Comparison of Customer Similarity Methods

Method 1: Cosine Similarity

Pros:

- Simple implementation
- Uses normalized customer profiles
- Focuses on specific customer features
- Computationally efficient

Cons:

- Limited feature set
- Only considers the first 20 customers
- Basic similarity scoring

Method 2: Nearest Neighbors

Pros:

- More comprehensive feature engineering
- Uses multiple data dimensions
- Handles categorical variables through one-hot encoding
- Considers the entire dataset
- More robust similarity calculation

Cons:

- More complex preprocessing
- Potentially higher computational complexity

Practical Recommendation

Preferred Method: Method 2 (Nearest Neighbors)

- More sophisticated feature representation
- Handles diverse data types
- Provides more nuanced customer similarities
- Scales better with larger datasets

Key Differences

Feature Selection:

- **Method 1:** Basic customer transaction features
- **Method 2:** Includes categorical encodings, temporal features

Similarity Calculation:

- **Method 1:** Pure cosine similarity on limited features
- **Method 2:** Cosine distance with richer feature matrix

Scalability:

- **Method 1:** Limited to first 20 customers
- **Method 2:** Processes the entire customer base

Recommendation for Production-Level Lookalike Modeling

- **Preferred Method:** Use Method 2 for its richer feature representation and robustness.
- **Further Refinements:**
  - Feature importance weighting
  - Advanced encoding techniques
  - Incorporating more sophisticated similarity metrics