The Sound of Union

Quantifying Union College's Music Preferences with Logistic Regression

Conor Fryer

STA 264 – Hoerl – Winter 2025

Abstract

This study explores whether Union College students have distinct musical preferences that can be quantified using logistic regression. The project builds on Union Rewind, an initiative by WRUC, the college's student-run radio station, that collects students' top five most-played songs through an in-person survey. By comparing these selections to a dataset of globally popular songs, the analysis aimed to determine whether measurable song attributes, such as danceability and energy, influence a song's likelihood of inclusion in a Union student's top five.

To address this question, a logistic regression model was developed using Spotify's audio features. The initial model included all available features, but popularity and loudness were each removed due to their overwhelming influence and redundancy, respectively. The final model assessed seven predictors: danceability, energy, acousticness, instrumentalness, speechiness, happiness, and live performance elements across 4,274 unique songs.

The results revealed stronger trends than initially expected, with five of seven predictors emerging as significant. Danceability had the strongest negative effect, suggesting that highly danceable music, despite its presence in social settings, was less likely to appear in students' personal listening habits. Acousticness and speechiness also had negative effects, reinforcing a preference for polished, structured music over raw acoustic tracks or spoken-word-heavy content. Instrumentalness was the only variable with a positive effect, suggesting that songs with instrumental elements had a slightly higher chance of inclusion.

Beyond identifying these trends, this study highlighted the challenges of modeling human taste using regression analysis. While measurable predictors emerged, the model's explanatory power remained limited, reinforcing the complexity and subjectivity of personal music selection. The

project also underscored the importance of data preparation, as structuring and cleaning the dataset was more time-consuming than running the regression itself.

Ultimately, logistic regression revealed meaningful patterns in Union students' music preferences, though fully capturing personal taste remains challenging.

Background

Music is both a personal and social experience. While individual preferences vary, patterns emerge within specific communities, shaped by cultural exposure and social environments. Research has shown that music plays a central role in identity formation, particularly among young adults, who use it to express values and align with peer groups (Bonneville-Roussy, Rentfrow, Xu, & Potter, 2013).

Union College in Schenectady, NY, is a liberal arts institution with a student body of approximately 2,000 students. Social life is heavily influenced by Greek life, and weekend parties serve as central gathering spaces where music selection helps shape the atmosphere. Certain songs dominate these environments, becoming unofficial anthems of campus culture. For example, *Mr. Brightside* by The Killers, despite being released in 2004, remains a staple at Union's social events, mirroring its enduring popularity at colleges nationwide. Similarly, country folk artist Zach Bryan has a strong presence in Union's social scene, particularly among fraternity members. His music, characterized by storytelling and a strong vocal presence, resonates with students who favor singable, lyric-driven tracks.

Given these anecdotal observations, this project examines whether Union students have identifiable musical preferences that differ from broader streaming trends. Specifically, it tests whether certain song characteristics significantly predict a song's inclusion in a student's top five most played tracks. This analysis provides a quantitative perspective on Union students' music preferences, offering insight into the role of song attributes in shaping personal listening habits.

Data Collection

This study relies on two datasets: one representing Union students' self-reported top five songs and another capturing a broader set of globally popular songs. The included dataset was compiled through Union Rewind, an annual initiative by Union College's radio club, WRUC, which collects students' most-played songs via an in-person survey. The non-included dataset was drawn from a publicly available streaming dataset, providing a wider benchmark for comparison.

Survey Design

The included dataset was built using Union Rewind, an in-person survey conducted by WRUC. Over six days during common hour (12:50–1:50 PM), when no classes are scheduled, students were surveyed at three primary dining locations: Reamer Dining Hall, West Dining Hall, and Rathskeller. To maximize participation, the survey was held at each location for two days.

In 2024, the survey gathered responses from 12.1% of the student body, totaling 246 students out of 2,031. A similar survey was conducted in 2023, where 13.2% of students participated, totaling 275 students out of 2,082.

To encourage participation, students were asked: "Do you have one minute to spare?" to emphasize the process's speed. Those who agreed pulled up their top five most-played songs from Spotify Wrapped or Apple Music Replay, and a quick photo of their list was taken to ensure accuracy. The entire process lasted no more than a minute or two per student.

Most students had access to a year-end streaming summary, but in rare cases (fewer than five students in 2024), participants who did not subscribe to a streaming service named five songs they felt best reflected their listening habits. While this introduced some subjectivity, the number of these cases was too small to meaningfully impact the results.

Data Processing

Once survey responses were collected, the songs were manually searched for and added to a Spotify playlist. If a song was a remix, live version, or alternate mix, that specific version was included rather than the standard release.

Initially, duplicates were retained in the playlist to determine the most frequently mentioned songs. After this analysis, Spotify Dedup, an external tool that removes duplicate tracks based on song name, artist, duration, and Spotify URI (Uniform Resource Identifier), was used to create a finalized unique list.

Since the survey was conducted in both 2023 and 2024, the unique songs from both years were combined into a single dataset. After merging the playlists, Spotify Dedup was used again to remove duplicates between years. The final included dataset consisted of 2,137 unique songs reported as top five favorites by Union students over the two-year period.

Comparison Dataset

The non-included dataset was built using the "Top Spotify Songs in 73 Countries" dataset from Kaggle, a publicly available dataset that tracks globally popular songs. This dataset was chosen because it provided a broad and diverse sample of frequently streamed music, ensuring that the study captured a wide variety of popular tracks rather than being limited to a single country or region.

Before being used for analysis, the dataset underwent several filtering steps. First, duplicate songs were removed based on name, artist, and duration to ensure each track appeared only once. Next, any songs released after the survey years were excluded, as they would not have been eligible for students' year-end playlists and could not be fairly compared to the included dataset. Finally, the dataset was shuffled to eliminate country-based grouping effects that could introduce bias.

While the full dataset contained over 18,000 unique songs, practical constraints required limiting the selection. Due to Spotify API playlist upload limits and the need for a balanced sample, only 5,000 songs from the filtered dataset were uploaded to Spotify. After confirming that no duplicates remained, this non-included set was split in half to create two separate playlists: one for the main regression analysis and another for an alternate regression used to verify dataset consistency. Chosic's Playlist Analyzer, an online tool that extracts Spotify audio features, was then used to retrieve the necessary song characteristics via CSV file for analysis.

To ensure a fair comparison, both playlists were finalized at 4,274 songs, maintaining a 50-50 split between included and non-included songs. This dataset structure ensured that regression results were not influenced by unequal sample sizes.

The response variable, UnionRewindInclude, is a binary indicator where 1 represents a song included in a Union student's top five, and 0 represents a song from the globally popular non-included dataset. This variable serves as the dependent variable in the regression model.

Song Characteristics

Spotify quantifies various song characteristics through its Audio Features. These features translate subjective musical elements into numerical data, making them useful for statistical analysis. The primary features used in this study include:

- Danceability: Measures how suitable a track is for dancing based on tempo, rhythm stability, and beat strength.
- Energy: Represents the intensity and activity level of a song, influenced by volume, tempo, and dynamic range.
- Acousticness: Estimates the likelihood that a song is primarily acoustic, with higher values indicating a more organic sound.
- Instrumentalness: Predicts whether a track contains vocals, with higher values indicating fewer or no vocals.
- Speechiness: Identifies the presence of spoken words, with higher values indicating more speech-like content.
- Happiness/Valence: Measures the overall positivity of a track. Higher values indicate happier, more cheerful, and euphoric songs, while lower values suggest sadness, melancholy, or tension.
- Live: Assesses whether a track was recorded live, with higher values suggesting more audience noise or live performance elements.

While Spotify originally scales these features from 0 to 1, this study uses data processed through Chosic's Playlist Analyzer, which normalizes the values to a 0 to 100 scale for easier interpretation.

In addition to these core features, popularity and loudness were also considered but differed in their scoring methods:

- Popularity: A dynamic score that changes over time based on a song's streaming activity rather than a fixed characteristic of the song itself.
- Loudness: Measured in decibels, loudness represents a song's average volume.

These features were chosen because they provide an objective way to evaluate song characteristics beyond subjective genre labels. Since the goal was to determine whether Union students favor specific types of songs, these attributes serve as the foundation for the regression analysis.

Bias Considerations

While every effort was made to ensure an unbiased dataset, certain factors may have influenced the final sample. Since the survey was conducted during common hour at dining halls, students who typically ate elsewhere or had commitments at that time may be underrepresented, potentially skewing the sample toward those with more flexible routines or social tendencies.

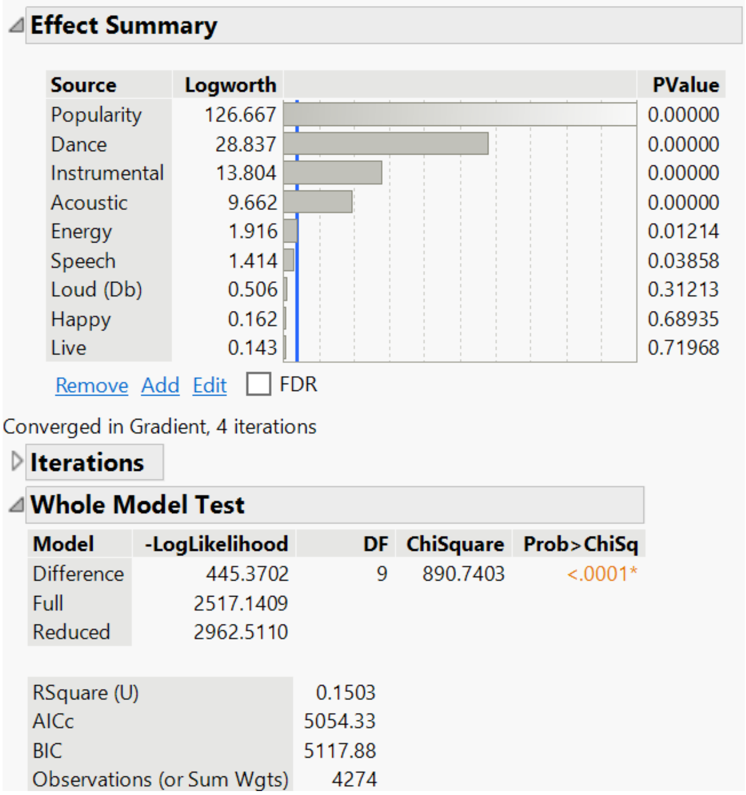
Participation was voluntary, and while some students declined, there is no evidence suggesting their preferences differed systematically from those who participated. Some students were initially hesitant to list non-English songs, possibly due to social influence, but they were encouraged to report their genuine preferences. Students who shared streaming accounts with family members were asked to select five songs that best reflected their listening habits. While shared accounts could introduce some distortion, participants were encouraged to choose tracks that genuinely represented their tastes.

While the non-included dataset captures a broad sample of popular music worldwide, it does not encompass all regional or niche listening habits. However, it serves as a reasonable benchmark for understanding mainstream trends and provides a structured comparison for the regression analysis.

Model Development and Validation

The goal of this analysis was to determine whether Union students' musical preferences, as captured through the WRUC Rewind survey, could be distinguished from a globally popular set of songs based on measurable audio features. The primary method used for this study was logistic regression, where the response variable indicated whether a song was included in a Union student's top five most played tracks. The model development followed a sequential process, beginning with an initial full model, assessing feature contributions, testing alternative datasets, and ultimately refining the model for interpretability. Each step aimed to balance explanatory power with statistical rigor.

Initial Model: Full Feature Set



◿ **Effect Summary**

Source	Logworth	PValue
Popularity	126.667	0.00000
Dance	28.837	0.00000
Instrumental	13.804	0.00000
Acoustic	9.662	0.00000
Energy	1.916	0.01214
Speech	1.414	0.03858
Loud (Db)	0.506	0.31213
Happy	0.162	0.68935
Live	0.143	0.71968

Remove Add Edit □ FDR

Converged in Gradient, 4 iterations

▷ **Iterations**

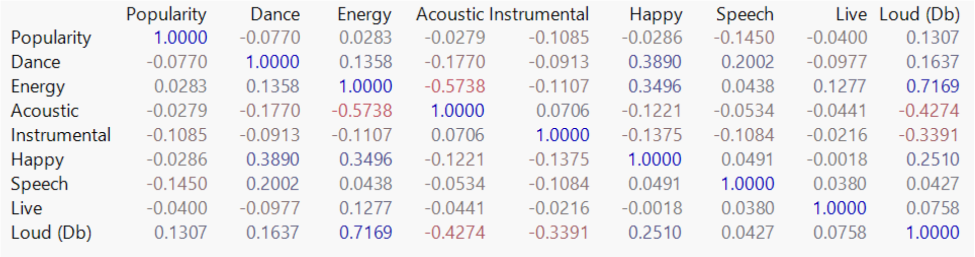
◿ **Whole Model Test**

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	445.3702	9	890.7403	<.0001*
Full	2517.1409			
Reduced	2962.5110			

RSquare (U)	0.1503
AICc	5054.33
BIC	5117.88
Observations (or Sum Wgts)	4274

Figure 1: Effect Summary & Whole Model Test of Full Feature Dataset Regression

The first regression model included all Spotify audio features along with popularity and loudness. The initial results demonstrated that popularity had an overwhelming influence on classification, with an extremely high log-odds coefficient and a dominant effect in the effect summary chart. While popularity is an important factor in music consumption, it is a dynamic measure that changes over time and reflects global listening habits rather than intrinsic song characteristics. Including popularity made it difficult to isolate the impact of other musical features, so it was removed in the next model.



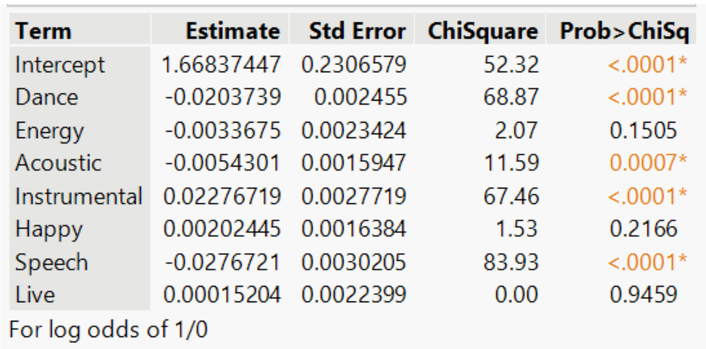
	Popularity	Dance	Energy	Acoustic	Instrumental	Happy	Speech	Live	Loud (Db)
Popularity	1.0000	-0.0770	0.0283	-0.0279	-0.1085	-0.0286	-0.1450	-0.0400	0.1307
Dance	-0.0770	1.0000	0.1358	-0.1770	-0.0913	0.3890	0.2002	-0.0977	0.1637
Energy	0.0283	0.1358	1.0000	-0.5738	-0.1107	0.3496	0.0438	0.1277	0.7169
Acoustic	-0.0279	-0.1770	-0.5738	1.0000	0.0706	-0.1221	-0.0534	-0.0441	-0.4274
Instrumental	-0.1085	-0.0913	-0.1107	0.0706	1.0000	-0.1375	-0.1084	-0.0216	-0.3391
Happy	-0.0286	0.3890	0.3496	-0.1221	-0.1375	1.0000	0.0491	-0.0018	0.2510
Speech	-0.1450	0.2002	0.0438	-0.0534	-0.1084	0.0491	1.0000	0.0380	0.0427
Live	-0.0400	-0.0977	0.1277	-0.0441	-0.0216	-0.0018	0.0380	1.0000	0.0758
Loud (Db)	0.1307	0.1637	0.7169	-0.4274	-0.3391	0.2510	0.0427	0.0758	1.0000

Figure 2: Correlation Matrix of Full Feature Dataset

Loudness was also assessed for its contribution to the model. The multicollinearity analysis revealed that loudness was highly correlated with energy, with a correlation coefficient of 0.7169. Since energy captures the intensity of a track and is standardized on the same scale as the other audio features, loudness was determined to be redundant. A comparative model without loudness showed minimal impact on overall model performance, confirming that energy alone was sufficient for capturing song intensity. Given this redundancy, loudness was removed in the next iteration.

To ensure that the model was not overly dependent on the structure of the dataset, two alternative non-included datasets were tested. These tests helped confirm that the identified patterns in Union students' preferences were not driven by dataset selection bias, reinforcing the robustness of the final model.

- *Western-Focused Non-Included Dataset*



Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1.66837447	0.2306579	52.32	<.0001*
Dance	-0.0203739	0.002455	68.87	<.0001*
Energy	-0.0033675	0.0023424	2.07	0.1505
Acoustic	-0.0054301	0.0015947	11.59	0.0007*
Instrumental	0.02276719	0.0027719	67.46	<.0001*
Happy	0.00202445	0.0016384	1.53	0.2166
Speech	-0.0276721	0.0030205	83.93	<.0001*
Live	0.00015204	0.0022399	0.00	0.9459

For log odds of 1/0

Figure 3: Parameter Estimates of Western Country Dataset Regression

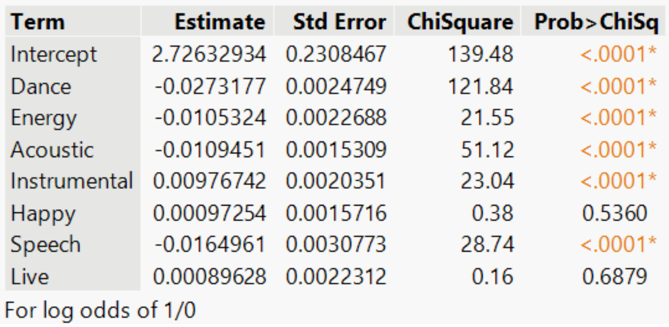
- To better reflect the listening environment of the majority of Union students, an alternative non-included dataset was created that focused on Western countries. This dataset included songs from the United States, Canada, the United Kingdom, Australia, Ireland, New Zealand, Sweden, Norway, Denmark, the Netherlands, Germany, and France. The goal was to reduce the influence of global music trends that may not be relevant to Union students.

  A more restrictive dataset consisting only of English-majority-speaking countries (United States, Canada, United Kingdom, Australia, Ireland, and New Zealand) was initially considered to further isolate cultural influences. However, this dataset contained too few unique songs for meaningful analysis. To ensure sufficient coverage while maintaining relevance, the broader Western-focused dataset was selected. The additional European countries were included because of their strong cultural ties to English-language music and their influence on Western streaming trends.

When the regression was run using this refined dataset, most key predictors remained consistent with the original model, reinforcing the stability of the main findings. However, energy, which was significant in the full model, lost significance in the Western-focused dataset when loudness was removed. This suggests that energy's effect may depend on its interaction with other variables, particularly loudness, and that its predictive power is somewhat dataset dependent.

Despite this variation, the overall trends remained intact. Danceability, acousticness, instrumentalness, and speechiness continued to play significant roles in distinguishing Union students' preferences from broader streaming trends. This dataset served as a secondary reference, confirming that Union students' preferences differ not just from global trends but also from Western streaming patterns.

- *Alternative Non-Included Dataset Comparison*



Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.72632934	0.2308467	139.48	<.0001*
Dance	-0.0273177	0.0024749	121.84	<.0001*
Energy	-0.0105324	0.0022688	21.55	<.0001*
Acoustic	-0.0109451	0.0015309	51.12	<.0001*
Instrumental	0.00976742	0.0020351	23.04	<.0001*
Happy	0.00097254	0.0015716	0.38	0.5360
Speech	-0.0164961	0.0030773	28.74	<.0001*
Live	0.00089628	0.0022312	0.16	0.6879

For log odds of 1/0

Figure 4: Parameter Estimates of Alternative Reduced Dataset Regression

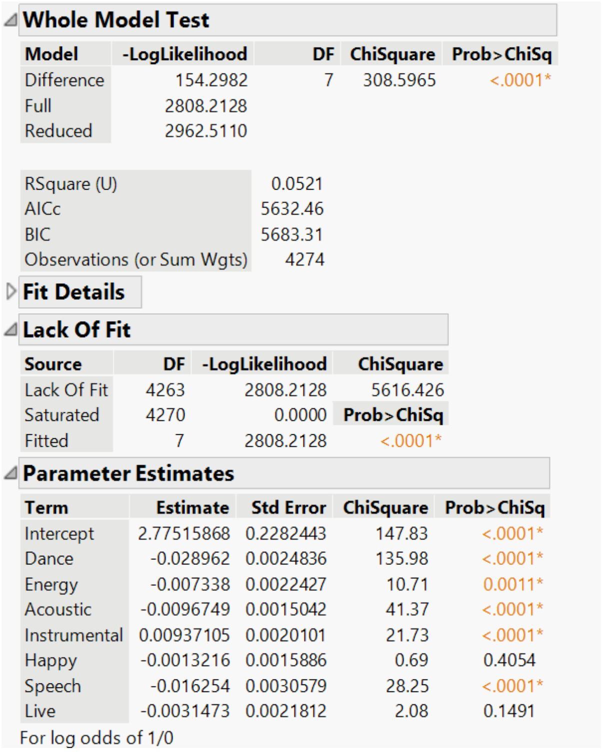
To further test the robustness of the findings, an alternative non-included dataset was used. This dataset was drawn from the same global top songs source but included a different random subset of tracks. Running the regression with this dataset ensured that selection effects were not driving the observed trends.

The results remained largely consistent with the full model, reinforcing its stability. One notable difference was in the significance of energy. In the full model, energy was weakly significant (p = 0.0122), suggesting it played a secondary role in predicting song inclusion compared to other predictors like danceability and acousticness.

However, in the alternative dataset, energy became more significant (p < 0.0001). This suggests that while energy is not the dominant factor, it still contributes meaningfully to Union students' music preferences, particularly in relation to other song characteristics.

Despite this shift, the overall trends in Union students' musical preferences remained intact, further supporting the validity of the model.

Final Model: Removing Popularity and Loudness



Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	154.2982	7	308.5965	<.0001*
Full	2808.2128			
Reduced	2962.5110			

RSquare (U)	0.0521
AICc	5632.46
BIC	5683.31
Observations (or Sum Wgts)	4274

Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	4263	2808.2128	5616.426
Saturated	4270	0.0000	Prob>ChiSq
Fitted	7	2808.2128	<.0001*

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	2.77515868	0.2282443	147.83	<.0001*
Dance	-0.028962	0.0024836	135.98	<.0001*
Energy	-0.007338	0.0022427	10.71	0.0011*
Acoustic	-0.0096749	0.0015042	41.37	<.0001*
Instrumental	0.00937105	0.0020101	21.73	<.0001*
Happy	-0.0013216	0.0015886	0.69	0.4054
Speech	-0.016254	0.0030579	28.25	<.0001*
Live	-0.0031473	0.0021812	2.08	0.1491

For log odds of 1/0

Figure 5: Results of Final Reduced Dataset Regression

With popularity and loudness removed, the final model included the following predictor variables:

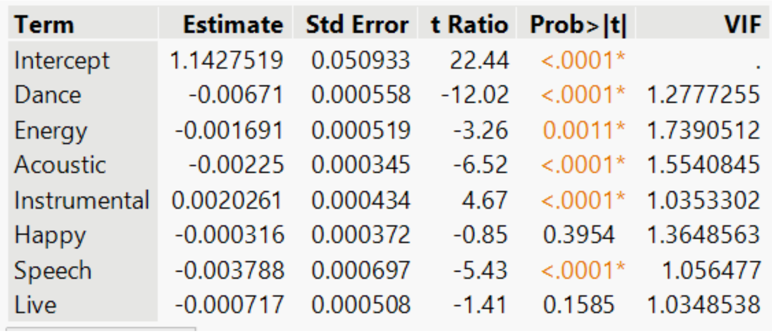
- Danceability

- Energy
- Acousticness
- Instrumentalness
- Happiness (Valence)
- Speechiness
- Live

Each of these features quantifies a fundamental aspect of a song’s composition. The final model was evaluated for goodness of fit and interpretability, with an emphasis on identifying meaningful trends in Union students' song preferences rather than maximizing predictive accuracy.

Model diagnostics confirmed that this refined model provided a reasonable balance between explanatory power and simplicity. While the pseudo-R-squared value (0.0521) was lower than that of the full model (0.1503 – See Figure 1), this was expected due to the removal of popularity, which accounted for a substantial portion of the variance. However, pseudo-R-squared is not directly comparable to R-squared in ordinary least squares regression and tends to be low in logistic regression, particularly when modeling complex human behaviors like music preference. Thus, it was not given significant weight in model selection.

In contrast, AIC and BIC values were higher in the reduced model (AIC = 5632.46, BIC = 5683.31) compared to the full model (AIC = 5054.33, BIC = 5117.88 – See Figure 1). Although lower AIC/BIC values typically indicate a better trade-off between model fit and complexity, these differences highlight how popularity dominated the full model’s explanatory power, inflating its overall fit statistics. By removing popularity and loudness, the reduced model offers a more interpretable structure that isolates the effects of intrinsic song characteristics rather than relying on external popularity-driven trends.



Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	1.1427519	0.050933	22.44	<.0001*	.
Dance	-0.00671	0.000558	-12.02	<.0001*	1.2777255
Energy	-0.001691	0.000519	-3.26	0.0011*	1.7390512
Acoustic	-0.00225	0.000345	-6.52	<.0001*	1.5540845
Instrumental	0.0020261	0.000434	4.67	<.0001*	1.0353302
Happy	-0.000316	0.000372	-0.85	0.3954	1.3648563
Speech	-0.003788	0.000697	-5.43	<.0001*	1.056477
Live	-0.000717	0.000508	-1.41	0.1585	1.0348538

Figure 6: Parameter Estimates of Reduced Dataset Least Squares Regression with VIF

Variance inflation factors (VIF) from a corresponding least squares regression confirmed that multicollinearity was not a concern. All predictor variables had VIF values below 2, well below the conventional threshold of 5. This suggests that the remaining features contribute independently to song inclusion likelihood, reinforcing the robustness of the reduced model.

Finally, the Lack of Fit test indicated statistical significance ($p < 0.0001$) in both the full and reduced models. While this suggests that additional unmeasured factors influence song preference, it does not undermine the model's ability to identify meaningful trends in the data.

Traditional classification-based diagnostics such as misclassification rate, confusion matrices, and classification accuracy are not well-suited to this analysis. Since the goal is to uncover patterns in music preference rather than predict song inclusion with high accuracy, these metrics are less informative. Unlike predictive modeling applications, where classification accuracy is a key objective, this study focuses on how specific song characteristics influence inclusion likelihood rather than building a predictive tool. Given the subjective nature of human taste, pseudo-R-squared and misclassification rates are inherently weak indicators of explanatory power in this context. While logistic regression outputs probabilities, applying an arbitrary classification threshold (e.g., 0.5) would not yield meaningful insights into Union students' listening habits. Instead, the emphasis remains on identifying the statistical significance and direction of song attributes that impact inclusion.

Given the results of these validation tests, the final reduced model was selected for interpretation as it provides a balance of statistical robustness, interpretability, and dataset stability, ensuring meaningful insights into Union students' preferences.

Interpretation of Results

Among the predictors, danceability has the strongest negative effect, with a coefficient of -0.02896 (p < 0.0001). Ceteris paribus, for every one-unit increase in danceability, the log-odds of a song being included in a student's top five decrease by 0.02896, corresponding to an approximate 2.85% decrease in the odds of inclusion. This suggests that highly danceable music, despite its prominence in social settings, is actively avoided in students' personal listening habits. This distinction may reflect a separation between music selected for social interaction and music enjoyed individually.

Acousticness also has a significant negative effect (-0.00967, p < 0.0001). Ceteris paribus, a one-unit increase in acousticness decreases the log-odds of inclusion by 0.00967, corresponding to an approximate 0.97% decrease in the odds of inclusion. This aligns with the observation that students favor more polished, produced tracks over raw acoustic compositions, despite artists like Zach Bryan maintaining a presence in Union's social spaces.

Speechiness has a negative effect as well (-0.01624 per unit increase, p < 0.0001). Ceteris paribus, a one-unit increase in speechiness decreases the log-odds of inclusion by 0.01624, reducing the odds of inclusion by 1.62%. This suggests that tracks with prominent spoken elements, such as rap or spoken-word passages, are less likely to be among a student's most played songs.

Energy has a weaker but still significant negative effect (-0.0073, p = 0.0011). Ceteris paribus, a one-unit increase in energy decreases the log-odds of inclusion by 0.0073, corresponding to an approximate 0.73% decrease in the odds of inclusion. This result suggests that while high-energy music is a key part of college party culture, Union students may favor more dynamically varied or moderately energetic music in their personal listening. The fact that energy's effect is weaker and dataset-dependent (as shown in alternative dataset testing) suggests that its influence on preference is more situational and may depend on interactions with other variables.

Instrumentalness is the only variable with a positive relationship to inclusion (0.00937, p < 0.0001). Ceteris paribus, a one-unit increase in instrumentalness increases the log-odds of inclusion by 0.00937, corresponding to an approximate 0.94% increase in the odds of inclusion. This suggests that students may be slightly more likely to include music with instrumental passages in their top five, indicating a slight preference for instrumental elements, particularly when contrasted with the strong negative effect of speechiness. This finding is particularly notable given the negative effect of speechiness, further reinforcing the trend that Union students favor structured, vocal-driven songs while avoiding speech-heavy tracks.

Limitations and Next Steps

While the model identifies statistically significant predictors of song inclusion, it does not fully capture the complexity of music preference. Factors such as genre, lyrical themes, and cultural background are not explicitly accounted for in Spotify's numerical audio features. Although removing popularity and loudness improved the interpretability of other predictors, the model remains limited in its ability to quantify subjective and contextual factors that shape personal music selection. Future work could integrate additional metadata, such as Spotify's genre classifications or user-generated tags, to refine preference modeling and provide better contextualization.

Further refinements to the regression model could explore interaction terms to assess whether specific audio features jointly influence song inclusion. Additionally, polynomial regression could help determine whether nonlinear relationships exist between variables such as danceability and inclusion likelihood. Expanding the dataset to include multiple years of WRUC Rewind surveys would provide insight into how Union students' musical preferences evolve over time while also increasing statistical power and reducing the influence of year-specific anomalies.

While popularity was removed due to its overwhelming influence, future models should explore ways to account for popularity without allowing it to dominate the analysis. One effective approach would be to categorize songs by popularity tiers (e.g., widely recognized hits vs. niche selections) to determine if its effect differs across levels of mainstream exposure. This would allow for a more nuanced understanding of how popularity influences listening habits without overpowering other factors.

Conclusion

This study examined whether Union College students have distinct musical preferences that can be quantified using logistic regression. The results identified clear trends, with danceability, energy, acousticness, instrumentalness, and speechiness emerging as statistically significant predictors. Danceability had the strongest negative effect, suggesting that highly danceable music, despite its prominence in social settings, is less favored in students' individual listening choices.

These findings highlight distinct listening patterns within the Union student body but also emphasize the challenge of modeling human taste. Even with multiple significant predictors, the model's explanatory power remained low, reflecting the inherent complexity and subjectivity of music preferences. While measurable song characteristics provide insight into listening habits, they cannot fully account for the social, emotional, and cultural factors that shape individual music choices.

From a methodological standpoint, this project reinforced the importance of data preparation in regression analysis. Structuring and cleaning the dataset required significantly more effort than running the regression itself, particularly due to challenges with deprecated Spotify API commands. This highlights a broader reality: data preprocessing is often the most time-consuming and critical stage of a regression-based project, particularly when working with real-world, unstructured datasets.

Ultimately, this study demonstrated that while logistic regression can identify meaningful patterns in music preference, fully capturing personal taste through statistical modeling remains a complex challenge. While this research focused on Union College, its findings provide broader insight into how measurable song attributes influence music preference in social settings. The results reaffirm that while statistical models can reveal trends, music preference remains deeply personal and difficult to reduce to numerical features alone.

References

Bonneville-Roussy, A., Rentfrow, P., Xu, K., & Potter, J. (2013). Music Through the Ages: Trends in Musical Engagement and Preferences From Adolescence Through Middle Adulthood. *Journal of Personality and Social Psychology*, *105*. https://doi.org/10.1037/a0033770

JMP documentation. (n.d.). Retrieved March 14, 2025, from https://www.jmp.com/en/support/jmp-documentation

Spotify Dedup—Remove duplicate songs from your Spotify library. (n.d.). Retrieved March 14, 2025, from https://spotify-dedup.com/

Spotify Playlist Analyzer. (n.d.). Chosic. Retrieved March 14, 2025, from https://www.chosic.com/spotify-playlist-analyzer/

Top Spotify Songs in 73 Countries (Daily Updated). (n.d.). Retrieved March 14, 2025, from https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated

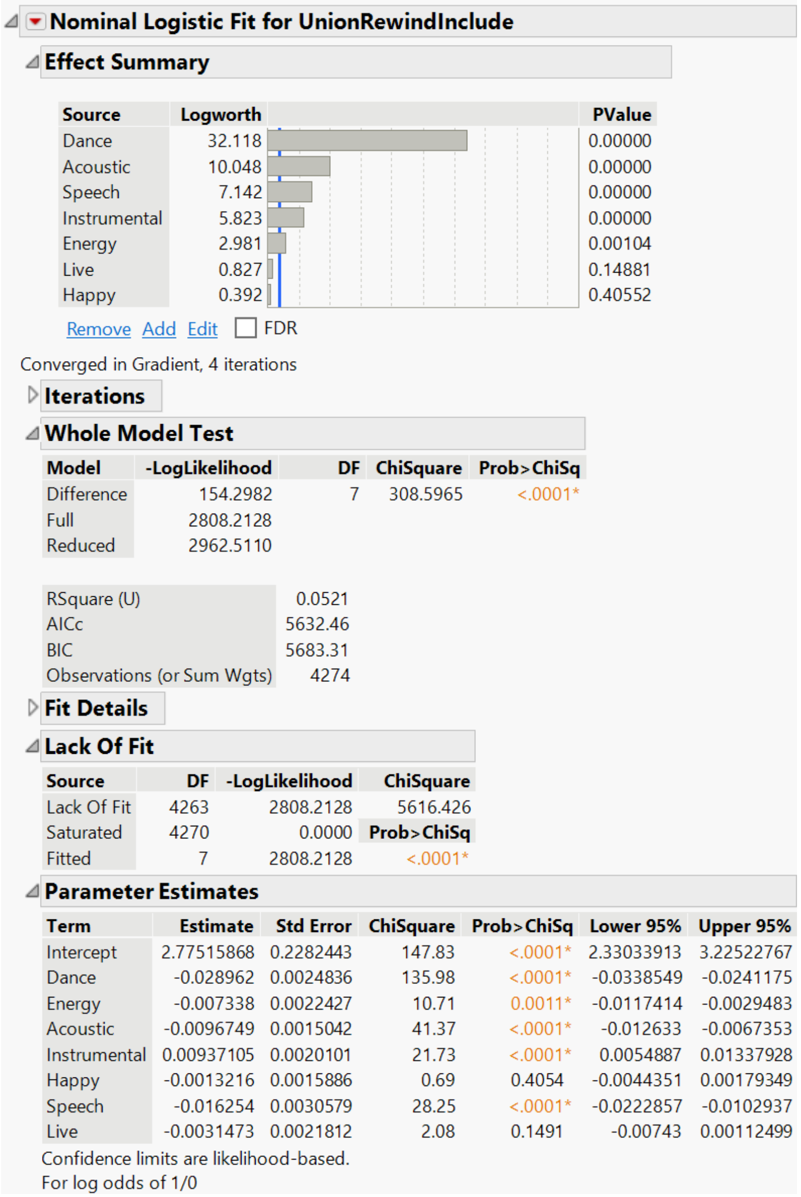
Union at a Glance | Admissions | Union College. (n.d.). Retrieved March 14, 2025, from https://www.union.edu/admissions/union

Web API Reference | Spotify for Developers. (n.d.). Retrieved March 14, 2025, from https://developer.spotify.com/documentation/web-api/reference/get-audio-features

WRUC 89.7FM @ Union College. (n.d.). Spotify. Retrieved March 14, 2025, from https://spotify.com/user/wruc89.7fm

Appendix

Final Reduced Regression Model Results



Nominal Logistic Fit for UnionRewindInclude

Effect Summary

Source	Logworth	PValue
Dance	32.118	0.00000
Acoustic	10.048	0.00000
Speech	7.142	0.00000
Instrumental	5.823	0.00000
Energy	2.981	0.00104
Live	0.827	0.14881
Happy	0.392	0.40552

Remove Add Edit □ FDR

Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	154.2982	7	308.5965	<.0001*
Full	2808.2128			
Reduced	2962.5110			

RSquare (U)	0.0521
AICc	5632.46
BIC	5683.31
Observations (or Sum Wgts)	4274

Fit Details

Lack Of Fit

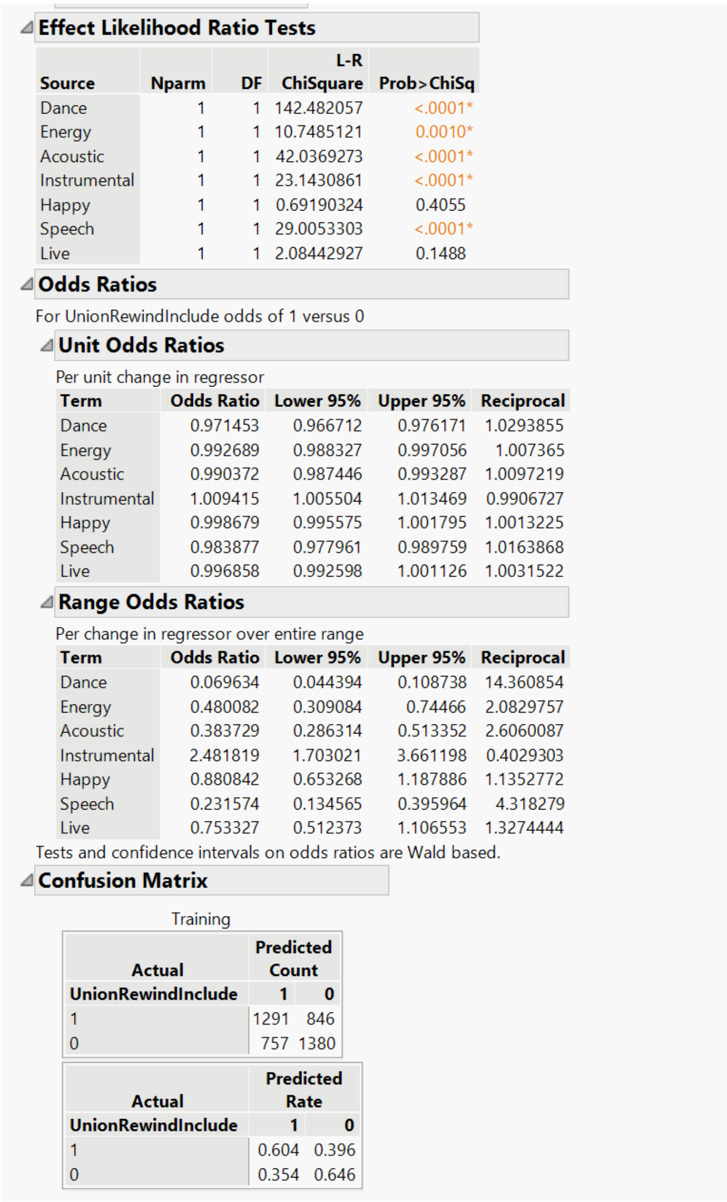
Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	4263	2808.2128	5616.426
Saturated	4270	0.0000	Prob>ChiSq
Fitted	7	2808.2128	<.0001*

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept	2.77515868	0.2282443	147.83	<.0001*	2.33033913	3.22522767
Dance	-0.028962	0.0024836	135.98	<.0001*	-0.0338549	-0.0241175
Energy	-0.007338	0.0022427	10.71	0.0011*	-0.0117414	-0.0029483
Acoustic	-0.0096749	0.0015042	41.37	<.0001*	-0.012633	-0.0067353
Instrumental	0.00937105	0.0020101	21.73	<.0001*	0.0054887	0.01337928
Happy	-0.0013216	0.0015886	0.69	0.4054	-0.0044351	0.00179349
Speech	-0.016254	0.0030579	28.25	<.0001*	-0.0222857	-0.0102937
Live	-0.0031473	0.0021812	2.08	0.1491	-0.00743	0.00112499

Confidence limits are likelihood-based.

For log odds of 1/0



Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Dance	1	1	142.482057	<.0001*
Energy	1	1	10.7485121	0.0010*
Acoustic	1	1	42.0369273	<.0001*
Instrumental	1	1	23.1430861	<.0001*
Happy	1	1	0.69190324	0.4055
Speech	1	1	29.0053303	<.0001*
Live	1	1	2.08442927	0.1488

Odds Ratios

For UnionRewindInclude odds of 1 versus 0

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Dance	0.971453	0.966712	0.976171	1.0293855
Energy	0.992689	0.988327	0.997056	1.007365
Acoustic	0.990372	0.987446	0.993287	1.0097219
Instrumental	1.009415	1.005504	1.013469	0.9906727
Happy	0.998679	0.995575	1.001795	1.0013225
Speech	0.983877	0.977961	0.989759	1.0163868
Live	0.996858	0.992598	1.001126	1.0031522

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Dance	0.069634	0.044394	0.108738	14.360854
Energy	0.480082	0.309084	0.74466	2.0829757
Acoustic	0.383729	0.286314	0.513352	2.6060087
Instrumental	2.481819	1.703021	3.661198	0.4029303
Happy	0.880842	0.653268	1.187886	1.1352772
Speech	0.231574	0.134565	0.395964	4.318279
Live	0.753327	0.512373	1.106553	1.3274444

Tests and confidence intervals on odds ratios are Wald based.

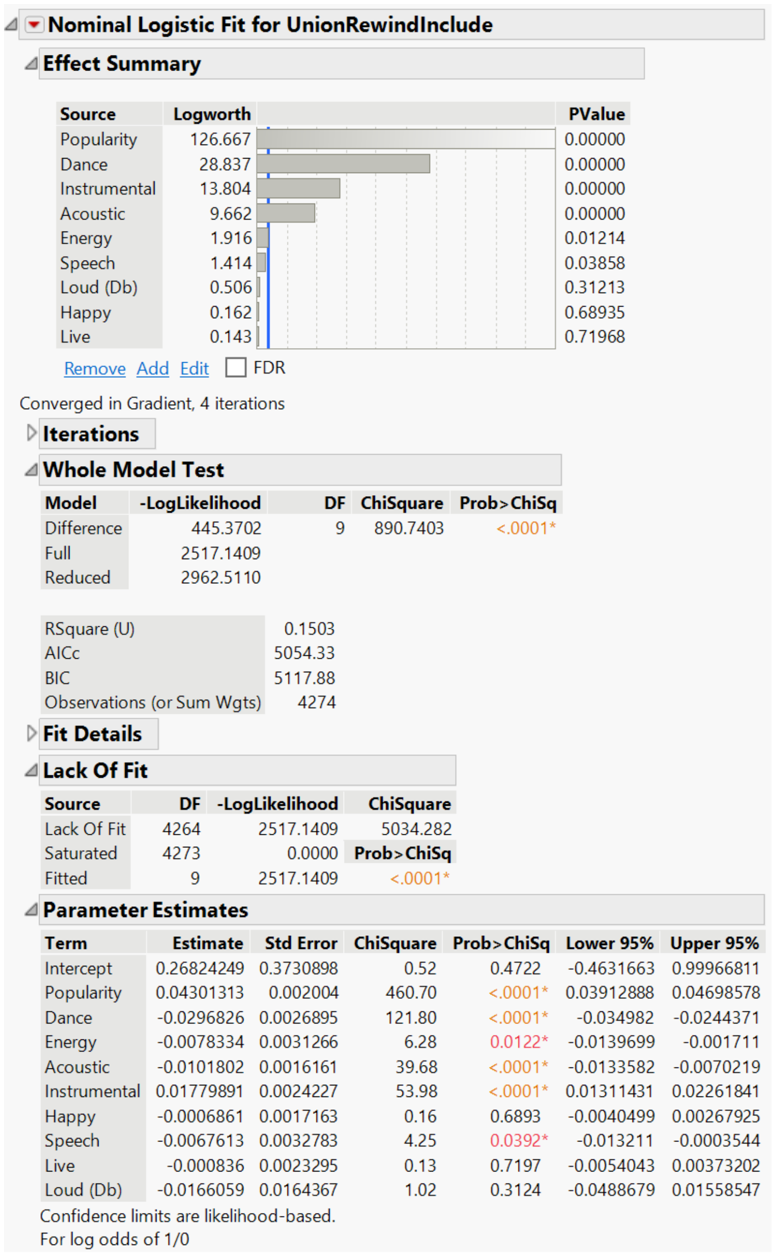
Confusion Matrix

Training

Actual UnionRewindInclude	Predicted Count 1	 0
1	1291	846
0	757	1380

Actual UnionRewindInclude	Predicted Rate 1	 0
1	0.604	0.396
0	0.354	0.646

Full Regression Model Results



Nominal Logistic Fit for UnionRewindInclude

Effect Summary

Source	Logworth	PValue
Popularity	126.667	0.00000
Dance	28.837	0.00000
Instrumental	13.804	0.00000
Acoustic	9.662	0.00000
Energy	1.916	0.01214
Speech	1.414	0.03858
Loud (Db)	0.506	0.31213
Happy	0.162	0.68935
Live	0.143	0.71968

Remove Add Edit □ FDR

Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	445.3702	9	890.7403	<.0001*
Full	2517.1409			
Reduced	2962.5110			

RSquare (U)	0.1503
AICc	5054.33
BIC	5117.88
Observations (or Sum Wgts)	4274

Fit Details

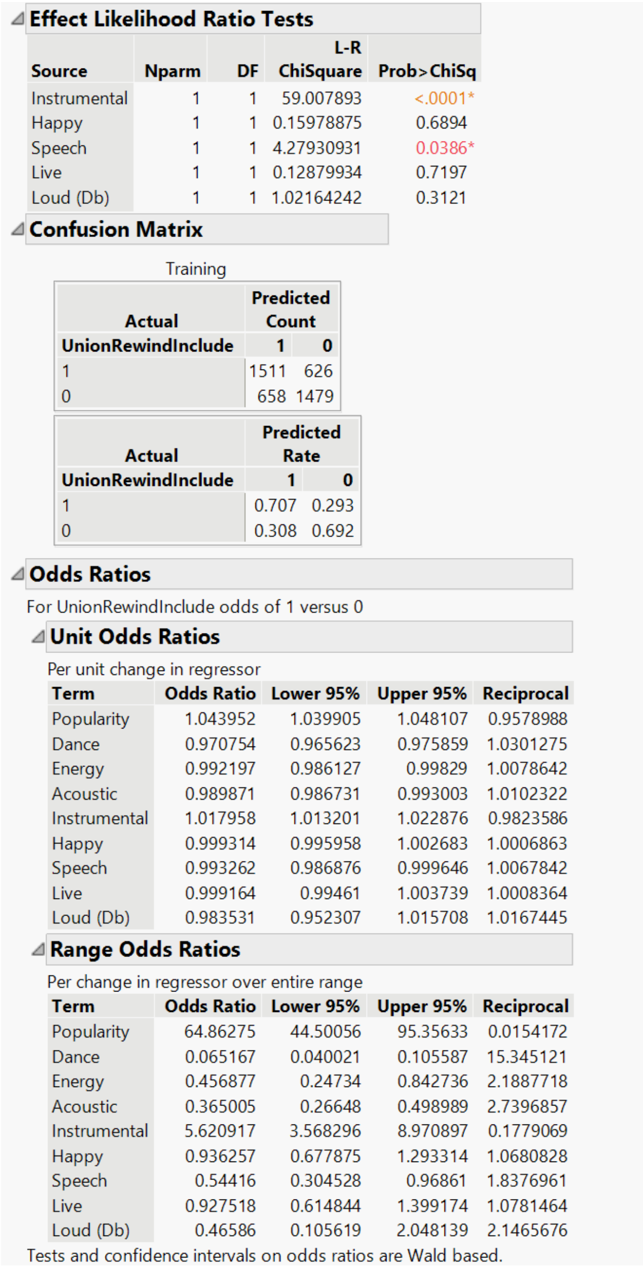
Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	4264	2517.1409	5034.282
Saturated	4273	0.0000	Prob>ChiSq
Fitted	9	2517.1409	<.0001*

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept	0.26824249	0.3730898	0.52	0.4722	-0.4631663	0.99966811
Popularity	0.04301313	0.002004	460.70	<.0001*	0.03912888	0.04698578
Dance	-0.0296826	0.0026895	121.80	<.0001*	-0.034982	-0.0244371
Energy	-0.0078334	0.0031266	6.28	0.0122*	-0.0139699	-0.001711
Acoustic	-0.0101802	0.0016161	39.68	<.0001*	-0.0133582	-0.0070219
Instrumental	0.01779891	0.0024227	53.98	<.0001*	0.01311431	0.02261841
Happy	-0.0006861	0.0017163	0.16	0.6893	-0.0040499	0.00267925
Speech	-0.0067613	0.0032783	4.25	0.0392*	-0.013211	-0.0003544
Live	-0.000836	0.0023295	0.13	0.7197	-0.0054043	0.00373202
Loud (Db)	-0.0166059	0.0164367	1.02	0.3124	-0.0488679	0.01558547

Confidence limits are likelihood-based.
For log odds of 1/0



Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Instrumental	1	1	59.007893	<.0001*
Happy	1	1	0.15978875	0.6894
Speech	1	1	4.27930931	0.0386*
Live	1	1	0.12879934	0.7197
Loud (Db)	1	1	1.02164242	0.3121

Confusion Matrix

Training

Actual	Predicted Count	
UnionRewindInclude	1	0
1	1511	626
0	658	1479

Actual	Predicted Rate	
UnionRewindInclude	1	0
1	0.707	0.293
0	0.308	0.692

Odds Ratios

For UnionRewindInclude odds of 1 versus 0

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Popularity	1.043952	1.039905	1.048107	0.9578988
Dance	0.970754	0.965623	0.975859	1.0301275
Energy	0.992197	0.986127	0.99829	1.0078642
Acoustic	0.989871	0.986731	0.993003	1.0102322
Instrumental	1.017958	1.013201	1.022876	0.9823586
Happy	0.999314	0.995958	1.002683	1.0006863
Speech	0.993262	0.986876	0.999646	1.0067842
Live	0.999164	0.99461	1.003739	1.0008364
Loud (Db)	0.983531	0.952307	1.015708	1.0167445

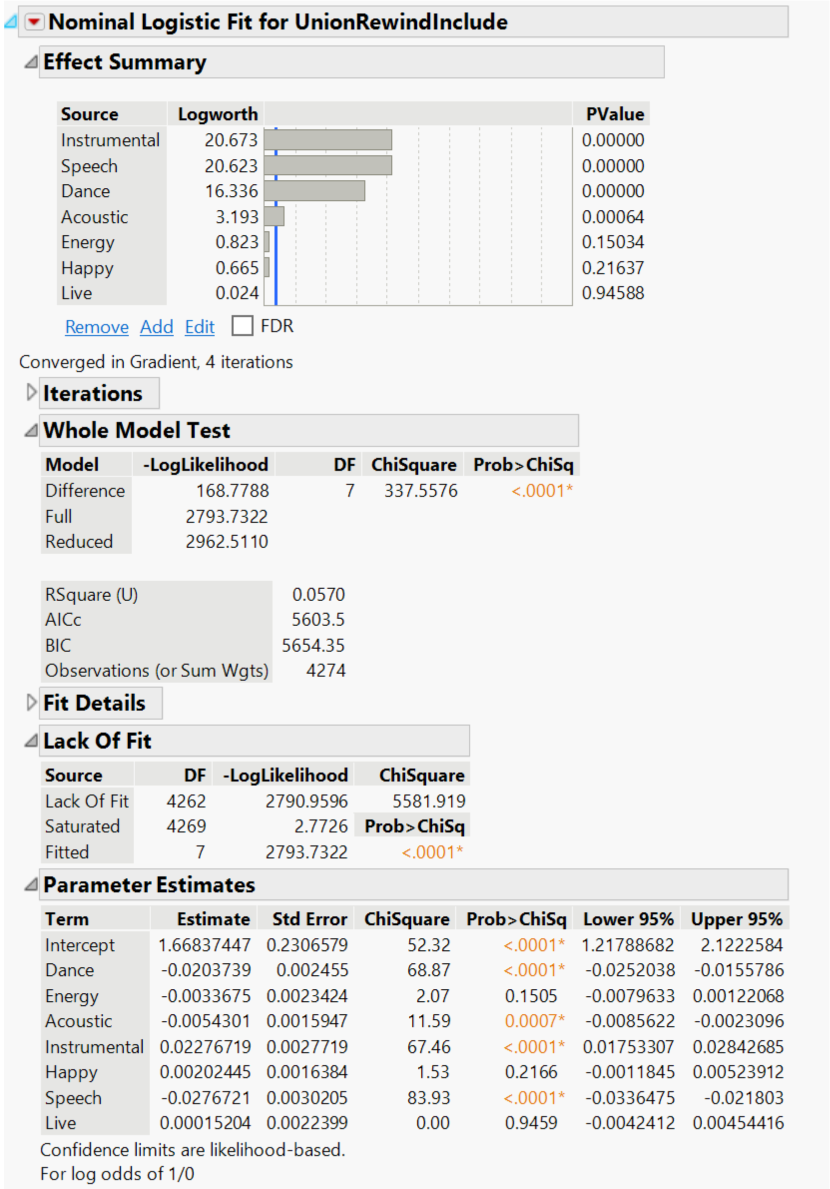
Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Popularity	64.86275	44.50056	95.35633	0.0154172
Dance	0.065167	0.040021	0.105587	15.345121
Energy	0.456877	0.24734	0.842736	2.1887718
Acoustic	0.365005	0.26648	0.498989	2.7396857
Instrumental	5.620917	3.568296	8.970897	0.1779069
Happy	0.936257	0.677875	1.293314	1.0680828
Speech	0.54416	0.304528	0.96861	1.8376961
Live	0.927518	0.614844	1.399174	1.0781464
Loud (Db)	0.46586	0.105619	2.048139	2.1465676

Tests and confidence intervals on odds ratios are Wald based.

Reduced Western Countries Regression Model Results



Nominal Logistic Fit for UnionRewindInclude

Effect Summary

Source	Logworth	PValue
Instrumental	20.673	0.00000
Speech	20.623	0.00000
Dance	16.336	0.00000
Acoustic	3.193	0.00064
Energy	0.823	0.15034
Happy	0.665	0.21637
Live	0.024	0.94588

Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	168.7788	7	337.5576	<.0001*
Full	2793.7322			
Reduced	2962.5110			

RSquare (U)	0.0570
AICc	5603.5
BIC	5654.35
Observations (or Sum Wgts)	4274

Fit Details

Lack Of Fit

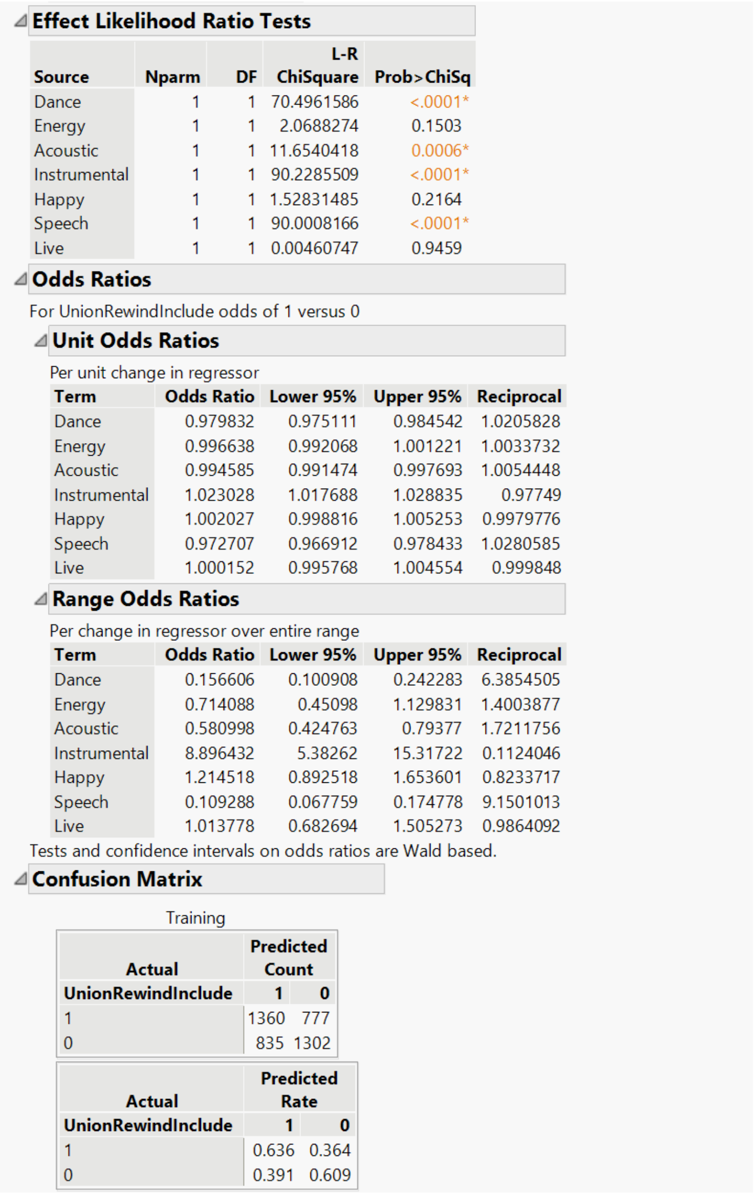
Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	4262	2790.9596	5581.919
Saturated	4269	2.7726	Prob>ChiSq
Fitted	7	2793.7322	<.0001*

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept	1.66837447	0.2306579	52.32	<.0001*	1.21788682	2.1222584
Dance	-0.0203739	0.002455	68.87	<.0001*	-0.0252038	-0.0155786
Energy	-0.0033675	0.0023424	2.07	0.1505	-0.0079633	0.00122068
Acoustic	-0.0054301	0.0015947	11.59	0.0007*	-0.0085622	-0.0023096
Instrumental	0.02276719	0.0027719	67.46	<.0001*	0.01753307	0.02842685
Happy	0.00202445	0.0016384	1.53	0.2166	-0.0011845	0.00523912
Speech	-0.0276721	0.0030205	83.93	<.0001*	-0.0336475	-0.021803
Live	0.00015204	0.0022399	0.00	0.9459	-0.0042412	0.00454416

Confidence limits are likelihood-based.

For log odds of 1/0



Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Dance	1	1	70.4961586	<.0001*
Energy	1	1	2.0688274	0.1503
Acoustic	1	1	11.6540418	0.0006*
Instrumental	1	1	90.2285509	<.0001*
Happy	1	1	1.52831485	0.2164
Speech	1	1	90.0008166	<.0001*
Live	1	1	0.00460747	0.9459

Odds Ratios

For UnionRewindInclude odds of 1 versus 0

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Dance	0.979832	0.975111	0.984542	1.0205828
Energy	0.996638	0.992068	1.001221	1.0033732
Acoustic	0.994585	0.991474	0.997693	1.0054448
Instrumental	1.023028	1.017688	1.028835	0.97749
Happy	1.002027	0.998816	1.005253	0.9979776
Speech	0.972707	0.966912	0.978433	1.0280585
Live	1.000152	0.995768	1.004554	0.999848

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Dance	0.156606	0.100908	0.242283	6.3854505
Energy	0.714088	0.45098	1.129831	1.4003877
Acoustic	0.580998	0.424763	0.79377	1.7211756
Instrumental	8.896432	5.38262	15.31722	0.1124046
Happy	1.214518	0.892518	1.653601	0.8233717
Speech	0.109288	0.067759	0.174778	9.1501013
Live	1.013778	0.682694	1.505273	0.9864092

Tests and confidence intervals on odds ratios are Wald based.

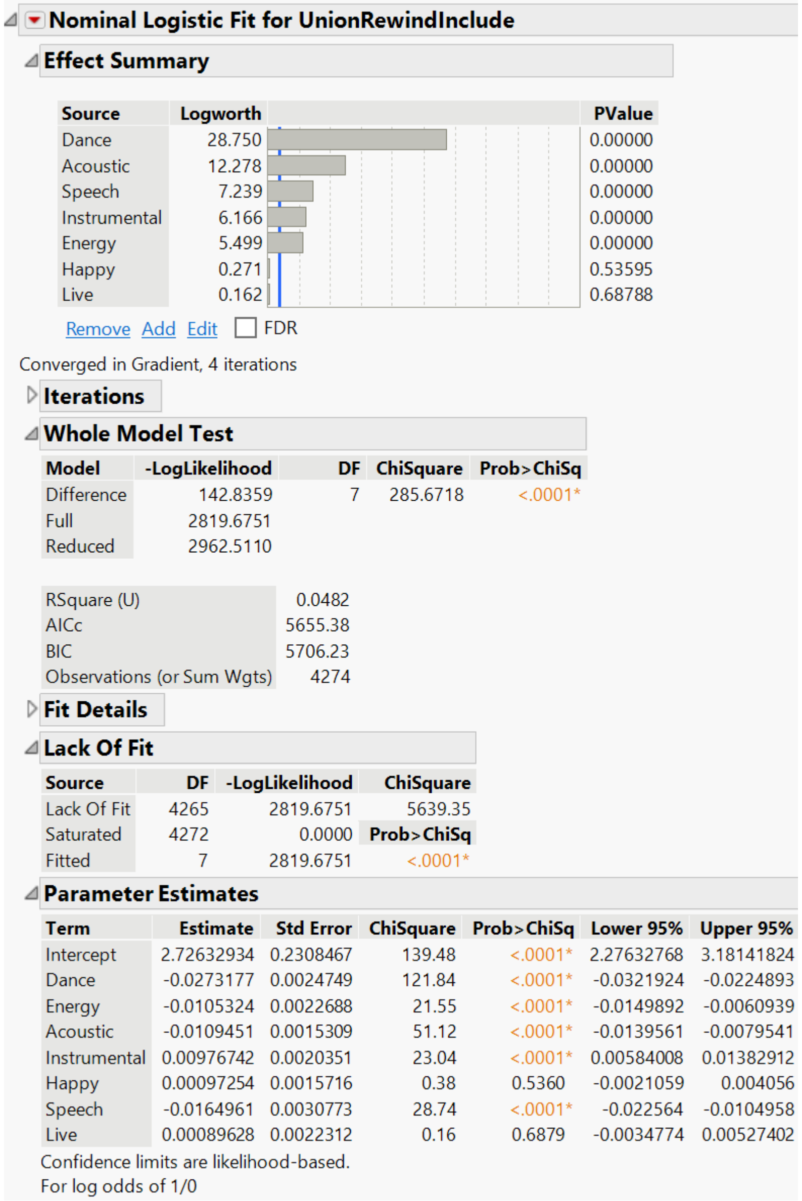
Confusion Matrix

Training

Actual	Predicted Count	
UnionRewindInclude	1	0
1	1360	777
0	835	1302

Actual	Predicted Rate	
UnionRewindInclude	1	0
1	0.636	0.364
0	0.391	0.609

Reduced Alternate Regression Model Results



Nominal Logistic Fit for UnionRewindInclude

Effect Summary

Source	Logworth	PValue
Dance	28.750	0.00000
Acoustic	12.278	0.00000
Speech	7.239	0.00000
Instrumental	6.166	0.00000
Energy	5.499	0.00000
Happy	0.271	0.53595
Live	0.162	0.68788

Remove Add Edit □ FDR

Converged in Gradient, 4 iterations

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	142.8359	7	285.6718	<.0001*
Full	2819.6751			
Reduced	2962.5110			

RSquare (U)	0.0482
AICc	5655.38
BIC	5706.23
Observations (or Sum Wgts)	4274

Fit Details

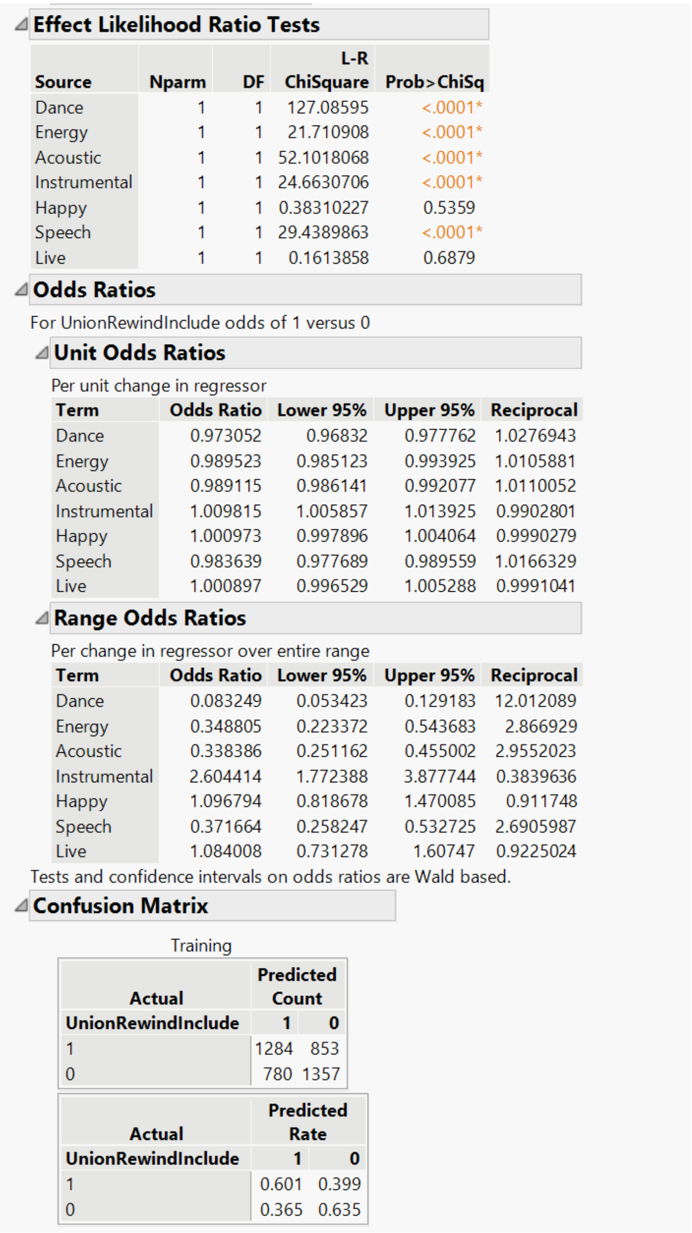
Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	4265	2819.6751	5639.35
Saturated	4272	0.0000	Prob>ChiSq
Fitted	7	2819.6751	<.0001*

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95%	Upper 95%
Intercept	2.72632934	0.2308467	139.48	<.0001*	2.27632768	3.18141824
Dance	-0.0273177	0.0024749	121.84	<.0001*	-0.0321924	-0.0224893
Energy	-0.0105324	0.0022688	21.55	<.0001*	-0.0149892	-0.0060939
Acoustic	-0.0109451	0.0015309	51.12	<.0001*	-0.0139561	-0.0079541
Instrumental	0.00976742	0.0020351	23.04	<.0001*	0.00584008	0.01382912
Happy	0.00097254	0.0015716	0.38	0.5360	-0.0021059	0.004056
Speech	-0.0164961	0.0030773	28.74	<.0001*	-0.022564	-0.0104958
Live	0.00089628	0.0022312	0.16	0.6879	-0.0034774	0.00527402

Confidence limits are likelihood-based.
For log odds of 1/0



Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Dance	1	1	127.08595	<.0001*
Energy	1	1	21.710908	<.0001*
Acoustic	1	1	52.1018068	<.0001*
Instrumental	1	1	24.6630706	<.0001*
Happy	1	1	0.38310227	0.5359
Speech	1	1	29.4389863	<.0001*
Live	1	1	0.1613858	0.6879

Odds Ratios

For UnionRewindInclude odds of 1 versus 0

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Dance	0.973052	0.96832	0.977762	1.0276943
Energy	0.989523	0.985123	0.993925	1.0105881
Acoustic	0.989115	0.986141	0.992077	1.0110052
Instrumental	1.009815	1.005857	1.013925	0.9902801
Happy	1.000973	0.997896	1.004064	0.9990279
Speech	0.983639	0.977689	0.989559	1.0166329
Live	1.000897	0.996529	1.005288	0.9991041

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
Dance	0.083249	0.053423	0.129183	12.012089
Energy	0.348805	0.223372	0.543683	2.866929
Acoustic	0.338386	0.251162	0.455002	2.9552023
Instrumental	2.604414	1.772388	3.877744	0.3839636
Happy	1.096794	0.818678	1.470085	0.911748
Speech	0.371664	0.258247	0.532725	2.6905987
Live	1.084008	0.731278	1.60747	0.9225024

Tests and confidence intervals on odds ratios are Wald based.

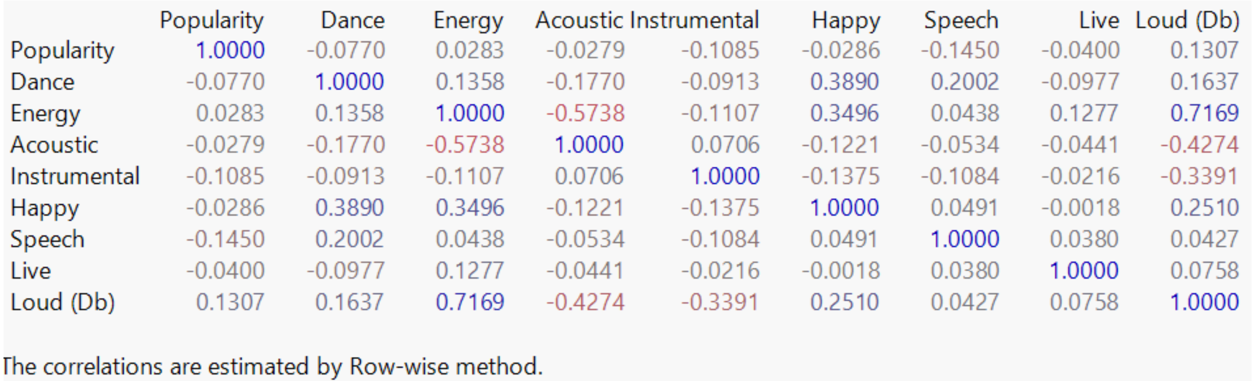
Confusion Matrix

Training

Actual	Predicted Count	
UnionRewindInclude	1	0
1	1284	853
0	780	1357

Actual	Predicted Rate	
UnionRewindInclude	1	0
1	0.601	0.399
0	0.365	0.635

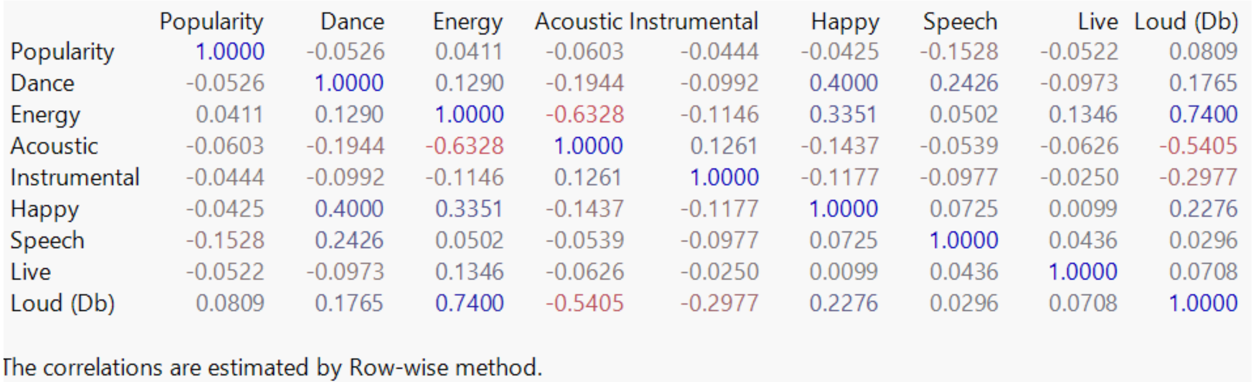
Final Reduced & Full Regression Model Correlation Matrix



	Popularity	Dance	Energy	Acoustic	Instrumental	Happy	Speech	Live	Loud (Db)
Popularity	1.0000	-0.0770	0.0283	-0.0279	-0.1085	-0.0286	-0.1450	-0.0400	0.1307
Dance	-0.0770	1.0000	0.1358	-0.1770	-0.0913	0.3890	0.2002	-0.0977	0.1637
Energy	0.0283	0.1358	1.0000	-0.5738	-0.1107	0.3496	0.0438	0.1277	0.7169
Acoustic	-0.0279	-0.1770	-0.5738	1.0000	0.0706	-0.1221	-0.0534	-0.0441	-0.4274
Instrumental	-0.1085	-0.0913	-0.1107	0.0706	1.0000	-0.1375	-0.1084	-0.0216	-0.3391
Happy	-0.0286	0.3890	0.3496	-0.1221	-0.1375	1.0000	0.0491	-0.0018	0.2510
Speech	-0.1450	0.2002	0.0438	-0.0534	-0.1084	0.0491	1.0000	0.0380	0.0427
Live	-0.0400	-0.0977	0.1277	-0.0441	-0.0216	-0.0018	0.0380	1.0000	0.0758
Loud (Db)	0.1307	0.1637	0.7169	-0.4274	-0.3391	0.2510	0.0427	0.0758	1.0000

The correlations are estimated by Row-wise method.

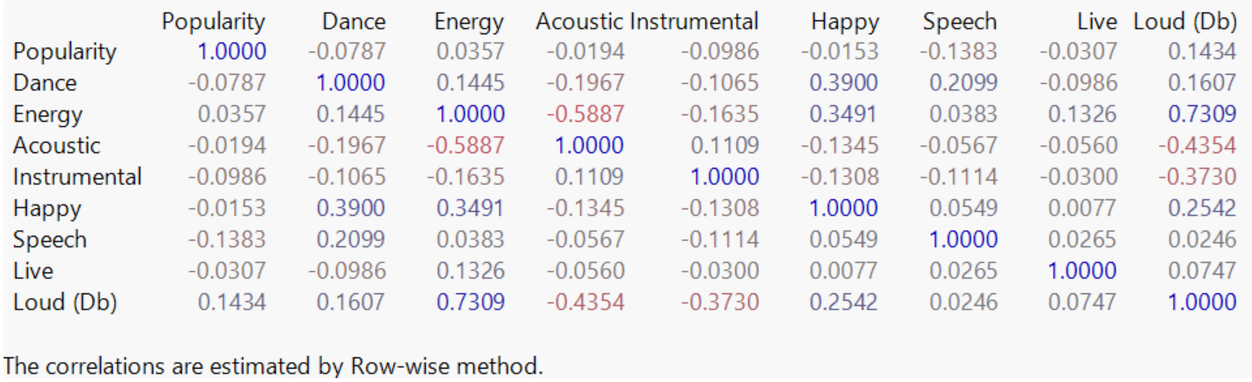
Western Countries Regression Model Correlation Matrix



	Popularity	Dance	Energy	Acoustic	Instrumental	Happy	Speech	Live	Loud (Db)
Popularity	1.0000	-0.0526	0.0411	-0.0603	-0.0444	-0.0425	-0.1528	-0.0522	0.0809
Dance	-0.0526	1.0000	0.1290	-0.1944	-0.0992	0.4000	0.2426	-0.0973	0.1765
Energy	0.0411	0.1290	1.0000	-0.6328	-0.1146	0.3351	0.0502	0.1346	0.7400
Acoustic	-0.0603	-0.1944	-0.6328	1.0000	0.1261	-0.1437	-0.0539	-0.0626	-0.5405
Instrumental	-0.0444	-0.0992	-0.1146	0.1261	1.0000	-0.1177	-0.0977	-0.0250	-0.2977
Happy	-0.0425	0.4000	0.3351	-0.1437	-0.1177	1.0000	0.0725	0.0099	0.2276
Speech	-0.1528	0.2426	0.0502	-0.0539	-0.0977	0.0725	1.0000	0.0436	0.0296
Live	-0.0522	-0.0973	0.1346	-0.0626	-0.0250	0.0099	0.0436	1.0000	0.0708
Loud (Db)	0.0809	0.1765	0.7400	-0.5405	-0.2977	0.2276	0.0296	0.0708	1.0000

The correlations are estimated by Row-wise method.

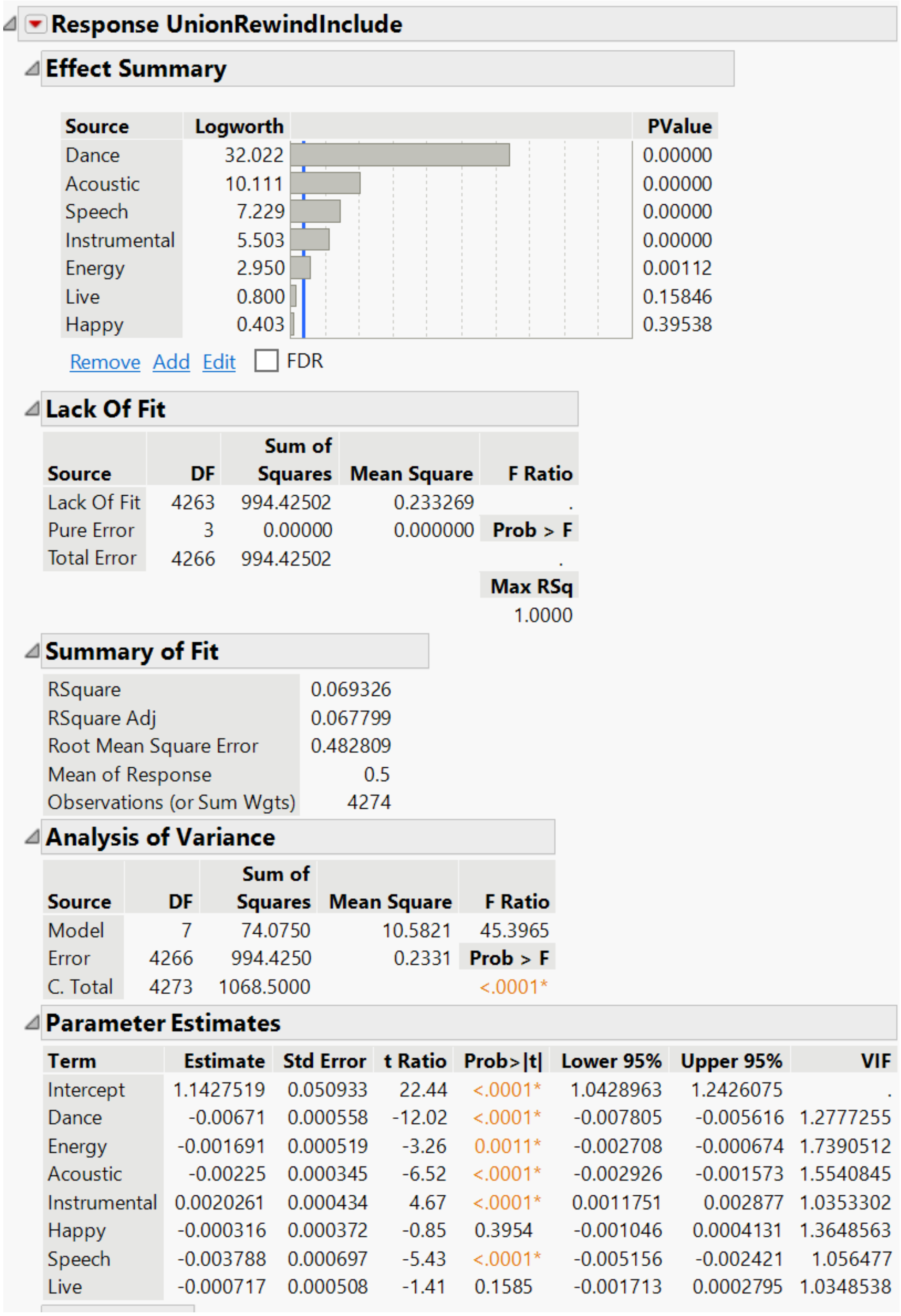
Alternative Regression Model Correlation Matrix



	Popularity	Dance	Energy	Acoustic	Instrumental	Happy	Speech	Live	Loud (Db)
Popularity	1.0000	-0.0787	0.0357	-0.0194	-0.0986	-0.0153	-0.1383	-0.0307	0.1434
Dance	-0.0787	1.0000	0.1445	-0.1967	-0.1065	0.3900	0.2099	-0.0986	0.1607
Energy	0.0357	0.1445	1.0000	-0.5887	-0.1635	0.3491	0.0383	0.1326	0.7309
Acoustic	-0.0194	-0.1967	-0.5887	1.0000	0.1109	-0.1345	-0.0567	-0.0560	-0.4354
Instrumental	-0.0986	-0.1065	-0.1635	0.1109	1.0000	-0.1308	-0.1114	-0.0300	-0.3730
Happy	-0.0153	0.3900	0.3491	-0.1345	-0.1308	1.0000	0.0549	0.0077	0.2542
Speech	-0.1383	0.2099	0.0383	-0.0567	-0.1114	0.0549	1.0000	0.0265	0.0246
Live	-0.0307	-0.0986	0.1326	-0.0560	-0.0300	0.0077	0.0265	1.0000	0.0747
Loud (Db)	0.1434	0.1607	0.7309	-0.4354	-0.3730	0.2542	0.0246	0.0747	1.0000

The correlations are estimated by Row-wise method.

Final Reduced Least Squares Regression Model Results



Response UnionRewindInclude

Effect Summary

Source	Logworth		PValue
Dance	32.022		0.00000
Acoustic	10.111		0.00000
Speech	7.229		0.00000
Instrumental	5.503		0.00000
Energy	2.950		0.00112
Live	0.800		0.15846
Happy	0.403		0.39538

Remove Add Edit □ FDR

Lack Of Fit

Source	DF	Sum of Squares	Mean Square	F Ratio
Lack Of Fit	4263	994.42502	0.233269	.
Pure Error	3	0.00000	0.000000	**Prob > F**
Total Error	4266	994.42502		.
				Max RSq
				1.0000

Summary of Fit

RSquare	0.069326
RSquare Adj	0.067799
Root Mean Square Error	0.482809
Mean of Response	0.5
Observations (or Sum Wgts)	4274

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	74.0750	10.5821	45.3965
Error	4266	994.4250	0.2331	**Prob > F**
C. Total	4273	1068.5000		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	1.1427519	0.050933	22.44	<.0001*	1.0428963	1.2426075	.
Dance	-0.00671	0.000558	-12.02	<.0001*	-0.007805	-0.005616	1.2777255
Energy	-0.001691	0.000519	-3.26	0.0011*	-0.002708	-0.000674	1.7390512
Acoustic	-0.00225	0.000345	-6.52	<.0001*	-0.002926	-0.001573	1.5540845
Instrumental	0.0020261	0.000434	4.67	<.0001*	0.0011751	0.002877	1.0353302
Happy	-0.000316	0.000372	-0.85	0.3954	-0.001046	0.0004131	1.3648563
Speech	-0.003788	0.000697	-5.43	<.0001*	-0.005156	-0.002421	1.056477
Live	-0.000717	0.000508	-1.41	0.1585	-0.001713	0.0002795	1.0348538

GitHub Repository

https://github.com/conorfryer/unionrewindregressionanalysisproject

This repository contains all digital files related to *The Sound of Union: Quantifying Union College's Music Preferences with Logistic Regression*. It includes:

- Datasets: WRUC Rewind survey data and the comparison dataset.
- JMP Files: Regression models and analysis output.
- Code: Scripts for data preprocessing and feature extraction.

This repository ensures transparency and reproducibility for further analysis.

Acknowledgment of AI Assistance

This paper was developed with assistance from ChatGPT, which was used to refine explanations, clarify technical concepts, and ensure statistical interpretations were presented accurately. All analysis, methodology, and conclusions are my own, with AI serving as a tool for improving clarity and coherence.