Capstone Proposal

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

Author: Dilay Fidan Ercelik

Date of Proposal: 1st September 2020

Programme of Study: Machine Learning Engineer Nanodegree
From: Udacity - School of Artificial Intelligence

Customer Segmentation – Arvato Financial Solutions

Capstone Proposal

Dilay Fidan ERCELIK

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.

In this project, we will be undertaking a ML task mimicking real-life projects 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].**

Dividing the current customer base of the company into smaller meaningful groups will enable us to gain insight into their different types of customers, which will then permit the company to target the German population at large in an informed way: for example, marketing teams would highly benefit from such grouping information, as they would be able to determine which promotional campaign would most appeal to which demographic group before even launching these campaigns **[2].** Adding to customer segmentation, we wish to develop a supervised ML model capable to predict whether a person (from the German population) will be a new customer.

Problem Statement

The problem we will be working on in this project is the following:

"How can the German mail-order company acquire new customers more efficiently, given the access to German demographics data?"

Essentially, given the demographics data of a single person, what can we do to predict, with sufficiently high/significant accuracy, whether this person will be a new customer to the mail-order company? Out of all of these people with their associated demographics information (third dataset), can we predict with confidence how many of them could be future customers with high probabilities of becoming customers?

The problem can be quantified in the following terms: number of current/established customer clusters (customer segmentation unsupervised problem) and probability of being a new customer to the company (supervised problem).

Machine Learning techniques can be employed in the two main subsections of the project:

- Using unsupervised learning methods on the data of established customers and the general population's demographics data, we can 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).

Solution Statement

Ultimately, 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 at large, by predicting in advance and with sufficient accuracy who would become a future customer.

To make that possible, after initial data exploration and cleaning, we will first employ unsupervised learning techniques to identify customer segments (*customer segmentation*): these include applying **PCA** (Principal Component Analysis) for Dimensionality Reduction, and an algorithm such as **K-Means Clustering** to obtain the meaningful 'clusters' of customers.

Then, we will make use of supervised learning techniques for the second part of the project, which consists of predicting future potential clients from the German Population dataset, based partially on insight gained by customer segments. For this task, we will try different supervised algorithms, such as **DecisionTreeRegressor** (from sklearn module in Python), **XGBClassifier**, **RandomForestClassifier** (sklearn), or **GradientBoostingClassifier** (sklearn).

At this premature stage of the project (Proposal), it is impossible to choose one supervised algorithm over another, so we keep a wide range of choices until actually working on the task. Keeping our options open will help us get to an approach (not necessarily the ones cited above) with satisfying results.

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. *Definition of an evaluation metric*

As part of this capstone project, once we have chosen and trained a model, we will be using it to make predictions on the campaign data from the associated Kaggle Competition **[3].** Following from that, we will use our position (or score) within the leader-board of the Kaggle Competition as our evaluation metrics for the performance of our model on test data.

To be more precise, the ranking/scoring is based on AUC, the curve being the ROC curve.

A receiver operating characteristic curve (ROC curve) is a graphical plot that illustrates the diagnostic ability of a binary classifier (here, new customer or not) system as its discrimination threshold is varied.
The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings.
TPR is also known as sensitivity, recall or probability of detection in machine learning. FPR is also known as probability of false alarm and can be calculated as 1 – specificity **[4],** with specificity referring to the proportion of negatives that are correctly identified **[5].**

AUC ("Area under the ROC Curve") measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). One interpretation of AUC is seeing it as the probability that the model ranks a random positive example more highly than a random negative example **[6].** AUC ranges in value from 0 to 1. A model whose predictions are 100% wrong has an AUC of 0; one whose predictions are 100% correct has an AUC of 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

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