Capstone Project Proposal

Büşra Oğuzoğlu
Udacity Capstone Project - Arvato Analytics

September 23, 2024

1 1. Project Domain 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.

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 the company. It contains 191,652 rows and 369 features.
3. **Udacity_MAILOUT_052018_TRAIN.csv**: Demographics data for a marketing campaign's targets, including a `RESPONSE` column indicating whether they converted into customers. It has 42,982 rows and 367 features.
4. **Udacity_MAILOUT_052018_TEST.csv**: Similar to the MAILOUT_TRAIN dataset but without the `RESPONSE` column, used for final model evaluation in a competition setup.

Each dataset contains information on individuals, households, and their respective neighborhoods, including categorical features (e.g., region, car ownership) and numerical features (e.g., income, age).

4 4. Solution Statement

The proposed solution involves two main steps:

1. **Unsupervised Learning for Customer Segmentation**: We apply **K-Means clustering** on the general population (AZDIAS dataset) to identify demographic clusters. We then map these clusters onto the customer dataset (CUSTOMERS) to determine which clusters are over-represented among customers. These clusters will form the target demographic for the marketing campaign.
2. **Supervised Learning for Conversion Prediction**: After segmenting the population, we build a **supervised machine learning model** using the MAILOUT_TRAIN dataset to predict which individuals are most likely to convert into customers. Multiple machine learning algorithms will be tested, including **Logistic Regression**, **Random Forest**, and **XGBoost**, with hyperparameter tuning for optimal performance. Additionally, we address class imbalance in the dataset using techniques like **random under-sampling**.

5 5. Benchmark Model

A basic **Logistic Regression model** will serve as the benchmark for this project. Logistic regression is widely used in classification tasks and provides a simple but effective baseline model for comparison. While logistic regression may perform reasonably well, more complex models like **XGBoost** or ensemble methods are expected to outperform the benchmark.

Baseline Evaluation:

- **ROC-AUC score**: 0.50–0.55 (initial logistic regression without class balancing or feature engineering).

6 6. Evaluation Metrics

To assess the effectiveness of the proposed models, we will use the following evaluation metrics:

- **ROC-AUC Score**: The primary evaluation metric for this project. ROC-AUC measures the trade-off between the true positive rate (sensitivity) and false positive rate (specificity) and is well-suited for imbalanced datasets.
- **F1-Score**: The harmonic mean of precision and recall. This metric will be used to evaluate how well the model balances false positives and false negatives, especially in the context of the highly imbalanced dataset.
- **Precision**: The proportion of positive predictions that are actually correct. Useful for minimizing false positives in targeted marketing campaigns.
- **Recall**: The proportion of actual positives that are correctly identified. This is important to capture as many potential customers as possible.

7 7. Project Design Outline

The project is divided into several phases:

Phase 1: Data Preprocessing

- Clean and preprocess the datasets: handle missing values, encode categorical variables, and normalize numerical features.

- Perform **dimensionality reduction** using **PCA** to reduce the complexity of the dataset.

Phase 2: Clustering for Segmentation

- Apply **K-Means clustering** to the general population (AZDIAS) and assign customer segments in the customer dataset.
- Compare the distribution of customer clusters with the general population to identify over-represented segments among customers.

Phase 3: Supervised Learning for Conversion Prediction

- Split the marketing campaign data (MAILOUT_TRAIN) into training and validation sets.
- Train **Logistic Regression**, **Random Forest**, and **XGBoost** models.
- Perform **hyperparameter tuning** using **GridSearchCV** to find the best model parameters.
- Address class imbalance using techniques such as under-sampling or SMOTE (Synthetic Minority Over-sampling Technique).

Phase 4: Model Evaluation

- Evaluate the performance of the models using the validation set.
- Compare the models using the **ROC-AUC score**, **F1-score**, **precision**, and **recall**.
- Select the best-performing model for deployment.

Phase 5: Final Model Deployment

- Apply the selected model to the **MAILOUT_TEST** dataset to predict customer conversions.