

International Journal of Knowledge Processing Studies (KPS)



Homepage: <http://kps.artahub.ir/>



ORIGINAL RESEARCH ARTICLE

A Prospectively Validated K-Means Framework for Demand Tiering and Acquisition Prioritisation in Academic Library Collections

Yuvraj G. Jadhav^{1,*}, Rajshri R. More²

¹ Associate Professor in the Department of Library and Information Science, Shivaji University, Kolhapur, Maharashtra, India. ygj.lib@unishivaji.ac.in, 0000-0001-9865-1861.

² Librarian, Gopal Krishna Gokhale College, Kolhapur, Maharashtra, India, rajshrirmore@gmail.com, 0009-0006-6250-7538.

ARTICLE INFO

Article History:

Received: 2026-03-12

Revised: 2026-04-21

Accepted: 2026-05-19

Published Online: 2026-06-01

Keywords:

Library circulation analysis, K-Means clustering, Demand tiering, Collection development, Temporal validation, Decision support.

Number of Reference: 16

Number of Figures: 5

Number of Tables: 1

DOI: 10.22034/kps.2026.595592.1296



ABSTRACT

Though academic libraries have access to wide-ranging circulation data, collection decisions are still typically based on descriptive use reports, which have shown an inability to identify sustained demand titles, or anticipate demand for titles. A data-driven approach to create ordered demand tiers, and an ordered acquisition priorities list, from standard circulation data is proposed and validated. The circulation data for the period June 2024 to June 2026 was collected from the Knowledge Resource Centre, SNDT Women's University, Pune. A total of 4984 transactions on 2615 titles and 915 borrowers were analyzed. Five title-level circulation indicators were employed: total circulation transactions, unique borrowers, active months, transaction per active month, and renewals. The consecutive design involved computing features on the date range of January to December 2025 based on 1,626 titles, and testing prediction on January to April 2026 circulation levels. We compared K-Means clustering with Ward hierarchical clustering and DBSCAN using four internal indices and a stability test based on 30 random initialisations. A composite demand score was calculated for the indicators after normalization using four weighting schemes. However, performance of the composite score was similar to a simple transaction-count baseline. Therefore, the main benefit of the composite score was its use of an interpretable tier structure, rather than improved ranking performance. It provides a framework for reproducible acquisition prioritization, duplicate-copy planning, and identifying underused holdings. ©authors.

► **Citation:** Jadhav, Y., & More, R. (2026). A Prospectively Validated K-Means Framework for Demand Tiering and Acquisition Prioritisation in Academic Library Collections. *International Journal of Knowledge Processing Studies (KPS)*, 6(2): 12-30. Doi: 10.22034/kps.2026.595592.1296

1. Introduction

In an academic library setting, integrated library management systems record every loan, renewal, and return transaction processed by that library, providing a continuous record of which titles are moving, who borrowed which, and when. The data provided by these systems are created by the system, cover the full population of users, and are not subject to response bias, making them increasingly adopted as an analytical resource rather than an administrative by-product. As a result, there is currently growing interest in the application of AI or analytics in library operations (Narendra et al., 2025).

There is a real-world pressure behind this interest: budgets for academic library acquisitions have not kept pace with the increase in output or the inflation of the prices of scholarly materials, creating opportunity costs for each acquisition decision. However, librarians in university libraries have been increasingly called to document collection decisions with evidence-based assessment rather than solely relying on professional judgement. Several analytics techniques to analyze data from university libraries have been proposed to help document evidence-based collection decision-making (Roy, 2025). For example, circulation figures are a better measure of what was borrowed than what was thought to be useful.

In India, surveys of librarians have shown enthusiasm for artificial intelligence, but limits to technical skill and infrastructure have limited its uptake (Kalbande et al., 2024). Similarly, the library literature has stressed the gap between the analytical possibilities offered by library systems and their routine use in practice (Kumar & Jyoti, 2024). Consequently, most libraries still use descriptive circulation reports, tallied monthly or by class number, which do not indicate the pattern of use (or lack of use) of each title, nor the likelihood that demand will persist.

Fortunately, there are several tools in data mining that can solve this problem. Data

mining is the process of discovering patterns and extracting knowledge from large data sets (Han et al., 2012). Of the many data mining techniques, clustering is the most common with transactional data. Many studies of library information resources have found that circulation data can be useful to distinguish collections (Zhang et al., 2024). K-Means is the most widely used approach for clustering due to its low computational complexity, compactness of clusters, and interpretable results for a general audience (Jain, 2010). Although there have been several applications of the algorithm to library circulation data, the general tendency is for circulation data to separate into interpretable usage bands.

There are three methodological questions in applying clustering to a circulation database: first, the number of clusters, which is determined by the human analyst at the beginning and affects all subsequent results; second, the interpretation of the cluster labels; K-Means clustering yields arbitrary integers as labels and any ordinal labeling thus requires a definition of rules; and third, validation, which is the most important concern. For example, the score computed from past circulation data orders items based on their circulation history. Whether this prediction holds is an empirical question, and requires evaluation against data not seen by the model. With recent work on applying adaptive boosting to demand-driven acquisition in an academic library environment, the importance of holding out data and reporting predictive accuracy explicitly also holds for clustering-based frameworks (Walker & Jiang, 2019).

These queries are taken up with the help of circulation data collected from Knowledge Resource Centre, SNDT Women's University, Pune and an attempt is made to evolve a reproducible methodology that converts simple circulation figures into demand and gives an acquisition priority right for the books. There are four objectives in this: To categorize the circulating titles into demand tiers according to an explicit ordering rule and an explicit

justification of the number of demand tiers. The second is creating a composite demand score, specifying a formula, specifying weights, and performing a sensitivity analysis. The third is comparing the resulting rankings against circulation during a later period and against simple baselines. Fourth, monthly circulation data spanning more than one academic year offer the opportunity to test rather than simply assert seasonal circulation patterns.

To the authors' knowledge, this is the first prospective validation of a clustering-based library demand framework for Tier assignments and Score rankings against circulation in unrelated later periods. Additionally, it shows that internal validity indices cannot adequately derive cluster counts when the data fundamentally represent a single discrete counting dimension, as is the case for circulation data, and provides a better alternative. In addition, it produces the composite demand score in fully reproducible form and some weight-sensitivity analysis showing that the ranking of demand is strong to the particular coefficients employed. Finally, it generates a negative finding that the five indicators don't rank titles reliably better than transaction counts, but provides an interpretable, tiered structure for the most and least relevant titles.

The rest of the article is divided as follows. Section 2 defines terms and reviews literature on circulation analysis, clustering to support collection management, and predictive validation. Section 3 describes the data, preprocessing, time structure, feature engineering, clustering method, composite score, and evaluation metrics. Section 4 describes data structure, cluster validity, tier structure, sensitivity of scores to weights, potential for prospective validation, and monthly circulation for different clusters over time. Section 5 applies these results to collection management practice, Section 6 discusses the limitations of this study, and Section 7 concludes and outlines potential future work on this problem.

2. Literature Review

2.1 Definitions of key concepts

Some of the terms used in this literature are at times used with very different meanings across studies, and it will become important to understand very precisely the way in which the terms will be used in this analysis. A circulation transaction is a single recorded event in the library management system. These are issue, renewal, and return. Some studies treat returns separately from issues and renewals, while others combine the initial issue and the return of matching loans. Therefore, these values are not comparable unless it is stated which convention is used.

Demand, or use, in a print collection is the number of times a title is borrowed over a time period. It is often used as a proxy for user need, even though circulation statistics may not capture in-library use, nor need which cannot be met because materials are not held. Where a study distinguishes between historical demand and demand in the future, the difference is whether the measure is applied to a time period in the past, or a future time period. A demand score constructed on the basis of a time period and applied to that time period is an observation-based summary of use. A demand forecast refers to a future period.

Clustering is an unsupervised task in which observations are grouped together such that members of a cluster are more similar to each other than to those of other clusters (Han et al., 2012). K-Means, a partitional clustering algorithm, is a heuristic that iteratively assigns each observation to the nearest of k centroids, such that the aggregate within-cluster variance is minimized (Jain, 2010). This algorithm requires the number of clusters k to be specified a priori, and the resulting cluster identifiers are unordered integers.

Cluster validity indices are internal indices when they rely on data used to fit the model. Examples of such indices include the silhouette coefficient, which compares within and between cluster cohesion, and the Calinski-Harabasz and Davies-Bouldin indices, which are ratio indices of similar measures of spread or dispersion. These

indices measure only the geometry of the underlying partitions, without distinguishing whether or not the partition has any meaning outside the model, while prospective validation is represented by a model on outcomes after the period on which it is conditioned. In cold-start cases where there is no available history, no such model can be computed.

2.2 Circulation records as evidence for collection decisions

As a standard measure of collection use, circulation statistics have a long history. Similarly, studies that analyze public library circulation patterns show that borrow histories are a stable measure of reading interests and useful in planning and service (Galyani-Moghaddam & Taheri, 2021). These records are generated automatically and continuously, avoiding the recall problems and low response rates of survey-based assessment and covering the entire borrowing population rather than a self-selected sample of the borrowing population.

Their limitations are plain, and well documented in the literature. Circulation, in contrast to reading, does not preserve the difference between whether the title was borrowed and read or borrowed and returned unread. In-house consultation leaves no documentary record of use, systematically under counting usage of reference works and whatever other materials users choose not to remove from the library building. Library use counts register only the collection the library owns and leave no evidence of demand for works the library never purchased.

This has not prevented circulation data from becoming the primary evidence used for collection assessment, but it has affected its use: the literature on analytics in academic libraries argues that systematic processing of transactional data can inform acquisitions and retention decisions, but this seldom happens (Roy, 2025). The techniques covered below are attempts to close the gap between what data is available and what evidence it can provide.

2.3 Clustering approaches to library collection management

K-Means clustering has been more popular in some of this literature. For example, one of the first applications of this was to cluster library holdings by number of times borrowed. Having shown that usage bands could be determined solely from loan records (Tamba et al., 2019), the next study using the method discussed identifying candidate titles for collection development and considered clustering as an use of the method, for acquisition support, rather than as a descriptive tool (Asminah, 2022). Both studies found structure in the circulation data, but both studies used limited input data and only briefly discussed the cluster quality.

Later work has broadened the application without altering the method. For example, library loan records have been clustered for items that are frequently or infrequently borrowed so that purchasing decisions can be made; these clusters can be interpreted as recommendations (Tanjung & Sriani, 2025). The other line of inquiry has focused on borrower groups of similar behavior, to tailor services to different types of user (Monika et al., 2025). Studies of university borrowing behavior have classified the usage of the collection as high, medium, or low-circulation groups (Sinulingga et al., 2025), with the aim of recognizing patterns in overall collection usage. Clustering output has also been used as input to book recommendation, where the cluster membership could be used to determine which titles to recommend (Safira & Mohammed, 2026).

K-Means is not only a library-centric method: it has also been employed for behavioral segmentation in other fields, such as modeling driver risk profiles based on transportation safety records (Badanina & Sudakov, 2022). These applications show that K-Means works well with count-based behavior data, and it is therefore a promising candidate for circulation records.

Three limitations were common in the library studies. First, the number of clusters is rarely justified, meaning that the report

contains no validation analysis that would allow the reader to ascertain the defensibility of the chosen partition or whether a different number would have led to different recommendations. Ordinal demand labels assigned to the clusters do not specify any ordering among the clusters, thus the order is arbitrary because the method produces unordered identifiers. These composite demand indices are not specified as to formula, weights or normalization. Therefore it is not possible to verify the results or to know if the weightings are reasonable for the constituent commodities.

2.4 Predictive modelling and validation

Other experimental work recasts the collection questions as prediction problems that measure predictive accuracy. An adaptive boosting model trained on the title-level demand-driven acquisition records to predict which titles will be triggered for purchase, produces an accuracy of 0.824 on held-out records compared to approximately ten per cent explained variance with logistic regression [8]. There are two relevant aspects of that design. The data was separated into a training and a testing set prior to modeling, so there was no risk of information leakage. The performance was reported as a single interpretable figure, which a librarian can compare to the cost of acting.

Clustering studies in the library domain have largely not used this standard, because clustering is an unsupervised method that does not have an associated outcome variable to solve to: validation of clustering models is conducted using internal validation measures (e.g. silhouette coefficient), which measure the geometric separation of clusters in the data on which the model is fitted, and not whether those clusters can be used to predict future behavior.

This can be problematic when the clustering output is for a forward-looking purpose. When the tiers or scores are put forward as a basis for acquisition, it is implicitly assumed that those rated highly now are going to be in demand. This can be validated by withholding information about

circulation at some later time from the model while it is created, and then using the model to predict the withheld future data. The quality of the model as evaluated in this way can be judged with the same measures as supervised work.

2.5 Research gap

The studies do make it clear that such data can be clustered into useful groupings, and those groupings can be used to inform collection management practices. They do not, however, prove that those cluster-based rankings actually match up to circulation data later on, or that these numbers of clusters or interpretations are statistically or operationally defensible, or that a demand index can be calculated. The present study addresses these three gaps under a single analysis framework using an integrated, complete academic library circulation historical data set spanning 2+ years.

3. Method

3.1 Study setting and data source

The Knowledge Resource Centre, SNDT Women's University, Pune's print circulation data was exported from the integrated library management system as a comma separated file for each borrower type. Each row records an item circulation with details: circulation transaction date, item barcode, title, author, call number, borrower's name, course, card number, borrower category and transaction type.

The eight export files, which contained transaction keys formed from date, barcode, borrower card number, and transaction type, were compared. It was found that three of the earlier files were complete subsets of the files exported afterwards for the same borrower categories, with all of the earlier file's transactions present in the later export. These three files were dropped to avoid duplication, and the following five non-redundant exports were subsequently analyzed: undergraduate students, postgraduate students, research staff, permanent teaching staff and temporary teaching staff.

3.2 Data preparation

The transaction dates in both file formats were day-first and separated by slashes or hyphens. Instead of inferring the date (which could silently interpret a day-first date as month-first, and thus shift the transaction records from month to month), explicit format strings were provided for date parsing. For text fields, extra whitespaces were removed and consecutive whitespaces were replaced with one space. Barcodes had their leading zeros removed and converted to upper case. The titles were normalized to lower case, with whitespace collapsed, in order to treat formatting variants of the same title as one item.

Records with a date that could not be parsed, blank title or blank barcode were discarded. Exact duplicate transactions were identified and removed by checking the composite key above. The title, not the

barcode, is the basic unit of acquisition and duplicate-copy decision-making; if physical copies were treated as separate titles, circulation data for a popular title would be split between copies and demand data for that item would be artificially low.

3.3 Operational definitions

These categories have different definitions in different library systems, and the statistics depend on how this event type is defined. Table 1 lists the definitions of every term in this paper. The issues figure accounts for 91.4% of the export records data. This is because of the way the source systems record data. Returns account for just 7.4% of transactions. And this should be borne in mind for comparisons with studies that treat issues and returns as a matched pair rather than a transaction reconciled with a return.

Table 1. Operational definitions of circulation terms used in this study.

Term	Definition applied
Transaction	One recorded circulation event of any type
Issue	Initial loan of an item to a borrower
Renewal	Extension of an existing loan; counted as a separate transaction
Return	Recorded return of an item; present for 7.4 per cent of records only
Title	Distinct normalised title string; the unit of analysis
Copy	Distinct barcode; multiple copies may share one title
Borrower	Distinct card number, deduplicated across all files
Active month	A calendar month, identified by year and month, in which a title recorded at least one transaction

3.4 Temporal design

Calculating the score over an interval and comparing it to the score calculated over the same interval is a historical use of the score, which differs from prediction, usually done by subdividing the data in time. All of the features were calculated over transactions between 1 January and 31 December 2025. The outcome measure was circulation between 1 January and 30 April 2026, which contributed to no features. Thus, the records of May and June 2026 are not included in the sample, as the export included only 13 transactions and 37 transactions of a monthly average above one hundred.

The overall pipeline is shown in figure 1 and the temporal partitioning step in figure 2. The 10 steps are raw exports acquired, pre-processed, temporally partitioned,

features constructed, features scaled, clustered, tier ordered, composite scored, prospective validated, and decision-support output generated. In stage 3, the outcome window is disjoint from all prior stages. No later stages, other than the validation step, use the outcome window after stage 3. The shaded stages of Figure 1 illustrate preprocessing, unsupervised, evaluation, and reporting stages of this framework.

Figure 1. System architecture of the proposed framework, with ten processing stages from raw circulation exports to decision support products. In Stage 3, time-based partitioning ensures that the feature window is distinct from the outcome window, such that only validation accesses outcome-window records.



Figure 1. System architecture of the proposed framework

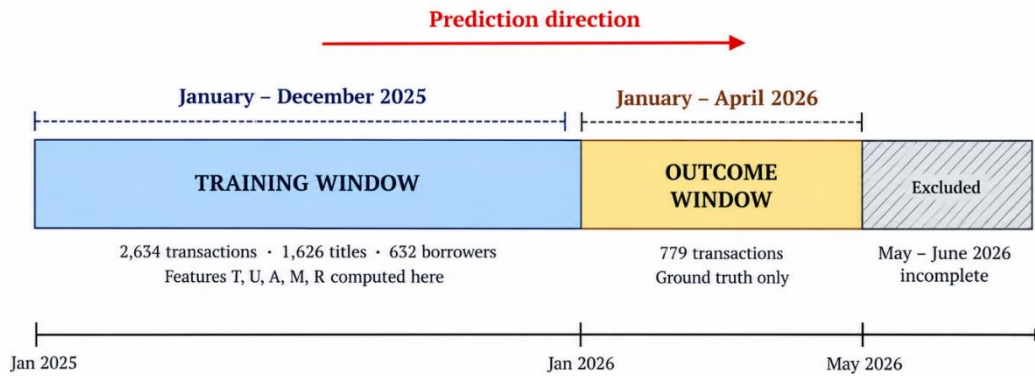


Figure 2. Temporal validation design.

3.5 Feature construction

Five indicators were calculated for each title, including its symbol, name, and definition, from its records in the training window. These are shown in Table 2. To calculate transactions per active month, total transactions are divided by the number of months during which the property was

Table 2. Title-level circulation indicators derived from the training window.

Symbol	Indicator	Definition
T	Total transactions	Count of all circulation events recorded for the title
U	Unique borrowers	Count of distinct borrower card numbers
A	Active months	Count of distinct year-months with at least one transaction
M	Transactions per active month	T divided by A; borrowing intensity while the title is in use
R	Renewals	Count of renewal events recorded for the title

A relatively small number of titles are borrowed very frequently while many are borrowed only once. Since the borrowing counts are highly right-skewed, the raw

Figure 2 shows the two windows in conjunction. The 1,626 titles that circulated in 2025 formed the evaluation universe. Titles that appear only in the outcome window represent newly acquired or previously dormant holdings which do not map to historical features and represent a form of cold start to be excluded from the model. Titles that are present in 2025 but not in the outcome window are assigned an outcome value of zero. Without these, all precision scores are inflated.

Features are derived exclusively from the January to December 2025 training window, while circulation observed between January and April 2026 serves as the outcome measure. May and June 2026 are excluded as incomplete.

active. If the entire time period was used in the denominator, the measure would simply be a constant proportional to total transactions and thus not independently informative.

counts were log transformed (Equation 1) before scaling.

$$(1) \quad \hat{x}_{ij} = \log(1 + x_{ij})$$

The transformed variables were then scaled again before clustering, so that the clustering indicators had a mean of zero and a variance of one (Equation 2). This was done because K-Means algorithm relies on distance, in this case Euclidean distance, which would mean the indicator with the largest range would prevail.

$$(2) \quad z_{ij} = (\hat{x}_{ij} - \mu_j) / \sigma_j$$

To support composite scores, indicators were rescaled to the unit interval using the minimum and maximum of the indicator after transformation (Equation 3). The bounded nature of transformed indicators ensures that the score lies within the final zero to one hundred range.

$$(3) \quad \tilde{x}_{ij} = (\hat{x}_{ij} - \min_j) / (\max_j - \min_j)$$

3.6 Clustering and validity assessment

When running K-Means, the following hyperparameters were used: k-means++ initialization, 50 runs, 300 maximum iterations, tolerance of 1×10^{-4} when determining convergence and a fixed random seed of 42. In addition, to avoid a pre-specified range of clusters beforehand, the analysis was computed for each number of clusters k from 2 to 15. At each value, four internal indices were computed: the within-cluster sum of squares for the elbow criterion, the silhouette coefficient, the Calinski-Harabasz index, and the Davies-Bouldin index.

A second test of the solution's stability as a cluster was done by fitting the selected solution 30 times with seeds chosen randomly, and then calculating the Adjusted Rand Index between the first set of partitions and the remaining fits. Two comparator algorithms were fitted to the same standardized matrix. Ward hierarchical clustering was performed with the number of

clusters set to equal the number of K-Means clusters. The Adjusted Rand Index was calculated between this clustering and the K-Means clustering. DBSCAN was run with the neighbourhood parameter set to twice the number of clustering features and the neighbourhood radius set to the ninety-fifth percentile of the k -distance graph as described previously, rather than manually tuning.

Due to the K-Means algorithm only returning the cluster identifiers, which are simply numbered, an artificial ordering was necessary in order to impose the ordinal demand labels to the clusters in the clustering results. The ordinal demand labels Low, Medium, High, Very High, Peak, and Extreme were assigned to the clusters in ascending order, according to the mean standardized centroid across the clustering indicators.

3.7 Composite demand score

To obtain the composite score, the five normalized indicators are aggregated using their specified weights to form a score that ranges from zero to one hundred (Equation 4).

$$(4) \quad S_i = 100 \times \sum_j w_j \tilde{x}_{ij}, \quad \sum_j w_j = 1$$

The normalization is necessary because a simple sum of raw indicators of different ranges would not give a correct weight, and the wide-ranged indicator would dominate the result regardless of its weight. Equal weighting was used as a initial default, because there is no universally accepted criterion for which circulation data should be used as superior. Three other schemes were tested to see if the ranking depended on which of the four schemes was used, which are all presented in Table 3.

Table 3. Weighting schemes evaluated for the composite demand score.

Scheme	w(T)	w(U)	w(A)	w(M)	w(R)	Rationale
Equal (primary)	0.20	0.20	0.20	0.20	0.20	Neutral; no indicator privileged
Transaction-led	0.40	0.20	0.20	0.10	0.10	Emphasises total borrowing volume
User-led	0.20	0.40	0.20	0.10	0.10	Emphasises breadth of readership
Breadth-led	0.25	0.25	0.30	0.15	0.05	Emphasises sustained use over time

Agreement between the rankings from each scheme and the primary specification was assessed using Kendall's rank correlation coefficient.

3.8 Baselines and evaluation measures

Five baselines were created to compare the composite score to simpler methods. The random ranking baseline is the simplest

baseline for this task. To see if the composite adds anything to a plain count, we rank on total transactions, and to see if recency alone is useful, we rank on transactions in the last three months. Ranking by unique borrowers and active months isolates each of those criteria from the others in turn.

Ranking quality was estimated by Spearman rank correlation between the predicted ranking and outcome-window circulation. Precision and recall were evaluated at 25, 50, 100, and 200 title cut-offs (Equation 5) using the sets of titles whose observed circulation situates them inside the top k of the true circulation rankings.

$$(5) P@k = |\{top-k \text{ predicted}\} \cap \{top-k \text{ observed}\}| / k$$

Normalized discounted cumulative gain at 100 titles was also reported, as the most highly-ranked titles receive the most attention and are those that decision-makers are most likely to acquire. Mean absolute percentage error was not reported. For much of the titles with a zero circulation, the result in the outcome window is undefined, and for all the rest, unstable with small denominators.

3.9 Implementation and reproducibility

Algorithm 1 summarizes the whole process. All analysis was conducted in Python 3.12.3 with the packages pandas 3.0.2, NumPy 2.4.4, scikit-learn version 1.8.0, SciPy 1.17.1, and Matplotlib 3.10.8. Clustering initialization, stability resampling, and the random baseline are all initialized using one random seed (42), thus these numbers can be exactly reproduced.

Algorithm 1. Demand tiering and prospective validation procedure.

Step	Operation
1	Load five non-redundant exports; parse dates with explicit day-first formats
2	Normalise text fields; remove blank and duplicate records
3	Partition records into training window and outcome window
4	For each title in the training window, compute T, U, A, M, R
5	Apply $\log(1 + x)$; standardise for clustering and min-max scale for scoring
6	For $k = 2$ to 15, fit K-Means and record four internal validity indices
7	Fit the selected solution 30 times with different seeds; compute pairwise ARI
8	Fit Ward hierarchical and DBSCAN on the same matrix; compare indices
9	Order clusters by mean standardised centroid; assign ordinal demand labels
10	Compute composite score under four weighting schemes; measure Kendall agreement
11	Aggregate outcome-window circulation per title as ground truth
12	Evaluate all rankings by Spearman correlation, precision, recall and nDCG

4. Findings

4.1 Dataset characteristics

Table 4 shows the composition of the dataset at each stage of preparation, and there are 5,011 rows of retained exports. Excluding transactions without title or barcode leaves a total of 4984 transactions.

Removing duplicate files leaves no records with identical transactions. The remaining transactions (3 June 2024 - 30 June 2026) cover 2615 titles with 3456 unique barcoded copies held, and the titles have been issued to 915 different borrowers.

Table 4. Dataset composition after each preparation stage.

Stage or attribute	Value
Export files retained	5
Export files excluded as redundant	3
Raw rows in retained files	5,011
Records with blank title or barcode removed	27
Exact duplicate transactions removed	0
Final transactions	4,984
Issue events	4,556 (91.4%)
Return events	371 (7.4%)
Renewal events	57 (1.1%)
Distinct titles	2,615
Distinct barcoded copies	3,456
Distinct borrowers	915
Date range	3 June 2024 to 30 June 2026

The imbalance between issues and returns is a property of the source system and not an issue with borrowing behavior.

Additionally, renewals are rare, consisting of only 57 events in over two years, so a renewal-based indicator offers limited information.

Table 5 shows how many transactions were made by each borrower category. Undergraduate and postgraduate students together account for 70.4 per cent of transactions. Teaching and research staff

make many fewer transactions, but borrow many more per person on average, with permanent teaching staff averaging 17.5 transactions per borrower, against undergraduates averaging 3.6. This matters for acquisition assessment because the small researcher population can disproportionately influence research-focused titles' usage statistics.

Table 5. Transactions and distinct borrowers by borrower category.

Borrower category	Transactions	Share (%)	Borrowers	Transactions per borrower
Regular Undergraduate	1,982	39.8	553	3.6
Regular Postgraduate	1,529	30.7	264	5.8
Teaching Permanent Staff	719	14.4	41	17.5
Research Staff	519	10.4	30	17.3
Teaching Temporary Staff	235	4.7	27	8.7
Total	4,984	100.0	915	5.4

According to the temporal partition, the training window consists of 2634 transactions on 1626 titles borrowed by 632 borrowers; the outcome window consists of 779 transactions on 562 titles borrowed by 252 borrowers. Out of the 1626 titles in the evaluation universe, 207 titles (12.7%) were also observed in the outcome window. An additional 355 titles were only observed in the outcome window and are considered cold-start cases.

4.2 Indicator distributions and collinearity

Table 6 shows the five indicators calculated for the 1626 titles in the evaluation universe. They are all right-skewed near the minimum value. Overall, the median total transactions, unique borrowers, and active months of the titles was one. Maximum total transactions was fifty. Renewals were rare: mean 0.008, not zero for just 13 of the titles.

Table 6. Descriptive statistics for the five circulation indicators ($n = 1,626$ titles).

Indicator	Mean	SD	Minimum	Median	Maximum
Total transactions (T)	1.620	2.111	1	1	50
Unique borrowers (U)	1.519	1.588	1	1	30
Active months (A)	1.327	0.815	1	1	9
Transactions per active month (M)	1.116	0.401	1	1	5.56
Renewals (R)	0.008	0.089	0	0	1

Particularly noticeable is that 1,231 of the titles (75.7 per cent of the evaluation universe) show exactly one transaction within the year 2025. Therefore, all further results have to be seen in this context.

As shown in Table 7, correlations between total number of transactions, number of unique borrowers, and number of active months for each title are high (between

0.830 and 0.970). This is not surprising since if a title is borrowed often, it will be borrowed by many borrowers in many months. Because renewals are uncorrelated with all of the other four indicators, these four count-based indicators describe closely related aspects of a single underlying quantity.

Table 7. Pearson correlation coefficients between circulation indicators.

	T	U	A	M	R
Total transactions (T)	1.000	0.970	0.830	0.707	-0.013
Unique borrowers (U)	0.970	1.000	0.871	0.689	-0.016
Active months (A)	0.830	0.871	1.000	0.425	-0.019
Transactions per active month (M)	0.707	0.689	0.425	1.000	-0.009
Renewals (R)	-0.013	-0.016	-0.019	-0.009	1.000

Because of these sustained but low usage records, the clustering algorithm would isolate these thirteen titles into their own

cluster. Therefore, clustering was performed on total transactions, unique borrowers, and

active months. All five of the measurements are used in the composite measure, as a bounded normalization prevents any one measure from being dominant.

4.3 Cluster validity and the choice of k

Table 8 lists the four internal validity indices for all values of k in $\{2, 3, \dots, 15\}$.

Table 8. Internal cluster validity indices for K-Means solutions from $k = 2$ to $k = 15$.

k	Within-cluster SS	Silhouette	Calinski-Harabasz	Davies-Bouldin
2	1516.26	0.8443	3600.6	0.4991
3	630.82	0.8725	5463.7	0.5128
4	405.19	0.8827	5968.3	0.6072
5	299.38	0.9070	6197.9	0.5323
6	222.72	0.9103	6772.1	0.6366
7	184.13	0.9155	6878.7	0.6545
8	156.63	0.9129	6967.4	0.7022
9	131.29	0.9192	7307.5	0.6776
10	113.51	0.9291	7536.8	0.6585
11	98.45	0.9335	7840.7	0.6278
12	85.26	0.9343	8248.3	0.6470
13	73.99	0.9376	8726.9	0.6340
14	67.45	0.9441	8843.1	0.5729
15	61.15	0.9441	9063.6	0.5539

Both criteria are presented in Figure 3. In particular, the Table 6 distribution has a monotonic silhouette curve. If three quarters of all cases fall into a single value, the silhouette coefficient's successive divisions become increasingly finer partitions of the one dimensional discrete integer variable, and thus increasingly more likely to yield a partition that separates the two values. Since no single final partition is superior on that dimension, it cannot be used to determine the number of clusters.

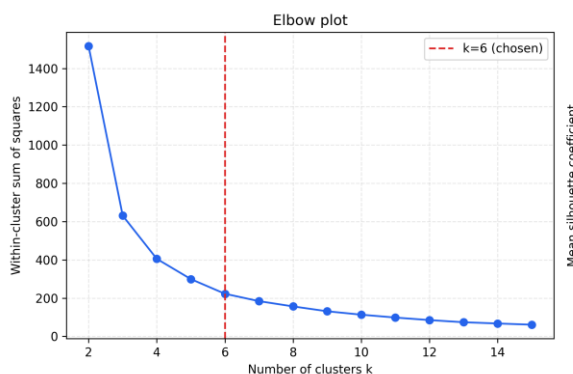


Figure 3. Within-cluster sum of squares and mean silhouette coefficient

Figure 3 shows Within-cluster sum of squares (left) and mean silhouette coefficient (right) for K-Means solutions from $k = 2$ to $k = 15$, computed on 1,626 titles. The dashed line marks the six-cluster solution adopted for reporting.

The within-cluster sum of squares drops from $k = 2$ to $k = 6$, and stabilizes for larger values of k . This is the curve that the elbow criterion is looking for. The increase in silhouette coefficient from 0.844 at $k = 2$ to 0.944 at $k = 14$ is much smoother.

The six-cluster solution was selected because, statistically, it was the most convenient (i.e. lying on the flattening point of the within-cluster sum of squares curve) and the silhouette coefficient 0.9103 was close to the maximum value for all solutions. Additionally, the six-cluster form was granular enough to allow the Knowledge Resource Centre to distinguish between acquisition priorities, duplicate copies and withdrawal candidates. Solutions with 12 or more clusters were seen to offer little additional benefit in terms of separation, and the cluster sizes were not large enough for distinct collection policies.

The output shows a high degree of stability: when 30 different random seeds are chosen as the center point, the Adjusted Rand Index between the partitions has an average of 1.000, minimum of 1.000, meaning that the six-cluster solution is reproduced exactly.

4.4 Comparison of clustering algorithms

Table 9 compares K-Means against Ward hierarchical clustering and DBSCAN on the same standardized matrix. Both K-Means and Ward Linkage yield almost identical partitions, with a silhouette coefficient of 0.9103 and 0.9060 respectively, and an Adjusted Rand Index of 0.9677. The choice

of either one of these two algorithms is therefore not important.

Table 9. Comparison of clustering algorithms on the standardised indicator matrix.

Algorithm	Clusters	Noise points	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-Means (k = 6)	6	0	0.9103	6772.1	0.6366
Ward hierarchical (k = 6)	6	0	0.9060	6068.2	0.6961
DBSCAN ($\epsilon = 0.462$, minPts = 6)	12	73	0.9953	126500.9	0.1974

DBSCAN yields the highest index. This would be a misleading result because it has labeled 73 titles as noise, which are excluded from silhouette calculation. This inflates DBSCAN's score. However, DBSCAN groups the collection into twelve clusters, whose boundaries are defined by the number of transactions per title. Thus, circulation demand is a continuous heavy-tailed distribution whose density does not have any natural structure for a density-based method to recover. Partitional methods are better suited for the tiering task.

4.5 Demand tier structure

Table 10. Cluster centroids in original units, ordering statistic, and assigned demand tier.

Tier	Titles	Share (%)	Mean T	Mean U	Mean A	Mean M	Centroid score
Low	1,243	76.45	1.010	1.000	1.000	1.010	-0.452
Medium	47	2.89	2.170	2.043	1.000	2.170	0.368
High	187	11.50	2.070	1.973	2.000	1.035	0.871
Very High	81	4.98	3.642	3.358	2.506	1.548	1.801
Peak	53	3.26	6.509	5.547	3.755	1.863	3.026
Extreme	15	0.92	16.667	13.067	6.067	2.701	5.001

The Medium and High tiers show the value of combining more than one indicator: while transaction totals in both tiers are the same (with an average of 2 transactions per month), the frequency of these transactions is different (a single month for Mediums vs multiple months for Highs). Although merchants would regard both tiers as identical, the ordering statistic separates them because sustained demand over months

Table 10 introduces the six clusters ordered by their mean standardized (z-)centroid value and a corresponding ordinal label. The cluster sizes differ strongly among the clusters. The lowest, or Low, sub-tier has 1243 titles or 76.45% of the total collection, with an average of 1.01 transactions per title from a single library cardholder in one month. The top, or Extreme, sub-tier has 15 titles with an average of 16.67 transactions from 13.07 library cardholders over 6.07 months.

indicates relevance, not just a one-time transaction.

Figure 4 shows the distribution of borrowings and provides a visual confirmation of the well-known, but rarely quantified fact that a small proportion of the books in a collection comprise the majority of borrowings. Of the 68 titles in the Peak and Extreme tiers, 4.18 per cent were circulating in the relevant year.

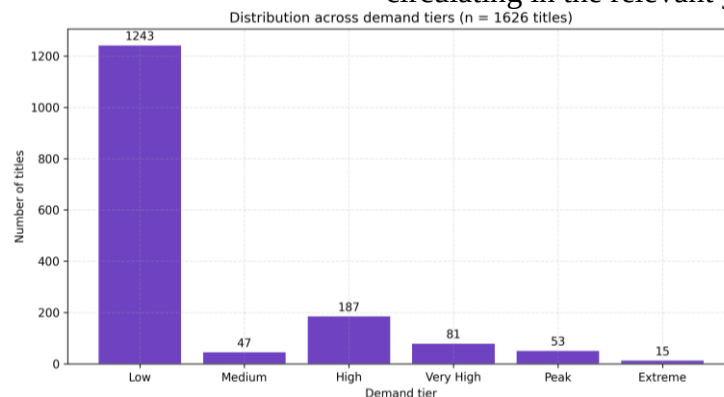


Figure 4. Number of titles assigned to each demand tier by the six-cluster K-Means solution ($n = 1,626$ titles). Tier labels follow the ascending centroid ordering reported in Table 10.

4.6 Sensitivity of the composite score to weighting

The effect of this weighting scheme on the resulting ranking is shown in Table 11 and the Kendall rank correlations with the equal-weight specification never fall below 0.959.

Table 11. Effect of weighting scheme on ranking agreement and predictive association.

Weighting scheme	Kendall τ against primary	Spearman ρ with outcome
Equal (primary)	1.0000	0.3314
Transaction-led	0.9605	0.3345
User-led	0.9590	0.3347
Breadth-led	0.9591	0.3349

Consequently, the ranking is strong to the specification of the weights, which means that a good theoretical justification need not be provided for any particular set of coefficients employed in the method. A library using the method may apply its own local weightings, but this will have a limited effect on the title rankings.

4.7 Prospective validation against baselines

Table 12 describes how well each ranking predicted circulation in the outcome

Table 12. Prospective ranking performance against observed outcome-window circulation.

Ranking method	ρ	P@25	P@50	P@100	P@200	nDCG@100
Random	-0.038	0.00	0.00	0.12	0.12	0.0353
Total transactions (T)	0.342	0.44	0.52	0.57	0.43	0.6716
Recent transactions (last 3 months)	0.171	0.48	0.38	0.39	0.28	0.5231
Unique borrowers (U)	0.337	0.36	0.56	0.58	0.43	0.6507
Active months (A)	0.332	0.32	0.48	0.56	0.40	0.6386
Composite score, equal weights	0.331	0.40	0.52	0.57	0.40	0.6534
Composite score, transaction-led	0.335	0.44	0.54	0.56	0.41	0.6496

In practice of the 100 highest data priorities that the system would recommend based only on the 2025 data, 57 of these titles were among the 100 titles that were

Spearman rank correlations with outcome window circulation range from 0.3314 to 0.3349 for each of the four schemes.

window. Random rankings had a Spearman correlation near 0 and precision at 100 titles of 0.12, which would be expected by chance. All circulation-based rankings exceed that floor, which results in a composite score of a Spearman correlation of 0.3314, precision at 100 results of 0.57, and normalized discounted cumulative gain at 100 results of 0.6534.

most-borrowed during the next four months. Figure 5 shows the Spearman correlations.

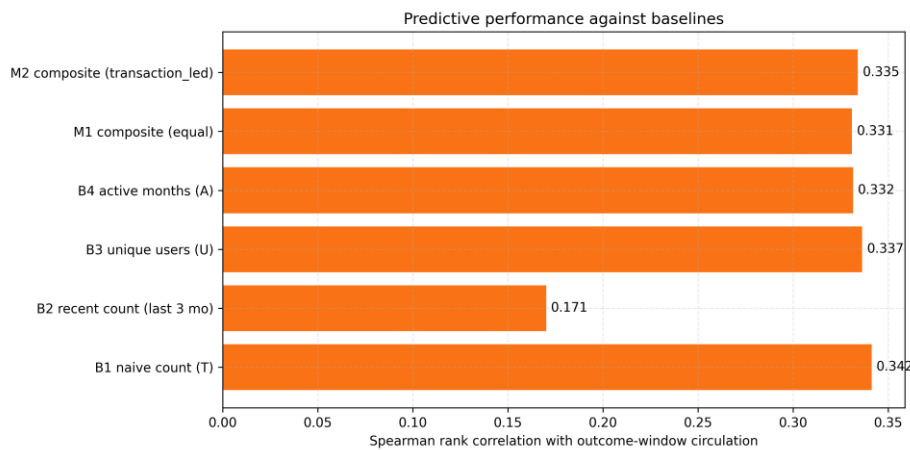


Figure 5. Spearman rank correlation between each ranking method and observed outcome-window circulation

The random baseline is omitted for legibility, having achieved a correlation of -0.038.

Despite the composite score performing poorly against the total number of

transactions ranking (Spearman correlation: 0.342; normalized discounted cumulative gain: 0.6716), the difference is negligible (0.011) considering the dataset sizes. This

is an important and honest qualification: the main benefit of using these five indicators might not actually be their superior ranking performance but rather the intuitively interpretable tier structure that they create. An individual recency measure performs noticeably worse with a Spearman correlation of 0.171. Borrowing over the preceding year does better than quarterly borrowing at forecasting future demand.

4.8 Prospective validation of the demand tiers

Table 13. Observed outcome-window circulation by demand tier assigned from training-window data.

Demand tier	Titles	Mean transactions	Total transactions
Low	1,243	0.099	123
Medium	47	0.277	13
High	187	0.262	49
Very High	81	0.630	51
Peak	53	1.736	92
Extreme	15	3.867	58

Five of the six levels are ordered monotonically. An inversion occurs between Medium (0.277 transactions per title) and High (0.262 transactions per title), but the inversion is so small and the level Medium

As shown in Table 13, the outcome-window circulation levels are therefore clearly separated into tiers, with the mean outcome-window circulation increasing from 0.099 transactions per title in the Low tier to 3.867 transactions per title in the Extreme tier, a difference of about thirty-nine fold. Fifteen Extreme-tier titles (0.92 per cent of titles) accounted for 58 of the 386 outcome-window transactions from previously in-circulation titles.

only contains 47 titles; therefore, the inversion is best understood as a product of sampling variation rather than a defect in the ordering. Figure 6 depicts the pattern.

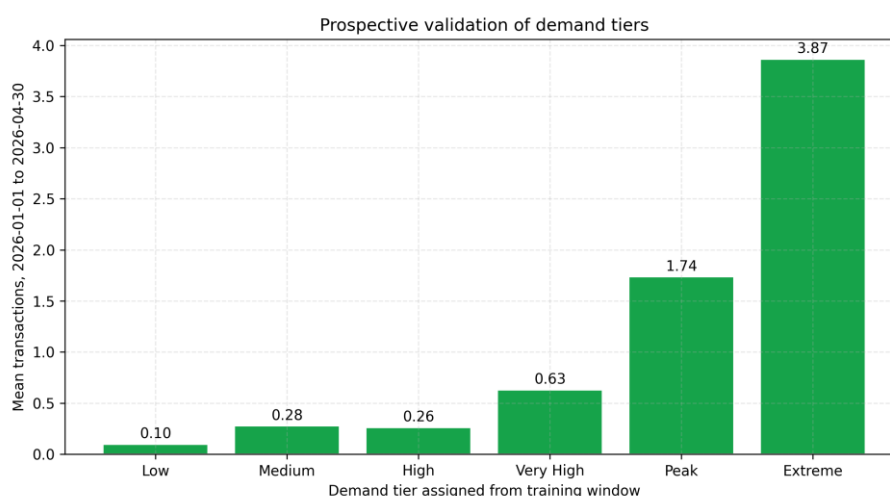


Figure 6. Mean circulation

Figure 6 shows mean circulation averaged over January through April 2026 for the titles assigned to each demand tier based on data from January through December 2025. Tiers were assigned without accessing data from the outcome window.

Table 14 lists the ten highest scoring titles and their observed outcome-window

circulation. Nine of the ten circulated again in total. The exception is 13 such transactions in 2025 after which there are no transactions for the next four months, suggesting even high-sales books fall out of use with a course change or syllabus revision.

Table 14. Ten highest-scoring titles with training-window indicators and observed outcome-window circulation.

Title	T	U	A	Score	Observed
Business communication	50	30	9	80.00	13
Management accounting	25	19	8	63.53	7
The inner courtyard	21	15	8	58.69	8
Introduction to Indian knowledge system	17	16	5	56.12	6
Bālyavasthā va bālavikās	15	11	4	51.87	4
Shāstra adhyapan paddhatī	15	13	8	51.85	6
Wise and otherwise	14	12	7	50.16	3
Cost and works accounting, Paper II	14	9	5	48.65	1
Upayojit mānasshāstra	13	10	5	48.01	0
Childhood and growing up for B.Ed., Paper I	12	12	5	47.81	1

4.9 Monthly circulation across academic years

The number of transactions per month for each of the years on record is shown in Table

Table 15. Monthly transaction totals by calendar year. Dashes indicate months outside the export range.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2024	—	—	—	—	—	159	171	397	204	241	143	206
2025	314	227	231	219	97	123	318	316	215	125	269	180
2026	214	240	200	125	13*	37*	—	—	—	—	—	—

An asterisk marks months for which the export is incomplete. Figure 7 plots the full series.

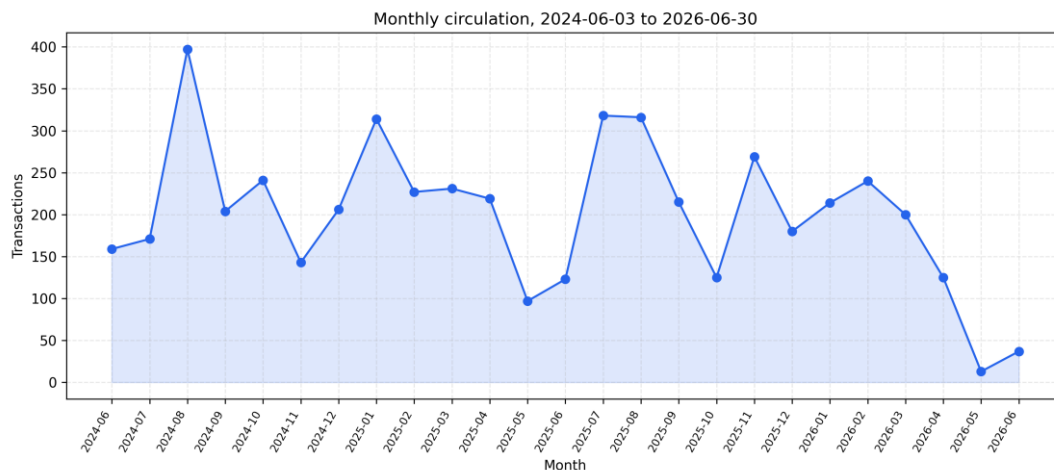


Figure 7. Monthly transaction totals

Figure 7 shows monthly transaction totals from June 2024 to June 2026. The decline in the final two months reflects incomplete data capture rather than reduced borrowing.

The monthly data can instead be tested for consistency with Spearman's correlation: comparing January through June 2025 with the six months of January through June 2026 gives a Spearman correlation of 0.829. This also means that while there is a reliable signal for the first half of the school year, the correlation between June 2024 and December 2025 is 0.036. Since only the last 6 months overlap, no reliable signal is found in the second half of the school year.

15, which shows 7 months of 2024, all of 2025, and 6 months of 2026 (the last two are partial months).

August is the only month with consistently high numbers, with 397 in 2024 and 316 in 2025, both of which would fall on the start of a new academic year; October varies from year to year. Interestingly, it has 241 transactions in 2024 (the second most at that point of 2024), but only 125 transactions in 2025, amongst the lowest in that year (there is no reasonable explanation, e.g. a matcake festival, to explain both of the phenomena). This month-to-month variation is only partly systematic, so monthly circulation counts within the same year should be treated with caution.

5. Discussion

The implications of these results are threefold for circulation-based demand modeling in the academic library. First, they suggest that the structure intrinsic in circulation data can be discovered, second, that the discovered structure is stable across multiple estimation attempts, and finally that tier membership in one period is correlated to measurable changes in circulation in another period. The last of these has not been demonstrated for a clustering-based library demand framework and involves taking a descriptive segmentation and transforming it into evidence for decision-making about future acquisitions.

What does a precision of 0.57 for the top hundred titles mean? It would correctly have predicted six out of every ten most-borrowed titles in the next period given only circulation data. This is far better than the random baseline, which is 0.12, and is good enough to decide on purchasing or discarding duplicate copies, for which the cost of a false positive means an extra copy of something that was already circulating. It is not sufficient to completely automate the acquisition process without professional review, and should be presented as a tool to support librarian judgement, rather than replace it.

Another important result is that the aggregated score built on the five different indicators does not dominate the ranking based on total number of transactions. This reflects the fact that the five count-based indicators are highly collinear, so that aggregating them does not meaningfully change the ordering of any individual indicator (see correlation matrix in Table 7). This finding slightly modifies a common assumption in the clustering literature: the virtue of composite measures over a typical single-item baseline was seldom compared with such baselines [12], [14]. The evidence here suggests that composite measures add interpretational structure, rather than accuracy. Yet, the very practical virtue of composite measures is that they assign each title to a named tier with a stated meaning,

unlike a bare count, which can guide differentiated collection policy.

This is also visible in the behavior of the validity indices. The reason for this is that whereas 75.7 per cent of the titles have a single transaction, the indicator space is very close to one-dimensional, and the silhouette coefficient increases almost monotonically between two and fifteen. An analyst who has looked only for the maximum between one and six will report six, whereas one who searched between one and fifteen will report fourteen. In this situation, reporting any internal indices would need to be done over a range of k values, as well as interpreted in the context of the underlying distribution due to the frequency of library circulation distributions being of this shape.

A comparison of the relative performance of the algorithms supported this. It was found that the output of the K-Means algorithm and the Ward linkage clustering had an Adjusted Rand Index of 0.9677, indicating that the tier structure was a property of the data and not simply an artifact of the algorithm. In spite of the poor suitability of DBSCAN as an algorithm (73 titles were classified as noise and most of the others clustered in series of discrete levels of transaction count), it produced the highest index value. If the only consideration had been the index value, this algorithm would have been chosen, but here the suitability and interpretability of the algorithm are key considerations.

However, looking at seasonal data showed that this was not the case: the first half of 2025 was similar to that of 2026, though the second half of 2025 did not correlate with the second half of 2024. October, for example, went from a high month to a low month. Otherwise, the single-year study finds a story on exam periods and festival holidays that the second year does not sustain, so to use the monthly circulation pattern as a guide to hiring or to acquisitions, you would need to wait for a multi-year cycle.

The Knowledge Resource Centre generated the analysis to determine that of the 68 Peak

and Extreme tier titles, those that had more than 1.7 subsequent transactions per title over the subsequent four months' circulation period would be candidates for duplicate-copy purchase and reserve. Of the 1,243 Low-tier titles, which make up over three quarters of the circulating collection, titles that would be expected to circulate only once for every ten titles over the same period would be candidates for closed storage. The Medium and High tiers are defined by the temporal spread of similar transaction counts, and identify titles whose demand is sustained rather than episodic, making them viable for retention in open access.

These findings point towards a broader potential in using library transactional data in an evidence-based management context [2], [9], where the problem is not a lack of data, but a lack of analytical ability to ground decision-making in evidence [3]. The pipeline we propose here addresses these concerns as it is reproducible given a single script and an appropriate random seed, is based exclusively on fields exported by any external library management system, and is reported at the level of granularity at which acquisition decisions are actually made.

5.1. Limitations

As this study examines one academic library, the percentages are based on this academic library's collection profile, curriculum, and user population. However, the study's method is applicable to other academic libraries, although the percentages will differ from those shown here.

The potential performance period was between January and April 2026, four months long. This would allow for tracking performance for part of an academic year, but not a full one. The next two months were not included because this data was cut off from the export, so the potential collection time was limited.

Electronic circulation, in-library use, or interlibrary loan are not counted, so these titles that are not circulated outside the library, but only used on location, are undercounted at all levels. However, since there are only 57 renewals and only 7.4 per cent of all events are return events, the

information we can learn from these events is limited to this dataset.

The framework predicts ranks among titles rather than raw transaction counts, and is unable to generate predictions for titles that have no prior circulation history. In the prediction window, 355 cold-start titles that have no prior circulation history in the prediction window are not covered by the framework. These titles would also need to be analyzed by an independent method based on bibliographic characteristics. Lastly, since the framework has not been incorporated into operational decision making, its practical usefulness for acquisition staff has yet to be tested.

6. Conclusion

We present and evaluate a method for translating regular academic library circulation transactions into demand tiers and a priority list for purchasing. We evaluate the method with 4984 circulation transactions between June 2024 and June 2026, showing that a 6-demand-tier model is reproducible across 30 different random restarts, and highly consistent between K-Means and Ward hierarchical clustering algorithms (ARI = 0.9677).

The main contribution was prospective validation. Demand tiers and scores based on 2025 circulation were validated with independent circulation data from January through April 2026. Mean subsequent circulation per title ranged from 0.099 in the Low tier to 3.867 in the Extreme tier. Of the top 100 books, 57 are in the top 100 of the composite ranking, whereas just 12 would be expected by chance. The composite ranking was strong to weighting schemes, with Kendall rank correlation coefficients of at least 0.959 for four schemes.

Two caveats to this framework, which are reported as they are relevant for practice, are that the composite score did not perform any better than ranking alone on total transactions, and it generated interpretable tier structure rather than higher hit rates. Since the silhouette coefficient increases

monotonically for all k values, the number of clusters was decided based on operational needs and the elbow criterion, since internal validity indices do not optimally determine the number based on the shape of the circulation data.

Inevitably, there are some limitations: one library, a four-month validation window, printed circulation only, and books that have never circulated before cannot be scored.

This leads to four extensions to the model. Checking the consistency of mixed performance across adjacent outcomes windows could establish whether the performance stability appears to be consistent across the academic year.

Extending validation to more institutions could establish whether the tier structure generalizes. Including electronic resource usage data would extend the paper's scope.

Modelling the bibliographic data, including the subject class and publication year, of cold-start titles would make the model cover the library's new acquisitions, where the library needs and would find acquisition guidance the most valuable.

Data Availability Statement

The raw circulation data cannot be released due to the privacy policies of the institutions which govern the release of individual borrower names and card numbers. The following material is available from the corresponding author on reasonable request: The code used to conduct this analysis including all preprocessing, clustering, scoring, and evaluation, an anonymized title level dataset of the five circulation indicators, composite score, assigned tier, and observation window circulation for each title (with all borrower identifiers removed), the cluster validity, algorithm comparison, weight sensitivity, and predictive validation tables, and the software and analyses run environment record comprising the versions used.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Asminah, A. (2022). Sistem penentuan penambahan koleksi buku di perpustakaan menggunakan metode k-means clustering. *Journal of Information System Research (JOSH)*, 4(1), 330–338. <https://doi.org/10.47065/josh.v4i1.2383>
- Badanina, N. D., & Sudakov, V. A. (2022). Driver clustering according to the ratio of dangerous behavior using machine learning algorithms. *Modelling and Data Analysis*, 12(1), 5–15. <https://doi.org/10.17759/mda.2022120101>
- Galyani-Moghaddam, G., & Taheri, P. (2021). Public library circulation records: What do they reveal about users' reading preferences? *Journal of Librarianship and Information Science*, 53(2), 328–337. <https://doi.org/10.1177/0961000620949643>
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Elsevier/Morgan Kaufmann.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666. <https://doi.org/10.1016/j.patrec.2009.09.011>
- Kalbande, D., Yuvaraj, M., Verma, M. K., A., S., Suradkar, P., & Chavan, S. (2024). Exploring the integration of artificial intelligence in academic libraries: A study on librarians' perspectives in India. *Open Information Science*, 8(1), 20240006. <https://doi.org/10.1515/opis-2024-0006>
- Kumar, P., & Jyoti. (2024). Reshaping the library landscape: Exploring the integration of artificial intelligence in libraries. *IP Indian Journal of Library Science and Information Technology*, 9(1), 29–36. <https://doi.org/10.18231/j.ijlsit.2024.005>
- Monika, W., Nasution, A. H., Syam, F. A., & Wijesundara, C. (2025). Clustering of library's patron behavior using machine learning. *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, 16(1), 1–11. <https://doi.org/10.31849/digitalzone.v16i1.19680>

- Narendra, A. P., Dewi, C., Gunawan, L. S., & Ardi, A. S. (2025). Artificial intelligence implementation in library information systems: Current trends and future studies. *Vietnam Journal of Computer Science*, 12(03), 209–233. <https://doi.org/10.1142/S2196888824300023>
- Roy, P. (2025). Big data analytics in university libraries on today's librarianship decision-making: A disruptive innovation perspective. *IFLA Journal*, 51(4), 1013–1033. <https://doi.org/10.1177/03400352251318753>
- Safira, W. E., & Mohammed, S. (2026). Supporting academic library collection decisions using k-means-based book recommendation. *Artificial Intelligence in Educational Decision Sciences*, 1(2), 22–33. <https://doi.org/10.66053/aieds.v1i2.25>
- Sinulingga, N., Sarif, M. I., Ramadhani, N. A., Nurfebria, K., & Sulistia, F. Y. (2025). Penerapan algoritma k-means clustering untuk mengetahui pola peminjaman buku di perpustakaan Universitas Imelda Medan. *Jurnal Komputer Teknologi Informasi Sistem Informasi (JUKTISI)*, 4(3), 1552–1560. <https://doi.org/10.62712/juktisi.v4i3.707>
- Tamba, S. P., Batubara, M. D., Purba, W., Sihombing, M., Mulia Siregar, V. M., & Banjarnahor, J. (2019). Book data grouping in libraries using the k-means clustering method. *Journal of Physics: Conference Series*, 1230(1), 012074. <https://doi.org/10.1088/1742-6596/1230/1/012074>
- Tanjung, M. I., & Sriani, S. (2025). Clustering library loan books using k-means clustering. *Journal of Information Systems and Informatics*, 7(1), 889–908. <https://doi.org/10.51519/journalisi.v7i1.1037>
- Walker, K. W., & Jiang, Z. (2019). Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach. *The Journal of Academic Librarianship*, 45(3), 203–212. <https://doi.org/10.1016/j.acalib.2019.02.013>
- Zhang, X., Zang, L., Bai, W., & Liu, H. (2024). Big data analysis and application of library information resources. *Applied Mathematics and Nonlinear Sciences*, 9(1), 20241212. <https://doi.org/10.2478/amns-2024-1212>