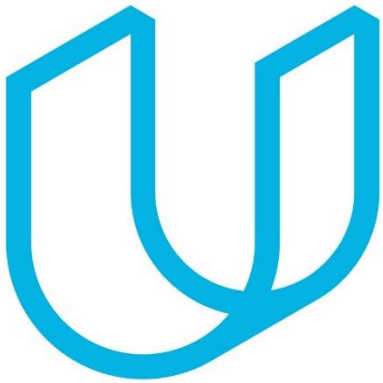

UDACITY

Capstone Proposal

Customer Segmentation & Optimizing the Customer Acquisition Process with Arvato Financial Solutions

Author: Hassan Essa Alghanim
Date of Proposal: 1st January 2022

Udacity - Machine Learning Engineer Nanodegree
Customer Segmentation – Arvato Financial Solutions

Domain Background

Arvato is a multinational services firm based in Germany that provides services such as customer service, information technology, logistics, and finance [1]. Arvato, for example, assists clients with questions about client profiles and acquisition as part of their services.

In this project, we will perform a machine learning task that is similar to real-world projects that data scientists at Arvato would work on.

The fundamental business issue that we will be concentrating on here is as follows:

A client mail-order company requests our assistance/services in order to more efficiently gain new clients. We will use customer segmentation to complete our project: this is the process of breaking a client base into groups of individuals based on well-defined specific attributes such as age, gender, interests, spending habits, and so on.

Dividing the company's current customer base into smaller meaningful groups will allow us to gain insight into their various types of customers, allowing the company to target the German population as a whole more effectively: for example, marketing teams would greatly benefit from such grouping information, as they would be able to determine which promotional campaign would appeal to which demographic group before even launching these campaigns. In addition to client segmentation, we want to create a supervised machine learning model that can predict whether or not a person (from the German population) will become a new customer.

Problem Statement

The problem we will be working on in this project is *"How can the German mail-order company acquire new customers more efficiently, given the access to German demographics data?"*

In other words, given a single individual's demographics, what can we do to forecast whether that person would be a new client to the mailorder firm with a high/significant degree of accuracy? Can we confidently forecast how many of these persons, with their related demographic data (third dataset), will become future customers?

The challenge may be expressed in terms of the number of current/established customer clusters (unsupervised customer segmentation problem) and the likelihood of becoming a new client to the firm (supervised problem).

Machine Learning methods can be used in the project's two primary sections:

- We may generate client segments using unsupervised learning algorithms on data from existing customers and demographic data from the broader population.

- We may train a model to forecast the possibility of a person becoming a new client using supervised learning methods on a third dataset (beyond a specific threshold, the model will designate the individual to be a very probable new customer), and then use this model to make future predictions.

Dataset 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

Finally, Arvato Financial Solutions wants to help their client firm acquire information from their existing customer base so that they may better target the German population as a whole by forecasting who will become a future customer in advance and with adequate precision.

To achieve this, we will first use unsupervised learning techniques to identify customer segments (customer segmentation), such as using PCA (Principal Component Analysis) for Dimensionality Reduction and an algorithm like K-Means Clustering to obtain meaningful 'clusters' of customers, after performing initial data exploration and cleaning.

Then, for the second half of the research, we'll employ supervised learning approaches to anticipate future possible clients from the German Population dataset, based in part on information garnered by customer segments. We'll use supervised algorithms like DecisionTreeRegressor (from the Python sklearn package), XGBClassifier, RandomForestClassifier (sklearn), and GradientBoostingClassifier for this work (sklearn)

It is hard to pick one supervised method over another at this early stage of the project (Proposal), thus we retain a wide variety of options until we begin working on the assignment. Keeping our

alternatives open will aid us in arriving at a satisfactory solution (not necessarily the ones listed above).

Benchmark Model

A Logistic Regression Model might be an ideal benchmark model to measure our model's performance in the last part of this project, when we will use supervised machine learning techniques to our binary classification issue (new customer 1 - not a new customer 0). As a result, our benchmark model will be a typical Logistic Regression model, with outcomes 1 indicating a new client and 0 indicating no new customer.

Evaluation Metrics

1. An assessment metric's definition
   Once we've chosen and trained a model for this capstone project, we'll use it to generate predictions based on campaign data from the Kaggle Competition [3]. Following that, we'll utilize our position (or score) on the Kaggle Competition leaderboard as our assessment measure for our model's performance on test data.

   To be more specific, the ranking/scoring is determined by AUC, with the ROC curve as the curve.

   A receiver operating characteristic curve (ROC curve) is a graphical representation that shows how a binary classifier (here, new customer or not) system's diagnostic capacity changes when the discriminating threshold is changed.
   The true positive rate (TPR) is shown versus the false positive rate (FPR) at various threshold levels on the ROC curve. In machine learning, TPR is also known as sensitivity, recall, or probability of detection. FPR stands for false alarm probability and is computed as 1 – specificity [4], with specificity referring to the fraction of correctly recognized negatives [5].

   AUC ("Area under the ROC Curve") is a two-dimensional measurement of the area beneath the complete ROC curve from (0,0) to (1,1). (1,1). AUC may be thought of as the likelihood that the model would rate a random positive example higher than a random negative example [6]. The AUC value varies from 0 to 1. The AUC of a model whose predictions are 100 percent incorrect is 0; the AUC of a model whose predictions are 100 percent right is 1.

2. Mathematical Formulae

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = 1 - Specificity = \frac{FP}{FP + TN}$$

with:
TP = true positive

FP = false positive

TN = true negative

FN = false negative

*Note: $Specificity = \frac{TN"}{TN+FP}$

** Note: TPR = Recall = Sensitivity

***Note: FPR = probability of False Alarm

From **[6].**

Project Design

The following is a high-level overview of the theoretical workflow:

1. **Data cleaning and exploration**
    - 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 s
2. **Data Visualization**
    - Visualize 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-au