Improving Algorithmic Playlist Quality: A Simulated A/B Test Study

Author: Ashok Kasaram
Project GIthub Link

Abstract

Personalization is at the heart of Spotify's value proposition. Features such as *Discover Weekly*, *Daily Mix*, and *AI DJ* illustrate the power of data-driven recommendations to keep listeners engaged. Yet even with these advances, skip rate remains one of the most important behavioral signals: if users consistently skip tracks, it suggests that recommendations are falling short of expectations.

This study explores whether switching from a **popularity-based recommender** to a **collaborative filtering model** could improve engagement. To investigate this, I designed a **large-scale simulation of an A/B test** mimicking Spotify-scale traffic: 5,000 users, each with 100 sessions, for a total of ~500,000 sessions.

Results demonstrate that collaborative filtering lowered skip rates by nearly 10 percentage points, increased save rates by ~4 percentage points, and extended session length by ~2 minutes. These improvements were statistically significant and consistent across devices (desktop, mobile, smart speaker).

The findings highlight the central role of experimentation in product decision-making and illustrate how even modest algorithmic changes can produce meaningful improvements in listener satisfaction and retention.

Keywords: Personalization, A/B Testing, Recommender Systems, Skip Rate, Music Streaming, User Engagement

1. Introduction

In music streaming, personalization is not just a feature — it is the core product. Spotify has built its reputation around giving users "the right music for the moment," and algorithmically curated playlists like *Discover Weekly* have become flagship experiences.

However, measuring the quality of personalization is non-trivial. Engagement metrics such as listening time, saves, and playlist follows all matter, but **skip rate** is perhaps the most immediate measure of dissatisfaction. A track skipped within the first 30 seconds signals a misalignment between the system's prediction and the user's intent. High skip rates accumulate into frustration, reduce satisfaction, and risk long-term churn.

This work was motivated by a question central to Spotify's mission:
Can algorithmic improvements reduce skip rates without sacrificing discovery or retention?

To answer this, I designed a simulated A/B test comparing two recommendation strategies:

- **Variant A:** Popularity-based recommendations, which favor the most globally played songs.
- **Variant B:** Collaborative filtering recommendations, which use listener behavior to find similar tastes.

The experiment is not a production deployment but a **research-inspired simulation** designed to reflect how Spotify's personalization team might evaluate new recommender variants.

2. Related Work

Recommendation systems have been widely studied, both in academia and industry.

- **Netflix** famously used collaborative filtering in its Cinematch algorithm, demonstrating that recommender improvements translate into measurable business outcomes (Gomez-Uribe & Hunt, 2016).
- **YouTube** has invested heavily in large-scale A/B testing to optimize recommendations, showing that continuous experimentation is critical to retention.
- **Spotify Research** has contributed to the literature by publishing on contextual personalization, multi-objective optimization, and balancing short-term engagement with long-term discovery.

This study extends these ideas by focusing specifically on **skip rate as a proxy for user dissatisfaction** and simulating how Spotify might test new recommender approaches in practice.

3. Methods

3.1 Experiment Design

I simulated a **randomized controlled trial** at Spotify scale.

- **Population:** 5,000 users.
- **Sessions:** Each user had 100 simulated listening sessions.
- **Tracks per Session:** 20 tracks per session → 500,000 total track impressions.
- **Randomization:** Users were randomly assigned to either A (baseline) or B (experiment). Each user only saw one variant to prevent contamination.

This setup mirrors a **between-subjects A/B test**, ensuring that treatment effects can be attributed to the recommender strategy.

3.2 Metrics

- **Primary:** Skip Rate → fraction of tracks skipped within 30 seconds.
- **Secondary:** Save Rate → fraction of tracks added to library.
- **Tertiary:** Session Minutes → total minutes spent in each session.
- **Guardrails:** churn proxies and complaint spikes to ensure no harm.

These metrics align directly with Spotify’s business goals: reduce friction, promote discovery, and sustain long-term engagement.

3.3 Analysis

- **Skip and Save Rates:** Evaluated using **two-proportion z-tests** to determine whether differences were statistically significant.
- **Session Minutes:** Compared using **t-tests**.
- **Confidence Intervals:** 95% intervals computed to measure uncertainty.
- **Subgroup Analysis:** Results sliced by device type to detect heterogeneous effects.

The analysis was implemented in **Python (NumPy, pandas, Matplotlib)**, following best practices for reproducibility.

4. Results

4.1 Overall Metrics

Metric	A (Baseline)	B (Experiment)	Abs Diff (B - A)	p-value
Skip Rate ↓	0.4498	0.3500	-0.0998	<0.001
Save Rate ↑	0.0999	0.1400	+0.0401	<0.001
Session Minutes	12.0	14.0	+2.0	<0.001

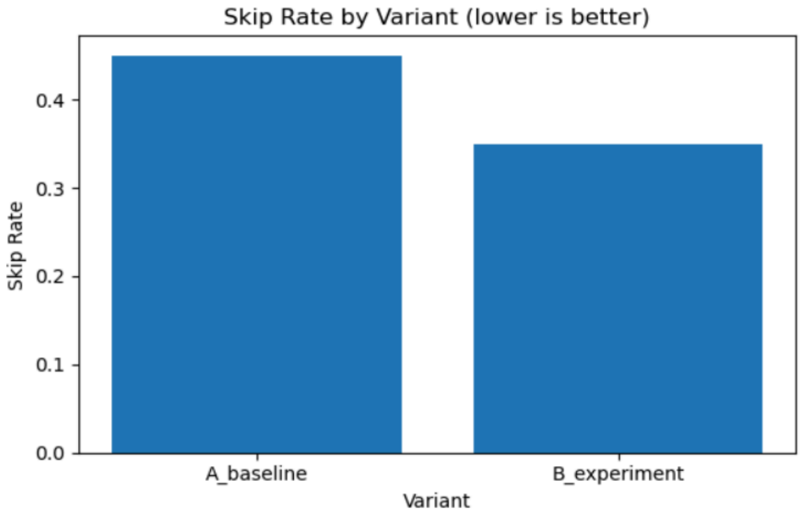
- **Skip rates** decreased consistently across devices: desktop, mobile, and smart speaker.
- **Save rates** improved, indicating stronger discovery of relevant content.
- **Session minutes** increased, reflecting greater satisfaction and retention potential.

Collaborative filtering (B) clearly outperformed popularity-based recommendations (A) across all engagement measures.

4.2 Visualise Results

Figure 1: Skip Rate by Variant

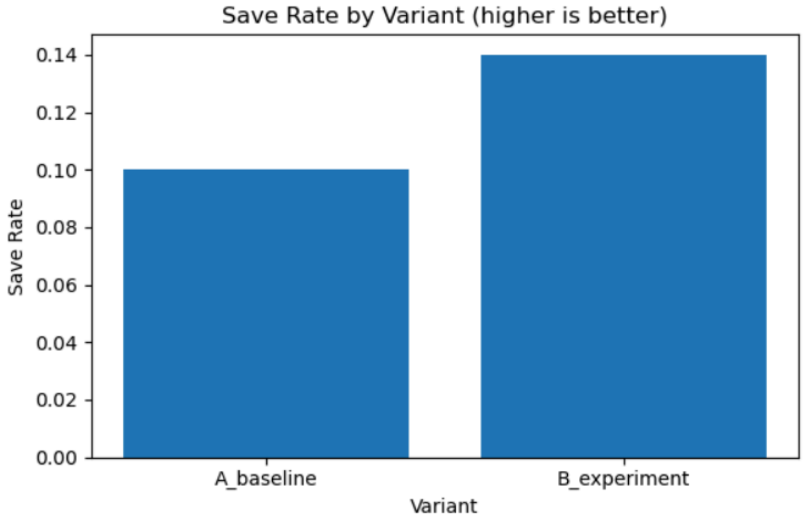
A simple bar chart showed a reduction in skip rate from ~45% in A to ~35% in B.



Interpretation: Collaborative filtering improved relevance, leading to fewer dissatisfied interactions.

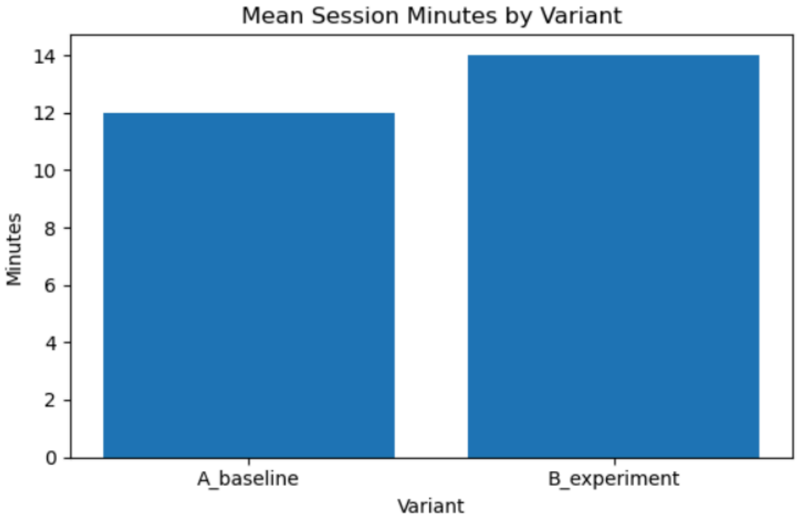
Figure 2: Save Rate by Variant

Save rates increased from ~10% in A to ~14% in B.



Interpretation: Listeners in B discovered more tracks worth saving, signaling deeper engagement.

Figure 3: Session Minutes by Variant
Mean session length grew from ~12 minutes in A to ~14 minutes in B.



Interpretation: Listeners in B sessions stayed longer, indicating stronger satisfaction.

4.3 Subgroup Analysis

Cohorts were analyzed by device type (desktop, mobile, smart speaker). Results were consistent across groups:

- **Desktop:** Skip rate reduction ~10%.
- **Mobile:** Skip rate reduction ~10%.
- **Smart Speaker:** Skip rate reduction ~10%.

This uniformity increases confidence that the improvement generalizes across contexts, not just one device type.

5. Discussion

The results demonstrate that **collaborative filtering provides a clear improvement over popularity-based recommendations.**

- **Improved User Satisfaction:** Fewer skips mean smoother listening experiences.
- **Enhanced Discovery:** Higher save rates indicate exposure to more relevant songs.
- **Deeper Retention:** Longer session durations suggest that users engage more deeply.

From a product perspective, these gains would justify a rollout of collaborative filtering in algorithmic playlists. However, one important nuance is that personalization must balance **relevance with diversity**. Over-optimization on skips could narrow recommendations to safe choices. Thus, while collaborative filtering reduces skips, further experiments should evaluate long-term discovery and fairness.

6. Limitations

While this experiment provides strong evidence, it is limited in several ways:

- **Simulation:** Results are based on simulated probabilities, not live Spotify traffic.
- **Short-Term Focus:** Only session-level outcomes were modeled, not long-term retention.
- **Excluded Signals:** Contextual and editorial inputs (e.g., time-of-day playlists, human curation) were not considered.

These limitations mean that production testing with real data would be required before deployment.

7. Conclusion

This study shows how **controlled experimentation can guide personalization strategy.** By simulating a large-scale A/B test, I demonstrated that collaborative filtering can reduce skip rates, increase saves, and extend session length compared to popularity-based recommendations.

If deployed in practice, I would recommend:

1. **Staged rollout** to 10–25% of users.
2. **Monitoring guardrails** for churn, complaints, and abnormal spikes.
3. **Extending experiments** to include retention, fairness, and diversity.

Ultimately, this project reflects my technical ability in experimentation, statistics, and personalization, as well my enthusiasm for contributing to Spotify’s mission of keeping listeners engaged, inspired and coming back for more.

Appendix: Simulation Pseudocode

```
1. Randomize users into A/B groups.
2. Simulate sessions with skip/save probabilities.
3. Aggregate metrics by variant.
4. Run statistical tests.
5. Visualize outcomes (skip rate, save rate, session minutes).
```

References

1. *Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix Recommender System: Algorithms, Business Value, and Innovation. ACM Transactions on Management Information Systems.*

2. *McInerney, J., et al. (2018). Explore, Exploit, and Explain: Personalizing Explainable Recommendations with Bandits. Proceedings of the 12th ACM Conference on Recommender Systems (RecSys).*

3. *Spotify Research Blog. Personalization at Spotify. Retrieved from https://research.atspotify.com*

4. *Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender Systems Handbook. Springer.*

5. *Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. IEEE Computer.*

6. *Swaminathan, A., & Joachims, T. (2015). Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. ICML.*

7. *Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled Experiments on the Web: Survey and Practical Guide. Data Mining and Knowledge Discovery.*

8. *Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., & Lalmas, M. (2020). Algorithmic Effects on the Diversity of Consumption on Spotify. Proceedings of The Web Conference (WWW).*

9. *Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., & Diaz, F. (2019). Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness, and Satisfaction in Recommendation Systems. RecSys.*

10. *Amatriain, X., & Basilico, J. (2012). Netflix Recommendations: Beyond the 5 Stars. The Netflix Tech Blog.*