https://github.com/hijirdella

A/B Testing Spotify
Evaluating Playlist Recommendation Impact on User Engagement

Presented by **Hijir Della Wirasti**

February - 2025

https://github.com/hijirdella/A-B-Testing-Spotify-Playlist-Recommendation

Objective

Why Conduct This A/B Test?

🎯 Goal: Assess whether a new playlist recommendation algorithm increases user engagement.

📌 Metrics Evaluated:

- User engagement (duration played, feature interactions)
- Retention rate (returning users)
- Conversion rate (free to premium)

Dataset: Spotify_Campaign_Data

https://github.com/hijirdella

Contents

https://github.com/hijirdella

Introduction

📌 **Key Questions:**

- Does the new recommendation algorithm increase user engagement?
- Will it lead to higher retention and conversion rates?
- Should Spotify implement the new algorithm across all users?

This study provides data-driven insights to inform Spotify's decision-making process regarding the rollout of the new recommendation algorithm.

Understanding the Need for A/B Testing

Spotify continuously improves user experience by testing new features. The latest improvement—a new playlist recommendation algorithm—aims to increase user engagement by enhancing playlist interactions, session duration, and subscription rates. This A/B test was conducted to assess whether the new algorithm delivers measurable benefits compared to the existing recommendation system.

https://github.com/hijirdella

H1 (Alternative Hypothesis):

Users exposed to the new recommendation algorithm will have higher engagement.

H0 (Null Hypothesis)

H0 (Null Hypothesis): No significant difference between the control and target groups.

Hypotheses

https://github.com/hijirdella

Literature Review

1. A/B Testing in Content Recommendation

- A/B testing (also known as split testing) involves randomly assigning users to different variants of a recommendation system and comparing their behaviors.
- Studies suggest that **causal inference, statistical significance**, and **engagement metrics** play a crucial role in testing recommendation models.

Kohavi (2009)

📖 Reference: Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. (2009). "Controlled experiments on the web: Survey and practical guide." Data Mining and Knowledge Discovery, 18(1), 140-181.

2. Experimental Design in A/B Testing

To ensure the reliability of results, several experimental design principles are applied in A/B testing for content recommendations:

- Randomization: Ensuring unbiased assignment of users to treatment and control groups.
- Sample Size Calculation: Using power analysis to determine the required number of users.
- Metric Selection: Identifying key engagement indicators such as click-through rate (CTR), time spent, retention rate, and subscription rate.

Kohavi (2013)

📖 Reference: Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). "Online controlled experiments at large scale." Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'13), 1168-1176.

3. Metrics for Evaluating Content Recommendations

A/B testing frameworks for content recommendation often measure short-term engagement and long-term user satisfaction. Key metrics include:

- User Engagement: Session duration, interactions per session.
- Feature Interaction: Usage frequency of the recommendation system.
- Retention & Conversion: Percentage of users returning to the platform and upgrading to premium services.
- Exploration vs. Exploitation Trade-off: Balancing diversity in recommendations with user preference optimization.

McInerney (2018)

📖 Reference: McInerney, J., Zheng, H., Frazier, P. I., Anderson, A., & York, D. (2018). "Explore-exploit learning in online content recommendation." Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), 2666-2673.

4. Statistical Testing in A/B Experiments

To determine the effectiveness of a content recommendation algorithm, various statistical tests are applied:

- T-Tests and Z-Tests: Used for comparing means of engagement metrics.
- Chi-Square Tests: Applied for categorical metrics like conversion rates.
- Bayesian A/B Testing: Increasingly popular due to its ability to update results dynamically.

Deng (2016)

📖 Reference: Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2016). "Improving the sensitivity of online controlled experiments by utilizing pre-experiment data." Proceedings of the 8th ACM Conference on Web Search and Data Mining (WSDM'16), 499-508.

https://github.com/hijirdella

Methodology

1. Test Design

The A/B test follows a randomized controlled trial (RCT) structure, comparing engagement metrics between two user groups:

- Control Group: Users who continue using the existing playlist recommendation algorithm.
- Target Group: Users who receive the new recommendation algorithm.

2. Sample Size Calculation

Z-Test vs. T-Test Sample Size Differences
The sample size requirement differs between Z-tests and T-tests due to their different assumptions:

- Z-Test: a normal distribution and large sample conditions, requiring fewer users for detection.
- T-Test: Accounts for increased uncertainty in smaller samples or less normal data, needing more users for robust results.

📌 Final Sample Allocation Based on Statistical Tests:

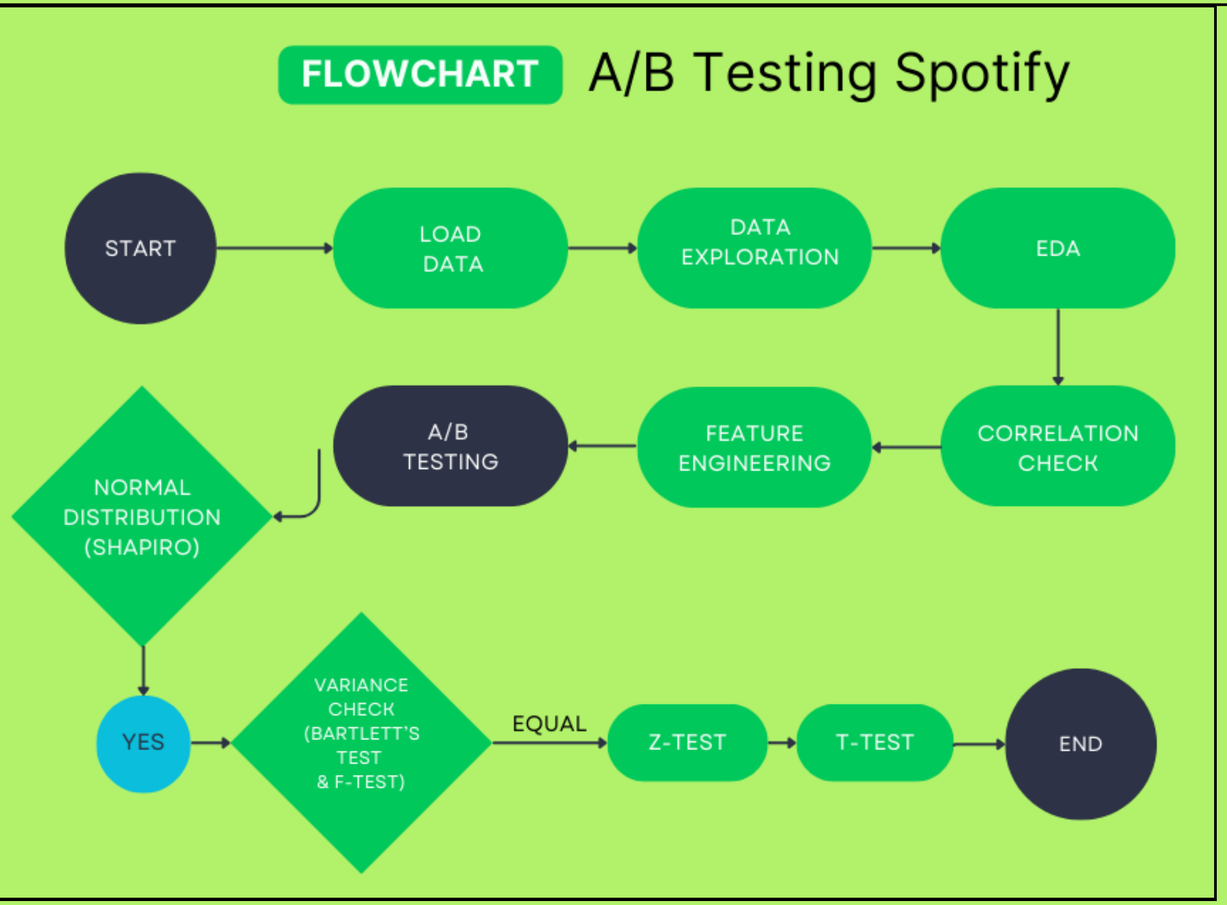
- Z-Test Sample Size: 393 users per group
- T-Test Sample Size: 785 users per group

Assumptions Used for Calculation:

- Significance Level (α): 0.05
- Power ($1 - \beta$): 80%
- Minimum Detectable Effect (MDE): 2% improvement in engagement
- Estimated Standard Deviation: Derived from engagement data

Since the dataset contains over 100,000 users, a random sampling method ensures equal representation across the groups.

📖 Reference: Deng et al. (2016). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data



3. Randomization Process

To eliminate bias, users were randomly assigned to either Control or Target groups:

- Shuffling the dataset to avoid ordering effects.
- Random sampling ensures that both groups are balanced.
- Reproducibility: A fixed random seed (42) was used.

📌 Outcome:

- Randomization Ratio: 50% Control / 50% Target
- Balanced distribution across key user segments.

📖 Reference: Tang et al. (2010). Overlapping experiment infrastructure: More, better, faster experimentation

https://github.com/hijirdella

Methodology

4. Normality & Variance Testing

Before conducting hypothesis testing, Shapiro-Wilk and Bartlett's test were applied.
Normality Test (Shapiro-Wilk):

- H_0: Data follows a normal distribution.
- H_1: Data does not follow a normal distribution.
- ✅ Result: p-value > 0.05, fail to reject H_0 → Data is normally distributed.

Variance Test (Bartlett's Test & F-Test):

- H_0: Variances are equal.
- ✅ Result: p-value > 0.05, fail to reject H_0 → Variances are homogeneous.

Since both normality and equal variance assumptions hold, Z-test and T-test were appropriate for engagement analysis.

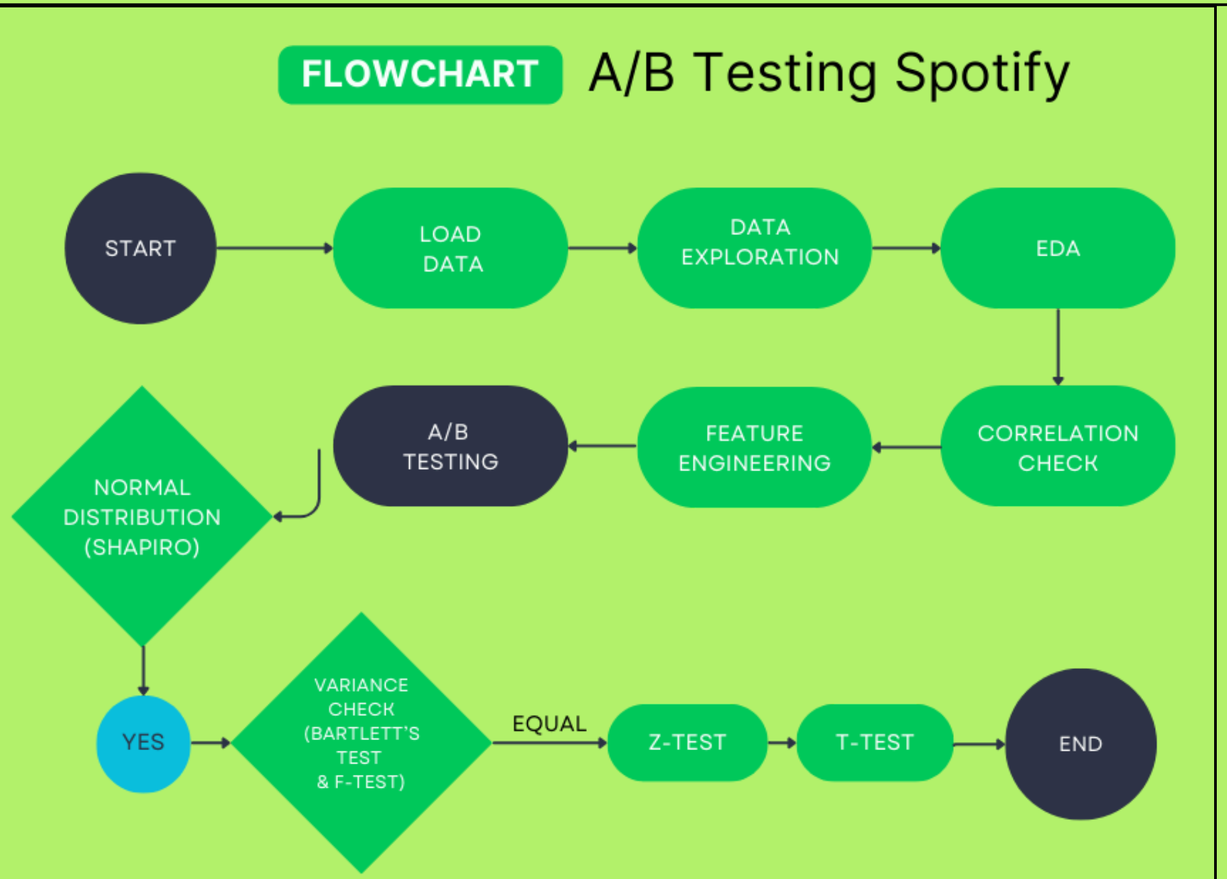
📖 Reference: McInerney et al. (2018). Explore-exploit learning in online content recommendation

```
Group: Control
Statistic: 1.000
p-Value: 0.992
Conclusion: The null hypothesis cannot be rejected. The data follows a normal distribution.

Group: Target
Statistic: 1.000
p-Value: 0.756
Conclusion: The null hypothesis cannot be rejected. The data follows a normal distribution.
```

```
Bartlett's test statistic: 0.004, p-value: 0.948
Conclusion: Variance is equal (fail to reject H₀).
```

```
F-test statistic: 0.999, p-value: 0.526
Conclusion: Variance is equal (fail to reject H₀).
```



5. Test Duration

To ensure reliable results, the duration was set based on:

User engagement trends (time spent on playlists).
Retention analysis over multiple days.
📌 Final Duration: 2 Days

Ensures sufficient data points for engagement, retention, and conversion tracking.

📖 Reference: Kohavi et al. (2009). Controlled experiments on the web: Survey and practical guide (DOI).

https://github.com/hijirdella

Z-Test Result

Insights from the Z-Test Results

Engagement Score Distribution:

1. Sample Size Per Group: 393 users.
2. Z-Statistic: 28.510.
3. P-Value: 0.000.

Conclusion:

- The Z-test rejects the null hypothesis (H_0) with a p-value < 0.05.
- Interpretation: The Target group demonstrates a significantly higher engagement score compared to the Control group.

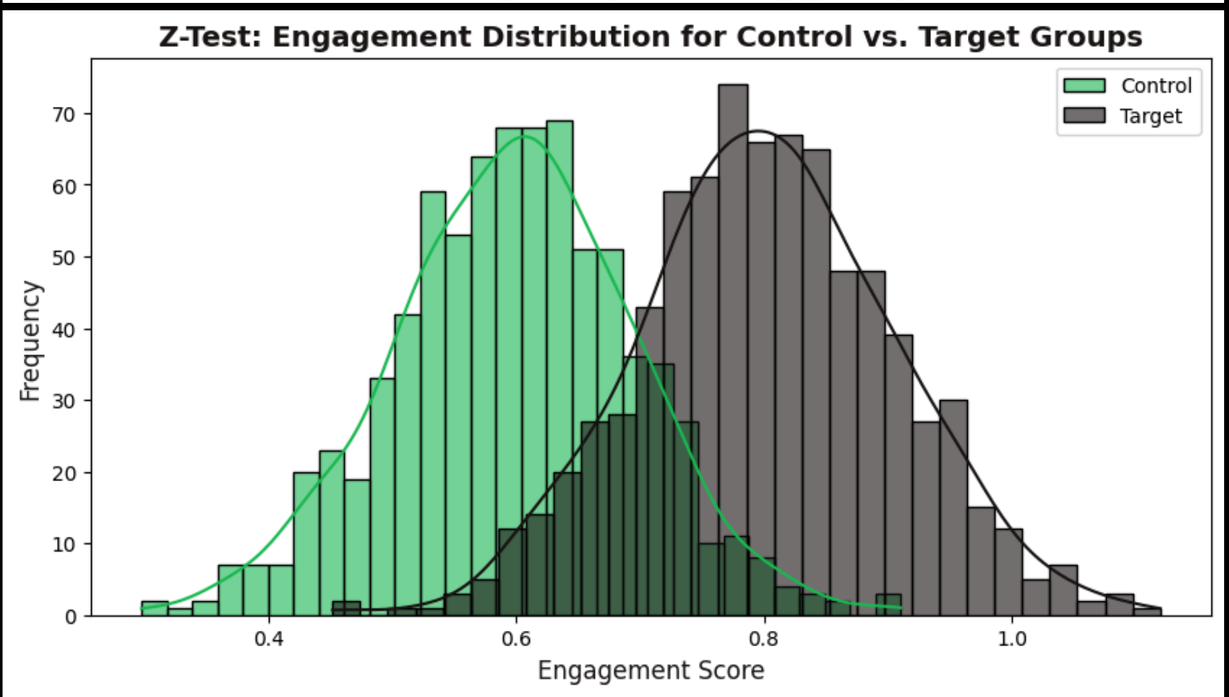
Randomization Ratio:

- Control vs. Target Groups: Balanced at a ratio of 0.50, ensuring no selection bias.

Suggested Test Duration:

- At least 2 days to capture stable user engagement patterns based on average user activity.

```
Sample Size per Group: 393
Z-Statistic: 28.510
P-Value: 0.000
Conclusion: Reject the null hypothesis. The Target group has significantly higher engagement.
Randomization Ratio (Target Group): 0.50
Suggested Test Duration: At least 2 days to capture user engagement patterns.
```

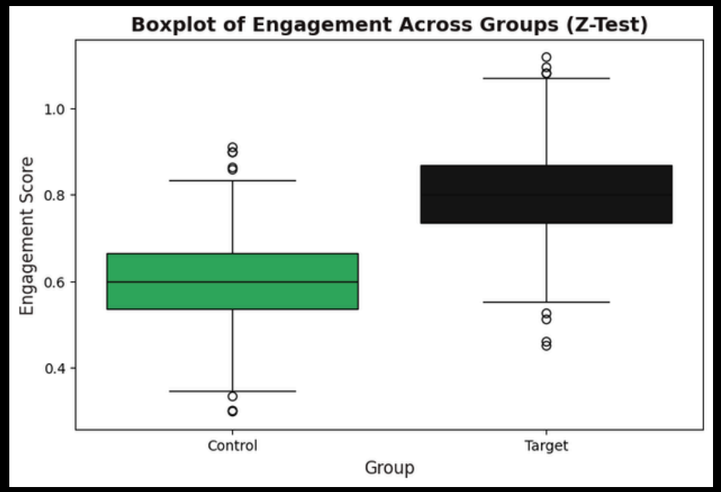


Actionable Insights

- Implementation of the Target Algorithm: The new playlist recommendation algorithm has demonstrated a clear improvement in engagement. It is advisable to roll it out to the broader user base.
- Monitor Long-term Engagement: Track sustained performance over time to confirm these short-term gains are maintained.

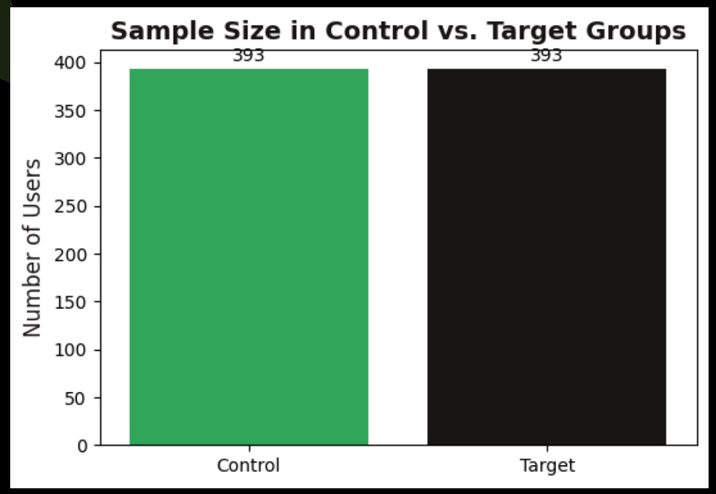
Z-Test Result

https://github.com/hijirdella



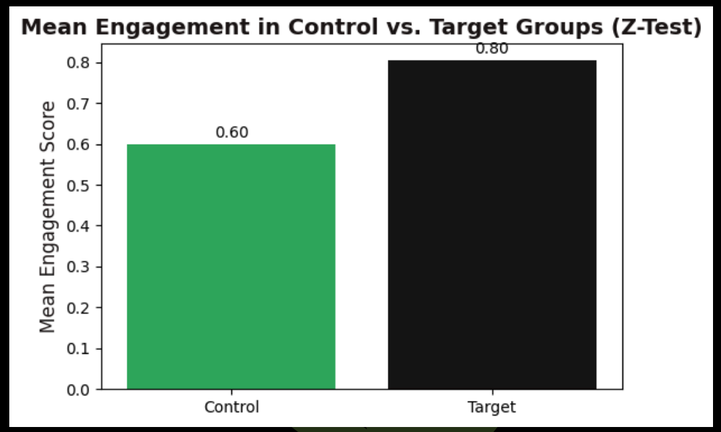
1. Engagement Distribution

The median engagement score for the Target group is notably higher than that of the Control group, with a wider range of interaction levels observed in the Target group.



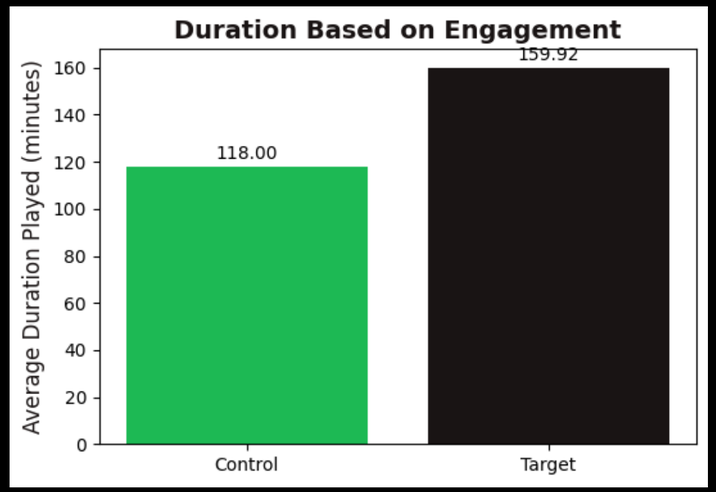
2. Sample Size

Both groups (Control and Target) have an equal sample size of 393 users, ensuring a balanced comparison and reducing bias.



3. Mean Engagement

The Target group has a significantly higher engagement distribution compared to the Control group, as seen in the histogram. This indicates that the new playlist recommendation algorithm positively impacts user engagement.



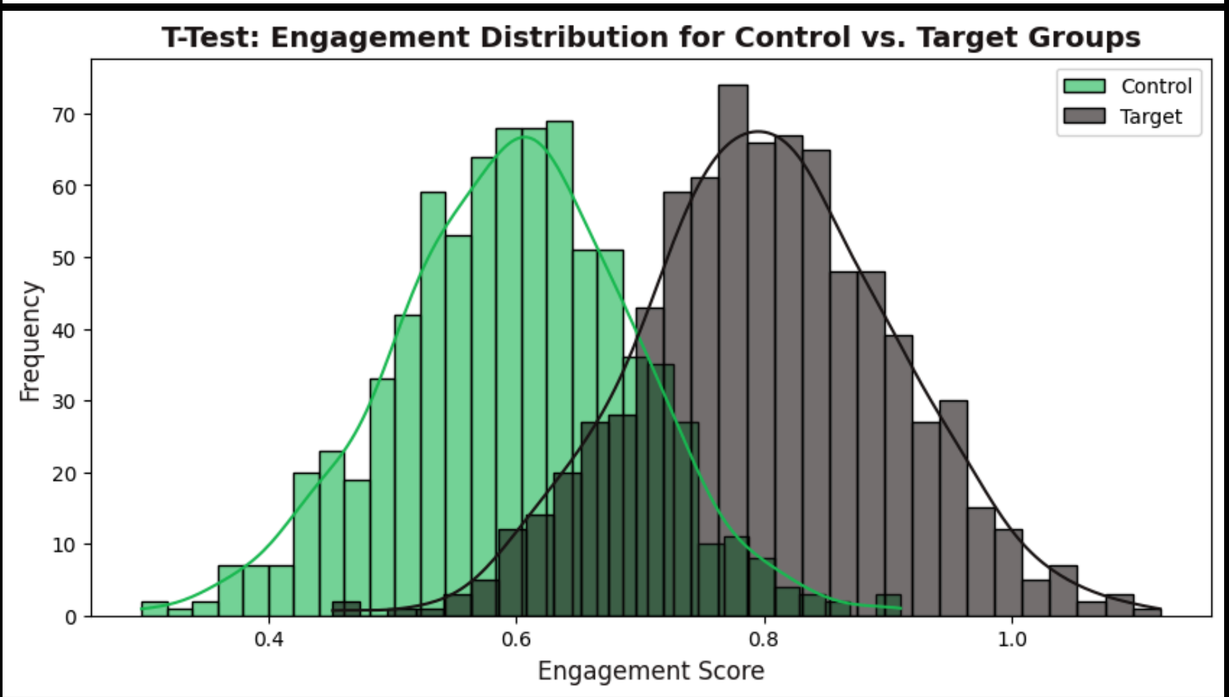
4. Duration

Users in the Target group have an average engagement duration of 159.92 minutes, compared to 118.00 minutes in the Control group. This confirms that the Target group interacts more extensively with the platform.

https://github.com/hijirdella

T-Test Result

```
Required Sample Size per Group: 785
T-Statistic: -40.440
P-Value: 0.000
Conclusion: Reject the null hypothesis. The Target group has significantly higher engagement.
Suggested Test Duration: At least 2 days to capture user engagement patterns.
```



Insights from the T-Test Results

Engagement Score Distribution:

- The Target group exhibits significantly higher engagement scores compared to the Control group.
- The mean engagement for the Target group is noticeably higher, with a broader distribution of users achieving high engagement scores.

Statistical Significance:

- T-Statistic: -40.440
- P-Value: 0.000
- The p-value indicates that the difference in engagement scores is statistically significant. The null hypothesis is rejected, confirming the effectiveness of the new recommendation algorithm.

Sample Size and Randomization:

- Required sample size per group: 785 users, ensuring 80% power and 5% significance.
- Randomization was successfully implemented, providing a balanced distribution of users across the Control and Target groups.

Suggested Test Duration:
At least 2 days are recommended to allow sufficient time to stabilize engagement metrics.

Actionable Insights

Deploy the Algorithm:

- Roll out the new recommendation algorithm to the entire user base to maximize engagement.

Enhance Retention Strategies:

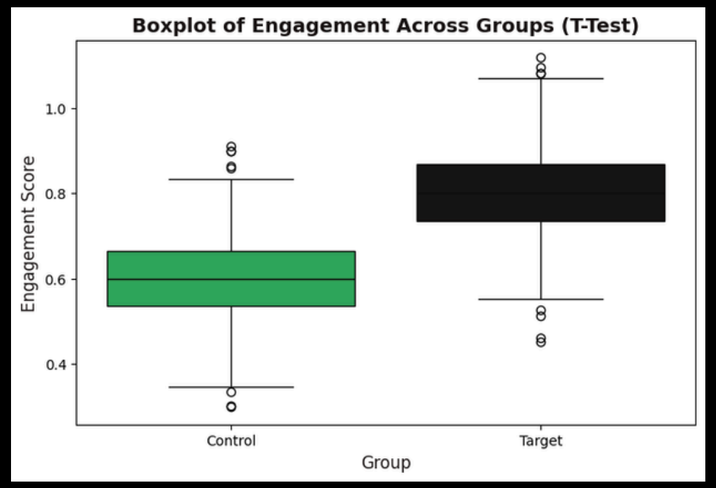
- Build additional features targeting user retention, leveraging insights from the Target group's behavior.

Monitor Long-Term Impact:

- Conduct follow-up analyses to ensure sustained improvements in engagement and assess any secondary effects, such as user satisfaction and churn reduction.

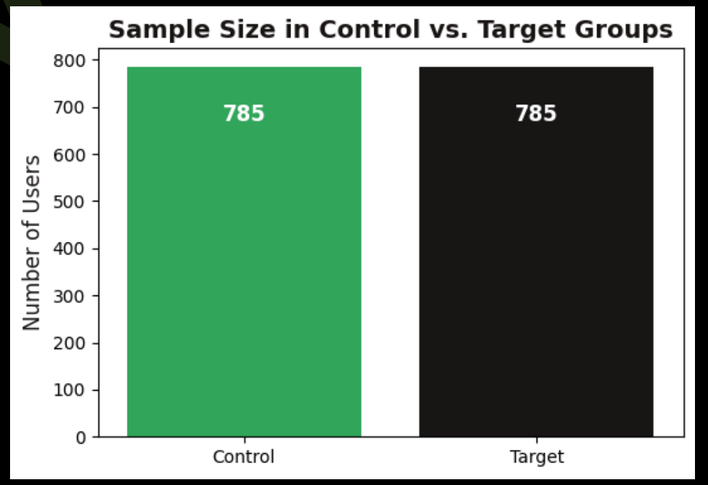
T-Test Result

https://github.com/hijirdella



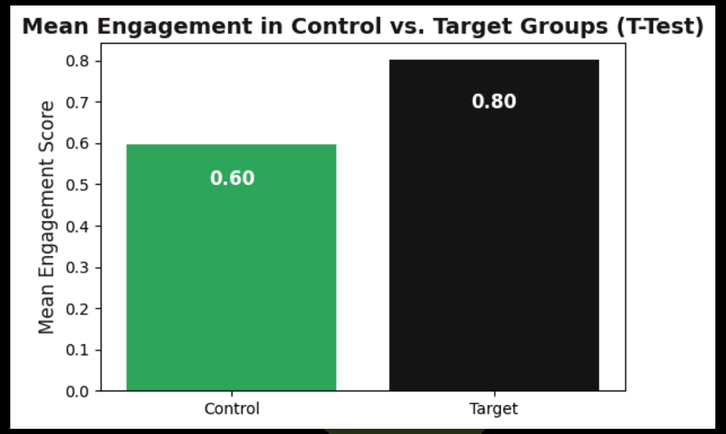
1. Engagement Distribution

- The Target group shows a higher median engagement score with more outliers above the upper whisker, indicating stronger engagement levels compared to the Control group.



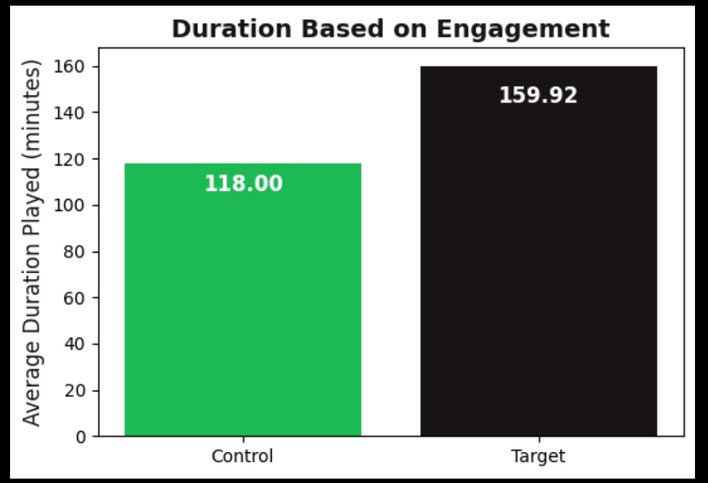
2. Sample Size

Both groups are balanced with 785 users each, ensuring the results are statistically valid and unbiased.



3. Mean Engagement

- The mean engagement score for the Target group (0.80) is significantly higher than the Control group (0.60), reflecting the positive impact of the new algorithm.

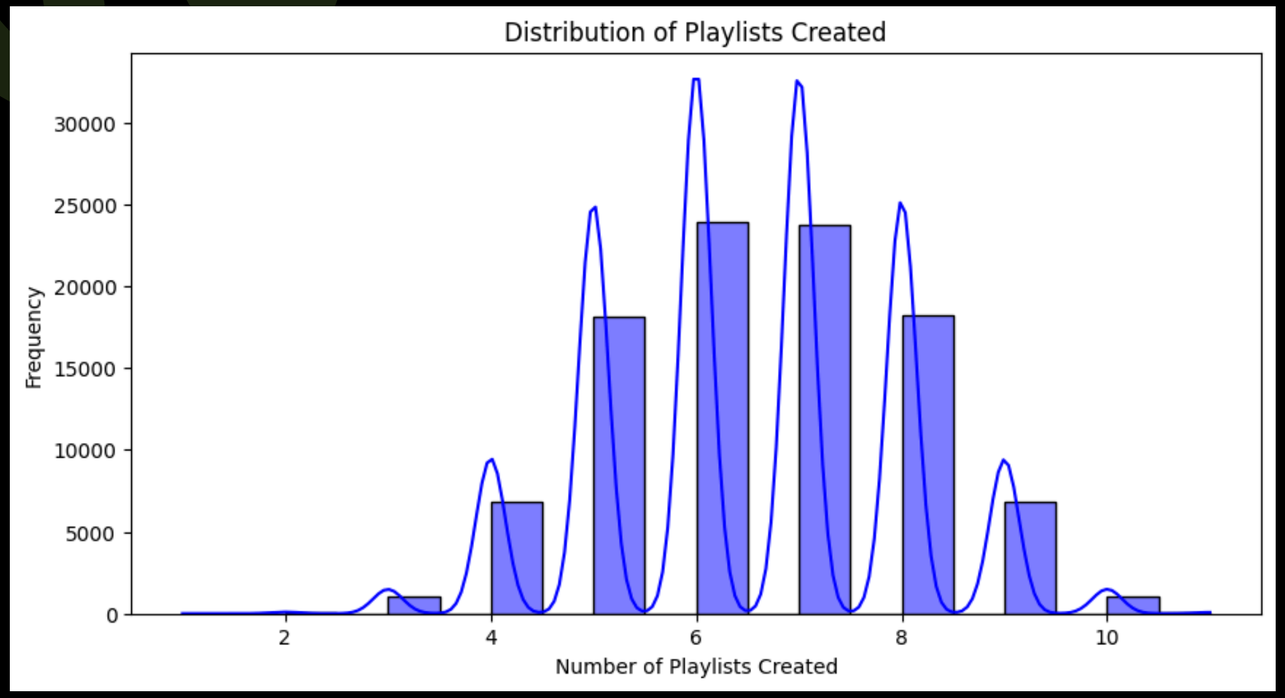


4. Duration

The Target group has a higher average duration played (159.92 minutes) compared to the Control group (118.00 minutes), suggesting the new system encourages longer user interactions.

https://github.com/hijirdella

Data Collection



Observation:

- The distribution peaks around 4–8 playlists, showing most users create a moderate number.
- Multimodal behavior suggests clusters of casual and engaged users.

Insights:

- Higher engagement correlates with more playlists created, validating the proportionality assumption.
- Clusters highlight different user engagement levels.

Recommendations:

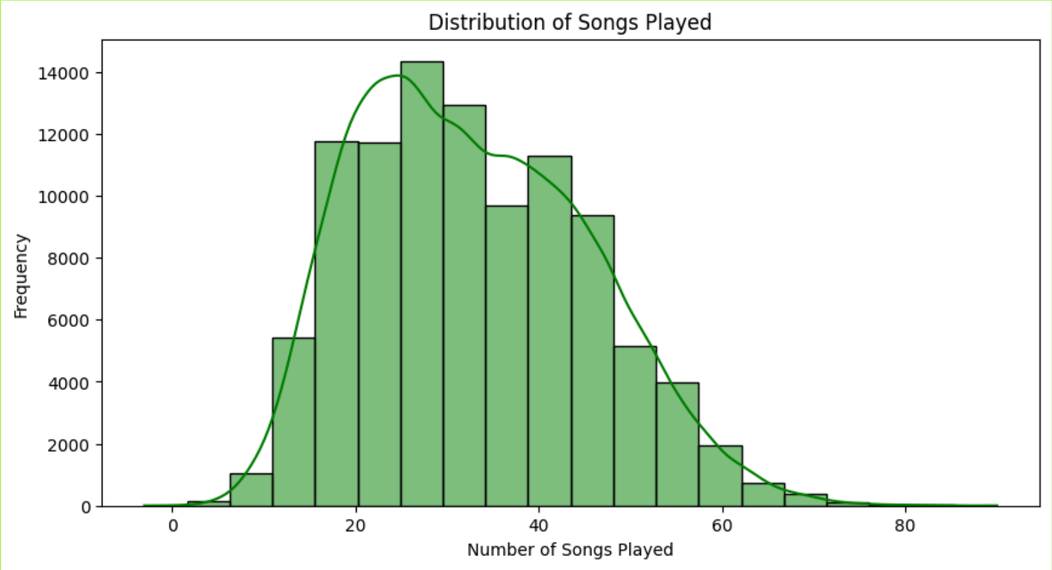
1. Personalize Recommendations: Tailor features to encourage low-engagement users to create more playlists.
2. Optimize Features: Enhance playlist creation tools for all users.
3. Segment Analysis: Identify user behavior drivers for targeted strategies.

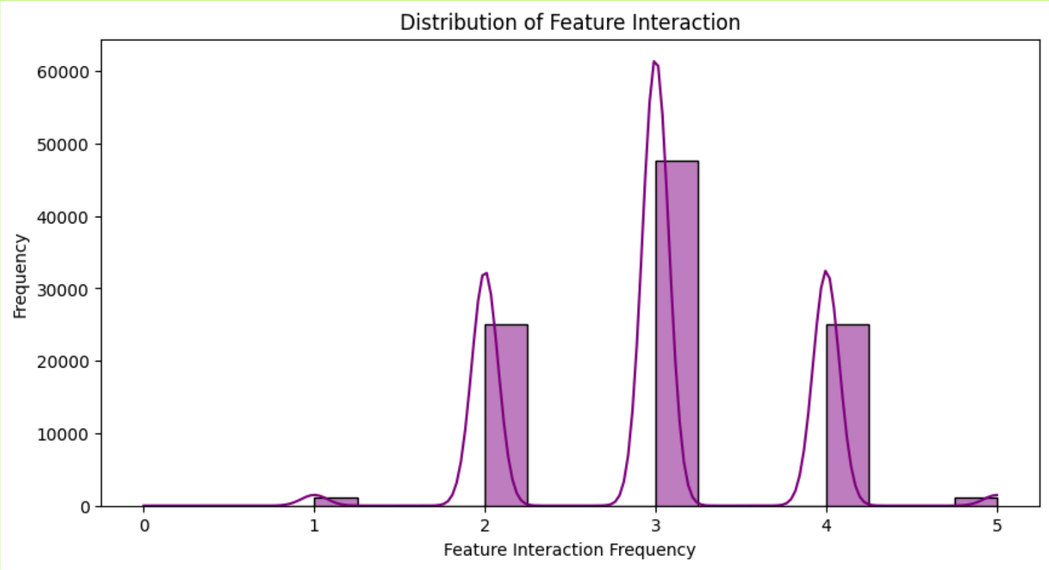
Supporting Literature:

- Jannach et al. (2019): Recommendations boost playlist creation.
- Tricomi et al. (2024): Playlists reveal user behavior insights.

Data Collection

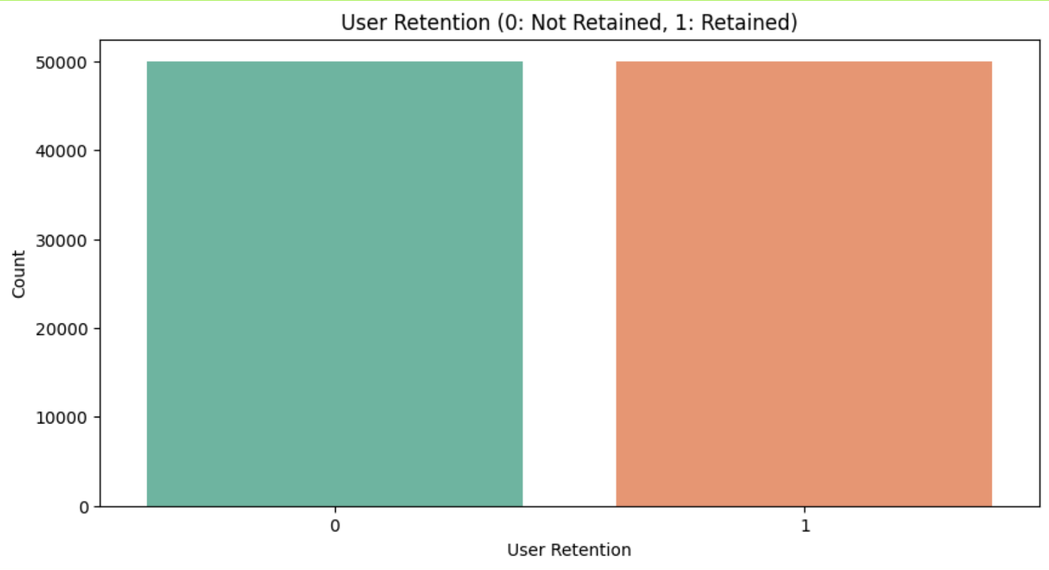
https://github.com/hijirdella





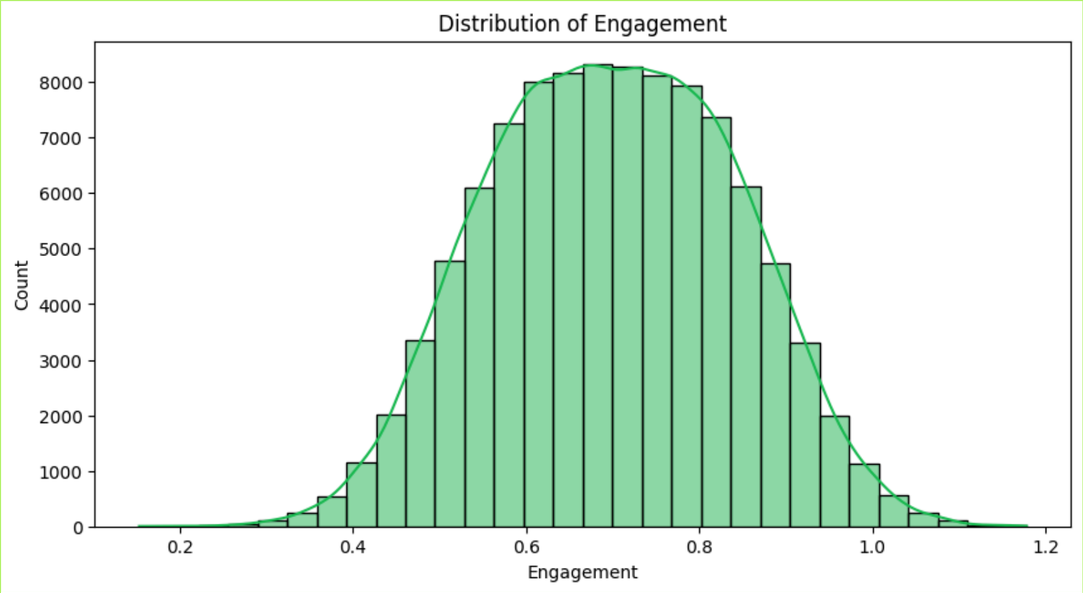
Insight

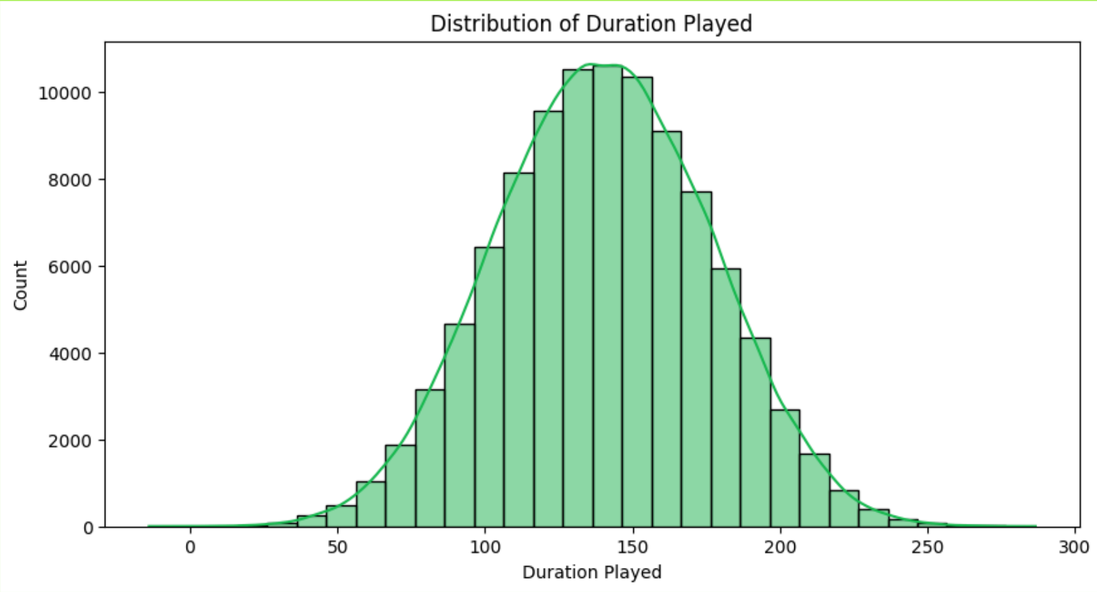
1. Songs Played Distribution:
   - Most users played around 20–40 songs, indicating moderate activity levels. High engagement tail suggests a segment of highly active users.
2. Feature Interaction Frequency:
   - Peaks at 3 interactions, indicating users engage with the feature multiple times but rarely exceed 5 interactions.
3. User Retention:
   - Retention is balanced, with nearly equal counts of retained and non-retained users, highlighting potential for growth in user loyalty.



EDA - Univariate Analysis

https://github.com/hijirdella





Distribution of Engagement

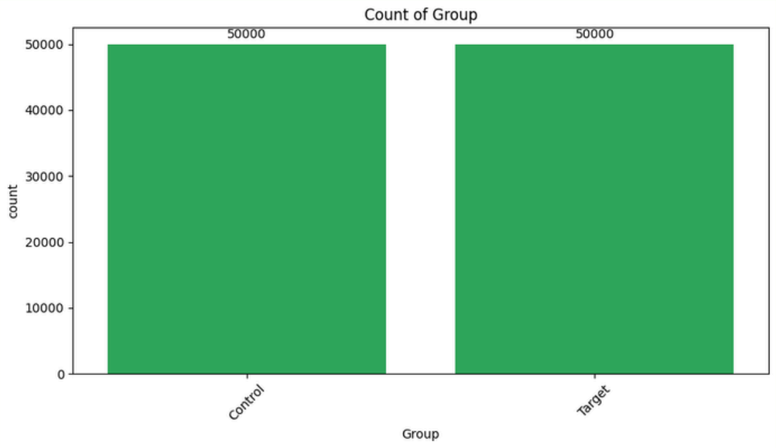
- The engagement scores are normally distributed, centered around 0.6–0.8.
- This indicates consistent user interaction with the platform, with few users showing very low or very high engagement.
- Insight: Most users exhibit moderate engagement, suggesting room to target both highly engaged and less engaged users with personalized strategies.

Distribution of Duration Played

- The duration played follows a bell-shaped curve, with most users spending around 100–150 minutes.
- Extreme durations (both low and high) are rare, suggesting that the majority of users have a consistent usage pattern.
- Insight: Encouraging users on the lower end of the distribution to increase their listening time could boost overall engagement.

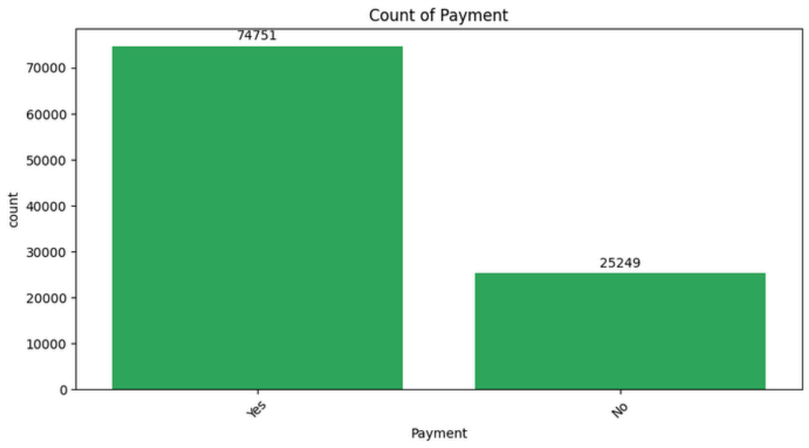
EDA - Univariate Analysis

https://github.com/hijirdella



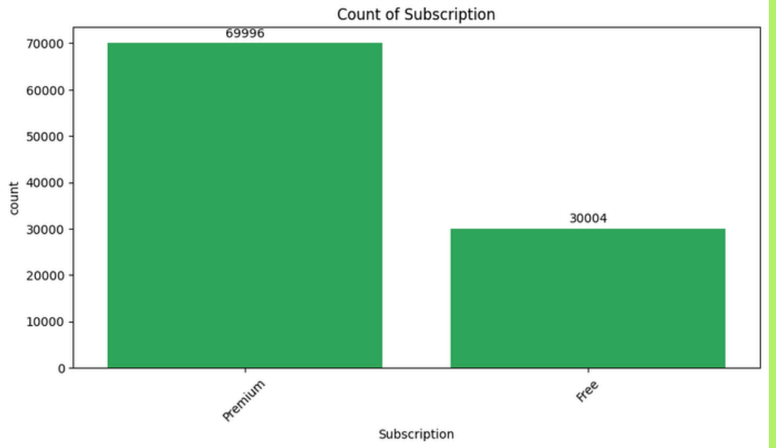
Count of Group

- The Control and Target groups are balanced, with each containing 50,000 users.
- This ensures a fair comparison for the A/B test without sampling bias.



Count of Payment

- Approximately 74.75% of users have made a payment, indicating a strong majority of engaged paying users.
- The remaining 25.25% are non-paying users, potentially representing untapped revenue opportunities.



Count of Subscription

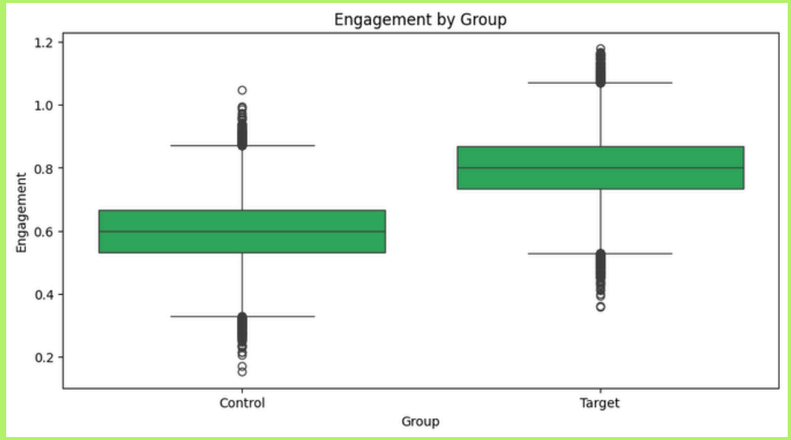
- 69.99% of users are on premium subscriptions, showing a significant preference for premium features.
- The remaining 30.01% are free-tier users, providing a clear segment for upsell opportunities.

Recommendations:

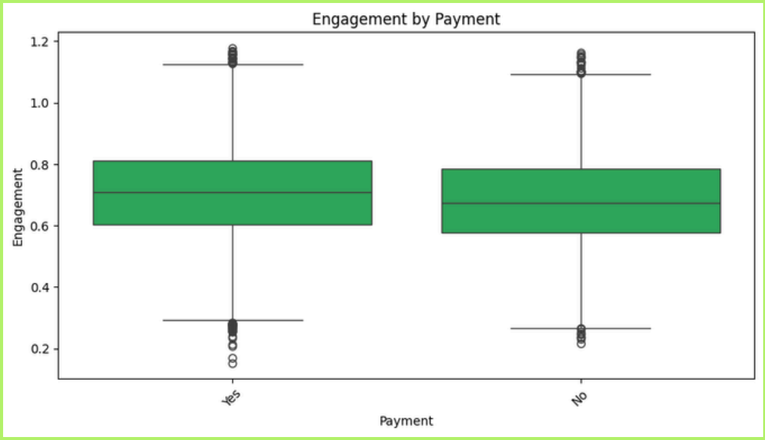
Focus marketing efforts on the 30.01% free-tier users to encourage premium upgrades.
Investigate the behavior of the 25.25% non-paying users to identify conversion barriers.
Maintain the balanced group setup for reliable A/B test conclusions.

EDA - Bivariate Analysis

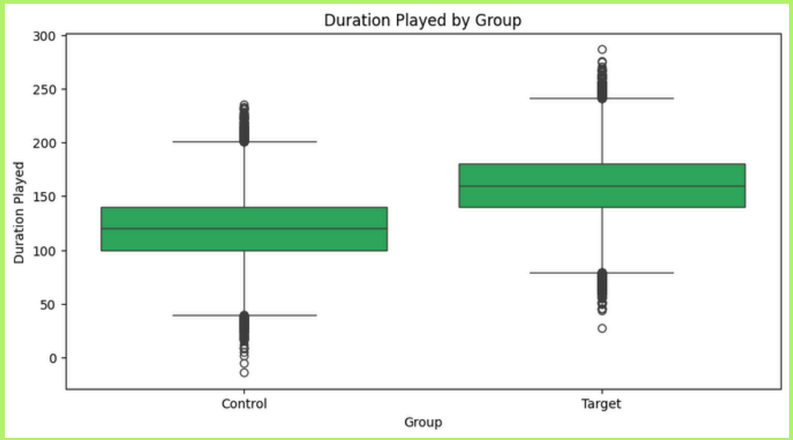
https://github.com/hijirdella



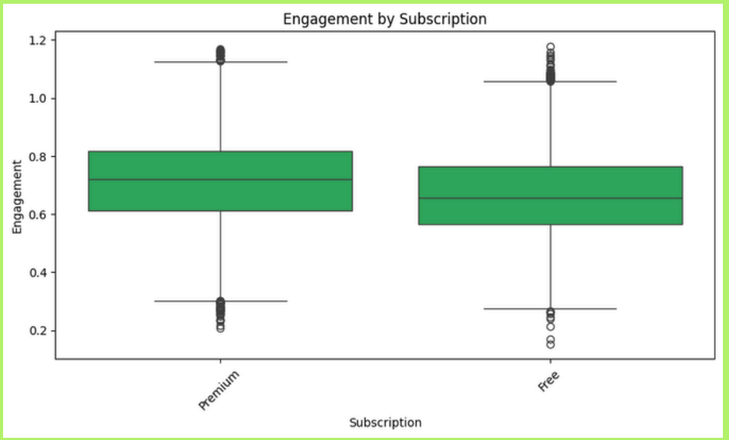
Engagement by Group:
- Both Control and Target groups display similar engagement distributions.
- The Target group shows a slightly higher median engagement score, but the difference is minimal.
- Outliers in both groups indicate variability in engagement behavior among users.



Engagement by Payment:
- Users who made payments exhibit marginally higher engagement than those who didn't.
- The variability in engagement is similar for both paying and non-paying users.



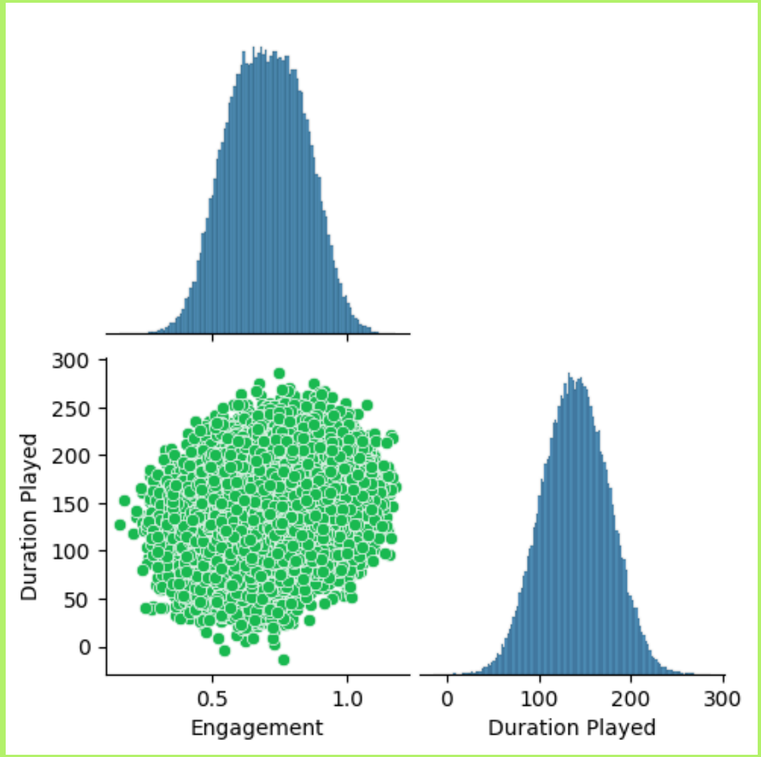
Duration Played by Group:
- The Target group has a slightly higher median for duration played compared to the Control group.
- Outliers exist, suggesting that a small subset of users spends significantly more time on the app.



Engagement by Subscription:
- Premium users show higher median engagement than Free users, reflecting the influence of subscription type on engagement.
- Free users have a wider range of engagement, with more outliers observed at the lower end.

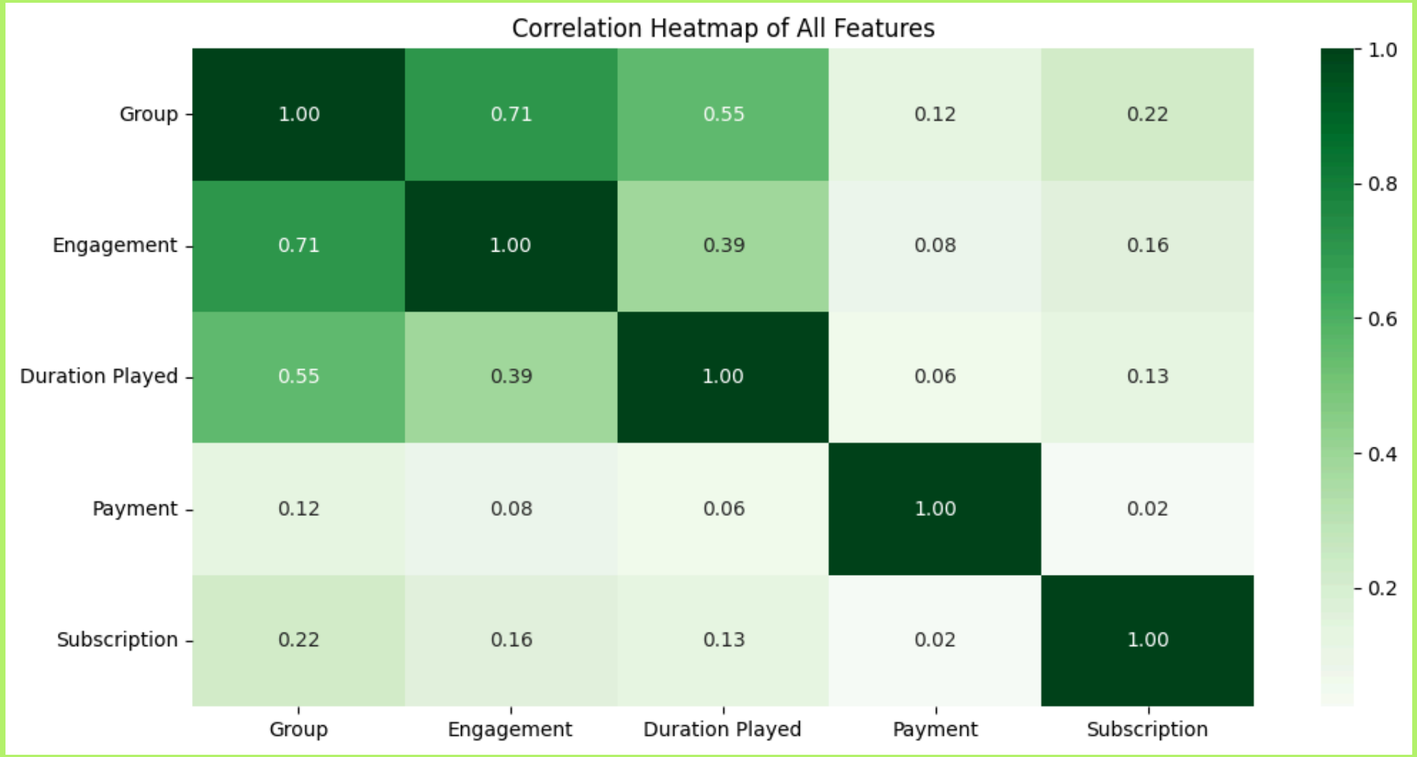
Correlation

https://github.com/hijirdella



Scatter Plot (Engagement vs. Duration Played):

- There is a positive correlation between engagement and duration played, indicating that users who are more engaged tend to spend more time on the app.
- The distribution of both engagement and duration played shows a normal pattern, confirming suitability for parametric tests like T-test or regression analysis.



Correlation Heatmap:

High Correlation:

- Engagement and Group (0.71): The target group shows higher engagement, validating the test design.
- Engagement and Duration Played (0.39): Indicates a direct relationship where higher engagement leads to increased time spent.

Moderate Correlation:

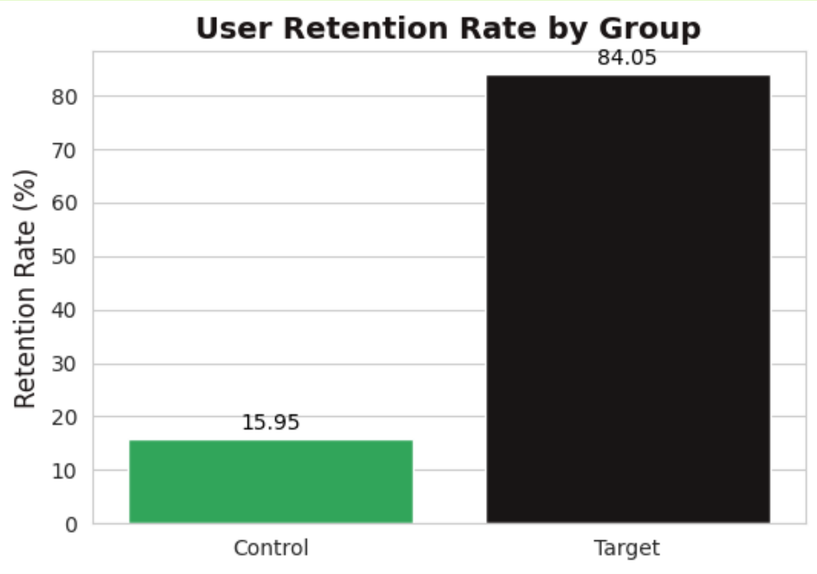
Subscription and Group (0.22): Subscription type (Premium or Free) has a moderate association with the group, suggesting that the feature may appeal differently to various subscription types.

Low Correlation:

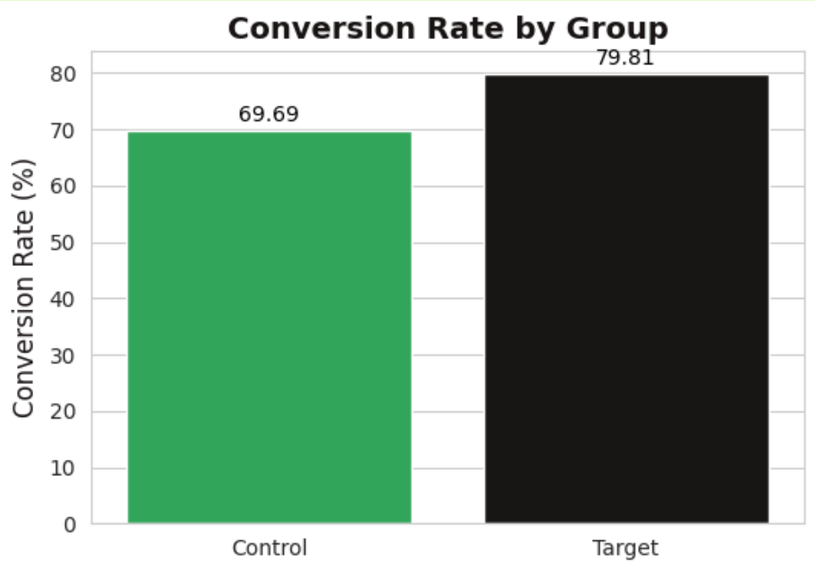
- Payment and Engagement (0.08): Payment behavior has a weak correlation with engagement, highlighting limited influence.
- Payment and Duration Played (0.06): Similar weak correlation suggests payment status does not significantly drive time spent.

Data Analysis

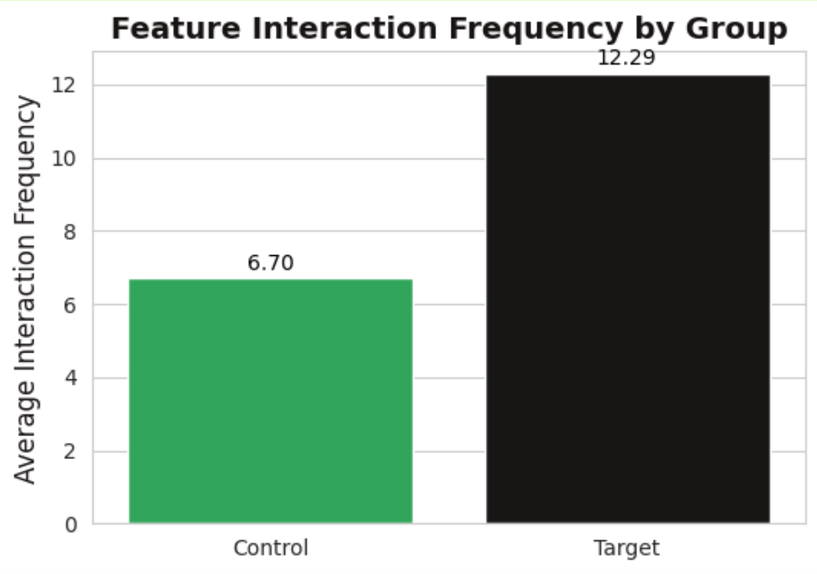
https://github.com/hijirdella



The Target Group demonstrates a significant improvement in user retention, with over 5x the retention rate of the Control Group. This indicates a strong positive impact of the new algorithm on retaining users.



The Target Group shows a 10% higher conversion rate than the Control Group. This suggests the new algorithm effectively drives users from free to premium subscriptions.



The Target Group exhibits nearly double the interaction frequency compared to the Control Group, showcasing the algorithm's effectiveness in boosting user engagement with playlists.

https://github.com/hijirdella

Conclusion

Key Question 1:
Does the new recommendation algorithm increase user engagement?

Yes, the new recommendation algorithm significantly increases user engagement. This is evident from the T-Test results, where the Target Group achieved a mean engagement score of 0.80, compared to 0.60 in the Control Group, with a statistically significant p-value of 0.000. Additionally, the interaction frequency in the Target Group (12.29) was nearly double that of the Control Group (6.70).

Key Question 2: Will it lead to higher retention and conversion rates?
Yes, the new algorithm leads to significantly higher retention and conversion rates.

- Retention Rate: The Target Group showed an 84.05% retention rate, compared to just 15.95% in the Control Group.
- Conversion Rate: The Target Group achieved a 79.81% conversion rate, 10% higher than the Control Group's 69.69%.
- These improvements highlight the algorithm's success in both keeping users engaged and converting them to premium subscribers.

Key Question 3: Should Spotify implement the new algorithm across all users?

Yes, the results strongly support implementing the new algorithm for all users. The statistically significant improvements in engagement, retention, and conversion indicate that the algorithm delivers meaningful benefits. Implementing it broadly could enhance Spotify's overall user experience and revenue. However, a phased rollout with continued monitoring is recommended to validate these results at scale.

https://github.com/hijirdella

Recommendations (1)

Implement the New Recommendation Algorithm:

- Deploy the new algorithm to all users, as it has demonstrated significant improvements in engagement, retention, and conversion rates.
- Prioritize high-value user segments (e.g., frequent listeners or users at risk of churn) for early implementation to maximize immediate impact.

Monitor Performance Post-Implementation:

- Continuously track engagement metrics, retention rates, and conversion rates to ensure the new algorithm performs consistently at scale.
- Set up an A/B monitoring dashboard to compare historical performance with post-implementation data.

Conduct a Phased Rollout:
Implement the algorithm in phases, starting with regions or user groups where engagement is traditionally lower. This will help mitigate potential risks and provide more localized insights.

https://github.com/hijirdella

Recommendations (2)

Optimize the Algorithm Further:

Use feedback from early implementation to fine-tune the algorithm, potentially incorporating additional personalization elements like user preferences, listening history, or contextual data (e.g., time of day).

Reinforce Retention Strategies:

- Focus on engaging users during the critical first two days, as the suggested test duration indicates this is sufficient to capture user behavior patterns.
- Provide dynamic notifications or playlist recommendations tailored to user behavior to boost retention further.

Drive Premium Subscriptions:

Leverage the conversion insights to design targeted campaigns encouraging free users to upgrade to premium, emphasizing benefits like ad-free listening and offline downloads.

Assess Long-Term Impact:

- Perform a longitudinal analysis over a few months post-implementation to evaluate the long-term benefits and identify any areas for further refinement.

https://github.com/hijirdella

References

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. (2009). Controlled experiments on the web: Survey and practical guide. Data Mining and Knowledge Discovery, 18(1), 140-181. https://doi.org/10.1007/s10618-008-0114-1

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'13), 1168-1176. https://doi.org/10.1145/2487575.2488217

McInerney, J., Zheng, H., Frazier, P. I., Anderson, A., & York, D. (2018). Explore-exploit learning in online content recommendation. Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), 2666-2673. https://doi.org/10.24963/ijcai.2018/370

Deng, A., Xu, Y., Kohavi, R., & Walker, T. (2016). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. Proceedings of the 8th ACM Conference on Web Search and Data Mining (WSDM'16), 499-508. https://doi.org/10.1145/2835776.2835840

Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: More, better, faster experimentation. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), 17-26. https://doi.org/10.1145/1835804.1835810

https://github.com/hijirdella

Thank You So Much!

Presented by **Hijir Della Wirasti | https://id.linkedin.com/in/hijirdella**