Capstone Project Proposal
Udacity Capstone Project - Arvato Analytics

Büşra Oğuzoğlu

September 23, 2024

1 1. Project Background

This project focuses on the field of customer analytics, specifically within the domain of **demographic data analysis and customer segmentation**. Customer segmentation is widely used in various industries, such as retail, telecommunications, banking, and marketing, to better understand consumer behavior and optimize marketing strategies. By segmenting customers based on their demographic profiles, businesses can target specific customer groups, tailor their offerings, and increase conversion rates.

In this project, we leverage machine learning algorithms to analyze and predict customer behavior for a mail-order company, **Arvato Bertelsmann**, based in Germany. Similar studies on customer segmentation have demonstrated success across different industries [1,2]. We aim to apply these concepts to the current dataset to provide actionable insights for the company.

2 2. Problem Statement

Arvato's mail-order sales company wants to understand which demographic segments are most likely to become customers based on historical data. The company needs a solution to:

- **Segment customers from the general population** to identify demographic clusters that are more likely to convert.
- **Predict which individuals in a marketing campaign are likely to become customers** based on demographic data, and thus improve the targeting of marketing resources.

The challenge arises due to the large and complex nature of the demographic data, as well as the imbalance in the dataset (where the number of actual customers is significantly smaller compared to the general population).

3 3. Datasets and Inputs

Four primary datasets are used in the project:

1. **Udacity_AZDIAS_052018.csv**: Demographics data for the general population of Germany. It contains 891,211 rows and 366 features.
2. **Udacity_CUSTOMERS_052018.csv**: Demographics data for customers of a mail-order company. It contains 191,652 rows and 369 features.
3. **Udacity_MAILOUT_052018_TRAIN.csv**: Demographics data for individuals targeted in a marketing campaign. It contains 42,982 rows and 367 features.
4. **Udacity_MAILOUT_052018_TEST.csv**: Demographics data for individuals targeted in a marketing campaign. It contains 42,833 rows and 366 features.

The data contains information on individuals, households, buildings, and neighborhoods, providing a rich source of demographic insights.

4 4. Solution Statement

We propose a solution based on unsupervised and supervised machine learning techniques:

- **Unsupervised Learning**: Perform customer segmentation using KMeans clustering to identify key demographic groups within the general population and compare them against customer segments.
- **Supervised Learning**: Use the demographic data from the general population and the labeled data from the marketing campaign to predict which individuals are most likely to convert into customers.

The solution will be deployed as a standalone machine learning model or hosted in the cloud using **AWS Sagemaker** or similar infrastructure. However, the initial development and training will be conducted locally on a **MacBook Pro (2023)** with the following configuration:

- Chip: Apple M2 Pro
- Memory: 32GB
- macOS: Ventura 13.3

This local setup provides sufficient resources to handle the dataset, run models, and perform hyperparameter tuning. After local testing, the solution can be deployed to a cloud-based service for scalability.

5 5. Benchmark Model

A simple benchmark model can be established by using logistic regression for the prediction task and KMeans clustering for segmentation. Logistic regression provides a strong baseline due to its simplicity and effectiveness on structured data.

6 6. Evaluation Metrics

We will evaluate the performance of our solution using the following metrics:

- **ROC-AUC**: This metric will help us evaluate the ability of the model to distinguish between customers and non-customers, which is especially important due to the class imbalance.
- **Precision, Recall, and F1-score**: These metrics will provide insights into the model's performance, especially its ability to recall true customers while maintaining precision.
- **Clustering Metrics (Silhouette Score, Inertia)**: These metrics will be used to assess the quality of the clusters in the unsupervised learning stage.

7 7. Project Design

The project will proceed through the following stages:

1. **Data Preprocessing and Cleaning**: Handle missing data, perform one-hot encoding for categorical features, and standardize numerical features. Identify and remove outliers if necessary.
2. **Feature Engineering**: Derive new features from the existing data to better represent the underlying customer behavior.
3. **Dimensionality Reduction**: Apply Principal Component Analysis (PCA) to reduce the high-dimensional dataset while retaining key information.
4. **Clustering Analysis**: Use KMeans to perform customer segmentation and compare cluster distributions between the general population and customers.
5. **Supervised Learning**: Train logistic regression and XGBoost models on the cleaned data, using hyperparameter tuning and cross-validation to optimize performance.
6. **Implementation Setup**: The models are going to be trained and tested on a local environment. Details of the setup are as follows:

References

[1] Kansal, Tushar, et al. "Customer Segmentation using K-means Clustering." 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). IEEE, 2018.

[2] Nandapala, E.Y.L, and Jayasena, K.P.N. "The practical approach in Customers segmentation by using the K-Means Algorithm." 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS). IEEE, 2020.