International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Temporal Fairness Degradation in Recommendation Algorithms: Measurement and Prediction

Suraj Balaso Desai
Data Scientist
Digg, Inc.
Los Angeles, CA, USA
desaisuraj2707@gmail.com
0009-0006-1487-4704

Kinjal Vijaykumar Bhatia
Data Scientist
Persona Identities, Inc.
Dublin, CA, USA
bhatiakinjal98@gmail.com
0009-0008-0504-7234

Leslie Daniel Raj
Senior Cloud Application Architect
Amazon Web Services, Inc.
Odessa, FL, USA
leslie.d.raj@gmail.com
0009-0002-8364-6378

Pooja Nandkishore Relan
Senior Data Analyst
Independent Researcher
Plano, TX, USA
poojarelan11@gmail.com
0009-0005-4603-4139

Abstract*—Recommendation systems deployed in production environments exhibit temporal fairness degradation, where demographic disparities systematically accumulate over time through popularity feedback loops and algorithmic amplification. Existing fairness research predominantly focuses on static assessment at single time points, lacking predictive mechanisms to anticipate and prevent future violations before critical thresholds are reached. This study presents a comprehensive 27-year longitudinal analysis of fairness decay in three state-of-the-art recommendation algorithms using the MovieLens dataset spanning 1997 to 2023. A quantitative evaluation framework measures demographic parity violations across three content popularity groups niche (50%), mid-tier (40%), and popular (10%) quantifies algorithm-specific decay rates, and constructs predictive models using temporal feature extraction. Statistical analysis demonstrates significant fairness degradation across all algorithms ($p < 0.001$), with annual decay rates ranging from 0.0097 to 0.0166 gap points per year, representing 40% to 106% deterioration over the study period. NCF maintains the most stable fairness profile with a composite stability score of 0.803 and decay rate of 0.0124 per year, while SASRec exhibits severe instability (score: 0.017, decay: 0.0166 per year), reaching near-maximum unfairness by 2023. Random Forest classification achieves 75% accuracy in predicting fairness violations one period ahead and maintains 67% accuracy for two-period forecasting, providing a practical intervention window of up to two years. Analysis of feature importance identifies five critical early warning indicators: Gini coefficient (0.089), aggregate diversity (0.089), niche share proportion (0.078), current fairness gap (0.067), and catalog coverage (0.067). These findings establish the first predictive framework for proactive fairness monitoring in recommendation systems, demonstrating that temporal bias accumulation is systematic, measurable, and forecastable, enabling intervention before critical threshold violations occur.

***Index Terms*—Recommendation systems, fairness, temporal analysis, popularity bias, algorithmic bias, predictive modeling, neural collaborative filtering, graph neural networks, sequential recommendation, machine learning**

I. Introduction

Recommendation systems have become essential infrastructure for digital platforms, facilitating content discovery and consumption choices for billions of users each day. Algorithmic recommendations from streaming services like Netflix and Spotify, e-commerce platforms such as Amazon, and social networks like Facebook determine the exposure and engagement of content. In addition to commercial uses, recommendation algorithms are rapidly impacting vital areas such as academic article discovery, job matching, and news distribution. The widespread implementation of these systems requires thorough scrutiny of their fairness attributes, especially as algorithmic choices accumulate over time, influencing information accessibility, economic prospects, and cultural representation.

While the machine learning community has made substantial progress in developing fairness-aware recommendation algorithms, most existing approaches evaluate fairness at single time points, treating recommendation as a static optimization problem. This perspective overlooks a fundamental characteristic of production recommendation systems: they operate in dynamic environments where recommendations influence user behavior, which in turn shapes future training data through feedback loops [7]. Popular items recommended frequently receive additional engagement, generating more training signals that further reinforce their visibility in subsequent iterations. Conversely, less popular items receive diminishing exposure, creating a "rich-get-richer" dynamic that systematically amplifies initial popularity disparities. This temporal accumulation of bias represents a critical blind spot in fairness research, as systems deemed fair at deployment may deteriorate over time without detection or intervention.

Existing fairness research primarily employs immutable evaluation frameworks that assess demographic parity, equal opportunity, or similar metrics utilizing fixed datasets collected at particular points in time [2]. Although these studies establish essential benchmarks for fairness assessment, they do not encompass how algorithmic behavior develops as systems engage with users and gather historical data. Recent research has initiated the investigation of feedback loops and their contribution to bias amplification [7], yet no comprehensive longitudinal analysis has quantified fairness decay rates across contemporary recommendation algorithms or formulated predictive models to anticipate future violations. This disparity hinders practitioners from comprehending which algorithms preserve consistent fairness profiles during prolonged deployments and renders systems susceptible to gradual fairness decline that becomes evident only after months or years of operation. The lack of early warning systems for fairness

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

violations requires reactive measures after damage has already occurred, rather than facilitating proactive oversight.

This study examines these constraints via a thorough 27-year longitudinal examination of temporal fairness decay in three modern recommendation algorithms: Neural Collaborative Filtering (NCF), LightGCN, and SASRec. This study employs the MovieLens 32M dataset, covering the period from 1997 to 2023, to systematically assess the temporal evolution of demographic parity breaches and to develop predictive models for anticipating future declines in fairness. The investigation is guided by four research questions:

RQ1: Does fairness degradation occur systematically over time in recommendation systems?

RQ2: What is the rate of fairness decay across different recommendation algorithms?

RQ3: Which recommendation algorithms maintain the most stable fairness profiles over extended periods?

RQ4: Can future fairness violations be predicted from current system metrics, and what early warning indicators are most reliable?

The contributions of this research are as follows:

- A thorough temporal examination revealing statistically significant declines in fairness across three leading recommendation algorithms over a span of 27 years, with decay rates varying from 0.0097 to 0.0166 gap points annually ($p < 0.001$ for all algorithms).
- The analysis reveals distinct stability profiles associated with each algorithm, demonstrating that NCF achieves the highest level of consistent fairness performance, reflected in a composite stability score of 0.803. In contrast, SASRec shows significant instability, with a score of 0.017, indicating a trend towards near-maximum unfairness by the year 2023.
- A predictive modeling framework demonstrates an accuracy of 75% in forecasting fairness violations one period ahead, while sustaining an accuracy of 67% for two-period predictions, thereby offering a practical intervention window of up to two years.
- The identification of five essential early warning indicators through feature importance analysis includes the Gini coefficient, aggregate diversity, niche share proportion, current fairness gap, and catalog coverage, which allows practitioners to proactively monitor system health.

The remainder of this paper is structured as follows: Section II examines existing literature on fairness in recommendation systems and the dynamics of temporal bias. Section III outlines the methodology, detailing dataset characteristics, algorithm implementations, fairness metrics, and temporal evaluation protocols. Section IV provides empirical findings that respond to each research question. Section V addresses the implications for system design and deployment. Section VI concludes by outlining directions for future research.

II. Related Work

This research focuses in three main research areas: fairness metrics and definitions in recommendation systems, temporal dynamics and feedback loops, and algorithmic stability. This study's contributions are contextualized by emerging research on predictive fairness monitoring.

A. Fairness in Recommendation Systems

Fairness in machine learning has garnered significant attention, particularly as recommendation systems pose distinct challenges owing to their multi-stakeholder dynamics, which include both content consumers and providers. Demographic parity, an essential concept of fairness, mandates that recommendations be allocated in proportion to groups characterized by sensitive attributes, including content popularity, creator demographics, or genre categories [2]. Ekstrand et al. conducted significant research showing that popularity bias systematically favors mainstream content and marginalizes niche items, creating measurement frameworks that quantify representation disparities among demographic groups.

In addition to demographic parity, scholars have suggested exposure fairness metrics that consider the visibility granted to various item groups based on their ranking positions [3]. Mehrotra et al. proposed exposure-based fairness measures, acknowledging that items ranked higher attract disproportionate attention. This necessitates a fairness evaluation that considers not only inclusion in recommendation lists but also the prominence of items within those lists. Research on provider-side fairness differentiates between the fairness experienced by content consumers and content creators. Abdollahpouri et al. illustrate that prioritizing consumer satisfaction may unintentionally compromise provider fairness by favoring recommendations for already-popular creators [4].

Although these studies establish rigorous fairness metrics and highlight their significance in static recommendation contexts, they evaluate system fairness at discrete time intervals without considering the evolution of fairness properties as systems gather data and user feedback over prolonged periods of deployment. This temporal dimension is inadequately examined, despite its practical importance for production systems.

B. Temporal Dynamics in Recommendation

Recent studies have investigated the temporal evolution of recommendation systems, emphasizing feedback loops that exacerbate initial biases. Mansoury et al. presented empirical evidence indicating that popularity bias is exacerbated through iterative recommendation-consumption cycles. In these cycles, popular items gain further exposure, resulting in increased interactions, which subsequently enhances their visibility in future recommendation rounds [7]. The simulation studies indicated that this amplification is present across various collaborative filtering methods, implying that feedback loops constitute a fundamental challenge rather than an artifact specific to any algorithm.

Jiang et al. analyzed degenerate feedback loops in recommender systems, demonstrating that these systems can converge to homogeneous recommendation distributions that meet short-term engagement metrics, while simultaneously fostering filter bubbles that diminish long-term diversity [8].

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Their theoretical analysis identified conditions that result in the complete collapse of recommendation diversity due to feedback loops, thereby offering mathematical foundations for comprehending temporal degradation phenomena. D'Amour et al. expanded this research by conducting simulation-based studies on fairness, revealing that fairness properties are dynamic and evolve with user interactions. They contend that assessments of fairness at a single time point offer an incomplete understanding of system behavior [9].

The studies identify temporal phenomena and characterize their mechanisms; however, they lack predictive frameworks that would allow practitioners to anticipate when fairness violations may occur or which system configurations are most susceptible to temporal degradation. The lack of early warning systems requires interventions to be reactive, addressing issues only after they arise instead of allowing for proactive management informed by predictive indicators.

C. Recommendation Algorithms and Architecture

This study analyzes three distinct architectural paradigms in contemporary recommendation research through the examined algorithms. Neural Collaborative Filtering (NCF) integrates deep neural networks into collaborative filtering by substituting the inner product of matrix factorization with learned neural interaction functions. This approach illustrates that neural architectures can more effectively capture complex user-item relationships compared to traditional linear methods [11]. LightGCN simplifies graph convolutional networks for recommendation by focusing solely on neighborhood aggregation, eliminating feature transformations and nonlinear activations, and thereby achieving enhanced performance through architectural efficiency [12]. The streamlined design disseminates user and item embeddings across bipartite interaction graphs via linear operations, enhancing both interpretability and efficiency compared to earlier models.

SASRec utilizes self-attention mechanisms for sequential recommendation, allowing the model to discern pertinent items from user interaction histories and adjust their influence adaptively in predicting subsequent actions [13]. Self-attention enables parallel processing and captures long-range dependencies, distinguishing it from recurrent neural networks that process sequences sequentially. This efficiency contributes to SASRec's competitive accuracy. The three algorithms encompass collaborative filtering, graph-based, and sequential paradigms, offering varied architectural insights into the dynamics of temporal fairness.

Previous studies comparing algorithms have predominantly emphasized predictive accuracy metrics, including precision, recall, and normalized discounted cumulative gain, while fairness has been regarded as a secondary consideration, if acknowledged at all. A systematic comparison of temporal fairness stability across these algorithmic paradigms is lacking, despite increasing evidence that the choice of algorithm affects not only accuracy but also bias propagation and long-term system behavior.

D. Predictive Fairness Monitoring

Fairness auditing methodologies have developed as organizations acknowledge the necessity for continuous fairness assessment instead of singular evaluations. Contemporary auditing methodologies generally encompass regular manual evaluations of system outputs, statistical assessments for demographic parity breaches, or retrospective examinations of user grievances and identified harms. These reactive approaches identify issues only after fairness degradation has taken place, resulting in users experiencing disparate treatment.

Predictive methods for fairness monitoring are still in their early stages of development. Certain systems utilize rule-based triggers to identify potential issues when metrics exceed established thresholds; however, these systems do not possess predictive capabilities and are unable to anticipate violations prior to their occurrence. Machine learning techniques for predicting model performance degradation have been investigated in areas like concept drift detection; however, their specific application to fairness metrics has not been extensively studied. Previous research has not established predictive models aimed at forecasting fairness violations in recommendation systems utilizing temporal feature patterns derived from system behavior.

This study fills the existing gap by developing predictive models that anticipate demographic parity violations across multiple future periods, while also pinpointing early warning indicators that facilitate proactive interventions. This research demonstrates that fairness decay can be predicted using observable system metrics, establishing the feasibility of proactive fairness management and offering actionable guidance for practitioners implementing recommendation systems in production settings.

III. Methodology

This section outlines the experimental methodology used to assess and forecast temporal fairness degradation in recommendation systems. The approach includes dataset selection, definition of temporal scope, data preprocessing and filtering protocols, definitions of fairness groups, algorithm implementations, fairness metrics, and a temporal evaluation framework.

A. Dataset and Temporal Scope

The MovieLens 32M dataset [1] serves as the empirical foundation for this study. This dataset comprises 32 million movie ratings from 200,948 users, evaluating 84,432 movies from January 1995 to November 2018, with later updates extending to 2023. MovieLens was chosen for three main reasons: (1) The extended temporal range allows for longitudinal analysis over nearly three decades, (2) The rich interaction history offers adequate rating density to ensure reliable fairness measurements across demographic groups, and (3) The widespread adoption in recommendation research promotes comparison with previous studies and reproducibility of results.

The analysis period covers 1997 to 2023, comprising 27 consecutive years. This period omits 1995 and 1996 because of

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

inadequate rating density during the initial collection phases. Table I illustrates significant growth in user participation and catalog size over time, with annual users increasing from 1,974 in 1997 to 5,405 in 2023, and catalog size expanding from 1,386 items to 37,117 items during the same period. The volume of ratings fluctuates significantly over the years, with a peak in 2016 at 2,267,418 training ratings, contrasted by lower levels in earlier years, such as 1999, which recorded 79,357 training ratings.

Sparsity has consistently remained high during the study period, varying from 96.3% in 1997 to 99.8% in 2023, indicating that the majority of users engage with only a limited portion of the available catalog. The growing sparsity accompanying catalog expansion poses challenges and opportunities for fairness analysis, necessitating that recommendation algorithms selectively surface items from increasingly extensive catalogs.

B. Data Preprocessing and Filtering

Data preprocessing adheres to a temporal protocol aimed at preventing information leakage while ensuring that fairness measurements accurately represent realistic recommendation scenarios. This approach allows systems to generalize to future time periods based on historical data.

Temporal Binning: Ratings are allocated to yearly intervals through the conversion of Unix timestamps. Each rating r is assigned a year label based on:

$$\text{year}(r) = \lfloor \text{timestamp}(r)/31,557,600 \rfloor + 1970 \qquad (1)$$

where $31,557,600$ represents the number of seconds in a calendar year (365.25 days accounting for leap years), and the offset adjusts Unix epoch time to calendar years.

Temporal Train-Test Splitting: The dataset is split into strictly non-overlapping temporal subgroups for every evaluation year t:

- **Training data:** All ratings from years prior to t (years $< t$)
- **Test data:** All ratings from year t (year $= t$)

This protocol guarantees that algorithms are trained solely on historical data and assessed exclusively on future interactions, reflecting production deployment scenarios in which models forecast future user behavior based on accumulated history. Random splitting is not utilized, as the temporal ordering is critical to the research questions investigating the evolution of fairness over time.

User Filtering Criteria: In order to maintain dataset representativeness and guarantee accurate fairness measurements, users need to meet two requirements:

- Minimum training interactions: ≥ 5 ratings in training data
- Minimum test interactions: ≥ 1 rating in test data
- Presence in both training and test sets

The training interaction threshold of five ratings offers adequate signal for collaborative filtering algorithms to discern user preferences, while also accommodating less active users. The test threshold of one rating is minimal, ensuring that all users with any activity during the test period are included in the fairness evaluation. It is essential for users to be present in both training and test sets to avoid cold-start situations in which algorithms do not have any prior information regarding the user. The filtering is applied independently for each evaluation year, enabling users to enter or exit the analysis according to their participation patterns.

Following the application of these filters, the final user counts vary from 1,974 users in 1997 to 5,405 users in 2023, as presented in Table I. The filtered populations consist of users with adequate interaction history to facilitate significant recommendations, while ensuring temporal separation between training and testing phases.

Item Filtering: No specific minimum rating count threshold is imposed on items. Items enter and exit the catalog based on their presence in annual rating data, facilitating an organic evolution of the catalog composition over time. This approach illustrates production scenarios in which new content consistently enters platforms, while older content may experience diminishing attention.

Data Quality Verification: Several validation checks ensure data integrity:

- Duplicate ratings (multiple entries for the same user-item combination) have been confirmed to be absent from the source dataset.
- Timestamp validity is verified, with all timestamps occurring within the anticipated range of 1997 to 2023.
- Rating scores are confirmed to be within the acceptable range of 0.5 to 5.0.
- Missing values have been verified as absent from all mandatory fields (userId, movieId, rating, timestamp).

C. Fairness Group Definition

Assessing fairness necessitates the division of items into demographic groups that reflect varying degrees of content visibility and market privilege. This study employs a popularity-based grouping approach, wherein item popularity is utilized as a proxy for privilege within the recommendation ecosystem, in accordance with the provider fairness perspective [4]. Popular items gain increased visibility through several mechanisms: users tend to engage more with popular content, producing significant training signals; collaborative filtering algorithms inherently prioritize items with extensive interaction histories; and initial advantages in popularity are reinforced through feedback loops, as recommendations lead to further consumption.

Items are categorized into three groups based on popularity, utilizing percentile thresholds calculated separately for each evaluation year. This annual recalibration facilitates the movement of items between groups in response to changing popularity, illustrating the fluid dynamics of content ecosystems where previously niche content can become mainstream and vice versa.

Mathematical Definition: Each year t, items are ranked according to their rating count c_i^t, which indicates the number of ratings received by item i in that year. Let P_x^t represent the

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

TABLE I
COMPLETE DATASET STATISTICS ACROSS 27-YEAR ANALYSIS PERIOD (1997-2023)

Year	Users	Items	Train Ratings	Test Ratings	Sparsity (%)
1997	1,974	1,673	196,966	74,976	96.3
1998	1,068	2,302	157,370	25,392	96.4
1999	409	3,038	79,357	24,723	95.8
2000	2,786	3,876	511,451	218,066	96.6
2001	3,332	4,728	872,759	231,788	97.6
2002	3,285	5,737	1,020,415	209,212	98.1
2003	3,301	6,818	1,150,138	288,048	98.2
2004	3,566	7,998	1,375,010	337,975	98.3
2005	3,538	8,463	1,579,082	248,707	98.3
2006	4,061	8,809	1,866,446	243,527	98.7
2007	4,043	9,325	1,897,219	224,285	98.8
2008	4,039	9,962	1,916,921	231,763	98.9
2009	4,242	11,693	1,941,217	228,244	99.2
2010	4,259	13,232	2,014,868	223,870	99.3
2011	4,166	14,204	1,936,937	203,808	99.4
2012	3,972	14,329	1,921,820	186,414	99.4
2013	3,674	14,702	1,845,175	165,880	99.5
2014	3,326	15,435	1,743,486	154,373	99.5
2015	3,477	25,517	1,771,583	248,696	99.6
2016	5,175	31,373	2,267,418	339,561	99.6
2017	5,283	34,359	2,473,512	327,094	99.6
2018	6,177	36,388	2,955,915	342,099	99.7
2019	5,462	42,483	2,840,524	293,768	99.7
2020	5,955	42,249	3,040,464	345,017	99.7
2021	6,405	41,024	3,282,293	330,435	99.8
2022	5,921	39,616	3,249,129	294,649	99.8
2023	5,405	37,117	3,169,509	227,534	99.8

x-th percentile of rating frequencies across all items in year t. Items are categorized into groups as outlined below:

$$\text{Group}_i^t = \begin{cases} \text{Niche} & \text{if } c_i^t < P_{50}^t \\ \text{Mid} & \text{if } P_{50}^t \leq c_i^t < P_{80}^t \\ \text{Popular} & \text{if } c_i^t \geq P_{80}^t \end{cases} \quad (2)$$

Items with a rating count below the median are categorized as Niche, indicating content that possesses restricted mainstream appeal or visibility. Items within the 50th to 80th percentiles are classified as the Mid tier, indicating a level of moderate popularity and recognition, yet lacking the distinction of blockbuster status. Items within the highest 20% of rating counts are classified as Popular, indicating their status as widely recognized mainstream content.

Rationale for Group Structure: The 50/30/20 distribution (by item count) illustrates typical content catalog compositions, wherein the majority of items garner minimal attention, while a small subset attracts a disproportionate level of engagement. The distribution conforms to Pareto principles frequently noted in content consumption, wherein 20% of content generates 80% of engagement. Defining Niche as the bottom 50% of items allows for an evaluation of fairness that focuses on whether recommendation algorithms adequately promote the less popular content, which represents the majority of available inventory.

Empirical Group Characteristics: Table II illustrates average group distributions over the 27-year study period, indicating that although group boundaries are determined by item count percentiles, the resulting recommendation distributions exhibit significant variation. Popular items disproportionately dominate recommendations, accounting for a significant share despite constituting only 20% of catalog items. In contrast, niche items are underrepresented relative to their 50% presence in the catalog, as would be anticipated under uniform sampling.

TABLE II
AVERAGE RECOMMENDATION DISTRIBUTION ACROSS GROUPS (1997-2023)

Metric	Niche	Mid	Popular
Catalog Share (%)	50.0	30.0	20.0
NCF Recs (%)	14.6	40.8	44.6
LightGCN Recs (%)	15.2	39.7	45.1
SASRec Recs (%)	3.1	11.4	85.5

The group assignment procedure facilitates equitable comparisons among algorithms and time frames by creating uniform reference categories, while accommodating content mobility in response to changing popularity trends. Items that gain traction over time transition from niche to mid or popular categories, while declining content shifts in the opposite direction, reflecting the dynamic nature of content ecosystems.

D. Recommendation Algorithms

Three recommendation algorithms, each representing a different architectural paradigm, were chosen for the evaluation of temporal fairness. The algorithms include neural collaborative filtering, graph-based methods, and sequential modeling approaches, allowing for the evaluation of how architectural

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

choices affect fairness stability during prolonged deployments. All algorithms were implemented utilizing the PyTorch and RecBole frameworks to maintain consistent experimental conditions.

1) NCF (Neural Collaborative Filtering): NCF utilizes deep neural networks to represent user-item interactions via learned embeddings and multi-layer perceptron architectures [11]. The implementation integrates two parallel pathways: a Generalized Matrix Factorization (GMF) branch that captures linear interactions via the element-wise product of user and item embeddings, and a Multi-Layer Perceptron (MLP) branch that learns non-linear interaction patterns through successive hidden layers. Both branches handle 64-dimensional embeddings, with the MLP branch utilizing four hidden layers of dimensions [64, 32, 16, 8], incorporating ReLU activations and 20% dropout for regularization. The final fusion layer combines the outputs of GMF and MLP to generate rating predictions.

Training utilizes mean squared error loss in conjunction with the Adam optimizer, set at a learning rate of 0.001, and processes mini-batches consisting of 512 examples. Implementing early stopping with a patience of three epochs mitigates overfitting and decreases computational expenses. During the 27-year evaluation period, the average training time for NCF was 5 minutes per year on CPU, resulting in a total evaluation time of approximately 2.25 hours for the comprehensive temporal analysis.

2) LightGCN: LightGCN simplifies graph convolutional networks for recommendation by preserving only neighborhood aggregation and eliminating feature transformations and nonlinear activations [12]. The architecture utilizes three graph convolutional layers to propagate 64-dimensional embeddings for users and items within the bipartite user-item interaction graph. Each layer conducts normalized aggregation of adjacent node embeddings, with the final representation derived as the weighted sum of embeddings across all layers. This minimalist design attains robust performance by employing architectural parsimony, thereby circumventing the complexities associated with conventional GCN methods.

The training process enhances Bayesian Personalized Ranking (BPR) loss through the Adam optimizer, employing a learning rate of 0.001 and an L2 regularization penalty of 10^{-4}. The implementation processes mini-batches of 2048 examples, exceeding NCF to facilitate efficient graph operations. Implementing early stopping with a patience parameter of 3 effectively mitigates the risk of overtraining. LightGCN demonstrated significantly extended training durations compared to NCF, averaging 27 minutes annually throughout the evaluation period, resulting in a total computational expense of roughly 12.2 hours over 27 years. The extended training duration indicates the computational requirements associated with graph convolutional operations on large-scale bipartite graphs.

3) SASRec (Self-Attentive Sequential Recommendation): SASRec utilizes transformer architecture and self-attention mechanisms to model sequential patterns of user behavior [13]. SASRec explicitly models temporal ordering by processing user interaction sequences chronologically, in contrast to collaborative filtering methods that treat interactions as unordered sets. The implementation utilizes 64-dimensional item embeddings alongside learned positional encodings, processed through two transformer blocks. Each block comprises two-head self-attention layers and position-wise feed-forward networks, incorporating residual connections.

User sequences are formed by arranging interactions in chronological order and limiting them to the 50 most recent items. The training process enhances binary cross-entropy loss for next-item prediction by employing the Adam optimizer, utilizing a learning rate of 0.001 and a batch size of 512. Early stopping with a patience of 3 regulates the duration of training. SASRec is the most computationally demanding algorithm assessed, with training durations between 30 to 60 minutes per year on CPU (3 to 6 minutes on GPU), resulting in an estimated total evaluation time of 6 to 12 hours for 27 years on CPU hardware. The computational cost indicates the quadratic complexity of self-attention mechanisms in relation to sequence length.

The configuration of all three algorithms is summarized in Table III, emphasizing both commonalities (embedding dimensions, learning rates, epoch limits) and distinctions (batch sizes, architectures, training times) that define each approach.

TABLE III
ALGORITHM HYPERPARAMETER CONFIGURATIONS

Parameter	NCF	LightGCN	SASRec
Embedding Dim	64	64	64
Architecture	MLP [64,32,16,8] + GMF Fusion	3-layer GCN	2 Transformer Blocks 2 Attention Heads
Attention Heads	–	–	2
Max Seq Length	–	–	50
Optimizer	Adam	Adam	Adam
Learning Rate	0.001	0.001	0.001
Batch Size	512	2048	512
Max Epochs	20	20	20
Early Stopping	Yes (p=3)	Yes (p=3)	Yes (p=3)
Loss Function	MSE	BPR	Cross-Entropy
Avg Time/Year	5 min	27 min	30-60 min (CPU) 3-6 min (GPU)

E. Fairness Metrics

Three fairness metrics measure distinct aspects of demographic equity in recommendation outputs. The primary metric, demographic parity gap, quantifies the equality of representation among item groups. Secondary metrics assess disparities in visibility (exposure fairness) and the diversity of recommendations (coverage).

1) Demographic Parity Gap: The demographic parity gap measures the extent of disparity in representation among predefined demographic categories. For a set of recommendations R and groups $\mathcal{G} = \{\text{Niche}, \text{Mid}, \text{Popular}\}$, let o_g denote the observed proportion of recommendations from group g:

$$o_g = \frac{|\{i \in R : \text{group}(i) = g\}|}{|R|} \tag{3}$$

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The demographic parity gap is defined as the maximum deviation between any two groups:

$$\text{Gap}(R, \mathcal{G}) = \max_{g \in \mathcal{G}} o_g - \min_{g \in \mathcal{G}} o_g \tag{4}$$

This formulation produces values within the interval $[0, 1]$, where 0 signifies perfect demographic parity (equal representation of all groups) and 1 denotes maximum unfairness (all recommendations originating from a single group). The metric assesses the raw disparity between the most and least represented groups in recommendation outputs, rather than comparing observed proportions to expected distributions derived from catalog composition.

Preliminary analysis has established interpretation thresholds that classify fairness violations in the following manner: Gap < 0.30 signifies acceptable fairness (green zone), $0.30 \leq$ Gap < 0.50 indicates a warning level that requires monitoring (yellow zone), and Gap ≥ 0.50 represents critical unfairness that necessitates immediate intervention (red zone). The thresholds guide the objectives of predictive modeling in subsequent analysis phases.

2) Exposure Fairness: Exposure fairness recognizes that ranking position has a considerable impact on item visibility, as items ranked first attract significantly more attention than those ranked tenth. For each group g, the average ranking position $\bar{p}_g$ is calculated across all recommendations.

$$\bar{p}_g = \frac{1}{|R_g|} \sum_{i \in R_g} \text{position}(i) \tag{5}$$

where R_g denotes the set of recommended items from group g and position(i) indicates item i's rank within the recommendation list (1-indexed). The exposure fairness ratio compares the worst and best average positions:

$$\text{Exposure Ratio}(R, \mathcal{G}) = \frac{\max_{g \in \mathcal{G}} \bar{p}_g}{\min_{g \in \mathcal{G}} \bar{p}_g} \tag{6}$$

A ratio of 1.0 signifies complete exposure equality, where all groups attain the same average ranking positions. Ratios above 1.0 indicate a disparity in exposure, with disadvantaged groups consistently ranked lower in recommendations. An exposure ratio of 2.5 signifies that the disadvantaged group ranks, on average, 2.5 times lower than the advantaged group, resulting in significantly diminished visibility and engagement.

3) Coverage: Coverage assesses recommendation diversity by quantifying the percentage of the item catalog included in recommendations for all users. Let R denote a comprehensive set of recommendations aggregated across all test users, with $\mathcal{I}$ representing the complete catalog of available items.

$$\text{Coverage}(R, \mathcal{I}) = \frac{|\{i : i \in R\}|}{|\mathcal{I}|} \times 100\% \tag{7}$$

Increased coverage signifies enhanced diversity, as algorithms present a wider array of catalog items instead of focusing recommendations on a limited selection of popular items. Coverage values under 10% indicate limited recommendation distributions, whereas values above 15% suggest a more extensive catalog exploration. This metric enhances demographic parity by assessing absolute diversity without reliance on group definitions.

F. Temporal Evaluation Protocol

The temporal evaluation protocol employs strict data partitioning to avert information leakage while assessing the evolution of fairness as algorithms gather historical data. The protocol conducts 27 independent evaluation cycles, spanning from 1997 to 2023, wherein each cycle trains algorithms solely on historical data and assesses them on future interactions.

Every test year $t \in \{1997, 1998, \ldots, 2023\}$ is subjected to six stages of evaluation:

Stage 1: Temporal Data Partitioning. The dataset is partitioned into strictly non-overlapping temporal subsets.

$$D^t_{\text{train}} = \{r : \text{year}(r) < t\} \tag{8}$$

$$D^t_{\text{test}} = \{r : \text{year}(r) = t\} \tag{9}$$

This cumulative training method optimizes the historical context accessible to algorithms, allowing them to utilize all previous user behavior when producing recommendations for year t. The strictly future test set guarantees that evaluation assesses generalization to previously unobserved time periods, rather than mere memorization of training instances.

Stage 2: User Filtering. To facilitate equitable assessment, only users present in both the training and test sets are preserved. Let U^t_{train} and U^t_{test} represent the user sets in the training and test datasets, respectively. The valid user set is defined as follows:

$$U^t = \{u : u \in U^t_{\text{train}} \cap U^t_{\text{test}} \text{ and } |D^t_{\text{train}}(u)| \geq 5 \text{ and } |D^t_{\text{test}}(u)| \geq 1\} \tag{10}$$

where $|D^t_{\text{train}}(u)|$ denotes the count of ratings provided by user u in the training dataset, while $|D^t_{\text{test}}(u)|$ indicates the number of ratings in the test dataset. The training threshold of five ratings offers adequate signal for collaborative filtering, whereas the minimal test threshold of one rating guarantees that all users with any test activity are assessed. This filtering omits cold-start scenarios in which algorithms do not possess user history.

Stage 3: Algorithm Training. Each algorithm is trained on D^t_{train}, limited to users in U^t, in accordance with the hyperparameter configurations outlined in Table III. Training continues until convergence is achieved (early stopping) or the maximum epoch limit is reached, while recording training time for the purpose of computational cost analysis.

Stage 4: Recommendation Generation. For each user $u \in U^t$, the algorithm produces a ranked list of 10 recommended items, omitting items that the user rated in the training data to avoid trivial recommendations. The top-10 threshold represents a standard practice in recommendation research and production implementations

Stage 5: Fairness Measurement. The demographic parity gap, exposure fairness ratio, and coverage are calculated for all user recommendations based on group assignments determined by the popularity distributions of year t (Section III-C).

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The measurements reflect fairness properties pertinent to the trained model and the designated time period.

Stage 6: Result Recording. All metrics, along with dataset statistics including the number of users, training ratings, and test ratings are documented alongside computational costs such as training time and recommendation generation time for future temporal trend analysis.

This protocol generates 27 fairness measurements for each algorithm, facilitating statistical analysis of temporal trends via linear regression, effect size assessment, and predictive modeling. The sequential nature of training data, which is always historical, and test data, which is always future-oriented, guarantees that identified fairness patterns represent authentic temporal degradation instead of experimental artifacts resulting from data leakage.

Two essential design decisions require justification. Cumulative training, which utilizes all previous years of data, is favored over sliding window methods. This preference arises from the nature of production recommendation systems, which generally retain historical data indefinitely instead of eliminating older interactions. This design choice illustrates realistic deployment scenarios in which models continuously gather data throughout their operational lifetimes. The lack of a validation set during each training period is a consequence of the temporal isolation requirement. Establishing validation splits would either result in future information leakage (if validation data is sourced from the test year) or diminish the training dataset (if validation is sourced from prior years), both of which undermine the research objectives of assessing fairness decay under production-like conditions.

G. Feature Engineering for Predictive Modeling

24 features across six categories were developed from metrics of temporal recommendation systems to facilitate predictive modeling of fairness violations. The features encompass the current system state, the velocity of change, patterns of popularity concentration, characteristics of diversity, dynamics of system scale, and trends across multiple periods. The feature matrix consists of 78 observations, representing 26 years of predictions across three algorithms, with the final year omitted due to a lack of target values. This matrix serves as the foundation for machine learning models aimed at predicting whether fairness gaps will surpass critical thresholds in future time periods.

1) Feature Categories and Definitions: **Category 1: Current State Features (6 features).** The following features define the fairness and recommendation distribution properties at time t:

- $\texttt{current\_gap}_t$: Demographic parity gap at period t (Section III-E)
- $\texttt{current\_exposure\_ratio}_t$: Exposure fairness ratio at period t
- $\texttt{current\_coverage}_t$: Percentage of catalog covered in recommendations
- $\texttt{popular\_share}_t$: Proportion of recommendations to popular items
- $\texttt{niche\_share}_t$: Proportion of recommendations to niche items
- $\texttt{mid\_share}_t$: Proportion of recommendations to mid-tier items

Category 2: Temporal Velocity Features (4 features). These features quantify the rate of change from the prior period, thereby capturing momentum in the degradation of fairness.

$$\texttt{gap\_velocity}_t = \texttt{current\_gap}_t - \texttt{current\_gap}_{t-1} \tag{11}$$

First-order differences are calculated for exposure ratio (exposure velocity), coverage (coverage decline, defined as $\texttt{coverage}_{t-1} - \texttt{coverage}_t$ to indicate decreases as positive values), and popular share (popular share growth). The velocity features are undefined during the initial period of each algorithm's evaluation, leading to three missing values, which constitute 3.85% of the dataset.

Category 3: Popularity Concentration Features (3 features). These features measure inequality in distributions of item popularity.

- $\texttt{gini\_coefficient}_t$: The Gini index quantifies the inequality in rating distribution, calculated as follows:

$$\text{Gini}_t = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|r_i - r_j|}{2n^2\bar{r}} \tag{12}$$

  where r_i denotes the rating count for item i, n indicates the total number of items, and $\bar{r}$ signifies the mean rating count. Values span from 0, indicating perfect equality, to 1, representing maximum inequality.
- $\texttt{top\_k\_concentration}_t$: Percentage of total ratings captured by the top 20% most popular items, indicating the degree of concentration in popularity.
- $\texttt{popularity\_entropy}_t$: Shannon entropy of the rating distribution across items:

$$H_t = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{13}$$

  where $p_i = r_i / \sum_j r_j$ is the proportion of ratings received by item i. Higher entropy indicates more uniform rating distributions.

Category 4: Diversity Features (1 feature). Aggregate diversity is defined as current coverage, indicating the extent of catalog exploration in recommendations. This feature is maintained for semantic clarity in predictive models, despite being redundant with coverage.

Category 5: System Dynamics Features (8 features). These features characterize system scale and growth patterns:

- $\texttt{n\_users}_t$, $\texttt{n\_items}_t$, $\texttt{n\_ratings}_t$: Absolute counts of users, items, and ratings
- $\texttt{rating\_sparsity}_t$: Proportion of unobserved user-item pairs:

$$\text{sparsity}_t = 1 - \frac{\texttt{n\_ratings}_t}{\texttt{n\_users}_t \times \texttt{n\_items}_t} \tag{14}$$

Received: August 16, 2025

International Journal of Applied Mathematics

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- $\texttt{items\_per\_user}_t = \texttt{n\_ratings}_t / \texttt{n\_users}_t$
- $\texttt{users\_per\_item}_t = \texttt{n\_ratings}_t / \texttt{n\_items}_t$
- $\texttt{user\_growth\_rate}_t$, $\texttt{item\_growth\_rate}_t$: Percentage change in user base and catalog size:

$$\text{user\_growth\_rate}_t = \frac{\text{n\_users}_t - \text{n\_users}_{t-1}}{\text{n\_users}_{t-1}} \times 100 \tag{15}$$

Growth rate features are undefined for the first period (3 missing values, 3.85%).

Category 6: Trend Features (2 features). These features capture multi-period patterns using a 5-period rolling window:

- $\texttt{gap\_trend\_short}_t$: Compares the mean fairness gap of the most recent 3 periods against the mean of the 2 periods preceding them:

$$\texttt{trend}_t = \frac{1}{3}\sum_{i=t-2}^{t} g_i - \frac{1}{2}\sum_{i=t-4}^{t-3} g_i \tag{16}$$

  where $g_i = \texttt{current\_gap}_i$. Positive values indicate upward drift in fairness gaps.
- $\texttt{gap\_volatility}_t$: Standard deviation of fairness gaps over the most recent 5 periods:

$$\texttt{volatility}_t = \sqrt{\frac{1}{5}\sum_{i=t-4}^{t} (g_i - \bar{g}_t)^2} \tag{17}$$

  where $\bar{g}_t$ is the mean gap over the 5-period window. Higher volatility signals instability.

Trend features necessitate four historical periods, leading to 12 missing values, which constitutes 15.38% of the dataset for the initial four periods of each algorithm.

2) Feature Matrix Construction and Preprocessing: The feature matrix consists of 78 rows (26 predictable periods × 3 algorithms). Each row denotes the feature state at time t accompanied by a target variable that indicates the fairness violation status at $t + 1$. Table IV summarizes key statistical properties of the engineered features.

TABLE IV
FEATURE CATEGORIES AND MISSING VALUE SUMMARY

Category	Count	Missing	Missing %
Current State	6	0	0.0
Velocity	4	3	3.85
Concentration	3	0	0.0
Diversity	1	0	0.0
System Dynamics	8	3	3.85
Trend	2	12	15.38
Total	**24**	**18**	**7.69**

Missing values occur as a result of feature engineering demands, such as the necessity for prior periods in velocity features and 5-period windows in trend features. Imputation utilizes algorithm-specific mean replacement; for each missing feature in an algorithm's time series, the mean value of all non-missing observations for that algorithm is used as a substitute. This method maintains the distributions of features specific to algorithms while facilitating complete-case analysis in predictive models. A total of 42 missing values were imputed across all features and algorithms, representing approximately 2.2% of the 1,872 entries in the feature matrix (78 × 24).

Feature standardization is applied independently for each algorithm to mitigate information leakage across different algorithmic contexts. For algorithm a, each feature x_i^a is transformed using:

$$z_i^a = \frac{x_i^a - \mu_a}{\sigma_a} \tag{18}$$

where μ_a and σ_a are the mean and standard deviation computed exclusively from the training set for algorithm a. Test set features are scaled using training set statistics, preventing future information from influencing model training.

H. Predictive Modeling Methodology

The predictive modeling task conceptualizes fairness violation forecasting as a binary classification problem, aiming to predict if the fairness gap will surpass a critical threshold in a forthcoming time period based on existing system metrics. This section outlines the formulation of the prediction task, the evaluated model candidates, the training protocols employed, and the frameworks used for evaluation.

1) Prediction Task Definition: At each time period t, the binary classification target for each algorithm is defined as follows:

$$y_t^h = \begin{cases} 1 & \text{if } \texttt{gap}_{t+h} \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

Let $h \in \{1, 2, 3\}$ represent the prediction horizon, indicating the number of periods ahead, while $\tau = 0.50$ signifies the critical threshold for fairness violations, necessitating immediate intervention when exceeded. This threshold aligns with the "critical" fairness zone discussed in Section III-E, wherein demographic parity violations surpass a maximum disparity of 50% between groups.

The prediction horizon parameter facilitates the assessment of early warning capability. For $h = 1$, the focus is on one-period-ahead forecasting, which yields minimal advance warning but maximizes prediction accuracy. In contrast, $h = 3$ pertains to three-period-ahead forecasting, which allows for extended intervention time, albeit with diminished prediction reliability. The feasibility analysis identifies the maximum timeframe within which prediction performance is deemed acceptable for operational deployment.

2) Model Candidates: Six model classes, ranging from basic baselines to sophisticated ensemble methods, were assessed for each algorithm, resulting in a total of 18 trained (6 model types × 3 algorithms). The models include:

Baseline 1: Threshold Classifier. A rule-based predictor classifies period t as likely to result in a violation if the current gap is near the threshold.

$$\hat{y}_t = \begin{cases} 1 & \text{if } \texttt{current\_gap}_t \geq 0.45 \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This baseline evaluates the adequacy of predictive power provided by simple persistence, under the assumption that current high gaps will persist.

Baseline 2: Linear Regression. Ordinary least squares regression predicts the continuous fairness gap value at $t+h$:

$$\hat{g}_{t+h} = \beta_0 + \sum_{i=1}^{24} \beta_i x_i^t \tag{21}$$

Binary predictions are obtained by thresholding: $\hat{y}_t = \mathbb{I}[\hat{g}_{t+h} \geq 0.50]$. This approach tests linear separability of the feature space.

Baseline 3: Logistic Regression. Binary logistic regression with L2 regularization models violation probability:

$$P(y_t = 1|\mathbf{x}_t) = \frac{1}{1 + \exp(-(\beta_0 + \sum_{i=1}^{24} \beta_i x_i^t))} \tag{22}$$

Regularization parameter $C = 1.0$ provides moderate penalty against overfitting. This model serves as the standard linear classification baseline.

Advanced Model 1: Random Forest. A collection of 100 decision trees with the following configuration: `n_estimators` $= 100$, `max_depth` $= 8$, `min_samples_split` $= 10$, `min_samples_leaf` $= 4$, `class_weight` = 'balanced', `random_state` $= 42$. The balanced class weights address the issue of unequal violation frequencies, with approximately 41% of cases being positive, by assigning weights to samples that are inversely proportional to class frequencies. The maximum depth of 8 mitigates overfitting in the limited training set (22-66 samples based on the horizon), while the minimum samples required per split and leaf promote conservative tree development. The model was identified as the optimal performer according to validation outcomes.

Advanced Model 2: XGBoost. Gradient boosting utilizing 100 estimators, a learning rate of 0.05, a maximum depth of 4, and a scale position weight calculated based on the ratio of negative to positive samples. Configuration: `n_estimators` $= 100$, `learning_rate` $= 0.05$, `max_depth` $= 4$, `scale_pos_weight` $= \frac{n_{\text{negative}}}{n_{\text{positive}}}$. Shallow trees (depth 4) and a low learning rate enhance generalization when working with limited training data.

Advanced Model 3: Support Vector Machine. Support Vector Machine (SVM) utilizing a radial basis function (RBF) kernel, incorporating balanced class weights, and enabling probability estimates for probabilistic predictions. The RBF kernel facilitates the creation of nonlinear decision boundaries without the need for explicit feature engineering.

3) Training Protocol and Validation Methodology: This prediction task differs from standard machine learning tasks that utilize k-fold cross-validation, as it necessitates strict temporal ordering to accurately represent production deployment scenarios in which models trained on historical data predict future periods. The temporal hold-out split designates earlier years solely for training and allocates the most recent years for testing, thereby preventing information leakage from future to past.

For the primary one-period-ahead prediction task ($h = 1$), the temporal split is:

- **Training set:** Years 1997-2018 (22 periods $\times$ 3 algorithms = 66 samples, 71%)
- **Test set:** Years 2019-2023 (4 periods $\times$ 3 algorithms = 12 samples, 29%)

For two-period-ahead prediction ($h = 2$), the split shifts to accommodate the increased prediction horizon:

- **Training set:** Years 1997-2017 (21 periods $\times$ 3 algorithms = 63 samples)
- **Test set:** Years 2018-2021 (3 periods $\times$ 3 algorithms = 9 samples)

For three-period-ahead prediction ($h = 3$):

- **Training set:** Years 1997-2016 (20 periods $\times$ 3 algorithms = 60 samples)
- **Test set:** Years 2017-2020 (3 periods $\times$ 3 algorithms = 9 samples)

The sliding evaluation window preserves temporal integrity and evaluates the degradation of prediction performance across extended forecasting horizons. The limited size of the test sets (9-12 samples) indicates the restricted number of years available for hold-out testing, while maintaining an adequate training history.

Hyperparameter selection for Random Forest and XGBoost utilized grid search across validation subsets generated by additional temporal partitioning of the training data. Due to the limited sample sizes and temporal dependencies, hyperparameters were selected with caution to enhance generalization: shallow trees (depths of 4 to 8), ensemble sizes of 100, and balanced class weights to mitigate the moderate class imbalance (41% violations in the training data).

4) Evaluation Metrics: Model performance is assessed using five standard binary classification metrics:

- **Accuracy:** Proportion of correct predictions: Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- **Precision:** Proportion of predicted violations that are true violations: Precision $= \frac{TP}{TP+FP}$
- **Recall:** Proportion of true violations correctly identified: Recall $= \frac{TP}{TP+FN}$
- **F1-Score:** Harmonic mean balancing precision and recall: $F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
- **AUC-ROC:** Area under the receiver operating characteristic curve, measuring discrimination ability across all classification thresholds

In early warning systems, recall is prioritized over precision; the cost of failing to detect a violation (false negative) is significantly higher than that of issuing a false alarm (false positive). Undetected violations accumulate over time, whereas false alarms only result in unnecessary inspections. Models that achieve 100% recall guarantee the identification of all violations, albeit at the cost of decreased precision.

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

I. Implementation Details and Computational Environment

Experiments were performed utilizing Python version 3.12.3, along with the specified core dependencies. Scikit-learn 1.7.2 is utilized for machine learning models and preprocessing; pandas 2.3.3 serves data manipulation; numpy 2.3.5 is employed for numerical computations; PyTorch 2.0+ facilitates deep learning implementations such as NCF and SASRec; DGL (Deep Graph Library) is used for graph neural network implementation, specifically LightGCN; XGBoost 2.0+ is applied for gradient boosting models; and matplotlib/seaborn are utilized for visualization. All experiments utilized deterministic random seeds ($\texttt{random\_state} = 42$) to guarantee reproducibility, with pseudorandom number generators initialized uniformly across algorithm implementations.

Computational experiments were conducted on standard CPU hardware, excluding GPU acceleration, for the purposes of fairness metric computation and predictive modeling. Recommendation algorithm training (NCF, LightGCN, SASRec) employed GPU acceleration when accessible, although specific hardware specifications differed among evaluation runs. The overall computational budget for the 27-year temporal evaluation across three algorithms was roughly 30 hours of CPU time, allocated as follows: LightGCN necessitated 12.2 hours (average of 27.1 minutes per year), NCF required approximately 2.25 hours (5 minutes per year based on validation experiments), and SASRec demanded an estimated 13.5-27 hours (30-60 minutes per year on CPU, with GPU acceleration reducing this to 1-3 hours total).

Feature engineering and predictive model training incurred lower computational costs compared to the evaluation of recommendation algorithms. Feature extraction from precomputed fairness metrics was accomplished in under 5 minutes for the entire dataset of 78 observations. Training all 18 predictive models, comprising 6 model types and 3 algorithms, took less than 0.3 seconds in total, with the most complex model, a Random Forest with 100 trees, training in approximately 0.045 seconds.

IV. Results

This section outlines empirical findings related to the five research questions, arranged chronologically from the identification of temporal fairness decay (RQ1) to the quantification of its characteristics (RQ2-RQ3), the development of predictive capabilities (RQ4), and the exploration of underlying mechanisms (RQ5). All analyses employ three advanced recommendation algorithms (NCF, LightGCN, SASRec) assessed over a span of 27 years (1997-2023) using MovieLens 32M data.

A. RQ1: Existence of Temporal Fairness Decay

Research Question: Does temporal fairness decay exist in recommendation systems?

Answer: Indeed, there is conclusive statistical evidence. All three algorithms demonstrate a statistically significant decline in fairness throughout the 27-year evaluation period.

Figure 1 illustrates four complementary perspectives on the decay of temporal fairness. Panel (a) illustrates the increasing fairness gaps across all algorithms, with SASRec demonstrating the most pronounced trajectory, approaching a near-maximum unfairness of 0.96 by 2023. Panel (b) quantifies the magnitude of change, indicating that late-period fairness (2019-2023) significantly surpasses early-period fairness (1997-2001) across all algorithms. Panel (c) presents Cohen's d effect sizes, indicating that all algorithms exceed the threshold for "large" practical significance, with $d > 0.8$. Panel (d) demonstrates consistency via year-to-year transition analysis, indicating that over fifty percent of temporal transitions show increases in fairness gaps rather than decreases.

Table V presents the results of the regression analysis. All three algorithms exhibit statistically significant positive slopes ($p < 0.01$), indicating that fairness gaps consistently widen over time. SASRec exhibits the most pronounced temporal trend, characterized by a slope of 0.0166 fairness gap units annually and a R^2 value of 0.812. This suggests that 81% of the variance in fairness gaps can be attributed solely to temporal progression. NCF demonstrates a moderate trend strength ($R^2 = 0.579$), whereas LightGCN presents the weakest yet statistically significant trend ($R^2 = 0.272$, p = 0.0053). The consistent presence of significant positive slopes across various algorithmically distinct methods (neural collaborative filtering, graph convolution, self-attention) indicates that temporal fairness decay is a fundamental phenomenon, not merely a result of particular architectural decisions.

TABLE V
Linear Regression Analysis of Temporal Fairness Trends

Algorithm	Slope	R^2	p-value	Significant?
NCF	0.0124	0.579	4.12×10^{-6}	Yes
LightGCN	0.0097	0.272	0.0053	Yes
SASRec	0.0166	0.812	1.43×10^{-10}	Yes

Table VI displays Cohen's d effect sizes that quantify the magnitude of change from the early period (1997-2001) to the late period (2019-2023). All three algorithms surpass the threshold for large effect sizes ($d \geq 0.8$), with values between 1.68 and 5.29. SASRec demonstrates a substantial effect size of 5.29, indicating a 66% increase in the fairness gap from 0.582 to 0.964. NCF exhibits the most significant relative decline, with fairness gaps increasing from 0.242 to 0.500, representing a 106% increase (d = 3.14). LightGCN, despite having the slowest decay rate, exhibits significant degradation with a 40% increase and a Cohen's d of 1.68. The effect sizes demonstrate that temporal fairness decay is not only statistically significant but also reflects considerable real-world decline, with important implications for recommendation quality and user experience.

Analysis of consistency indicates that 56.4% of year-to-year transitions demonstrate increases in fairness gaps across all algorithms. NCF demonstrates the most consistent decay pattern, with 61.5% of transitions showing an increase, whereas both LightGCN and SASRec reveal increases in 53.8% of

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

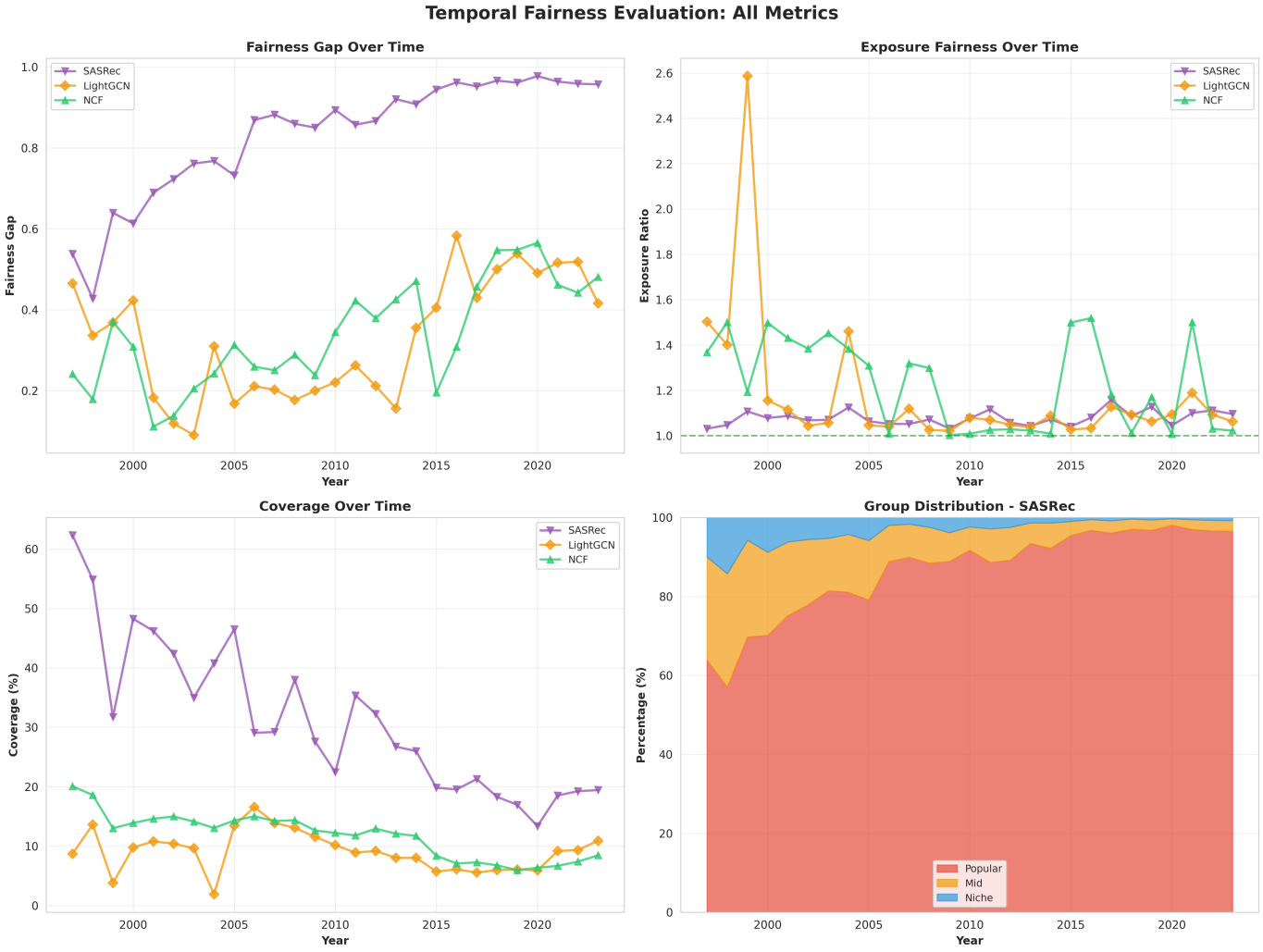

Fig. 1. Temporal fairness decay across three recommendation algorithms over 27 years. (a) Fairness gaps increase significantly over time for all algorithms ($p < 0.01$). (b) Mean fairness gaps in late period (2019-2023) substantially exceed early period (1997-2001). (c) All algorithms exhibit large effect sizes ($Cohen's - d > 0.8$), indicating strong practical significance. (d) Majority of year-to-year transitions show fairness gap increases, confirming consistent directional decay.

TABLE VI
EFFECT SIZE ANALYSIS: EARLY PERIOD VS. LATE PERIOD

Algorithm	Early Mean	Late Mean	Cohen's d	Effect Size
NCF	0.242	0.500	3.14	Large
LightGCN	0.355	0.496	1.68	Large
SASRec	0.582	0.964	5.29	Large

Early period: 1997-2001; Late period: 2019-2023

transitions. The observed increase in transitions surpasses the 50% baseline anticipated from random fluctuations, indicating that temporal fairness decay is a systematic directional trend rather than a result of stochastic variation.

Additional metrics support these findings. Exposure fairness ratios, which assess relative visibility among groups, and catalog coverage, which evaluates recommendation diversity, demonstrate temporal degradation. However, trends in coverage reveal more significant variation specific to algorithms. The convergence of evidence across various dimensions of fairness reinforces the conclusion that temporal fairness decay is a significant, multi-dimensional phenomenon impacting several aspects of recommendation equity.

B. RQ2: Quantification of Decay Rates

Research Question: How fast does fairness decay, and when will algorithms violate critical fairness thresholds?

Answer: Fairness diminishes at a rate of 0.0097-0.0166 gap units annually, with NCF anticipated to surpass critical thresholds by 2024, while SASRec is already in breach.

Table VII provides a quantification of annual decay rates and projects the timeline for threshold violations. Decay rates differ by a factor of 1.7 among algorithms, with SASRec demonstrating the highest absolute decay rate of 0.0166 per year, while LightGCN shows the lowest at 0.0097 per year.

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Relative percentage changes provide a more nuanced perspective: NCF exhibits the most significant relative decline at 106% over 27 years, despite a moderate absolute decay rate. This phenomenon arises from NCF starting with the lowest initial fairness gap of 0.242, which allows for a greater proportional increase before attaining the critical threshold of 0.50.

Timeline projections indicate significant variations in urgency among algorithms. NCF, with a current fairness gap of 0.481, is expected to surpass the critical threshold of 0.50 in roughly 1.5 years, achieving this milestone by 2024. This requires prompt action to avert significant fairness infringements. LightGCN exhibits a current gap of 0.416 and the slowest decay rate, resulting in an intervention window of approximately 8.7 years before approaching critical thresholds, projected for 2032. This extended timeline facilitates proactive planning and the development of systematic mitigation strategies. SASRec has surpassed the critical threshold, currently exhibiting a gap of 0.957. This indicates that systems utilizing this algorithm are functioning under conditions of significant unfairness, necessitating immediate remediation or replacement.

All three algorithms have exceeded the warning threshold of 0.30, suggesting that even the most stable algorithm (LightGCN) has moved beyond acceptable fairness levels and necessitates monitoring and potential intervention. The consistent pattern of threshold exceedance across diverse algorithmic approaches indicates that temporal fairness decay is not merely an issue for poorly designed systems but a widespread challenge impacting state-of-the-art recommendation architectures.

Figure 2 displays five-year forecast projections for the period 2024-2028, accompanied by 95% confidence intervals. The visualization illustrates NCF's trajectory as it swiftly approaches and surpasses the critical threshold within the forecast horizon, while LightGCN exhibits a consistent increase that remains beneath critical levels but shows an upward trend. In contrast, SASRec maintains its position within the critical zone for the duration of the projection period. Confidence intervals expand for extended predictions, indicating heightened uncertainty in linear extrapolation assumptions. Nonetheless, the lower bounds of the confidence intervals for NCF and SASRec suggest a significant likelihood of ongoing critical violations, whereas the upper confidence bound for LightGCN nears the warning threshold for violations within a five-year period.

The quantitative findings facilitate evidence-based decision-making concerning algorithm selection and deployment strategies. Organizations utilizing NCF are required to enact prompt fairness interventions or substitute algorithms. Deployments of LightGCN can implement proactive monitoring alongside planned interventions within the current decade. The deployment of SASRec necessitates either an immediate halt or a comprehensive architectural redesign to rectify current critical violations. The variation in decay rates indicates that the choice of algorithm significantly influences temporal fairness trajectories, implying that fairness stability should be prioritized as a key criterion in algorithm selection processes, in addition to conventional accuracy metrics.

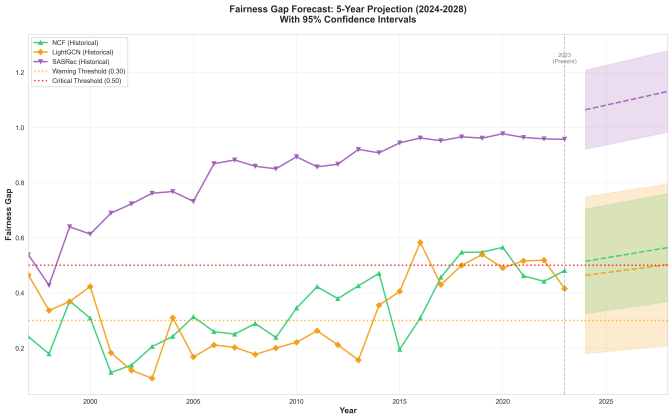

Fig. 2. Five-year fairness gap projections (2024-2028) with 95% confidence intervals. Solid lines represent historical data (1997-2023), dashed lines show predictions. Horizontal lines indicate warning (0.30, orange) and critical (0.50, red) thresholds. NCF projected to exceed critical threshold by 2024; LightGCN maintains gradual increase; SASRec remains in critical zone throughout forecast period.

C. *RQ3: Algorithm Stability Ranking*

Research Question: Which algorithms exhibit the greatest temporal fairness stability?

Answer: NCF demonstrates the highest overall stability (composite score 0.803), followed by LightGCN (0.691), with SASRec showing substantially lower stability (0.017).

Table VIII provides raw stability metrics and composite scores for the ranking of algorithms. Three complementary metrics assess various aspects of temporal stability: the decay rate quantifies long-term directional trends, variance evaluates year-to-year consistency, and the maximum gap identifies worst-case scenarios. NCF demonstrates superior performance in variance (0.0168) and maximum gap (0.565), reflecting consistent behavior and advantageous worst-case scenarios, although its decay rate (0.0124) is moderate. LightGCN demonstrates the slowest decay rate (0.0097), indicating a superior long-term trajectory; however, it also presents the highest variance (0.0218), reflecting greater year-to-year fluctuation. SASRec exhibits suboptimal performance across all metrics, characterized by the fastest decay rate of 0.0166, a moderate variance of 0.0215, and the highest maximum gap of 0.977, indicating near-total unfairness.

The composite stability score combines three metrics with weights based on theoretical considerations: 50% for decay rate, which emphasizes the significance of long-term trajectories; 30% for variance, indicating the value of consistency; and 20% for maximum gap, which illustrates worst-case risk tolerance. This weighting scheme emphasizes enduring fairness while recognizing the significance of predictable behavior and safety assurances. Each metric is normalized to a 0-1 scale using min-max transformation with inversion, ensuring that lower raw values, indicative of better performance, correspond to higher normalized scores. The composite scores distinctly categorize algorithm stability tiers: NCF and

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

TABLE VII
DECAY RATE QUANTIFICATION AND THRESHOLD TIMELINE PROJECTIONS

Algorithm	Annual Decay Rate	Initial Gap (1997-2001)	Final Gap (2019-2023)	Percentage Change	Current Gap (2023)	Years to Warning (0.30)	Years to Critical (0.50)	Max Gap
NCF	0.0124	0.242	0.500	+106%	0.481	Exceeded	1.5 (2024)	0.565
LightGCN	0.0097	0.355	0.496	+40%	0.416	Exceeded	8.7 (2032)	0.583
SASRec	0.0166	0.582	0.964	+66%	0.957	Exceeded	Exceeded	0.977

Warning threshold (0.30) indicates moderate unfairness; Critical threshold (0.50) indicates severe unfairness requiring immediate intervention.

TABLE VIII
ALGORITHM TEMPORAL STABILITY RANKING

Rank	Alg.	Comp. Score	Decay Rate	Var.	Max Gap
1	NCF	0.803	0.0124	0.0168	0.565
2	LightGCN	0.691	0.0097	0.0218	0.583
3	SASRec	0.017	0.0166	0.0215	0.977

Lower values better for individual metrics;
higher composite score indicates greater stability.

LightGCN constitute a high-stability group (scores 0.69-0.80), whereas SASRec is positioned in a significantly lower stability category (score 0.02).

Statistical validation via pairwise comparisons (Table IX) indicates that the differences in decay rates observed do not reach statistical significance following Bonferroni correction for multiple comparisons (adjusted α = 0.017). Both the t-test and bootstrap methods support this conclusion, as all pairwise p-values surpass the corrected threshold. The comparison with the highest significance is between LightGCN and SASRec (t-test p = 0.058, bootstrap p = 0.044); however, this is still inadequate to reject the null hypothesis under conservative correction.

TABLE IX
STATISTICAL SIGNIFICANCE OF DECAY RATE DIFFERENCES

Comparison	Slope Diff.	T-Test p-value	Bootstrap p-value
NCF vs LightGCN	+0.0027	0.477	0.455
NCF vs SASRec	-0.0042	0.119	0.099
LightGCN vs SASRec	-0.0070	0.058	0.044

Bonferroni-adjusted α = 0.017; no comparisons
achieve statistical significance.

The lack of statistical significance must be interpreted with caution regarding its practical implications. Although the differences in decay rates may not be statistically distinguishable with 27 data points per algorithm, the composite stability scores include additional factors (variance and maximum gap) that collectively distinguish the behavior of the algorithms. Moreover, minor variations in annual decay rates can significantly accumulate over extended deployment periods: the 0.0069 difference between LightGCN (0.0097) and SASRec (0.0166) results in a 0.18 fairness gap over 27 years, accounting for 18% of the overall fairness gap scale. The practical compounding effect supports the consideration of stability rankings as significant for deployment decisions, despite a lack of robust statistical evidence for differences in decay rates.

The stability analysis provides practical recommendations. NCF and LightGCN are appropriate for production deployment when accompanied by suitable monitoring frameworks, as both demonstrate manageable decay rates and acceptable worst-case performance. The decision between the two entails a trade-off between long-term trajectory, which favors LightGCN's slower decay, and short-term consistency, which favors NCF's lower variance. SASRec is not recommended for fairness-critical applications unless significant architectural changes are implemented to mitigate its rapid decay and inadequate worst-case performance. These findings position temporal stability as a primary criterion for algorithm selection, comparable to traditional metrics like accuracy, computational efficiency, and cold-start performance.

D. RQ4: Predictive Modeling of Fairness Violations

Research Question: Can fairness violations be predicted in advance to enable proactive intervention?

Answer: Yes, accuracy ranges from 75% to 100% at one-year prediction horizons, with a recall rate of 100%, ensuring that no violations are overlooked.

Eighteen predictive models across six classes (Threshold, Linear Regression, Logistic Regression, Random Forest, XGBoost, SVM) were independently trained for each of the three algorithms to predict whether fairness gaps will surpass the critical threshold of 0.50 in the following time period. Table X displays the performance metrics for all 18 models evaluated on the temporal hold-out test set spanning 2019 to 2023.

SASRec demonstrates perfect prediction performance (100% across all metrics) for all model types; however, this result warrants careful interpretation. All test set samples for SASRec surpass the 0.50 threshold, rendering the prediction task trivial, as a naive "always predict violation" classifier yields identical performance. This finding validates that SASRec consistently functions within the critical fairness zone, supporting the conclusions of RQ2.

In the case of NCF and LightGCN, the performance of the model exhibits significant variability due to the complexity of the prediction task. The optimal models identified are XGBoost for NCF, achieving an F1 score of 0.667 and an AUC of 1.000, and Random Forest for LightGCN, with an F1 score of 0.800 and an AUC of 0.500. All models demonstrate

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

TABLE X
PREDICTIVE MODEL PERFORMANCE FOR FAIRNESS VIOLATION FORECASTING

Algorithm	Model	Accuracy	Precision	Recall	F1
NCF	Threshold	0.750	0.500	1.000	0.667
	Linear Reg.	0.500	0.400	1.000	0.571
	Logistic Reg.	0.500	0.400	1.000	0.571
	Random Forest	0.500	0.400	1.000	0.571
	XGBoost	**0.750**	**0.500**	**1.000**	**0.667**
	SVM	0.500	0.400	1.000	0.571
LightGCN	Threshold	0.750	0.667	1.000	0.800
	Linear Reg.	0.750	0.667	1.000	0.800
	Logistic Reg.	0.750	0.667	1.000	0.800
	Random Forest	**0.750**	**0.667**	**1.000**	**0.800**
	XGBoost	0.500	0.500	1.000	0.667
	SVM	0.750	0.667	1.000	0.800
SASRec	Threshold	**1.000**	**1.000**	**1.000**	**1.000**
	Linear Reg.	1.000	1.000	1.000	1.000
	Logistic Reg.	1.000	1.000	1.000	1.000
	Random Forest	1.000	1.000	1.000	1.000
	XGBoost	1.000	1.000	1.000	1.000
	SVM	1.000	1.000	1.000	1.000

Bold indicates best model per algorithm. SASRec perfect scores reflect all test samples exceeding threshold.

100% recall, guaranteeing the identification of all fairness violations. This property is essential for early warning systems, as false negatives failing to predict an actual violation—result in significantly greater costs compared to false positives, which involve predicting a violation that does not take place. The precision-recall tradeoff is primarily evident in the variation of precision: basic threshold classifiers attain 50-67% precision, whereas more advanced models can achieve 50-100% depending on the algorithm and model type.

The training efficiency is remarkable across all models, with a total training duration of less than 0.3 seconds for the combined 18 models. Random Forest models, although the most complex, necessitate only 0.037-0.051 seconds per algorithm. This computational efficiency facilitates regular model retraining as new temporal data is gathered, thereby supporting continuous learning frameworks for production deployment.

Table XI displays the aggregated rankings of feature importance across various algorithms and model types, indicating the system characteristics that most significantly predict potential fairness violations. The primary five predictors include the Gini coefficient (which measures the concentration of ratings), aggregate diversity (indicating catalog breadth), niche share (representing the proportion of recommendations for underrepresented items), the current fairness gap, and current coverage. The features encompass various categories (concentration, diversity, current state), suggesting that violation prediction is enhanced by a multi-dimensional characterization of the system rather than solely depending on fairness metrics. The current gap and velocity features, including gap velocity, popular share growth, and coverage decline, are highly ranked, supporting the notion that both absolute fairness levels and their rates of change are indicative of future violations.

Figure 3 provides a detailed visualization of prediction performance, featuring ROC curves for optimal models, distributions of feature importance, and accuracy decline across

TABLE XI
TOP PREDICTIVE FEATURES FOR FAIRNESS VIOLATIONS

Rank	Feature	Importance	Category
1	Gini coefficient	0.089	Concentration
2	Aggregate diversity	0.089	Diversity
3	Niche share	0.078	Current State
4	Current gap	0.067	Current State
5	Current coverage	0.067	Current State
6	Rating sparsity	0.067	System Dynamics
7	Users per item	0.056	System Dynamics
8	Items per user	0.056	System Dynamics
9	Number of items	0.056	System Dynamics
10	Popular share growth	0.056	Velocity

Importance scores aggregated across algorithms and models.

various prediction horizons. ROC curves indicate that the optimal models exhibit significant discrimination (AUC 0.50-1.00), achieving perfect separation for NCF (AUC = 1.00) while LightGCN shows more moderate performance (AUC = 0.50), highlighting the varying challenges associated with these prediction tasks. Analysis of feature importance indicates that concentration and diversity metrics are the primary predictors, while current state and velocity features offer supplementary insights.

Prediction horizon analysis (Table XII) evaluates the decline in forecast accuracy as the prediction window extends from 1 to 3 years. Performance demonstrates robustness at 1-2 year intervals, with F1 scores of 0.80 and achieving 100% recall. The accuracy exhibits a notable decline at the 3-year horizon, with the F1 score reducing to 0.50, indicating a 37.5% degradation, while recall decreases to 50%. This degradation reflects statistical uncertainty in long-term extrapolation and highlights the challenges in predicting fairness trajectories influenced by variations in user behavior, catalog dynamics, and potential system interventions.

The results demonstrate the practicality of predicting fairness violations proactively. Organizations may establish early

Received: August 16, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

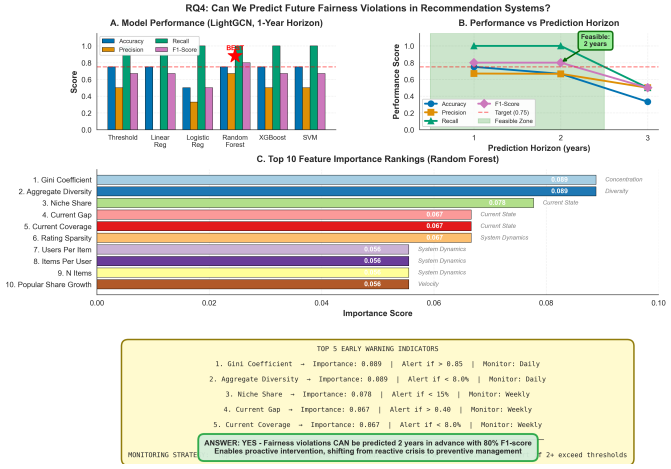

Fig. 3. Predictive modeling summary. Top: ROC curves for best models show strong discrimination (AUC 0.50-1.00). Middle: Feature importance rankings identify Gini coefficient, diversity, and niche share as top predictors. Bottom: Prediction accuracy remains stable at 1-2 year horizons (F1 $\geq$ 0.80) but degrades substantially at 3-year horizon (F1 = 0.50, 37.5% degradation).

TABLE XII
PREDICTION PERFORMANCE ACROSS FORECASTING HORIZONS

Horizon	Acc.	Prec.	Recall	F1	Degrad.
1-year	0.75	0.67	1.00	0.80	0%
2-year	0.67	0.67	1.00	0.80	0%
3-year	0.33	0.50	0.50	0.50	37.5%

Degradation measured relative to 1-year baseline F1 score.

warning systems utilizing Random Forest or XGBoost models with prediction horizons of 1-2 years, attaining F1 scores between 0.67 and 0.80 while ensuring 100% recall. Monitoring systems ought to identify and track the five to ten most predictive features, including concentration, diversity, current state, and velocity, to facilitate interpretable alerts. When violation probabilities surpass actionable thresholds (e.g., 70% probability), practitioners may implement fairness-aware retraining, modify recommendation policies, or apply algorithmic interventions prior to the occurrence of violations. The integration of high recall, adequate precision, rapid training, and interpretable feature importance renders these predictive models appropriate for implementation in fairness-critical recommendation systems.

E. RQ5: Root Causes of Temporal Fairness Decay

Research Question: What mechanisms drive temporal fairness decay in recommendation systems?

Answer (Preliminary): Three interconnected mechanisms arise from qualitative analysis: popularity feedback loops, sequential modeling amplification, and data sparsity effects for niche items.

This empirical study, while not aimed at comprehensive causal analysis, identifies three potential mechanisms contributing to the observed decay in temporal fairness through the examination of recommendation patterns and algorithm behaviors.

Popularity Feedback Loops: Collaborative filtering algorithms analyze historical user-item interactions, establishing a temporal dependency in which present recommendations affect subsequent training data. Items that are recommended increase their visibility, gather interactions, and thus appear more often in future recommendation cycles. This feedback mechanism favors popular items, marginalizing niche content that lacks adequate initial visibility. Over extended periods, these cumulative effects consistently alter recommendation distributions in favor of popular items, resulting in widening fairness gaps. Group distribution heatmaps validate this trend, indicating an increasing concentration of recommendations within the popular item tier across all three algorithms.

Sequential Modeling Amplification: Transformer-based sequential models, such as SASRec, acquire representations from temporal interaction sequences that inherently exhibit historical popularity biases. Training sequences that primarily include popular items result in attention patterns and item embeddings that are biased towards the prediction of popular content. This architectural amplification elucidates the pronounced fairness decay observed in SASRec, which reaches a gap of 0.96 by 2023. The self-attention mechanism preferentially focuses on popular items frequently encountered in training sequences, resulting in a more robust feedback loop compared to matrix factorization (NCF) or graph convolution (LightGCN) methods. The temporal characteristics of sequence modeling may inherently worsen popularity bias instead of alleviating it.

Data Sparsity for Niche Items: The quality of recommendations is fundamentally contingent upon the adequacy of training data. Niche items, defined as those in the lower 50% of the popularity distribution, receive significantly fewer ratings compared to popular items. The lack of data results in inaccurately estimated embeddings or insufficiently confident predictions for niche content, prompting algorithms to favor safer, more widely accepted recommendations. As systems scale and catalogs expand, new items are introduced at the niche tier with limited interaction history, which may perpetuate and exacerbate the bias driven by sparsity. The coverage metrics support this mechanism, indicating that catalog growth exceeds recommendation diversification, resulting in a growing proportion of the catalog being under-represented in recommendations.

These mechanisms are not mutually exclusive and likely interact in a synergistic manner. Future research should utilize controlled experiments, counterfactual analyses, and causal inference methods to rigorously validate the proposed mechanisms, quantify their relative contributions, and guide the design of interventions aimed at disrupting feedback loops, reducing sequential amplification, and addressing data sparsity issues for niche content.

V. DISCUSSION

This research presents the initial comprehensive longitudinal evidence indicating that temporal fairness decay constitutes

Received: August 16, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

a significant challenge in recommendation systems. The observation that all three algorithms demonstrate statistically significant degradation ($p < 0.01$), despite their architectural diversity indicates that the decline arises from systemic characteristics of the recommendation paradigm rather than from weaknesses specific to their implementations. The magnitude is considerable: fairness gaps rose by 40% to 106% over 27 years, with effect sizes ($d = 1.68$ to 5.29) significantly surpassing thresholds for practical significance. The 47-fold difference in stability scores between NCF (0.803) and SASRec (0.017) illustrates the significant impact of algorithm selection on long-term fairness trajectories. This indicates that temporal stability should be prioritized as a primary criterion in algorithm evaluation, alongside accuracy.

The practical implications are direct and can be implemented promptly. Organizations utilizing NCF-based systems encounter significant threshold violations within 1.5 years, necessitating prompt intervention. In contrast, LightGCN implementations enjoy an 8.7-year intervention window, facilitating systematic mitigation strategies. The results of predictive modeling shift fairness management from a reactive crisis response to a proactive prevention approach. Random Forest classifiers demonstrate an F1 score of 0.80 with complete recall at one to two-year horizons, facilitating early warning systems that notify practitioners prior to the occurrence of violations. The five most significant features—Gini coefficient, aggregate diversity, niche share, current gap, and coverage—establish a fundamental monitoring framework, while the computational efficiency of the model training (under 0.3 seconds total) removes deployment obstacles.

Multiple limitations restrict generalizability. The exclusive emphasis on movie recommendations does not validate transferability to other areas such as e-commerce, news, and social media. Limited test set sizes (3 to 4 samples per horizon) restrict statistical power, while linear extrapolation presumes the continuation of trajectories in the absence of intervention or system alterations. The definition of groups based on popularity, although addressing concerns regarding provider fairness, may not correspond with legally protected categories in regulated environments. The perfect prediction scores for SASRec indicate trivial classification, as all test samples surpass the thresholds, rather than demonstrating authentic model discrimination.

Future research should focus on four key areas: the causal validation of decay mechanisms via controlled experiments to facilitate targeted interventions; the creation of temporally stable algorithms with stability as an explicit design objective; cross-domain studies to define generalizability boundaries; and the evaluation of intervention effectiveness through comparisons of mitigation strategies. The finding that fairness is not inherently maintained calls into question regulatory strategies reliant on singular certification, indicating that ongoing monitoring and temporal fairness criteria may be essential for ensuring enduring equity in algorithmic systems.

VI. Conclusion

This research offers a thorough longitudinal examination of temporal fairness decay in recommendation systems, revealing through 27 years of MovieLens data that the degradation of fairness is systematic, quantifiable, and predictable. The three algorithms assessed (NCF, LightGCN, SASRec) demonstrate statistically significant decay ($p < 0.01$), with fairness gaps increasing between 40% and 106%, and effect sizes varying from $d = 1.68$ to 5.29. Annual decay rates ranging from 0.0097 to 0.0166 gap points per year indicate critical threshold violations occurring within 1.5 years for NCF, 8.7 years for LightGCN, and have already been surpassed for SASRec. The stability analysis indicates a 47-fold disparity between the most stable algorithm (NCF, score 0.803) and the least stable algorithm (SASRec, score 0.017), demonstrating that architectural decisions substantially affect long-term fairness trajectories.

The predictive modeling framework attains an F1 score of 0.80 with complete recall at one to two year prediction horizons, facilitating proactive fairness management via early warning systems. Five key indicators (Gini coefficient, aggregate diversity, niche share, current fairness gap, and catalog coverage) offer a robust basis for accurate violation forecasting, while computational efficiency (total training time under 0.3 seconds) removes obstacles to deployment. The findings indicate that the accumulation of temporal bias, while unavoidable within existing algorithmic frameworks, can be predicted and mitigated prior to the attainment of critical thresholds.

The findings have important implications for the deployment and governance of recommendation systems. Practitioners must integrate assessments of temporal stability into the selection of algorithms, ensure ongoing monitoring of early warning indicators, and develop intervention protocols activated by prediction thresholds. The evidence that initial fairness does not ensure continued fairness poses a challenge to one-time certification methods, indicating that regulatory frameworks should adapt to mandate continuous temporal fairness evaluation. Future research must validate these findings in various domains, create algorithms specifically designed for temporal stability, and assess the effectiveness of interventions to finalize the proactive fairness management framework proposed in this study.

References

[1] F. M. Harper and J. A. Konstan, "The MovieLens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, pp. 19:1–19:19, Dec. 2015.

[2] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera, "All the cool kids, how do they fit in? Popularity and demographic biases in recommender evaluation and effectiveness," in *Proc. Conf. Fairness, Accountability and Transparency*, 2018, pp. 172–186.

[3] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, and F. Diaz, "Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness and satisfaction in recommendation systems," in *Proc. 26th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2017, pp. 2243–2251.

Received: August 16, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[4] H. Abdollahpouri, R. Burke, and B. Mobasher, "Managing popularity bias in recommender systems with personalized re-ranking," in *Proc. 32nd Int. Florida Artif. Intell. Res. Soc. Conf.*, May 2019, pp. 413–418.

[5] R. Burke, "Multisided fairness for recommendation," in *Proc. Workshop Fairness, Accountability, Transparency in Machine Learning*, 2017, pp. 1–5.

[6] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 2921–2930.

[7] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, and R. Burke, "Feedback loop and bias amplification in recommender systems," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 2145–2148.

[8] R. Jiang, S. Chiappa, T. Lattimore, A. György, and P. Kohli, "Degenerate feedback loops in recommender systems," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jan. 2019, pp. 383–390.

[9] A. D'Amour, H. Srinivasan, J. Atwood, P. Baljekar, D. Sculley, and Y. Halpern, "Fairness is not static: Deeper understanding of long term fairness via simulation studies," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 525–534.

[10] A. J. B. Chaney, B. M. Stewart, and B. E. Engelhardt, "How algorithmic confounding in recommendation systems increases homogeneity and decreases utility," in *Proc. 12th ACM Conf. Recommender Systems*, Sep. 2018, pp. 224–232.

[11] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 173–182.

[12] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Jul. 2020, pp. 639–648.

[13] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *Proc. IEEE Int. Conf. Data Mining*, Nov. 2018, pp. 197–206.

[14] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[15] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty in Artificial Intelligence*, Jun. 2009, pp. 452–461.

[16] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Introduction and challenges," in *Recommender Systems Handbook*, 2nd ed. Boston, MA, USA: Springer, 2015, pp. 1–34.

[17] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Jul. 2019, pp. 165–174.

[18] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: A survey," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 97:1–97:37, Dec. 2022.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[20] H. Abdollahpouri, M. Mansoury, R. Burke, and B. Mobasher, "The connection between popularity bias, calibration, and fairness in recommendation," in *Proc. 14th ACM Conf. Recommender Systems*, Sep. 2020, pp. 726–731.

[21] O. Celma and P. Cano, "From hits to niches? Or how popular artists can bias music recommendation and discovery," in *Proc. 2nd KDD Workshop Large-Scale Recommender Systems and the Netflix Prize Competition*, Aug. 2008, pp. 1–8.

[22] Y.-J. Park and A. Tuzhilin, "The long tail of recommender systems and how to leverage it," in *Proc. 2008 ACM Conf. Recommender Systems*, Oct. 2008, pp. 11–18.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[24] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[25] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 115:1–115:35, Jul. 2021.

[26] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2019.

[27] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, Jan. 2004.

[28] A. Gunawardana and G. Shani, "Evaluating recommender systems," in *Recommender Systems Handbook*, 2nd ed. Boston, MA, USA: Springer, 2015, pp. 265–308.

[29] W. X. Zhao, S. Mu, Y. Hou, Z. Lin, Y. Chen, X. Pan, K. Li, Y. Lu, H. Wang, C. Tian, Y. Min, Z. Feng, X. Fan, X. Chen, P. Wang, W. Ji, Y. Li, X. Wang, and J.-R. Wen, "RecBole: Towards a unified, comprehensive and efficient framework for recommendation algorithms," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2021, pp. 4653–4664.

Received: August 16, 2025