LRFS: Efficient Customer Segmentation in E-commerce

By: Daniel Mehta

Motivation: Why Better Segmentation?

- **E-commerce is booming**
  - Online shopping has become the default for millions of consumers.
- **Not all customers behave the same**
  - Businesses need to understand who buys, how often, and why they leave.
- **Traditional RFM models are limited**
  - Focus only on Recency, Frequency, and Monetary value, but ignore time spent, page behavior, and exit patterns.
- **Need for smarter, behavior-aware segmentation**
  - Better grouping = better targeting = more revenue.

Existing Models & Their Limitations

Widely used models:

- **RFM**: Recency, Frequency, Monetary value
- Enhanced variants: **LRFM, WRFM, CLV, LRFMP**

What they improve:

- Add dimensions like Length, Cost, Churn, or Periodicity
- Use weights (e.g. AHP) and cluster methods like K-Means, SOM, Fuzzy C-Means

Key limitations:

- Rely on **monetary or transaction data**, often missing in web analytics
- **Ignore behavioral features** like time on site, exit intent, or bounce rates
  Can’t capture **session-level dynamics** or new user behavior

Identified gap:

- Few models use **Google Analytics session features** for segmentation. LRFS addresses this by integrating "**Staying Rate for Revenue**" from bounce and exit data.

What is LRFS?

A behavior-based segmentation model using:

Component	Meaning
L (Length)	Months of association with the site (based on Month + VisitorType)
R (Recency)	Time since last visit (12 - current month + 1)
F (Frequency)	Total page visits during session
S (Staying Rate for Revenue)	Engagement + contribution to purchase likelihood

Understanding the 'S' Component

$S = PageValues * (1 - ExitRates)$

- **Page Value**: Average value of a transaction
- **Exit Rate**: Frequency of users leaving without action
- **(1 - Exit Rate)** = Probability user stays and engages

Why S matters:

- Captures session-level **intent and engagement**
- **Moderately correlated with Revenue** (r = 0.49)
- **Low correlation** with L, R, F -> adds unique signal
- Crucial for segmenting **new or low-recency buyers** who still convert

Dataset & Preprocessing

Dataset Overview

Source: UCI Online Shoppers Intention Dataset (from Google Analytics)

Sessions: 12,330 unique user sessions

Target: Revenue (binary - purchase or not)

Class Imbalance: Only 15.6% of sessions led to purchases

Feature Engineering

Visit counts (Admin, Info, Product pages) -> aggregated into total_page_view

Durations merged into total_page_duration

Created key features: L, R, F, and S

S derived from: S = Page Value × (1 - Exit Rate)

Data Cleaning

Removed 125 duplicate rows

Dropped irrelevant columns:

- SpecialDay, OperatingSystems, Browser, Region, TrafficType, Weekend

Encoded Month and VisitorType for use in modeling

Feature Engineering Details

Estimating Time Without Timestamps

- No DateTime field available
- Used Month + VisitorType to approximate:
  - **L (Length):**
    - ReturningVisitor -> months since January
    - NewVisitor → 1 month
  - **R (Recency):**
    - 12 - current month + 1 (so December = 1)

Frequency and Revenue Proxy

- **F (Frequency): total page visits**
- **S (Staying Rate for Revenue):**
  - Page Value × (1 - Exit Rate)
  - High S = high engagement and likely purchase

Dimensionality Reduction & Clustering

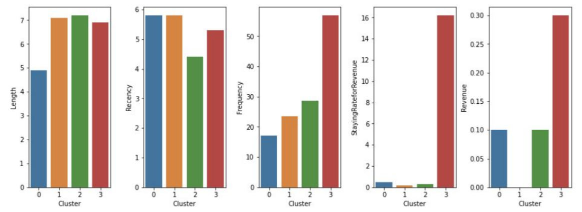

FIGURE 14. K-Means clustering analysis of LRFS model (t-SNE).

Why Reduce Dimensions?

- Improve clustering performance and visualization
- LRFS data is behavioral and multi-dimensional

Methods Used

- **PCA**: linear, captures variance
- **t-SNE**: non-linear, preserves local patterns
- **Autoencoder**: neural network, extracts deep structure

Clustering

- **K-Means**: centroid-based, simple but sensitive to noise
- **K-Medoids**: more robust, uses real points as centers

Result Highlights

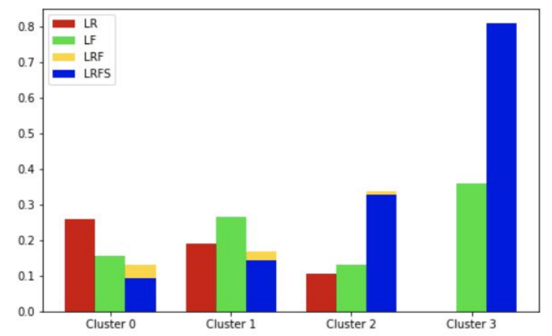

FIGURE 19. Differences of revenue among different clusters generated by LR, LF, LRF, and LRFS.

Model Comparison: Revenue by Cluster

Model	**Revenue Separation**
LR, LF, LRF	Moderate to weak cluster separation
LRFS	Clear revenue peak in Cluster 3

Why LRFS Wins

- **Only LRFS** (blue bars) shows a **sharp revenue spike** in Cluster 3.
- Indicates **stronger segmentation** of high-value customers.
- Confirms the added value of the "**S" (Staying Rate for Revenue)** feature.

Customer Typing

CPA Matrix

- Carriage Trade
- Passive
- Transaction
- Bargain Basement

CRM Matrix

- Loyal
- Potential
- New
- Uncertain

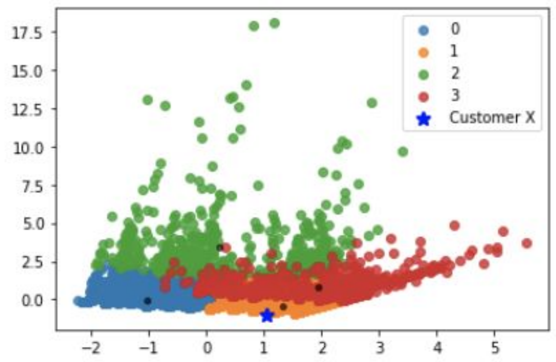

FIGURE 20. **Customer X test case.**

Customer X: Plotted on t-SNE map (Figure 20)

- **Cluster 1**: Transaction / Uncertain
- **Cluster 3**: Passive / Loyal

Limitations & Future Work

Limitations:

- **No timestamp**: prevents true recency or time-based patterns
- **Revenue is binary (0/1)**: limits granularity in customer value analysis
- **Session-level data only**: no multi-session behavior tracking

Future Work:

- Apply LRFS to richer datasets with timestamps and revenue amounts
- Explore real-time segmentation with dynamic LRFS updates
- Test generalizability across industries and traffic sources

Conclusion

- **LRFS enhances customer segmentation** by introducing the **"S" (Staying Rate)** feature
- **Delivers improved clustering** performance compared to LR, LF, and LRF models
- **Enables better targeting** through personalized marketing and retention strategies

"Thank you

Works Cited

R. Hayat Khan, D. Fabian Dofadar, M. G. R. Alam, M. Siraj, M. Rafiul Hassan and M. Mehedi Hassan, "LRFS: Online Shoppers’ Behavior-Based Efficient Customer Segmentation Model," in IEEE Access, vol. 12, pp. 96462-96480, 2024, doi: 10.1109/ACCESS.2024.3420221.