CAPSTONE PROJECT PROPOSAL

Customer Segmentation and Acquisition Optimization for Direct Marketing

Student Name: Aminat Shotade

Date: 14th,October 14, 2025
Program: Udacity Data Science Nanodegree

---

1. Domain Background

Direct marketing through mail-order campaigns has been a cornerstone of customer acquisition for decades. This project addresses a real-world business problem faced by Bertelsmann Arvato Analytics, helping a German mail-order company optimize marketing efficiency.

Traditional direct marketing campaigns operate with response rates of 1-3% (Stone & Jacobs, 2008), meaning significant resource waste. Machine learning enables companies to target high-propensity customers with greater precision. Market segmentation theory (Smith, 1956) suggests that dividing heterogeneous markets into homogeneous sub-markets improves marketing efficiency.

This project combines unsupervised learning for customer segmentation with supervised learning for response prediction, demonstrating how data science transforms business operations from reactive to proactive.

2. Problem Statement

How can a mail-order company efficiently identify individuals from the German population who are most likely to become customers?

The company currently employs broad, untargeted campaigns with 1-2% response rates, wasting 98-99% of marketing expenditure. Two fundamental gaps exist:

1. **Segmentation Gap:** Lack of systematic understanding of which demographic segments align with existing customers
2. **Prediction Gap:** Inability to reliably predict campaign responses before investment

Quantifiable Aspects:

- Current response rate: ~1-2% (baseline)
- Target: 2-3x improvement through targeting
- Dataset: 891,221 population, 191,652 customers, 42,982 campaign targets

The problem is measurable (ROC-AUC, response rates), quantifiable (percentage improvements), and replicable (standardized data and methodology).

3. Solution Statement

A two-stage machine learning pipeline combining unsupervised customer segmentation with supervised response prediction:

Stage 1: Customer Segmentation

- PCA for dimensionality reduction (366 → ~85 components)
- K-Means clustering (10-15 segments)
- Output: Over/under-represented customer segments

Stage 2: Response Prediction

- Ensemble classification (Random Forest, Gradient Boosting)
- Input: Campaign targets with engineered features
- Output: Probability scores for targeting

Quantifiable Goals:

- PCA: Retain 85% variance
- Clustering: Silhouette score >0.3
- Classification: ROC-AUC ≥0.70
- Business: Identify top 20% containing 40-60% of responders

The solution is measurable, replicable (fixed random seeds), and appropriate for direct marketing applications.

4. Datasets and Inputs

Four datasets from Bertelsmann Arvato Analytics:

1. AZDIAS (General Population): 891,221 × 366 features

- Represents German demographic distribution
- Person, household, building, neighborhood attributes

2. CUSTOMERS: 191,652 × 369 features

- Existing customer demographics
- Additional: customer group, online purchase, product preferences

3. MAILOUT_TRAIN: 42,982 × 367 features

- Campaign recipients with RESPONSE labels (0/1)
- Highly imbalanced (~1-2% positive)

4. MAILOUT_TEST: 42,833 × 366 features

- Test data for Kaggle evaluation
- RESPONSE withheld

Feature Documentation: Two Excel files detail attribute meanings and encoded values.

Data Challenges:

- Missing values encoded as -1, 0, 9, 'X'
- High dimensionality (366 features)
- Class imbalance (1-2% response rate)
- Mixed data types

Preprocessing Requirements: Missing value handling, feature selection (drop >80% missing), imputation, scaling, dimensionality reduction.

5. Benchmark Model

Primary Benchmark: Random Selection

- Method: Random individual selection
- Expected ROC-AUC: 0.50
- Expected response rate: 1-2%
- Represents current state without targeting

Secondary Benchmarks:

- Demographic filters (heuristics): ROC-AUC 0.55-0.60
- Basic logistic regression: ROC-AUC 0.65-0.70

Industry Standards:

- Good: 0.70-0.75 ROC-AUC
- Excellent: 0.75-0.80 ROC-AUC
- Outstanding: >0.80 ROC-AUC

Success Criteria: Achieve ROC-AUC ≥0.70, demonstrating 2-3x improvement in targeted response rates with interpretable customer segments.

6. Evaluation Metrics

Primary: ROC-AUC (Area Under ROC Curve)

- Threshold-independent evaluation
- Robust to class imbalance
- Range: 0.5 (random) to 1.0 (perfect)
- Target: ≥0.70

Secondary Metrics:

For Classification:

- **Precision@20%:** Response rate among top-scored individuals
- **Lift@20%:** Improvement over random selection (target: ≥2.0x)
- **Recall@20%:** Percentage of responders captured

For Segmentation:

- **Silhouette Score:** Cluster quality (target: >0.3)
- **Explained Variance:** PCA information retention (target: ≥85%)
- **Distribution Ratio:** Customer over-representation (>1.5 indicates target segment)

All metrics are interpretable, business-relevant, and appropriate for imbalanced classification and unsupervised learning evaluation.

7. Project Design

Phase 0: Data Preparation

1. Load datasets (AZDIAS, CUSTOMERS, MAILOUT_TRAIN)
2. Exploratory analysis: missing patterns, distributions, balance
3. Preprocessing: convert missing codes, drop high-missing features/rows, impute, remove categorical/constant features
4. Validation: ensure numeric, no NaN/inf, alignment

Phase 1: Customer Segmentation

1. Feature scaling (StandardScaler)
2. PCA: test variance retention, select ~85 components
3. Optimal K selection: elbow method, silhouette scores (K=10-15)
4. K-Means clustering on AZDIAS, predict for CUSTOMERS
5. Segment analysis: compare distributions, identify high/low-value clusters
6. Interpretation: characterize top segments

Phase 2: Response Prediction

1. Preprocess MAILOUT_TRAIN (same pipeline)
2. Apply fitted transformations (scaler, PCA, K-Means)
3. Feature engineering: PCA components, cluster membership, high/low-value flags, one-hot encoded clusters
4. Handle imbalance: SMOTE + RandomUnderSampler
5. Train models: Logistic Regression, Random Forest, Gradient Boosting
6. Evaluate: ROC-AUC, precision-recall, confusion matrix
7. Select best model, analyze feature importance

Phase 3: Test Predictions

1. Load MAILOUT_TEST
2. Apply full pipeline
3. Generate predictions
4. Create Kaggle submission file

Algorithm Selection:

- PCA: Fast, interpretable, handles correlation
- K-Means: Scalable, well-defined centroids
- Gradient Boosting: Handles complexity, robust to imbalance

Expected Challenges:

- High dimensionality → PCA
- Missing data → Systematic imputation
- Class imbalance → SMOTE + class weights
- Interpretability → Focus on cluster insights

Success Indicators:

- ROC-AUC ≥0.70
- Clear customer segments
- Actionable business recommendations
- Reproducible pipeline

References:

- Smith, W. R. (1956). Product differentiation and market segmentation. *Journal of Marketing*, 21(1), 3-8.
- Stone, B., & Jacobs, R. (2008). *Successful direct marketing methods*. McGraw-Hill.