Machine Learning Engineer Nanodegree

Capstone Report

Customer Segmentation Report for Arvato Financial Services

Name Surname: Demirhan Demirkol

Date: 25.10.2024

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 [1][2].

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.

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.

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.

Solution Statement

Part 1: Customer Segmentation and Alignment with General Population

1. Data Exploration and Preprocessing: The first step involves a thorough examination of the provided dataset to identify any missing or incorrectly recorded values. These issues will be addressed to ensure data integrity. Categorical features will be transformed into numerical formats using techniques such as label encoding. Additionally, feature scaling will be performed to prevent any single feature from dominating the analysis due to differing scales.

2. Feature Selection and Dimensionality Reduction: Given that there are 366 features representing an individual, not all of them will significantly contribute to the segmentation process. A dimensionality reduction technique, such as Principal Component Analysis (PCA), will be applied to identify a minimal set of features that sufficiently explain the dataset's variation.

3. Segmentation using Unsupervised Learning: The next step involves segmenting both the general population and the customer dataset based on the selected features. This will be achieved using an unsupervised learning algorithm, with K-means clustering being a suitable choice. This algorithm assigns each data point to a cluster based on its distance from the cluster centers.

Part 2: Customer Acquisition Prediction

1. Data Preprocessing for Training and Testing: The first two steps of Part 1 will be repeated on the training and testing datasets to ensure consistency in preprocessing.

2. Supervised Learning Model Development: A supervised learning model will be trained on the preprocessed training data to predict customer acquisition. The performance of the model will be evaluated using appropriate metrics.

3. Prediction on Test Data: The trained model will be used to make predictions on the provided test dataset.

Proposed Supervised Learning Algorithms:

- Logistic Regression: A basic binary classification model.

- Decision Tree Classifier: A tree-based algorithm that uses rule-based classification.

- Random Forest Classifier and XGBoost Classifier: These advanced models are derived from decision trees and provide greater accuracy and robustness.

A grid search algorithm may be employed to optimize the hyperparameters for the selected algorithm.

Project Design

1. Data Cleaning and Visualization:

The initial step involves checking the dataset for missing or incorrectly recorded values. Misrecorded values will be verified and corrected using information from the metadata files. An analysis of missing values per feature will be conducted to decide which features, if any, should be excluded. Additionally, visualizations will be employed to explore patterns within the data, providing insights into its structure and distribution.

2. Feature Engineering:

This step focuses on understanding the explained variance of features and determining the optimal number of features required to capture the majority of the dataset's variance. A dimensionality reduction technique such as Principal Component Analysis (PCA) will be used for this purpose. Correlation analysis will also be conducted to identify and eliminate redundant features, thus improving the efficiency and interpretability of the model.

3. Modelling:

The first part of this step involves segmenting customers using unsupervised learning techniques, with K-means Clustering as the primary algorithm to classify the data into appropriate clusters. The second part involves training and evaluating various supervised learning algorithms to predict whether a person is likely to become a customer. Algorithms such as Logistic Regression, Decision Trees, Random Forests, and Gradient Boosted Trees (like XGBoost) will be applied to build predictive models. The models' performances will be assessed using previously defined evaluation metrics to select the best-performing model.

4. Model Tuning:

Once the models have been evaluated, the algorithm with the best performance will be selected for tuning. Hyperparameter tuning will be performed using techniques like Grid

Search to find the optimal set of parameters, thereby enhancing the model's predictive accuracy and reliability.

Benchmark

In this scenario, a Logistic Regression model is chosen as the benchmark because of its simplicity and efficiency in training and testing within a short period. The performance of this model will be set as the baseline, providing a reference point for evaluating other algorithms. By comparing their performance against this benchmark, we can assess whether alternative algorithms offer significant improvements, guiding the decision on whether to explore and implement them further. This approach helps in systematically identifying and prioritizing more complex models that justify additional computational efforts.

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.

References

[1] Arvato-Bertelsmann, "Arvato," Bertelsmann, [Online]. Available: https://www.bertelsmann.com/divisions/arvato/#st-1. [Accessed April 2020].

[2] Bertelsmann, "Company," Bertelsmann, [Online]. Available: https://www.bertelsmann.com/company/ [Accessed April 2020].