Spotify Unwrapped: Deconstructing the Music Personalization Process

Mark Ayiah

December 17, 2024

1 Abstract

This report analyzes the factors of a song that influence music enjoyment using Spotify listening data received through a request to the platform. The dataset includes variables such as track popularity, loudness, danceability, acousticness, and other audio features. We employ exploratory data analysis (EDA), LASSO regression, logistic regression, and ensemble models like Random Forest and XGBoost to predict whether a song was "enjoyed" or not. While the models showed only modest accuracy (with AUC values ranging from 0.55 to 0.58), the results suggest that Spotify likely leverages far more granular, proprietary data about individual users that was not made available for this analysis. The key takeaway is that companies like Spotify collect and utilize highly detailed user metrics to curate personalized recommendations, but these methods and the extent of data collection remain opaque to the public. This raises important considerations about data privacy and the effort platforms invest in understanding user behavior.

2 Executive Summary

This report explores what makes a song enjoyable by analyzing Spotify listening data. The goal was to uncover the key features of songs that influence listening preferences and determine whether publicly available data can predict which songs people enjoy. Using audio features such as track popularity, energy, danceability, loudness, and acousticness, the analysis applied both simple statistical models and more advanced machine learning techniques.

The results showed that certain song characteristics are linked to enjoyment. Songs that are popular, energetic, and easy to dance to were more likely to be enjoyed, while quieter or highly acoustic songs were less favored. However, despite identifying these trends, the models' ability to predict enjoyment was limited, with accuracy hovering around 55-57%. This means that, while some patterns could be found, the publicly available data alone is not enough to fully explain why someone enjoys a particular song.

The modest accuracy of these models suggests that Spotify's ability to deliver highly personalized recommendations depends on far more detailed and specific data than what is publicly shared. Beyond basic song features, Spotify likely uses extensive behavioral data, such as when users listen, their mood, habits, and patterns over time. It is also possible that third-party data and other forms of personal identification help refine their algorithms, giving the platform a deeper understanding of individual preferences.

This raises important questions about transparency and data privacy. Users benefit from accurate, personalized music recommendations, but they often do not realize how much data is being collected and analyzed to achieve this. Platforms like Spotify operate sophisticated systems that rely on hidden, proprietary data, which gives them a significant advantage in understanding user behavior but also leaves users in the dark about how their information is used.

In conclusion, while this report highlights some factors that influence song enjoyment, it also underscores the limits of publicly available data. The findings point to the hidden power of proprietary data held by streaming services, which allows them to create far more accurate and tailored recommendations. Future work could explore richer datasets that include listening behavior, time-based trends, and contextual factors to better understand what truly drives music preferences.

3 Introduction

Music streaming platforms like Spotify have transformed the way people consume music, offering personalized recommendations based on listening habits. These platforms promise to understand individual tastes, curating playlists and suggestions tailored to each user. The ability to predict what someone will enjoy listening to is an impressive technical feat and imperative to the success of an entity such as Spotify, but it also raises questions about the extent of the data they collect and the methods used to generate these recommendations.

The motivation behind this project stems from a curiosity about the factors that influence music enjoyment and how much can be inferred from Spotify's publicly accessible data. While Spotify claims to provide a personalized experience, it is unclear how much of that personalization is driven by user listening patterns and song information versus proprietary data and algorithms that remain opaque. This project seeks to examine whether publicly available audio features and metadata can effectively predict song enjoyment and, in turn, to highlight the limitations of accessible data compared to what platforms like Spotify may actually use.

The dataset for this analysis was obtained by request from Spotify and contains various audio and song-level features, such as track popularity, danceability, loudness, speechiness, and acousticness. Each row represents a song listened to, with an outcome variable denoting whether the song was "enjoyed" based on metrics like play count and completion rate.

The remainder of this report is organized as follows:

- Section 4 presents exploratory data analysis (EDA) and visualizations, revealing relationships between features like loudness, danceability, and enjoyment.
- Section 5 describes the predictive models applied, including LASSO regression, logistic regression, and ensemble methods such as Random Forest and XGBoost. The models' performance is evaluated, and key predictors are identified.
- Section 6 focuses on the interpretation of the results, comparing models and discussing their limitations. The results underscore the modest predictive power of the available data.
- Section 7 concludes the report with reflections on the findings and recommendations for future work. It highlights the hidden depth of Spotify's proprietary data and raises important considerations about data privacy and transparency.

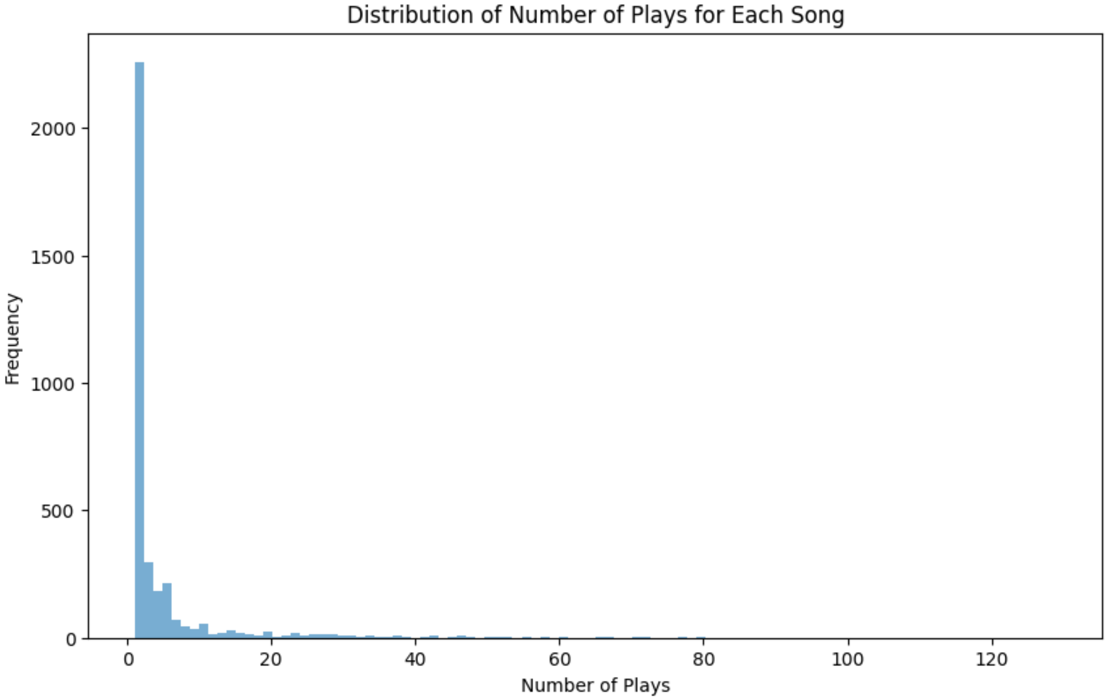
Through this project, the aim is to shed light on how much can be inferred from publicly available data, as well as its limitations.

4 Exploratory Data Analysis

To understand the data and prepare it for modeling, we conduct an exploratory data analysis (EDA). This section details the steps taken to investigate the data, uncover key patterns, and ensure the dataset was ready for further analysis.

4.1 Distribution of Number of Plays

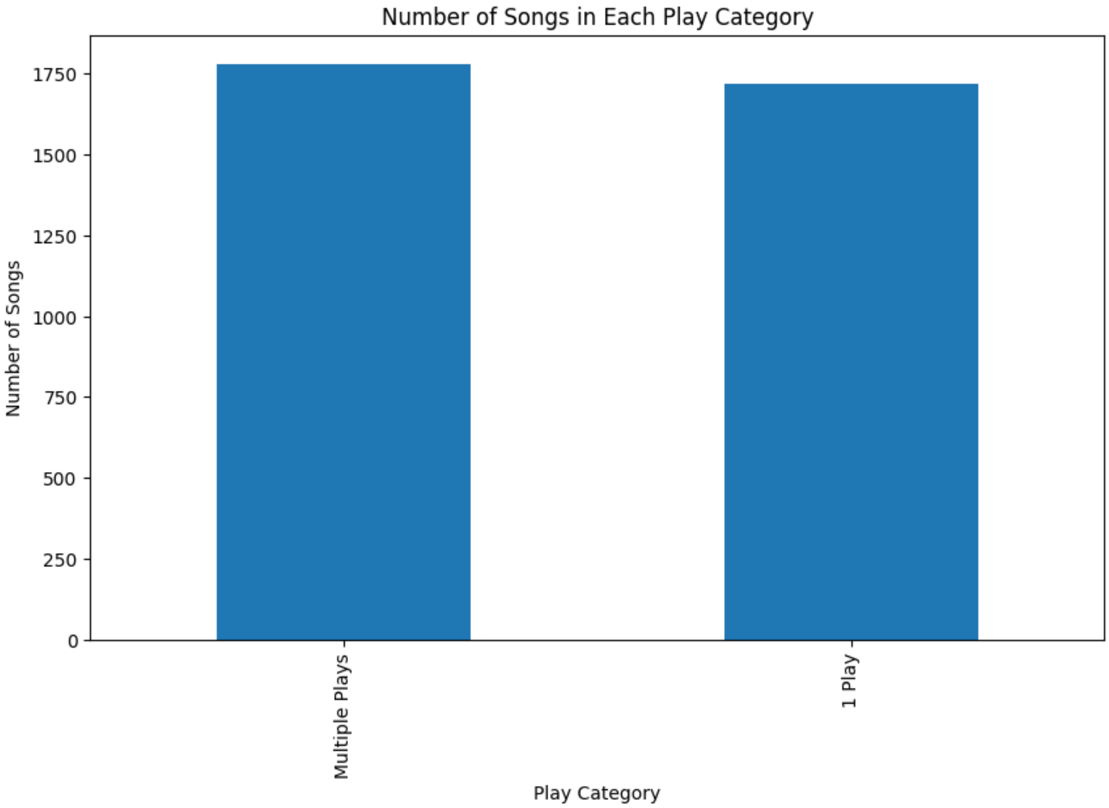
The first step was to examine the number of plays for each song. Understanding the spread of this variable helps identify whether listening frequency is a useful signal for enjoyment.



The histogram reveals that the majority of songs were played only once, with a small subset receiving multiple plays. This initial insight suggested the need to categorize songs into “1 Play” versus “Multiple Plays”.

4.2 Categorizing Songs by Play Count

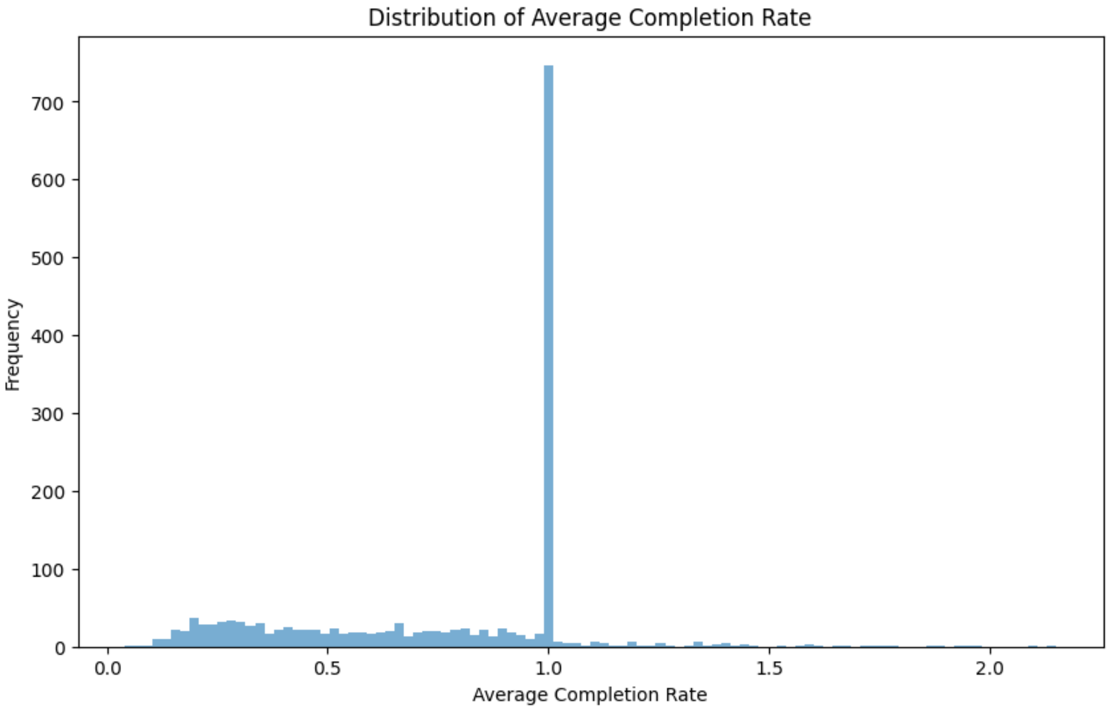
To simplify the analysis, we divide songs into two categories, with one representing single-play songs and the other being songs played multiple times



The bar chart shows a pretty even split between the two categories.

4.3 Completion Rates

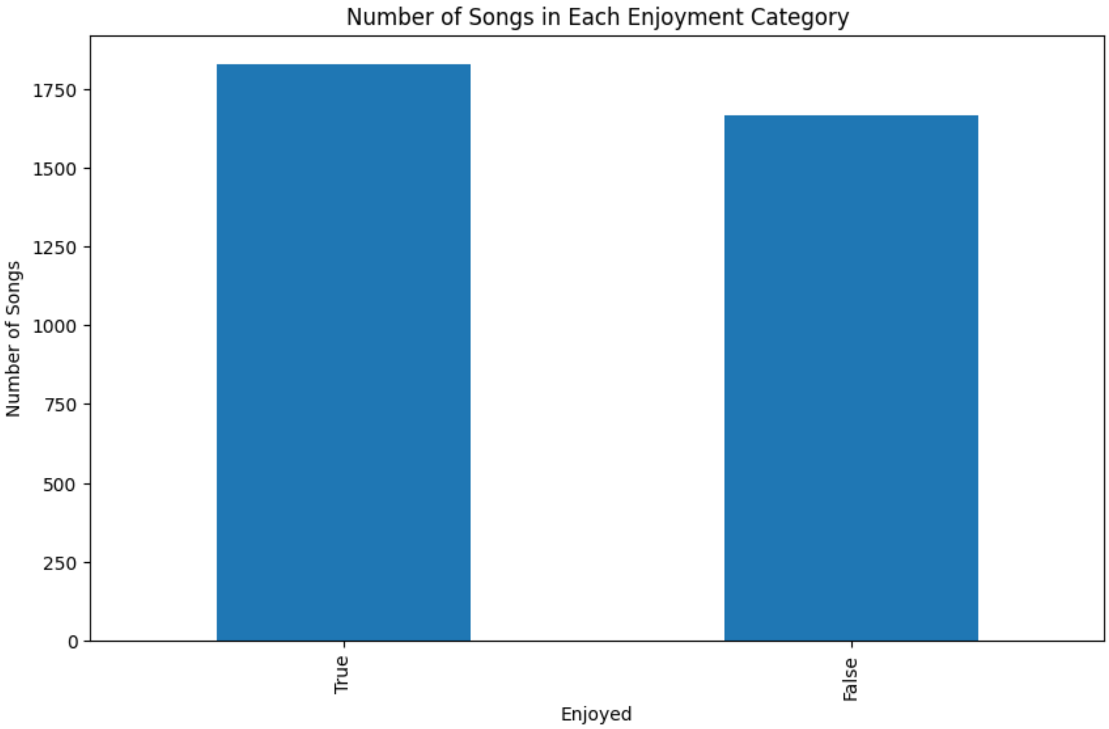
We next analyze the average completion rate for songs to see how fully they were listened to.



The histogram highlights a skewed distribution, with some songs having very high completion rates even when only played once. This suggests that high completion rates might indicate enjoyment.

4.4 Defining "Enjoyed" Songs

Using the completion rate and play count, we create a binary column enjoyed to identify songs that were either replayed or fully listened to (completion rate 1.25).



This bar chart confirmed a near-even split between songs classified as "enjoyed" and "not enjoyed," providing a balanced dataset for modeling.

4.5 Exploring Audio Features

We next the distribution of key audio features between "enjoyed" and "not enjoyed" songs. Features like danceability, energy, and loudness were investigated.

The resulting plots (shown in Figure 1 in the appendix) showed clear differences for certain features like danceability, where enjoyed songs tended to score higher. Meanwhile, acousticness appeared lower for songs classified as enjoyed, suggesting that quieter or highly acoustic songs were less likely to be favored.

4.6 Standardizing and Transforming Features

To address skewness (shown in Figure 2 in the appendix), we standardize and transform key features like loudness, acousticness, and speechiness. These transformations ensure that outliers do not disproportionately influence the analysis. Loudness was standardized to have a mean of 0 and a standard deviation of 1. Features like acousticness, speechiness, and liveness were log-transformed to reduce skewness. The updated distributions revealed smoother, more comparable patterns between enjoyed and not enjoyed songs across all features.

4.7 Checking for Multicollinearity

Finally, we check for multicollinearity among the predictors, as it is an assumption made by both LASSO regression and Logisic Classification which are used later. We calculate the Variance Inflation Factor (VIF) for each feature to detect highly correlated variables.

```
               Feature       VIF
12             key_1.0  3.088966
1               energy  2.684232
11  time_signature_4.0  2.554757
2             loudness  2.431251
10  time_signature_3.0  2.310737
16             key_5.0  2.064412
13             key_2.0  1.978676
21            key_10.0  1.959357
22            key_11.0  1.933872
19             key_8.0  1.889869
17             key_6.0  1.867304
18             key_7.0  1.814806
20             key_9.0  1.739779
15             key_4.0  1.651740
4         acousticness  1.489064
0         danceability  1.392902
14             key_3.0  1.349805
5              valence  1.338606
9   time_signature_1.0  1.193122
8     track_popularity  1.110048
3          speechiness  1.071115
6             liveness  1.059466
7                tempo  1.049386
```

All features have VIF values of below 5, indicating no severe multicollinearity. With the data cleaned, transformed, and explored, the next section will focus on building predictive models to understand which features are most strongly associated with song enjoyment.

5 Modeling & Analysis

This section describes the models and analyses conducted to determine the factors influencing song enjoyment.

5.1 Problem Setup and Assumptions

The goal is to predict whether a song is "enjoyed" based on its audio features and metadata. Key assumptions include: 1. Binary (Logistic) Classification Assumption: A song is either "enjoyed" or "not enjoyed," largely determined by play count and completion rate. 2. Independence of Observations: Each song in the dataset is treated as an independent observation. 3. No Multicollinearity: Predictors should not be highly correlated. Variance Inflation Factors (VIF) were calculated to confirm this. 4. Linear and Non-linear Relationships: Models were chosen to capture both linear and non-linear associations.

The predictors (X) include: - Audio Features: danceability, loudness, speechiness, acousticness, liveness, tempo
- Song Metadata: track popularity, time signature, key, genre The outcome variable (y) is binary, where $y = 1$ indicates an "enjoyed" song, and $y = 0$ indicates otherwise.

5.2 LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) regression was applied to identify important predictors and reduce overfitting by shrinking coefficients. The model can be represented as:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

Where t is a tuning parameter. The regularization helps identify the most relevant predictors.

- Top Predictors:
  - `track_popularity`: Strongest positive impact on enjoyment
  - `speechiness` and `danceability`: More moderate positive impact
  - `acousticness` and `liveness`: Negative impact
- Performance:
  - Mean Squared Error (MSE): 0.243
  - R^2: 0.02

LASSO's extremely low R^2 reflects its limited predictive power, highlighting the complexity of music enjoyment.

5.3 Logistic Regression

Logistic Regression was used for binary classification. The probability of a song being enjoyed is modeled as:

$$P(y = 1|X) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)}}$$

- AUC: 0.574
- Interpretability: Coefficients provide insights into feature importance. For example, *track_popularity* and *danceability* have positive associations with enjoyment, while `acousticness` has a negative association.

5.4 Random Forest

Random Forest, an ensemble method, was applied to capture non-linear relationships. It constructs multiple decision trees and averages their outputs.

- Performance: AUC = 0.5596

- Feature Importance:
  1. `track_popularity`
  2. `speechiness`
  3. `acousticness`

Logistic Regression outperformed Random Forest marginally, suggesting that the data cannot currently model the complex relationship between the music and the listener's preferences.

5.5 Gradient Boosting and XGBoost

Boosting methods, such as Gradient Boosting and XGBoost, were applied to further improve performance by iteratively learning from errors.

- Performance Comparison:
  - Logistic Regression: AUC = 0.574
  - Random Forest: AUC = 0.5596
  - XGBoost: AUC = 0.5558
  - Gradient Boosting: AUC = 0.5550

The results show minimal differences across models, again indicating that the predictors are insufficient for accurately modeling song enjoyment.

5.6 Evaluation of Models

The table below summarizes the performance metrics:
[]@lll@ Model AUC Interpretability
Logistic Regression 0.574 High
Random Forest 0.5596 Moderate
Gradient Boosting 0.5550 Low
XGBoost 0.5558 Low

Logistic Regression provided the best trade-off between performance and interpretability.

The analysis reveals that while models like LASSO and Logistic Regression identify key predictors (e.g., `track_popularity` and `danceability`), their predictive power remains limited. Ensemble methods like Random Forest and XGBoost do not significantly improve performance. This highlights the need for more granular or contextual data, such as user mood or temporal patterns, to better understand music enjoyment.

6 Visualization and Interpretation of the Results

With the data prepared and explored, the next step was to build and evaluate predictive models to determine which features were most strongly associated with song enjoyment. In this section, we apply multiple modeling techniques, including LASSO regression, Logistic Regression, and ensemble methods like Random Forest and XGBoost. These models were assessed for their performance, and their results were compared.

6.1 LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) regression was applied to identify important predictors while controlling for overfitting. LASSO uses regularization to shrink less informative coefficients to zero, effectively performing feature selection.

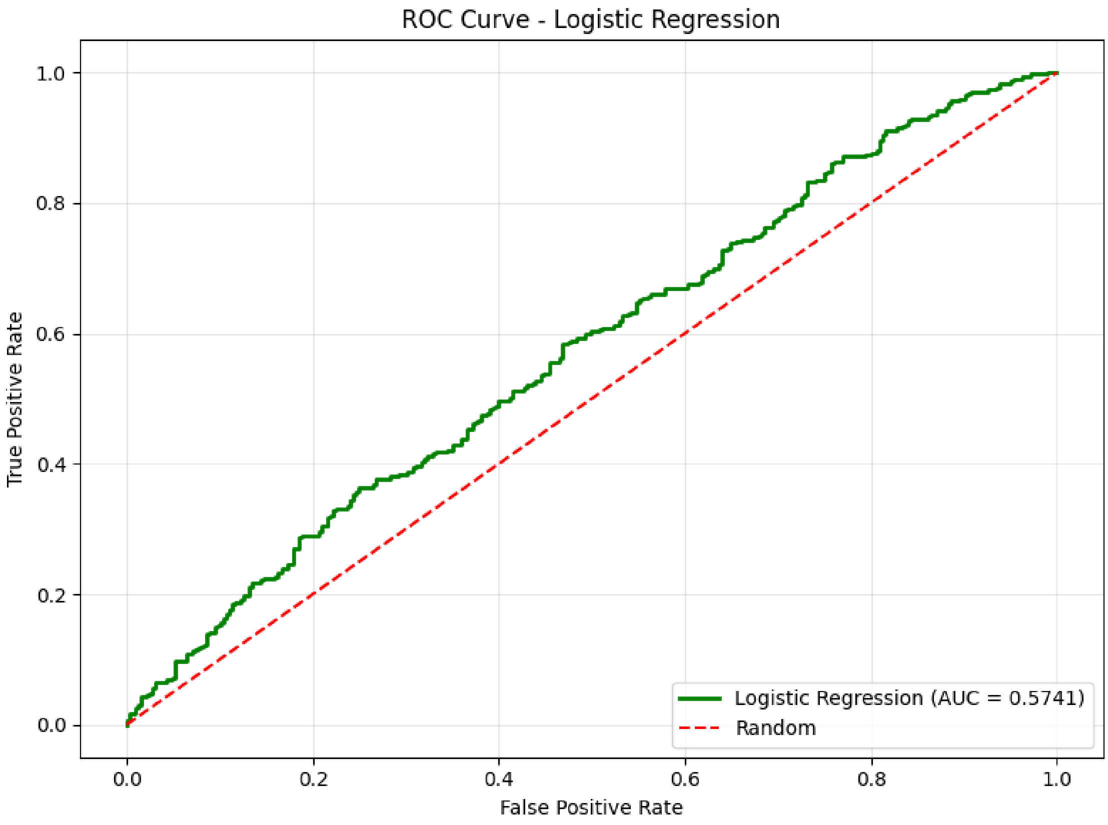
```
                    LASSO Feature Coefficients
track_popularity                      0.036461
speechiness                           0.021502
danceability                          0.011868
tempo                                 0.007267
key_4.0                               0.006978
time_signature_3.0                    0.006128
loudness                              0.001941
key_1.0                               0.000699
key_6.0                              -0.001781
time_signature_4.0                   -0.004950
liveness                             -0.017703
acousticness                         -0.034998

        MSE  R-Squared
0  0.243579   0.020686
```

The model indicates that higher values of features like track_popularity, speechiness, and danceability increase the likelihood of a song being "enjoyed," while features such as acousticness, liveness, and time_signature_4.0 have a negative impact. Among these, track_popularity emerges as the strongest predictor, suggesting that more popular tracks are more likely to be enjoyed. Additionally, songs with moderate speech content and higher danceability are favored, whereas high acousticness reduces enjoyment. The model's performance, with an R-squared of 2% and a mean squared error of 0.24, reflects modest predictive power, which is expected given the complexity of behavioral preferences like song enjoyment. However, a classification model may prove to be more effective at prediction.

6.2 Binary Classification (Logistic Regression)

Logistic Regression was used as a classification model to predict whether a song was enjoyed (binary outcome). This model is simple yet interpretable and serves as a baseline for comparison with more complex methods.

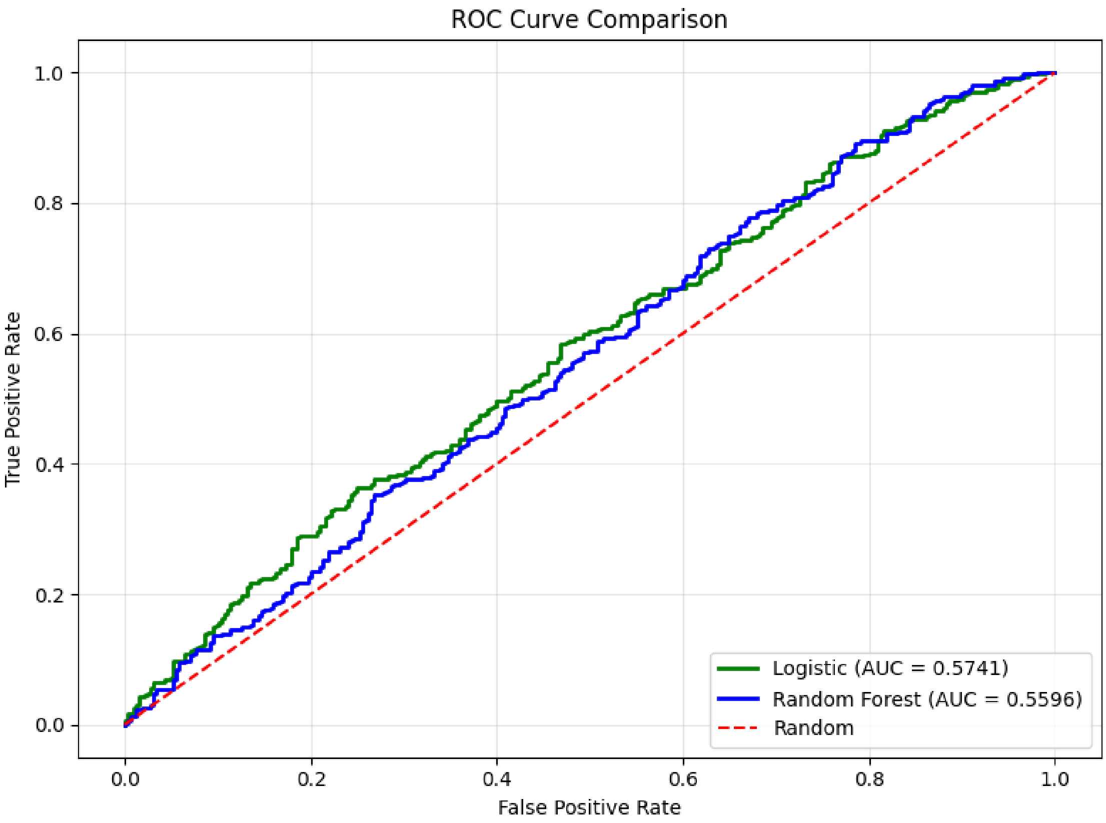
```
   Accuracy       AUC
0  0.557143  0.574137
```



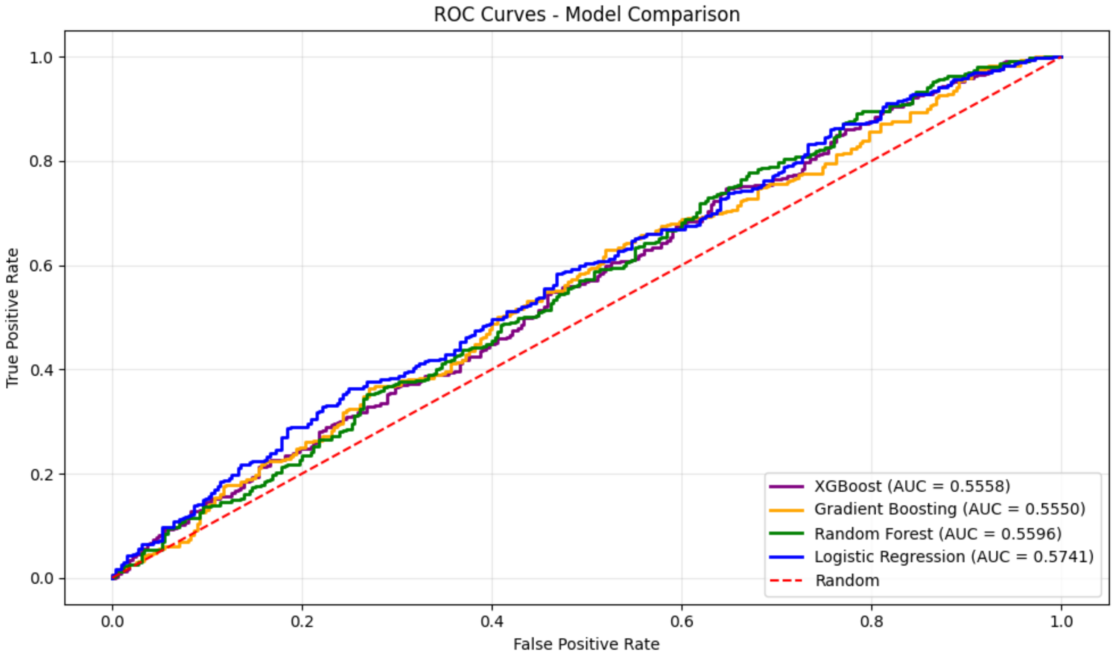
The ROC curve shows an AUC of 0.574, indicating that the logistic regression model performs only slightly better than random guessing (AUC = 0.5). The curve's proximity to the diagonal line reflects the model's limited ability to distinguish between "enjoyed" and "not enjoyed" songs. This suggests that the current features may not sufficiently capture the patterns in song enjoyment, and the linear nature of logistic regression may not be well-suited for the complexity of the data. We will next try models that do not assume linear relationships between the larget variable and the features.

6.3 Random Forest Classifier

The Random Forest algorithm, a non-linear ensemble method, was applied to capture more complex relationships between features and the outcome. This method builds multiple decision trees and averages their predictions to improve robustness.



While Random Forest typically outperforms Logistic Regression due to its ability to capture complex patterns, in this case of predicting music enjoyment, the simpler Logistic Regression model performed better. The Logistic Regression model achieves an AUC of 0.5599, while the Random Forest model slightly improves upon this with an AUC of 0.5644. To thoroughly explore the options for prediction, however, we will evaluate two other powerful ensemble methods: Gradient Boosting and XGBoost. These boosting algorithms might capture patterns that both Random Forest and Logistic Regression missed by learning from previous predictions.



The comparison of the four models reveals remarkably similar performance, with all models achieving AUC scores between 0.55 and 0.58. Logistic Regression achieved the highest AUC of 0.5749, followed by Random Forest (0.5596), XGBoost (0.5558), and Gradient Boosting (0.5550). While all models performed better than random chance (AUC = 0.5), the marginal improvement suggests that predicting song enjoyment based on audio features and metadata alone is a task that requires more granular data. The similar performance across different model complexities, from the simpler Logistic Regression to more complex ensemble methods like XGBoost, suggests that the inherent subjectivity of music preference may limit the predictive power of purely technical audio features.

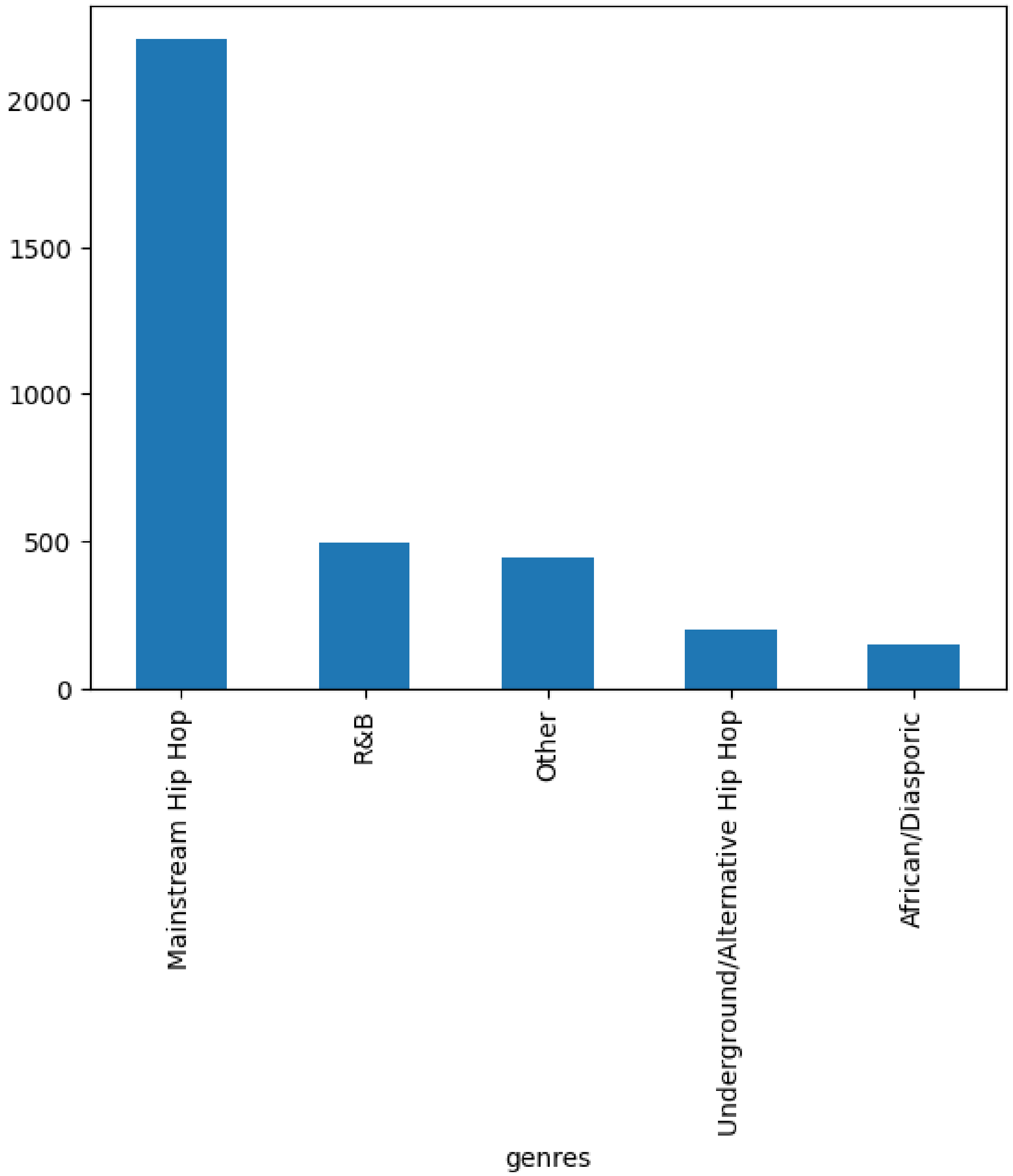
6.4 Genre Analysis

This section explores the distribution of songs across consolidated genres and their relationship with user enjoyment.

6.4.1 Consolidating Genres

The first step in the genre analysis was to simplify the raw genre tags into six overarching categories: Mainstream Hip Hop, Underground/Alternative Hip Hop, R&B, African/Diasporic, Lo-Fi, and Other.

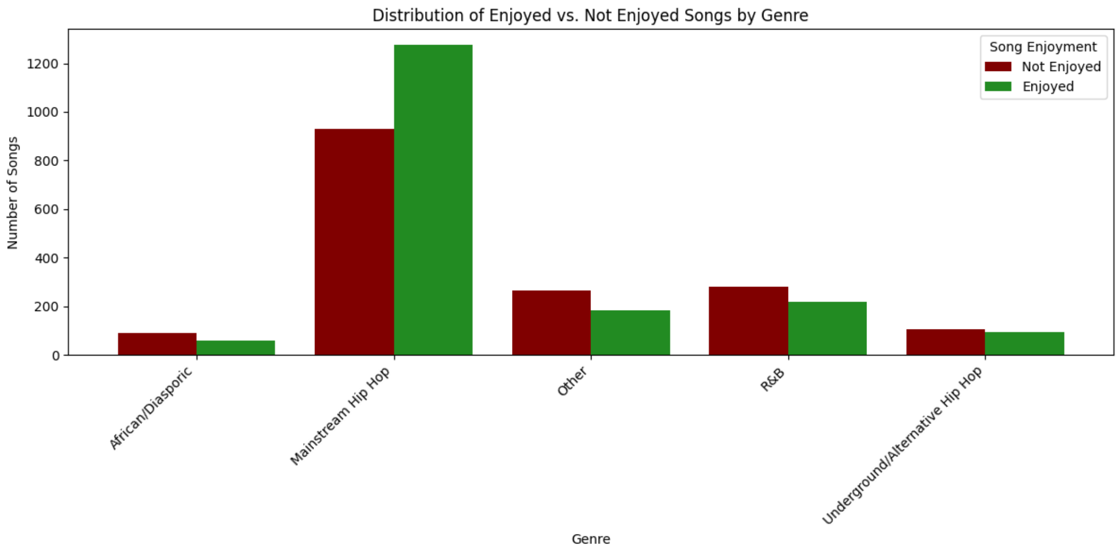
After consolidating the genres, the distribution of songs across these six categories was visualized using a bar chart. The resulting plot highlights which genres are most prevalent in the dataset.

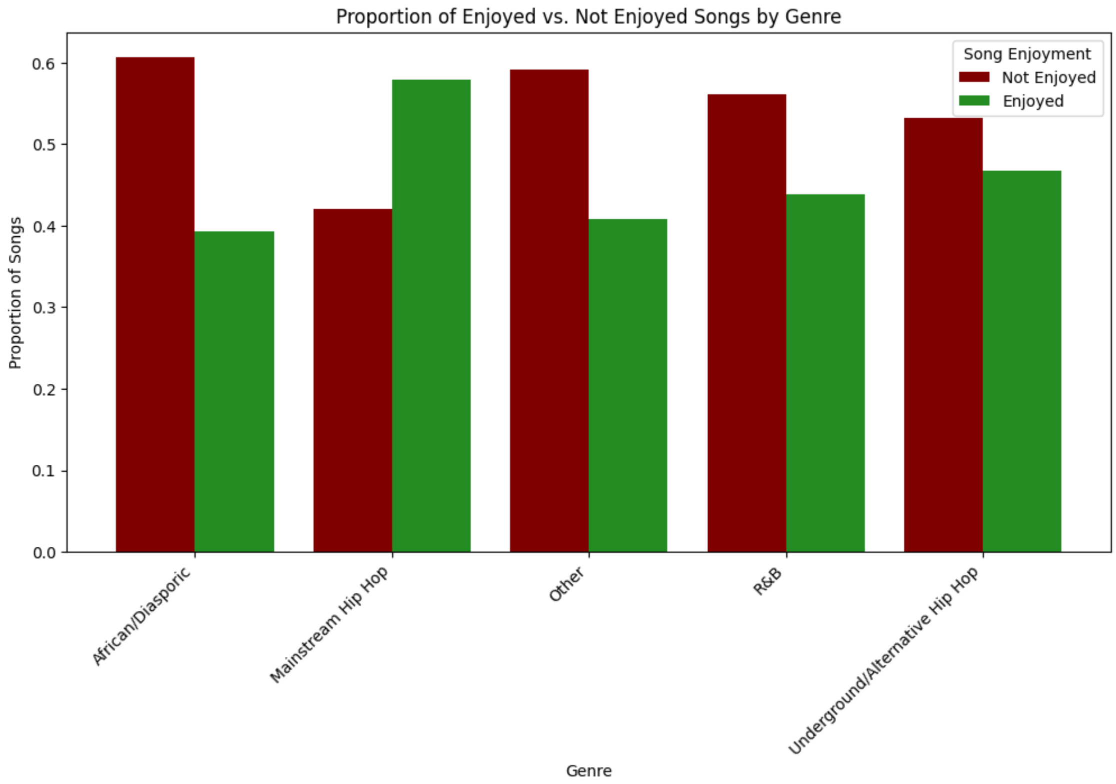


From the chart, it is evident that Mainstream Hip Hop overwhelmingly dominates my listening history, with significantly smaller contributions from genres like R&B, Other, and African/Diasporic.

6.4.2 Enjoyment Across Genres

Next, we explore the relationship between song enjoyment and genre. The distribution of enjoyed versus not enjoyed songs was plotted side-by-side for each genre, followed by a proportional analysis.





The first plot shows the raw counts of enjoyed and not enjoyed songs across genres. Notably, Mainstream Hip Hop has the largest number of enjoyed songs, but it also contains a substantial portion of not enjoyed tracks. In contrast, genres like African/Diasporic and Underground/Alternative Hip Hop have fewer total songs but a higher proportion of not enjoyed tracks. This may prove to be a challenge for the predictive models, as there is clear imbalance across the different genres.

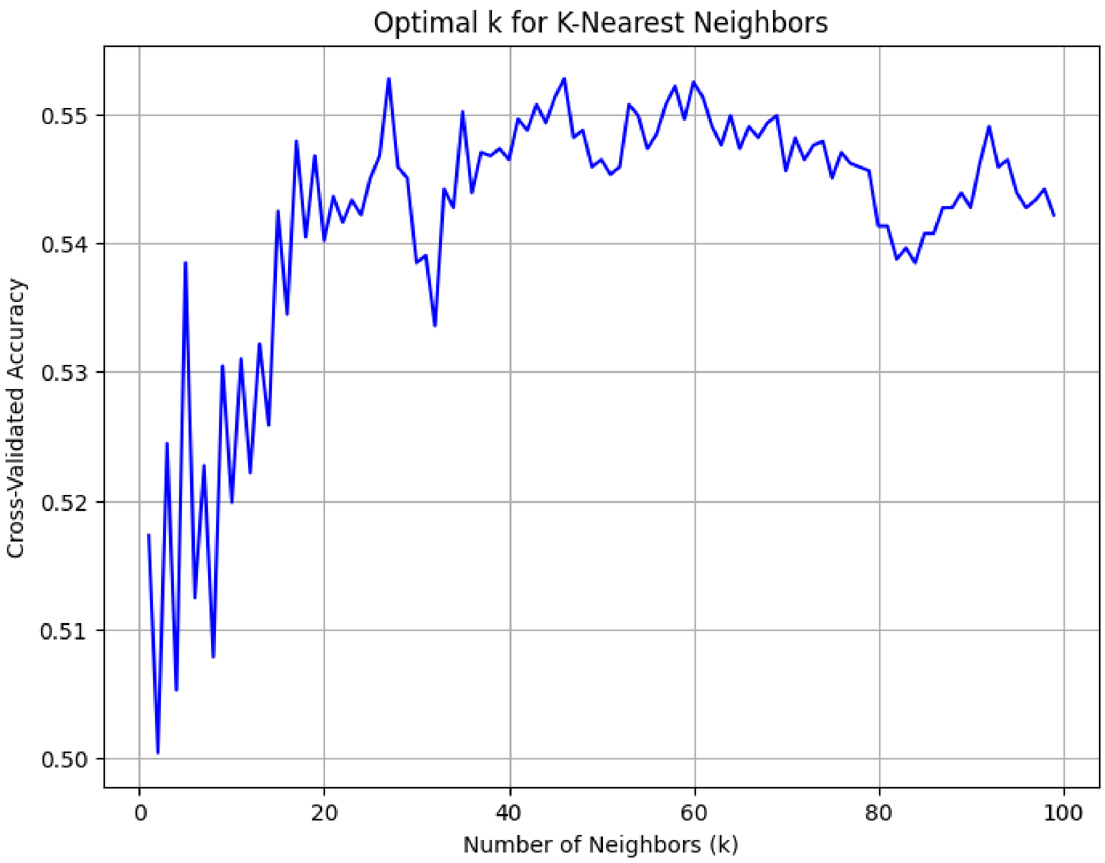
```
                              accuracy  precision    recall  f1-score  \
African/Diasporic             0.466667   0.333333  0.333333  0.333333
Mainstream Hip Hop            0.566893   0.620301  0.647059  0.633397
Other                         0.600000   0.510638  0.648649  0.571429
R&B                           0.540000   0.480769  0.568182  0.520833
Underground/Alternative Hip Hop  0.475000   0.461538  0.631579  0.533333

                                   auc
African/Diasporic             0.328704
Mainstream Hip Hop            0.560869
Other                         0.555839
R&B                           0.592532
Underground/Alternative Hip Hop  0.516291
```

The logistic regression models, with SMOTE applied to address class imbalance, showed varying performance across genres. Mainstream Hip Hop and Other achieved the best results, with accuracy around 57-60% and AUC values of 0.56, indicating moderate predictive ability. Recall was generally high, particularly for Underground/Alternative Hip Hop (0.63), reflecting the model's strength in identifying "enjoyed" songs, though precision remained lower. In contrast, African/Diasporic performed poorly (AUC = 0.33), likely due to limited data. Overall, performance improved for minority class predictions, but smaller genres still pose challenges due to data scarcity.

6.5 K-Nearest-Neighbors for Enjoyment Prediction

To wrap up the genre analysis, we employ the K-Nearest Neighbors (KNN) algorithm to explore how well song enjoyment can be predicted using a combination of audio features and genre information. KNN is a simple yet effective classification method that assigns labels to data points based on the majority class of their k closest neighbors in feature space.

By incorporating both genre data and audio characteristics like danceability, energy, and tempo, we aim to determine whether these attributes collectively influence the likelihood of a song being enjoyed. This step helps assess if genre plays a significant role in enjoyment prediction or if other song features are more critical.



The plot shows the cross-validated accuracy of a K-Nearest Neighbors (KNN) model as the number of neighbors k increases. The accuracy improves consistently as k rises from small values, stabilizing around k=27, where it reaches one of its peaks. This suggests that k=27 offers an optimal trade-off: it balances the model's ability to generalize without overfitting to noise, while still maintaining strong interpretability compared to higher k-values. After k=27, the accuracy fluctuates slightly but remains relatively stable around 54-55%. This stabilization indicates diminishing returns as k increases further. Larger k-values smooth out predictions but may reduce the model's sensitivity to more nuanced patterns.

Despite incorporating audio features and genre data, the model's best accuracy remains modest (around 55%). This suggests that listening behavior may not be fully captured by the available data. Important factors such as mood, context (e.g., time of day), personal associations with songs, or unexplored song metadata (like lyrics or artist familiarity) likely influence enjoyment and are absent from the current dataset. Further enhancements, such as including temporal trends, song sequencing, or contextual features, could better model these behaviors.

6.6 Experimenting With More Granular Notions of Enjoyment

6.6.1 Granular Enjoyment Score

Lastly, we propose an enjoyment score is designed to capture the degree of enjoyment for each song based on three key components:

- Completion Rate: The proportion of the song that was listened to, indicating engagement.
- Play Count: The number of times a song was replayed, reflecting repeat enjoyment.
- Track Popularity: Adjusts for song popularity to account for bias towards mainstream music.

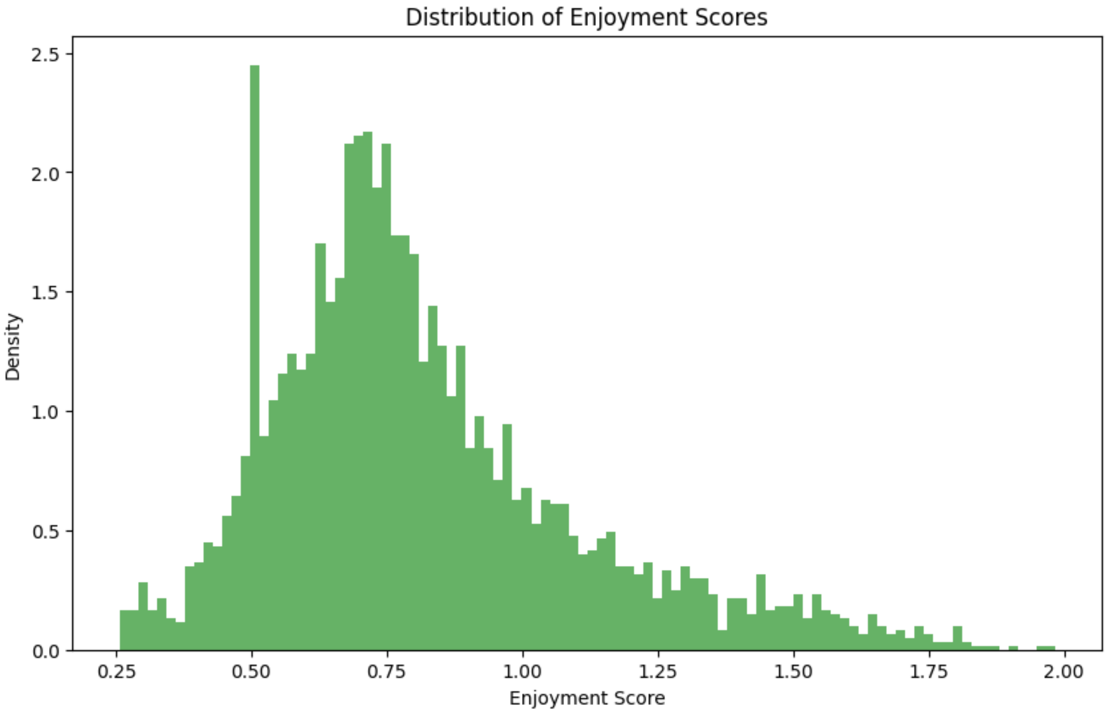
These components are combined into a weighted score to reflect enjoyment more precisely, defined as follows:

$$\text{Enjoyment Score} = w_1 \times \text{Completion Rate} + w_2 \times \log(\text{Play Count} + 1) + w_3 \times \text{Track Popularity (normalized)}$$

- w_1, w_2, w_3 are weights to control the contribution of each component.
- Logarithmic scaling for play count ensures diminishing returns for songs played many times.
- Normalization of track popularity keeps it in line with other features for balanced influence.

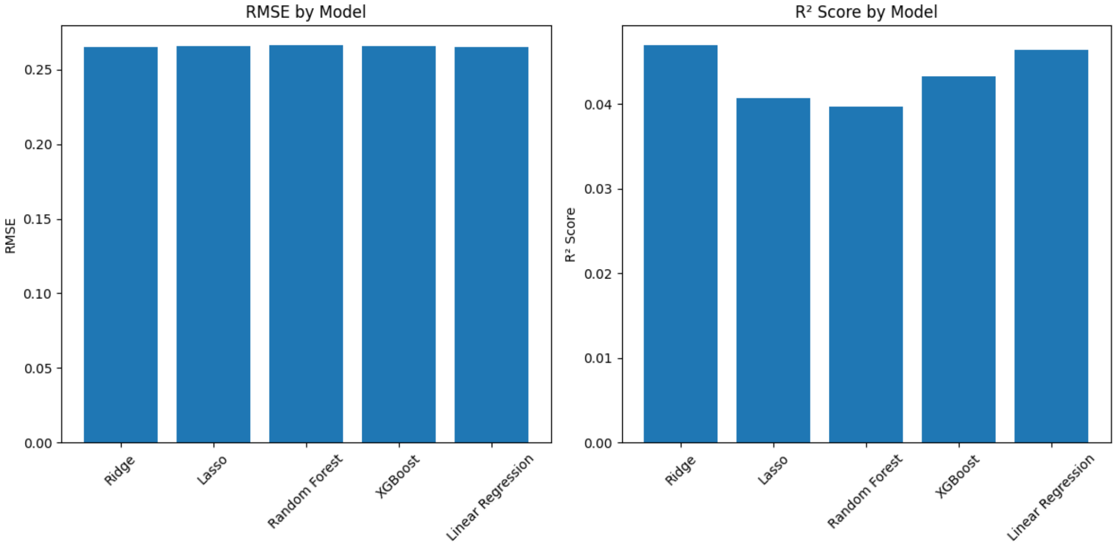
This granular score provides a richer, continuous measure of song enjoyment compared to a binary "Enjoyed" or "Not Enjoyed" outcome.

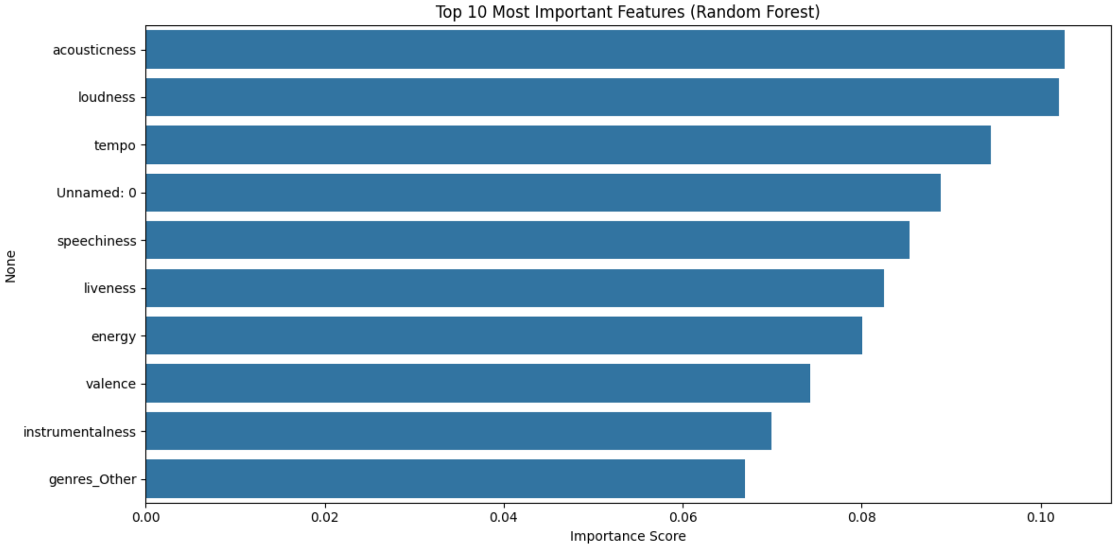
To optimize the weights for the enjoyment score components, a logistic regression model was used to predict whether a song was "enjoyed." The weights for completion rate, play count, and track popularity were optimized using `scipy`'s minimize function, which aimed to maximize the predictive performance measured by ROC-AUC. The objective function calculated the negative ROC-AUC for each combination of weights, allowing the optimizer to adjust the contributions of each component iteratively. The resulting optimal weights of $w_1 = 0.3$, $w_2 = 0.3$, and $w_3 = 0.4$ provide the best balance to predict enjoyment based on the data. The distribution of enjoyment scores can be found below:



Having previously explored binary enjoyment classification (enjoyed vs. not enjoyed), we now shift to predicting enjoyment scores as a continuous variable for a more nuanced understanding of user preferences by capturing varying levels of enjoyment rather than reducing the outcome to a binary decision.

By applying regression models, we aim to identify the specific song attributes and features that influence enjoyment scores the most. We again employ Ridge and Lasso Regression, Random Forest, and XGBoost to determine their predictive accuracy and feature importance.





The machine learning models' performance in predicting music enjoyment scores reveals several interesting patterns. Looking at the evaluation metrics, all models showed relatively consistent

RMSE values around 0.26, indicating similar prediction accuracy across different approaches. However, the R^2 scores, which measure the proportion of variance explained by the models, show more notable differences.

Linear Regression and Ridge Regression performed best with R^2 scores around 0.045 (4.5%), followed by XGBoost at approximately 0.042 (4.2%). The Random Forest and Lasso models showed slightly lower performance with R^2 scores around 0.03-0.04 (3-4%). Granted, these R^2 values are relatively low, providing more evidence that predicting music enjoyment is a complex task that may involve factors beyond the available features. However, they still provide valuable insights into feature importance, revealing the following:

1. Audio Characteristics: The top three predictors are loudness, acousticness, and tempo, suggesting that the technical aspects of a song's production and composition significantly influence listener enjoyment.

2. Vocal Elements: Speechiness and liveness rank highly (4th and 5th), indicating that the presence and quality of vocal components play a substantial role in listener engagement.

3. Emotional and Energy Factors: Valence (emotional positivity) and energy appear as the 6th and 7th most important features, showing that a song's emotional content and energy level are meaningful predictors of enjoyment.

4. Musical Structure: Danceability and instrumentalness round out the top features, demonstrating that the overall musical structure and composition style contribute to listener enjoyment.

5. Genre Impact: The presence of `genres_Other` in the top 10 suggests that genre classification has some predictive power, though relatively lower than the audio features.

7 Conclusion

This analysis investigated the factors influencing music enjoyment using publicly available Spotify audio features and metadata. By applying a range of statistical and machine learning models, including LASSO regression, logistic regression, Random Forest, and XGBoost, we seek to predict whether a song would be "enjoyed" based on attributes like track popularity, danceability, loudness, and acousticness. Despite the sophistication of the methods employed, model performance plateaued at around 55-57% AUC, demonstrating only modest predictive power. The results revealed that while features such as track popularity, danceability, and speechiness were positively associated with enjoyment, and acousticness and liveness were negatively associated, these attributes alone fail to fully capture the complexity of listening preferences.

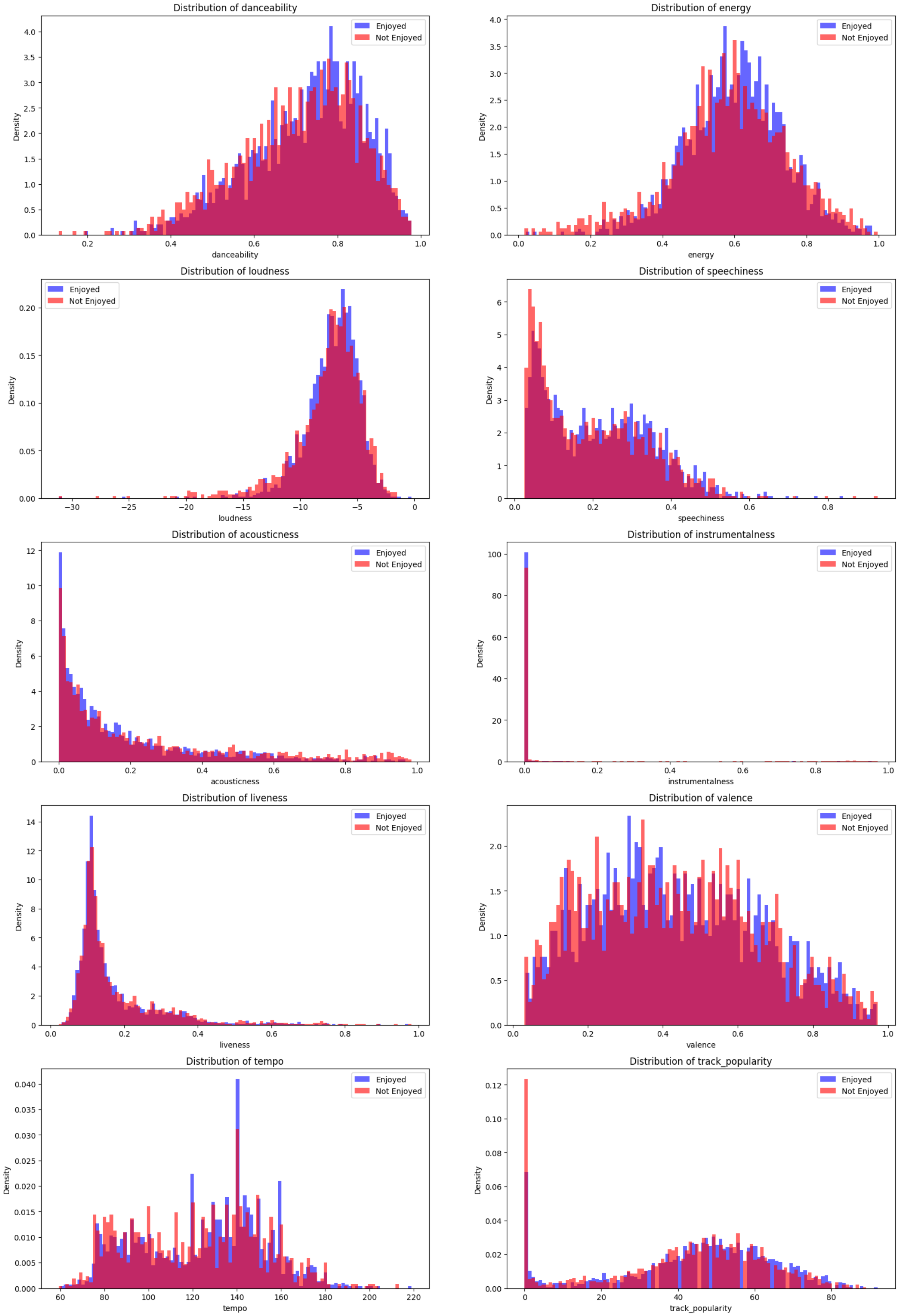
The limited accuracy achieved in this analysis raises a critical point about the nature of data available to the public. Streaming platforms like Spotify are known for their exceptional ability to curate highly personalized recommendations, which cannot be explained solely by the audio features and metadata analyzed here. This discrepancy suggests that Spotify—and likely other streaming services—possess far more granular data about individual users. Such data likely includes behavioral patterns, temporal listening trends, mood associations, device information, and possibly third-party data sources that build comprehensive personal profiles. These platforms may also integrate data from other means of personal identification, such as social media behaviors, location history, or interactions with advertisements, to construct robust models that predict user preferences with far greater accuracy.

The plateau observed in this study serves as indirect evidence of the extent and sophistication of the proprietary data Spotify leverages. The limited predictive power of publicly accessible data highlights its constraints, while simultaneously underscoring the opaque nature of data collection practices employed by these platforms. Users benefit from highly tailored recommendations, but this comes at the cost of extensive data collection, often with limited transparency regarding what data is collected, how it is used, and whether it is shared with third parties.

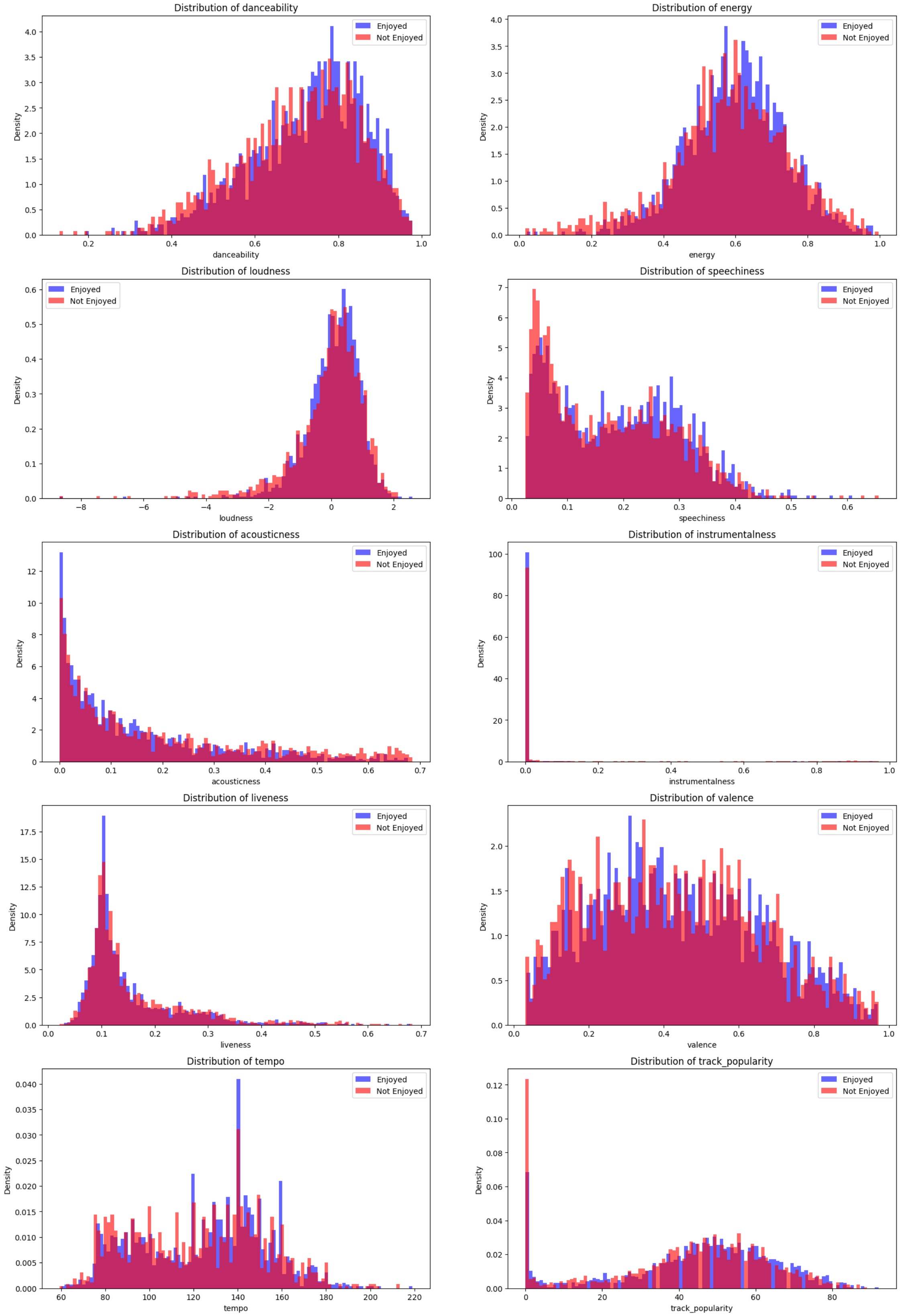
In conclusion, while this analysis sheds light on the relationship between song attributes and user enjoyment, it also highlights the critical role of proprietary, granular data in driving the success of music recommendation systems. Future research could incorporate richer behavioral and contextual data—such as listening times, sequential song patterns, and emotional sentiment—to better approximate the models used by platforms like Spotify. Additionally, this raises broader questions about data privacy and transparency: as consumers, we must remain aware of how much streaming services know about us and the extent to which our personal data fuels their algorithms.

8 Appendix

8.1 Figure 1



8.2 Figure 2



Distribution of danceability
Enjoyed
Not Enjoyed
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.2
0.4
0.6
0.8
1.0
danceability
Distribution of energy
Enjoyed
Not Enjoyed
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.0
0.2
0.4
0.6
0.8
1.0
energy
Distribution of loudness
Enjoyed
Not Enjoyed
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−8
−6
−4
−2
0
2
loudness
Distribution of speechiness
Enjoyed
Not Enjoyed
Density
0
1
2
3
4
5
6
7
0.0
0.1
0.2
0.3
0.4
0.5
0.6
speechiness
Distribution of acousticness
Enjoyed
Not Enjoyed
Density
0
2
4
6
8
10
12
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
acousticness
Distribution of instrumentalness
Enjoyed
Not Enjoyed
Density
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
instrumentalness
Distribution of liveness
Enjoyed
Not Enjoyed
Density
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
liveness
Distribution of valence
Enjoyed
Not Enjoyed
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
valence
Distribution of tempo
Enjoyed
Not Enjoyed
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
60
80
100
120
140
160
180
200
220
tempo
Distribution of track_popularity
Enjoyed
Not Enjoyed
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
20
40
60
80
track_popularity