Capstone Project Proposal: Customer Acquisition for Mail-Order Organic Products

1. Domain Background

Direct marketing and customer acquisition have long been central to retail and mail-order businesses. Traditionally, companies have relied on broad campaigns, intuition, and past experience to identify potential customers—an approach that is expensive and inefficient. With the rise of data-driven decision making, organizations increasingly seek to replace "gut feel" with analytical evidence to optimize marketing spend.

This project draws from the real-world work of Arvato Financial Services and Bertelsmann, who assist clients in making data-backed decisions. The domain spans customer analytics, demographic segmentation, and predictive modeling for targeted marketing. Related research includes customer segmentation via clustering (K-means, hierarchical methods) combined with PCA for dimensionality reduction [1], and supervised classification for response prediction in direct marketing (logistic regression, tree-based methods, gradient boosting) [4]. For imbalanced binary outcomes such as campaign response, evaluation metrics and learning strategies from imbalanced-data literature are relevant [2]. Cluster quality is commonly assessed using the silhouette coefficient [3]. The mission of such engagements is to enable senior managers to use reports, data points, and rationale to support their decisions rather than intuition alone.

---

2. Problem Statement

How can a mail-order company selling organic products acquire new customers more efficiently?

Specifically, the company has: - Demographic and lifestyle attributes for its existing customers - A large dataset of demographic attributes for the general population of Germany

The challenge is to identify which people in the general population are most likely to become new customers, so that marketing campaigns can be targeted at high-potential individuals instead of the entire population. This reduces cost, increases return on marketing investment, and supports a shift from broad outreach to data-driven customer acquisition. The problem is quantifiable (response rates, conversion, cost per acquisition), measurable (model performance, campaign outcomes), and replicable (same methodology can be applied to future campaigns).

---

3. Datasets and Inputs

The project uses four datasets provided by AZ Direct GmbH / Arvato for the Udacity Bertelsmann Capstone:

Dataset	Description	Role
Udacity_AZDIAS_052018.csv	Demographic and lifestyle attributes for the general population of Germany	Baseline population to segment and compare against customers
Udacity_CUSTOMERS_052018.csv	Same attribute structure for established customers of the mail-order company	Define customer profile and segments
Udacity_MAILOUT_052018_TRAIN.csv	Individuals targeted in a prior campaign, with response labels (1 = responder, 0 = non-responder)	Train supervised model to predict campaign response
Udacity_MAILOUT_052018_TEST.csv	Individuals targeted in a test campaign (no labels)	Evaluate model via Kaggle competition

The data is obtained through the Udacity Bertelsmann Capstone program and is governed by AZ Direct GmbH's terms (use limited to this project; deletion required within two weeks of completion). Attributes cover demographics, financial behavior, lifestyle, household, and regional information. Data use is appropriate for unsupervised segmentation (general population vs. customers) and supervised response prediction (campaign targets).

---

4. Solution Statement

The solution combines unsupervised and supervised learning in two stages:

Stage 1 – Customer Segmentation: Use unsupervised learning (e.g., PCA for dimensionality reduction followed by K-means clustering) to analyze attributes of established customers and the general population. This identifies segments that over- or under-represent customers and clarifies which demographic profiles align with the customer base.

Stage 2 – Response Prediction: Build a supervised classification model using the campaign training data to predict whether each individual will respond to a campaign. The model will use the same attribute structure as the segmentation step, with feature engineering and selection informed by the segmentation results. Predictions for the test set will be submitted to the Kaggle competition to evaluate performance on unseen data.

The approach is quantifiable (model metrics), measurable (AUC-ROC, precision, recall), and replicable (defined preprocessing, feature engineering, and modeling pipeline).

---

5. Benchmark Model

Benchmark 1 – Random targeting: Assume a random 50% of the population is targeted. The expected response rate equals the overall response rate in the training data. This reflects the "no intelligence" baseline.

Benchmark 2 – Majority class predictor: A classifier that always predicts the majority class (non-responder). This establishes a baseline for imbalanced classification and illustrates the gain from a real model.

Benchmark 3 – Simple rule-based heuristic: Target individuals whose attributes most closely match the average customer profile (e.g., based on top PCA components or a few key demographics). This provides a simple, interpretable baseline before introducing more complex models.

These benchmarks can be compared against the proposed solution using the same evaluation metrics (e.g., AUC-ROC, area under the precision-recall curve).

---

6. Evaluation Metrics

- **Supervised model (response prediction):**
  - **AUC-ROC:** Preferred for imbalanced binary classification; measures ranking quality across thresholds.
  - **Area Under Precision-Recall Curve (AUC-PR):** More informative than ROC when the positive class (responders) is rare.
  - **Kaggle metric:** The competition will specify the scoring metric (often AUC-ROC); this will be the primary external evaluation.
- **Unsupervised segmentation:**
  - **Silhouette score:** Measures cluster cohesion and separation.
  - **Over/under-representation analysis:** Compare segment proportions in the customer dataset vs. general population to quantify which segments are most customer-like.

These metrics are suitable for the problem and data and allow direct comparison of the benchmark and proposed models.

---

7. Project Design

Workflow overview:

1. **Exploratory Data Analysis (EDA):**
   - Load and inspect all datasets; document schema and attribute meanings.
   - Analyze missing values, distributions, and correlations.
   - Assess class imbalance in the campaign training labels.
2. **Data Preprocessing:**
   - Handle missing values (imputation or encoding).
   - Encode categorical variables.
   - Normalize or standardize numerical features as needed.
   - Align preprocessing across AZDIAS, CUSTOMERS, and MAILOUT datasets.
3. **Customer Segmentation:**
   - Apply PCA to reduce dimensionality and filter noisy features.
   - Use K-means (or similar) to cluster the general population.
   - Map customers to clusters and compute over/under-representation.
   - Visualize segments and summarize which profiles are most customer-like.
4. **Supervised Model:**
   - Use MAILOUT_TRAIN with the same preprocessed features.
   - Implement and tune models (e.g., logistic regression, random forest, XGBoost).
   - Use cross-validation and the chosen metrics (AUC-ROC, AUC-PR).
   - Apply the best model to MAILOUT_TEST and submit predictions to Kaggle.
5. **Documentation and Deliverables:**
   - Jupyter notebook(s) with code, visualizations, and commentary.
   - README with setup, dependencies, and usage instructions.
   - Blog post or report summarizing the process, findings, and business implications.

Software: Python 3.10 with libraries such as pandas, scikit-learn, XGBoost, and matplotlib/seaborn. All dependencies will be listed in a requirements file for reproducibility.

References

[1] Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. *Journal of City and Development*, 3(1), 12–30. https://pubs.sciepub.com/jcd/3/1/3

[2] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.

https://doi.org/10.1109/TKDE.2008.239

[3] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

[4] Ling, C. X., & Li, C. (1998). Data mining for direct marketing: Problems and solutions. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)* (pp. 73–79). AAAI Press. https://aaai.org/papers/00073-kdd98-011-data-mining-for-direct-marketing-problems-and-solutions