Machine Learning Engineer Nanodegree

Capstone Report

Customer Segmentation Report for Arvato Financial Services

Name Surname: Demirhan Demirkol

Date: 25.10.2024

Contents

Definition

Project Overview

This project focuses on analyzing demographic data of the German population in conjunction with customer information to support customer segmentation and acquisition efforts for Arvato Financial Solutions. As a global service provider, Arvato delivers solutions in financial services, IT, and supply chain management for businesses.

The objective of the project is to aid a mail-order company in identifying potential customers for its organic product line. By comparing demographic characteristics of existing customers to those of the general population, the project aims to identify the most promising individuals for targeted marketing campaigns.

The project is divided into two key phases:

1. **Customer Segmentation using Unsupervised Learning**: In this phase, thorough data analysis and feature engineering are conducted to prepare the dataset. Principal Component Analysis (PCA) is employed for dimensionality reduction, and K-Means clustering is performed on the resulting PCA components to segment both the general population and customer data into distinct groups. These clusters are then examined to identify unique features and patterns that set potential customers apart from the rest.
2. **Customer Acquisition using Supervised Learning**: In this phase, historical customer data with labeled outcomes (indicating previous customer responses) is used to train supervised machine learning models. These models are then applied to new, unseen data to predict which individuals are most likely to become future customers.

Domain Background

Arvato is a global services company specializing in financial services, IT solutions, and supply chain management (SCM) for business clients across diverse industries, including insurance, e-commerce, energy, and internet providers. The company focuses on developing innovative, automated solutions using data analytics. Arvato is a wholly owned subsidiary of Bertelsmann, a company operating in media, services, and education.

Arvato aids its clients in deriving valuable insights from data to inform business decisions, particularly in customer-centric marketing. This field has been expanding as companies look to uncover hidden patterns and understand customer behavior. Data Science and Machine Learning are increasingly being leveraged to achieve business objectives and enhance customer satisfaction.

In this project, Arvato is assisting a mail-order company in Germany that sells organic products. The objective is to analyze existing customer data alongside demographic data of the German population to identify distinct customer segments. The goal is to develop a system that predicts whether an individual is likely to become a customer based on demographic characteristics.

Dataset and Inputs

This project involves four main data files:

- **Udacity_AZDIAS_052018.csv**: Contains demographic data for the general population of Germany with 891,211 individuals (rows) and 366 features (columns).
- **Udacity_CUSTOMERS_052018.csv**: Holds demographic data for customers of a mail-order company, comprising 191,652 individuals (rows) and 369 features (columns).
- **Udacity_MAILOUT_052018_TRAIN.csv**: Provides demographic data for individuals targeted in a marketing campaign, with 42,982 individuals (rows) and 367 features (columns).
- **Udacity_MAILOUT_052018_TEST.csv**: Contains demographic data for individuals targeted in a marketing campaign, with 42,833 individuals (rows) and 366 features (columns).
- **DIAS Information Levels - Attributes 2017.xlsx**: A high-level list of attributes and descriptions categorized by information type.
- **DIAS Attributes - Values 2017.xlsx**: A detailed mapping of data values for each feature, listed alphabetically.

All these files were supplied by Arvato in the context of the Machine Learning Nanodegree program to enable analysis and customer segmentation. The four CSV files contain demographic information for everyone, including data about their household, building, and neighborhood. The customer dataset has three extra columns providing specifics related to the mail-order company. The training and test datasets are designed for evaluating supervised learning algorithms.

Problem Stamement

The problem can be reformulated as:

“How can a mail-order company efficiently target potential customers using demographic information?”

The approach involves two key steps. First, unsupervised learning methods are applied to analyze the demographic data of both the general population and the company’s existing customers. This helps in identifying distinct segments within these groups and determining which demographic traits are linked to being a customer.

Next, a supervised learning algorithm is implemented to predict whether a person is likely to become a customer, using their demographic profile.

Evaluation Metrics

This project involves customer segmentation and acquisition using both unsupervised and supervised learning techniques. In the first part, customer segmentation is achieved using Principal Component Analysis (PCA) for dimensionality reduction, where the explained variance ratio helps select the minimum number of dimensions that capture most of the dataset's variation. K-Means clustering is then applied to segment customers into clusters, with the optimal number of clusters determined using an elbow plot based on squared error. In the second part, customer acquisition prediction is framed as a binary classification problem, where the objective is to decide if a mail-order company should approach a potential customer. Given the class imbalance in the data (42,430 observations labeled as '0' and 532 as '1'), the model is trained on a split dataset and evaluated using the Area Under Receiver Operating Characteristic (AUROC) metric. The AUROC metric, which considers both the true positive rate and false positive rate, is selected for its ability to effectively assess model performance in distinguishing between customers and non-customers, with a higher AUROC indicating a better-performing model.

Data Exploration and Preprocessing

The datasets provided were carefully loaded and checked to ensure they had the expected number of rows and columns based on the given descriptions. The preprocessing was approached step-by-step, using specific helper functions for each task. This modular

approach made it easier to combine all the functions into one main data preprocessing function at the end.

Handling Columns with Mixed Data Types

During data loading, some warnings indicated that certain columns had mixed data types or incorrect entries. For example, columns 18 and 19 (‘CAMEO_DEUG_2015’ and ‘CAMEO_INTL_2015’) contained values that were inconsistently recorded. Using an attribute guide as a reference, we identified that values like 'X' or 'XX' were errors and replaced them with empty (NaN) values.

Fixing 'Unknown' Values

Many columns had special markers indicating missing or unknown data. With the help of a reference sheet, we identified these markers and replaced them with NaN. This affected 232 columns in total.

Checking for Overlapping Features

We compared the datasets to see which features were common across all of them and which were unique. We identified 272 shared features with clear descriptions, 3 specific to customer data, and 42 with no clear description available.

Dealing with Invalid Values in ‘LP_*’ Columns

Certain columns like ‘LP_FAMILIE_FEIN’ and ‘LP_STATUS_FEIN’ had issues with values like '0', which didn’t correspond to any category described in the reference sheet. We replaced these with NaN values. Some columns also contained very detailed information that was broken down into simpler categories for better clarity. Duplicate or overly detailed columns were dropped.

Re-Encoding Features

Some columns required reformatting for consistency:

The ‘EINGEFUGT_AM’ column, which contained dates, was reformatted to only include the year.
The gender column ‘ANREDE_KZ’ was re-encoded to use 0 for male and 1 for female.
The ‘CAMEO_INTL_2015’ column was split into two separate columns for better clarity.

The column ‘WOHNLAGE’ had incorrect values that were replaced with NaN.

Handling Missing Values

After cleaning, we assessed missing data at both the column and row levels:

Column-wise: We examined each column's percentage of missing values. If more than 30% of a column’s data was missing, it was removed. This led to 11 columns being dropped.
Row-wise: We removed rows with more than 50 missing features. This step resulted in dropping 153,933 rows from the general population data and 57,406 rows from the customer data.

Imputing Missing Values

For the remaining missing values, we used the most frequently occurring value in each column to fill the gaps. This method was chosen because the dataset represented the general population.

Feature Scaling

Finally, we applied a standard scaling technique to bring all the features within the same range. This helps to avoid certain features overpowering others when performing further analysis or applying machine learning techniques.

Algorithms, Techniques and Methodology

Customer Segmentation

The primary goal of this project is to categorize the general population and existing customers into different segments to compare and identify potential future customers. The company's current customer data was used to gain insights and make comparisons with the general population data. However, analyzing every feature to understand customer behavior is a time-intensive process, especially since not all features may significantly influence behavior. Additionally, complex interactions between various features could play a role in distinguishing customers, making manual analysis inefficient and less productive.

Dimensionality Reduction

To address this, an approach utilizing unsupervised learning algorithms was adopted to segment both the general population and customers. Principal Component Analysis (PCA) was applied to the dataset to reduce its dimensions, given that there were 353 features after data cleaning and feature engineering. The purpose of PCA was to identify which features best explained the variance within the dataset. Although there were originally 353 features, nearly 90% of the variance could be explained using only 150 PCA components. Consequently, the number of features was reduced from 353 to 150.

PCA Component Analysis

These 150 components can be further interpreted by examining the feature weights assigned by the PCA algorithm. For instance, Component 2 indicates individuals who exhibit high mobility and are likely to reside in neighborhoods dominated by smaller, single-family homes rather than larger apartment complexes. Similarly, the features in this component also correlate with specific types of car ownership. Another example is Component 4, which corresponds to individuals who are financially stable but less inclined toward saving or investing. Similar analyses were conducted for other components.

Clustering

After dimensionality reduction, the next step was to segment the general population and customer population using the K-Means clustering algorithm. This algorithm was chosen due to its simplicity and its suitability for measuring distances between observations to assign clusters. The goal was to leverage the reduced features to create distinct segments and use these clusters to analyze similarities between the general population and customer data.

The number of clusters was treated as a hyperparameter, and an elbow plot was utilized to determine the optimal number of clusters. The elbow plot displayed the sum of squared distances for each cluster number, helping identify a point where the sum stops decreasing significantly. Based on this analysis, the number of clusters was set to eight.

Cluster Analysis

The general population and customer population were segmented into eight clusters. While the general population was evenly distributed across these clusters, the customer population predominantly belonged to Clusters 2, 4 and 6. To confirm this, the ratio of customer segments to general population segments was calculated. A ratio greater than 1

indicated clusters with a higher number of existing customers and a potential for future customers. Conversely, clusters with a ratio below 1 had lower prospects for future customers.

K-Means Cluster Analysis

Like the analysis performed on PCA components, each cluster was examined to identify the main components and features that contributed to it. For example, Cluster 3 primarily consisted of Components 0 and 9, indicating a preference for smaller residential neighborhoods with a lower density of homes and cars. This analysis helped to uncover the defining characteristics of each cluster and their connection to existing customers.

Customer Acquisition

The second phase of the project involves using supervised learning algorithms to predict whether an individual is likely to become a customer based on demographic data. The provided dataset, 'Udacity_MAILOUT_052018_TRAIN.csv,' contains the same features as the general population and customer demographic data, with an additional column labeled 'RESPONSE.' This column indicates whether a person is a customer or not. The dataset underwent the same cleaning and processing steps as the general population and customer datasets.

Benchmark

The first step in the supervised learning process was to establish a benchmark using the simplest model to gauge future results. The dataset was split into training and validation sets, and a logistic regression model was trained on unscaled training data and evaluated on the unscaled validation set. The benchmark score achieved with logistic regression was an AUROC score of 0.63.

Baseline Performance

After establishing the benchmark, the data was scaled using a standard scaler and split into training and validation sets. Several algorithms were trained on the training set and evaluated on the validation set. The chosen algorithms included:

- Logistic Regression

- Decision Tree Classifier

- Random Forest Classifier

- Gradient Boosting Classifier

- AdaBoost Classifier

- XGBoost Classifier

These algorithms are suitable for classification tasks. The performance of all models was compared with each other and with the benchmark. The logistic regression model retained its initial performance, while the decision tree classifier underperformed. In contrast, ensemble algorithms outperformed other models. The Random Forest Classifier showed good results, and the Gradient Boosting Classifier from the sklearn library had the highest score, although it required longer training times. AdaBoost and XGBoost exhibited similar performances, with high scores and lower training times. Consequently, these two algorithms were chosen for hyperparameter tuning.

Hyperparameter Tuning

The AdaBoost and XGBoost classifiers were fine-tuned using Grid Search, selecting a set of hyperparameters for each algorithm to identify the best-performing models.

Feature Importances

Since the chosen models are tree-based, feature importance analysis was performed to understand which features were most influential.

AdaBoost: The feature importance analysis revealed that the feature 'D19_SOZIALES' had the highest importance, followed by other features.

XGBoost: Similarly, XGBoost indicated 'D19_SOZIALES' as the most significant feature. Although the exact description for 'D19_SOZIALES' is not provided in the attribute files, the feature's name suggests a connection to social transactions, inferred from other features starting with 'D19_'. This interpretation, however, may differ from the actual feature description.

In comparison, XGBoost’s feature importance distribution appeared more balanced than AdaBoost’s. This could be due to differences in algorithm design: AdaBoost improves upon weak learners by focusing on data points with higher weights, whereas XGBoost leverages gradients from an objective function to enhance weak learners.