Capstone Proposal

Customer Segmentation and Optimization of Customer Acquisition with Arvato Financial Solutions

Funing Tian
7th June 2021
Machine Learning Engineer Nanodegree
School of Artificial Intelligence

Domain Background
Arvato, wholly owned by Bertelsmann, is a services company that actively develops and implements innovative solutions for customers on a global scale. It provides services including customer support, information technology, logistics and finance [1]. As part of its services, Arvato is helping client companies get invaluable insights into client profiling and marketing.

In this project, we will employ the use of machine learning to deal with real-life data provided by Bertelsmann Arvato Analytics. More specifically, a client mail-order company seeks our help to better target next probable customers for their products.

To fulfill the goal, we will focus on customer segmentation, that is, characterizing customers segment of population based upon well-defined specific features [2]. Diving customers into groups based on common characteristics will enable our client company to market each group effectively and properly. Here, we will analyze demographics data of customers and the general German population. This will be followed by developing a supervised machine learning model to make predictions on whether a person will be a new customer.

Problem Statement
The problem that we will be investigating can be formulated as: “Given the access to German demographic profiles, how can the German mail-order company acquire new customers efficiently?”

This problem requires us to consider what we can do to predict with high accuracy whether a person with associated demographics data will be a new customer to our client company. Furthermore, how can we predict with confidence the probability of people with demographic profiles turning into future customers?

Proposed Solution with machine learning techniques being employed
1) Analyzing demographics data of general population and customers and using unsupervised learning techniques for customer segmentation

2) Using supervised learning techniques on response of marketing campaign and train the constructed model to make predictions on the probability of individuals being converted into becoming customers

Datasets and Inputs
The project makes use of four files:

1) Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).

2) Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).

3) Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).

4) 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
The goal of Arvato Financial Solutions is to more efficiently acquire new customers in the German population by predicting with sufficient accuracy who would be a becoming customer from demographic profiles of customers of the company.

To achieve that, we start with data exploration. Incomplete records will not be considered for downstream analysis. In addition, categorical features will be converted into numerical features. To consistently compare the values of different features, the scale numerical values will be standardized.

Before feeding data into a machine learning model, a dimensionality reduction step is a necessity with the aim being forming a smaller set of features to better help separate our data. The technique that will be used to reduce the number of features is principal component analysis (PCA). After having PCA attributes set up, we will use unsupervised clustering algorithm, k-means, to segment customers.

Next, we will use supervised learning techniques to predict potential customers from the German Population dataset by taking into account customer segments. To handle this task, we propose the following supervised learning algorithms: 1) Logistic Regression 2) Decision Tree Classifier 3) Random Forest Classifier and XG Boost Classifier.

Benchmark Model
A benchmark model for the binary classification problem would be a Logistic Regression Model. The performance of this model can be used to compare against different algorithms to help choose one supervised algorithm over another.

Evaluation Metrics
The selected model will be used to make predictions on the mailout campaign data in competition through Kaggle. The evaluation metric for this Kaggle competition is AUC for the ROC curve.

A receiver operating characteristic (ROC) curve is a graphic plot used to display the true positive rate (TPR, known as recall, proportion of correctly labeled actual customers) against the false positive rate (FPR, proportion of non-customers labeled as customers) [3].

Mathematical formulas of TPR and FPR are defined as:

$$\text{TPR} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{FPR} = \frac{FP}{FP+TN} \quad (2)$$

where
TP = ture positive
FP = false postive
TN = true negative
FN = false negative [4].

The area under the ROC curve (AUC) measures the entire two-dimensional area underneath the entire ROC curve from (0, 0) to (1, 1) [5]. If a model does not discriminate between classes at all, the score will be 0.5. In contrast, if a model perfectly captures all customers, the maximum score can possibly be 1.0.

Project Design
A theoretical workflow of the project can be summarized as following:

1) Data Exploration and Pre-processing
Exploring data helps gain an insight into data points and features. However, since it is required to use complete data points to train a model, any missing or mis-recorded values requires to be cleaned.

2) Data Visualization
To identify distribution patterns in the data, a visualization analysis on features will be performed.

3) Feature Engineering
A dimensionality reduction method will be applied to understand explained data variance by features and combine similar or redundant features to form a smaller feature set.

4) Model Selection and Training
The first step is to segment customers using unsupervised learning techniques. The k-means clustering algorithm will be used to identify clusters of customers based upon PCA attributes. Next, several different supervised learning algorithms will be trained and evaluated on predicting new customers. The algorithms include Logistic Regression, Decision Tree, Random Forest and XG Boost. The performance of these algorithms will be compared to choose the best model for this problem.

5) Model Tuning
Hyperparameter-tuning strategies will be implemented on the selected algorithm. The selected algorithm will be tuned to improve performance, including using a range of values for hyperparameter, and accounting for class imbalance.

6) Model Testing and Predictions
The best model will be used to make predictions on the test data and will be evaluated with the evaluation metrics.

References

[1] Arvato. *Wikipedia*. https://en.wikipedia.org/wiki/Arvato
[2] Customer Segmentation. *Wikipedia*. https://en.wikipedia.org/wiki/Market_segmentation
[3] Udacity+Arvato: Identify Customer Segments. *Kaggle*. https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/evaluation
[4] Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol*. 2008;56(1):45-50.
[5] Classification: ROC Curve and AUC. *Google Developers*. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc