem (new customer 1 – not a new customer 0), an appropriate benchmark model to compare our model's performance could be a Logistic Regression Model. Thus, our benchmark model will be a standard Logistic Regression model with outcomes 1 = new customer, 0 = not a new customer.

Evaluation Metrics

1. *First Part: Customer Segmentation using unsupervised learning algorithms*

This first task uses a dimensionality reduction technique called PCA to reduce the number if dimensions. The explained variance ratio of each feature could be used as reference in selection the number of dimensions for the last part(supervised learning).

Additionally, in case of segmenting the customers into different clusters, an unsupervised learning algorithm like K-Means Clustering is proposed. In this case, the number of clusters will be a hyperparameter and it will be selected based on the square error i.e. distance between all the clusters

- Explained Variance Ratio
- Distance between clusters

2. *Second Part: Predicting Future Potencial Customer using supervised learning algorithms*

    - *ROC*
    - *AUC*

Project Design

A broad outline of the theoretical workflow is summarised below:

1. **Data Exploration & Cleaning**
    - Explore the data
    - Clean the raw input datasets (missing values, features to keep or drop, revisions of data formats)
    - Create a function with the pre-processing steps

2. **Data Visualisation**
    - Visualise the data
    - Identify correlations between features or other data-specific patterns

3. **Feature Engineering**
    - Make informed decisions about features to drop/keep with a PCA implementation

4. **Model Selection**
    - Experiment with different algorithms in Step 1 (unsupervised learning, e.g. K-Means Clustering)
    - Experiment with different algorithms in Step 2 (supervised learning, e.g. DecisionTreeRegressor)
    - Select the best-suited algorithms for the problem

5. **Model Training & Tuning**
    - Train the model defined previously
    - Implement Hyperparameter-Tuning strategies (e.g. using a range of values instead of one value for hyperparameters, strategy to counteract the effect of class imbalance in the dataset of the supervised learning algorithm…)

6. **Model Testing/Predictions**
    - Test the model with the testing data against the benchmark model and with the evaluation metrics we defined earlier.

References

[1] Arvato. In *Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Arvato#cite_note-3

[2] Customer Segmentation (Online Definition). In *SearchCustomerExperience*. Retrieved from: https://searchcustomerexperience.techtarget.com/definition/customer-segmentation

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

[4] Receiver Operating Characteristic. In *Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[5] Sensitivity and Specificity. In *Wikipedia*. Retrieved from: https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[6] Classification: ROC Curve and AUC. In *Google Developers*. Retrieved from: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc