

Customer Profitability Analytics in Retail Banking: An RFM-Based Clustering and Customer Lifetime Value Approach

Nagamani K¹, Dr. R N Kulkarni²

¹Department of Management Studies, Ballari Institute of Technology and Management, Ballari.

²Department of Computer Science and Engineering, Ballari Institute of Technology and Management, Ballari.

Abstract

Retail banks accumulate large, heterogeneous customer bases, but existing literature and industry practice often treat customer value measurement, behavioural segmentation, and profitability estimation as separate exercises rather than a single connected pipeline. This study addresses that gap by building an integrated, data-driven framework that combines Recency-Frequency-Monetary (RFM) analysis, K-Means clustering, and a customer profitability/Customer Lifetime Value (CLV) model, applied to a publicly available, large-scale retail banking transaction dataset. Using 10,48,567 anonymised transaction records sourced from Kaggle, the data was cleaned to a verified 10,40,794 records covering 8,78,763 unique customers. RFM scoring classified customers into six behavioural segments, and K-Means clustering ($K = 5$) grouped them into five value-based segments, which were then compared with the independent RFM classification. A profitability model combining a transaction-fee revenue proxy, a deposit-based Net Interest Margin (NIM) proxy, and a servicing-cost estimate produced a total estimated gross profit of Rs 345.88 Crore and a projected 36-month CLV of Rs 2,175.02 Crore, with 84.9 per cent of customers profitable. The analysis found that deposit-based revenue accounts for 99.8 per cent of total revenue, that profit is heavily concentrated (7.27 per cent of customers generate 70.49 per cent of profit), and that age distributions were broadly similar across behavioural clusters — a finding with direct implications for how retention and cross-selling should be targeted. The study translates these findings into segment-specific retention and cross-selling recommendations and discusses their implications for customer relationship management in retail banking, while documenting the assumptions and limitations of the modelling approach used.

Keywords

Customer Profitability Analysis; RFM Analysis; K-Means Clustering; Customer Lifetime Value; Retail Banking Analytics; Customer Segmentation

1. Introduction

Retail banking generates revenue from a very large and behaviourally diverse customer base, and it is well understood in both practice and research that not every customer contributes equally to a bank's profitability. A small number of relationships can account for a disproportionate share of profit, while a much larger group barely covers the cost of servicing their accounts. Despite this, many banks continue to allocate retention effort, relationship management attention, and cross-selling campaigns fairly uniformly across their customer base, largely because the transactional and account data needed to build a structured, customer-level profitability view is rarely converted into one.

This project was designed to close that gap. It uses a publicly available, anonymised retail banking transaction dataset to build an end-to-end analytical pipeline that measures customer behaviour through RFM (Recency, Frequency, Monetary) analysis, groups customers into behaviourally meaningful segments through K-Means clustering, and estimates each customer's profitability and projected lifetime value through an explicit, assumption-documented

model. The resulting segment-level view is then used to develop concrete retention and cross-selling recommendations.

Specifically, the study cleans and prepares a large-scale retail banking transaction dataset, computes RFM metrics to classify customers into behavioural segments, and applies K-Means clustering to group customers by similarity. It then builds a profitability model to estimate Gross Profit and CLV at the customer and cluster level, examines how profitability varies across segments, age groups, gender, and location, and translates these findings into segment-specific retention and cross-selling recommendations for retail banking management.

2. Literature Review

The analytical foundation of this study draws on four streams of prior work: RFM-based customer analysis, the relationship between RFM and Customer Lifetime Value, customer profitability modelling, and profitability estimation specifically within retail banking.

The idea of scoring customers on how recently, how often, and how much they have transacted goes back to Hughes (1994) and his work on database marketing, and it has stayed popular for a simple reason: it needs nothing beyond a transaction history, and it correlates well with customer engagement and value across a range of retail and financial-services contexts. A later review of RFM applications by Wei, Lin and Wu (2010) points to an important limitation, however — RFM is descriptive of past behaviour rather than predictive of future value on its own, which is one reason this study does not stop at RFM scoring and instead uses it as an input to clustering.

A more direct link between RFM and Customer Lifetime Value was established by Fader, Hardie and Lee (2005), who showed that a customer's position on RFM dimensions can be used to project value forward rather than only describe what has already happened. This distinction — using a backward-looking score to build a forward-looking estimate — is essentially what this study does as well, using RFM scores as clustering inputs and then projecting CLV separately.

Profitability itself was examined by Reinartz and Kumar (2000) in a noncontractual retail setting, where they found that the commonly assumed link between how long a customer stays and how profitable they are is weaker, and more segment-dependent, than is often assumed. Their conclusion — that profitability needs to be measured directly rather than inferred from loyalty or tenure — is the reason this study computes profitability explicitly from transaction and balance data rather than treating recency as a stand-in for value.

Most relevant to the present context is Haenlein, Kaplan and Beeser's (2007) model for determining Customer Lifetime Value specifically in retail banking. Their key point is a practical one: a bank's revenue comes from structurally different sources — transaction fees and interest margin on deposits and loans — and these need to be modelled separately rather than combined into one undifferentiated figure. The profitability model in this study follows exactly that logic, splitting customer revenue into a transaction-fee component and a deposit-margin component before combining them.

Taken together, this literature establishes RFM analysis, cluster-based segmentation, and CLV/profitability modelling as individually well-supported techniques, and specifically flags the importance of modelling banking revenue as a composite of fee and margin income rather than a single figure. What is less commonly demonstrated is these techniques applied together, end to end, on one dataset, with the resulting segments and profitability figures then carried through into specific retention and cross-selling recommendations — which is the contribution this study makes.

3. Research Methodology

3.1 Data Source

This study uses the “Bank Customer Segmentation” dataset published on Kaggle by Bansal (n.d.), comprising anonymised Indian retail banking transaction records: TransactionID, CustomerID, CustomerDOB, CustGender, CustLocation, CustAccountBalance, TransactionDate, TransactionTime, and TransactionAmount (INR). The raw dataset contains 10,48,567 transaction records.

3.2 Data Cleaning

Records with missing values in CustomerDOB, CustGender, CustLocation, or CustAccountBalance (6,953 records) were removed. Two systematic date-quality issues were corrected: a two-digit-year parsing error that placed some birth years in the 21st century instead of the 20th (1,17,127 records corrected), and a placeholder date of birth (1 January 1800) used in 56,292 records, which was treated as missing and imputed with the median date of birth. Customer age was derived using a fixed reference date of 31 October 2016. A further 820 transactions recorded at zero value were removed as not representing genuine monetary activity. This process reduced the dataset to 10,40,794 records (a removal of 0.74 per cent of raw records), covering 8,78,763 unique customers.

3.3 RFM Analysis

Using 31 October 2016 as the reference date, Recency (days since last transaction), Frequency (transaction count), and Monetary value (total transaction amount) were computed for every customer and independently scored on a 1–5 scale using quintile-based binning. Customers were classified into six standard RFM-based segments — Champions, Loyal Customers, Potential Loyalists, At Risk, Lost Customers, and New Customers — based on their combined R, F, and M scores.

3.4 Customer Segmentation via K-Means Clustering

To complement the rule-based RFM segments with a data-driven grouping, K-Means clustering was applied to each customer's standardised R, F, and M scores, using K-Means++ initialisation with 10 initialisations to arrive at a stable solution. The number of clusters was evaluated using the Elbow Method ($K = 2$ to 10) and the Silhouette Score ($K = 2$ to 8, computed on a fixed 50,000-customer sample). $K = 5$ was retained as the final solution, chosen primarily for business interpretability — aligning naturally with the 1–5 RFM scale and producing a practically actionable number of segments — rather than because it was the strongest statistical fit: the Silhouette Score for $K = 5$ (0.2974) was neither the highest in the tested range nor a local peak, with $K = 8$ producing the highest score (0.3113). This is reported transparently because it affects how much statistical weight the resulting cluster boundaries should be given, even though the clusters remain interpretable and, as later compared with the independent RFM segmentation, behaviourally coherent.

3.5 Customer Profitability and CLV Model

Following the logic of modelling retail banking revenue as a composite of fee and margin income (Haenlein, Kaplan, & Beeser, 2007), each customer's Gross Profit was estimated as:

$$\text{Gross Profit} = \text{Transaction Revenue} + \text{Deposit Revenue} - \text{Servicing Cost}$$

where Transaction Revenue = Monetary value $\times$ 0.5% (a transaction-fee proxy), Deposit Revenue = Average Account Balance $\times$ 3.5% (a Net Interest Margin, or NIM, proxy consistent with disclosed margins for large Indian private-sector banks in recent years), and Servicing Cost = Rs 50 base cost + Rs 30 $\times$ Frequency. Customer Lifetime Value (CLV) was projected over a 36-month horizon by extending each customer's observed monthly revenue rate

(Monetary value divided by the 81-day, or 2.7-month, observation window) forward: $CLV = (\text{Monetary} \div 2.7 \text{ months}) \times 36 \text{ months}$. Because the dataset does not include a bank's actual internal cost and margin data, these rates are used as documented proxies rather than bank-specific figures, and the resulting Gross Profit and CLV values are best interpreted as a demonstrative, internally consistent framework rather than audited financial figures.

3.6 Analytical Tools

Data cleaning, RFM computation, clustering, and profitability modelling were carried out in Python using the pandas library (McKinney, 2010) for data manipulation and scikit-learn (Pedregosa et al., 2011) for K-Means clustering and Silhouette Score evaluation.

4. Results and Findings

4.1 RFM Segment Distribution

Applying the RFM scoring methodology to all 8,78,763 customers produced six behavioural segments (Table 1). Potential Loyalists and Loyal Customers together account for close to 69 per cent of the customer base, indicating a large pool of customers with positive but not maximal engagement. Champions, the highest-value segment, represent 12.55 per cent of customers, while At Risk and Lost Customers combined represent 17.57 per cent — a tangible but not overwhelming retention challenge.

Table 1: Distribution of Customers Across RFM Segments

RFM Segment	Customers	% of Total
Potential Loyalists	3,46,753	39.46%
Loyal Customers	2,59,485	29.53%
At Risk	1,21,199	13.79%
Champions	1,10,313	12.55%
Lost Customers	33,204	3.78%
New Customers	7,809	0.89%

4.2 Cluster Profiles

K-Means clustering ($K = 5$) produced five reasonably balanced clusters, ranging from 18.20 per cent to 21.95 per cent of customers each (Table 2). Each cluster's dominant RFM segment composition compares sensibly with the independent RFM classification — for example, Cluster 4 is 66.7 per cent Champions, while Cluster 2 is 63.8 per cent At Risk — supporting the interpretability of the clustering solution.

Table 2: K-Means Cluster Profiles (K = 5)

Cluster	Business Label	Customers	% of Total	Avg Recency (days)	Avg Frequency	Avg Monetary (Rs)
0	High-Value Inactive	1,59,900	18.20%	63	1.00	3,185
4	High-Value Active	1,65,469	18.83%	52	1.70	3,878
3	Moderate-Value Frequent	1,81,108	20.61%	79	1.19	2,229
1	Low-Value Recent	1,92,850	21.95%	53	1.06	198
2	Low-Value Inactive	1,79,436	20.42%	79	1.00	214

4.3 Overall Profitability

The profitability model produced a total estimated Gross Profit of Rs 345.88 Crore and a total projected 36-month CLV of Rs 2,175.02 Crore across the portfolio. Of 8,78,763 customers, 7,46,441 (84.9%) were profitable and 1,32,322 (15.1%) were not. Average Gross Profit per customer was Rs 3,936.02, against a median of only Rs 573.85 — a considerable gap indicating a right-skewed distribution in which a relatively small number of high-balance customers pull the average upward.

Deposit-based revenue (the NIM proxy) accounted for 99.8 per cent of total revenue, with transaction-fee revenue contributing only 0.2 per cent.

4.4 Cluster-wise Profitability

Table 3 ranks the five clusters by average Gross Profit per customer. Cluster 0 (High-Value Inactive) generates the highest average profit per customer (Rs 6,139) despite having the lowest transaction frequency of the two high-value clusters — its profitability is driven almost entirely by its high average account balance (Rs 1,77,229) rather than transactional activity. Cluster 4 (High-Value Active), while second in average profit, holds the highest average CLV (Rs 51,701) due to its higher transaction frequency. Clusters 0 and 4 together represent only 37.0 per cent of customers but contribute 53.16 per cent of total profit.

Table 3: Cluster-wise Profitability Summary

Cluster	Business Label	Avg Gross Profit	Total Profit (Rs Cr)	Profit Share	Avg CLV
0	High-Value Inactive	Rs 6,139	98.16	28.38%	Rs 42,461
4	High-Value Active	Rs 5,179	85.70	24.78%	Rs 51,701
3	Moderate-Value Frequent	Rs 4,373	79.19	22.90%	Rs 29,726
2	Low-Value Inactive	Rs 2,265	40.65	11.75%	Rs 2,855
1	Low-Value Recent	Rs 2,187	42.18	12.19%	Rs 2,645

4.5 Profitability Concentration

Classifying customers into profitability tiers (Unprofitable, Low, Moderate, High, Very High) shows an even sharper concentration than the cluster-level view: the Very High tier, just 7.27 per cent of customers, generates 70.49 per cent of total profit. The Unprofitable tier (15.06 per cent of customers) contributes a small net negative to overall profit (−0.21%), indicating these customers cost marginally more to service than they generate in revenue under this model.

4.6 Age, Gender, and Location Patterns

Average profitability rose consistently with customer age, from Rs 1,358 for the 18–25 age group to Rs 16,877 for the 56+ age group — a more than twelvefold increase with no reversals across the five age bands. The 26–35 age group was the single largest contributor to total profit (47.45 per cent of profit from 52.63 per cent of customers), while the 56+ age group, despite comprising only 2.25 per cent of customers, generated the highest per-customer profit of any age band.

A cross-tabulation of cluster membership against age group produced a finding that qualifies the cluster analysis above: the 26–35 age band was the largest single age group within every one of the five clusters (ranging narrowly from 50.9% to 54.0%). A modest skew was still observed, with Cluster 0 showing the highest combined 36+ age share (27.8%) of any cluster, roughly double that of Clusters 1 and 2 (13.8% and 14.3%).

Profitability differences by gender were modest (average Gross Profit of Rs 3,984 for male customers versus Rs 3,807 for female customers), while profit contribution by location was concentrated in Mumbai, New Delhi, and Gurgaon, with New Delhi and Gurgaon customers showing a higher average profit per customer than Mumbai despite smaller customer counts.

5. Discussion

The finding that deposit-based revenue accounts for 99.8 per cent of total revenue is the single most consequential result of this study, and it reframes what “customer value” should mean for this portfolio. It suggests that transaction activity, while useful as a behavioural signal for segmentation, is not where retail banking profit in this dataset actually originates — balance retention is. This is consistent with Haenlein, Kaplan and Beeser's (2007) emphasis on modelling deposit-margin income separately from fee income, and it explains a result that would otherwise look counter-intuitive: Cluster 0 (High-Value Inactive), whose customers transact only about once on average, is more profitable per customer than Cluster 4 (High-Value Active), whose customers transact more frequently. Engagement, in this model, is not synonymous with value.

This has a direct bearing on retention strategy. A bank guided only by activity metrics would likely direct the least attention toward Cluster 0, since low frequency and declining recency (63 days, the third-highest in the portfolio) typically read as disengagement signals. Read alongside the profitability figures, however, this cluster represents the most urgent retention priority in the portfolio: high current value combined with quietly weakening engagement is a materially different risk profile from a customer who is both low-value and low-engagement.

Age distributions being broadly similar across clusters — with the 26–35 band dominant in every cluster — has a similar reframing effect on how segmentation should be used. It indicates that behavioural clustering and demographic targeting answer different questions and should be applied together rather than assumed to overlap: a bank cannot infer a cluster's likely age profile from its behavioural label, and equally cannot infer a customer's likely cluster from their age. This matters practically for cross-selling, where product relevance is often reasoned about in age terms (for example, home loans for a particular life stage) — the results here suggest that such reasoning needs to be applied as a targeting filter within a cluster, not as a description of the cluster itself.

The sharp concentration of profit — 70.49 per cent from just 7.27 per cent of customers — is broadly consistent with Pareto-style concentration patterns commonly observed in retail financial services, and reinforces the practical case for the kind of segment-level, rather than uniform, resource allocation this study argues for in its introduction. In short, these findings suggest that customer profitability analytics adds the most value not by replacing behavioural segmentation, but by correcting the assumptions a bank might otherwise make from behavioural data alone.

6. Managerial Implications and Strategic Recommendations

The findings above translate into a segment-prioritisation logic for retention effort, ranked by average Gross Profit, average CLV, and the proportion of profitable customers within each cluster: Cluster 0 (High-Value Inactive) as the top priority, followed by Cluster 4 (High-Value Active), Cluster 3 (Moderate-Value Frequent), Cluster 1 (Low-Value Recent), and Cluster 2 (Low-Value Inactive).

6.1 Retention Recommendations

- Cluster 0 (High-Value Inactive): dedicated relationship management for high-balance customers within this cluster, a churn early-warning system triggered by balance decline, and re-engagement communication that does not demand increased transaction activity — since deposit retention, not transaction growth, is what protects this cluster's profit contribution.
- Cluster 4 (High-Value Active): a loyalty programme rewarding transaction frequency and balance maintenance, with proactive service-quality monitoring. This cluster is 66.7% Champions by RFM classification, and Champions are generally understood in the CRM literature to be less price-sensitive and more relationship-sensitive; the project itself did not test satisfaction or churn drivers directly, so this point is offered as a managerial hypothesis to monitor rather than a finding of the analysis.
- Cluster 3 (Moderate-Value Frequent): balance-growth incentives and a digital re-activation campaign, given its combination of the highest average recency (79 days) in the portfolio with a still-meaningful average balance.
- Clusters 1 and 2 (Low-Value Recent / Low-Value Inactive): a value-graduation programme for Cluster 1, given its size (21.95% of customers, the largest cluster) and low but recent engagement; and lower-cost, digital-only servicing for Cluster 2, given its thin or negative profit margins.

6.2 Cross-Selling Recommendations

Because the dataset does not indicate which products customers currently hold, cross-selling recommendations here are grounded in customer financial profile (balance, monetary value) and portfolio-wide age-profitability patterns rather than actual product-holding data. High-balance clusters (0, 3, 4) are reasonable candidates for Fixed Deposit conversion, wealth management, and mutual fund distribution products; the highly active Cluster 4 is a reasonable candidate for premium credit cards and pre-approved personal loans; and the younger-skewing, lower-balance Clusters 1 and 2 are more suited to micro-investment products (recurring deposits, digital gold) that build a savings habit at low commitment.

Because the 56+ age group shows the highest average profitability of any age band (Rs 16,877) despite representing only 2.25 per cent of customers, and because this age group is not concentrated in any single cluster, a dedicated senior-citizen relationship programme applied across clusters — rather than confined to one segment — is a reasonable response to this finding. Similarly, the 26–35 age group, as the largest single contributor to total profit (47.45%) and the dominant age band in every cluster, represents the base that retention and cross-selling investment should protect and grow across the portfolio as a whole.

7. Limitations of the Study

- The transaction-fee (0.5%) and NIM (3.5%) rates used in the profitability model are industry-benchmark proxies, not a specific bank's actual internal figures, since such data was not available.
- The dataset covers only an 81-day window (1 August to 21 October 2016), which constrains the reliability of Recency and Frequency measures and requires the 36-month CLV projection to extrapolate substantially beyond the observed period.
- $K = 5$ was selected primarily for business interpretability; the Silhouette Score analysis does not identify $K = 5$ as the statistically strongest option among the values tested.
- The dataset does not include information on which specific banking products customers hold, historical churn outcomes, or marketing campaign response rates, limiting the cross-selling and retention recommendations to a profile-based rather than an outcome-validated approach.
- All findings are based on a single, anonymised, publicly available dataset from one time period and have not been cross-validated against data from another bank, period, or geography.

8. Conclusion

This study set out to build a structured, customer-level view of profitability for a retail banking customer base, bringing RFM analysis, K-Means clustering, and an explicit profitability/CLV model together into a single applied pipeline rather than treating them separately, as much of the existing literature does. Working through a large, publicly available transactional dataset, the analysis shows that behavioural segmentation and profitability measurement genuinely need each other: segments that look similar in size or engagement can differ sharply in the value they actually hold, and dimensions such as customer age can matter for profitability without being visible in the segmentation at all. The retention and cross-selling recommendations that follow from this analysis are offered as a replicable framework for retail banking customer management — one built and documented with its assumptions and limitations made explicit, rather than as a substitute for a bank's own internal financial and customer data.

References

1. Bansal, S. (n.d.). Bank Customer Segmentation [Dataset]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation>
2. Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4), 415–430.
3. Haenlein, M., Kaplan, A. M., & Beeser, A. J. (2007). A model to determine customer lifetime value in a retail banking context. *European Management Journal*, 25(3), 221–234.
4. Hughes, A. M. (1994). *Strategic database marketing*. Chicago: Probus Publishing Company.
5. McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*.
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
7. Reinartz, W. J., & Kumar, V. (2000). On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing. *Journal of Marketing*, 64(4), 17–35.

8. Wei, J. T., Lin, S. Y., & Wu, H. H. (2010). A review of the application of RFM model. African Journal of Business Management, 4(19), 4199–4206.