Teagan Britten, Aileen Kent, Ruth Melese, Jessica Ni, Nyla Upal, Maya Uwaydat, and Gracie Williams
DS 3021: Machine Learning I
Final Paper
9 May 2025

Abstract

Our paper investigates what variables predict the presence of the word "love" in the title of a Billboard Hot 100 hit. Using data derived from both Billboard magazine itself and from the Spotify API, we applied both logistic regression and random forest models to identify compositional and popularity-related predictors of the word "love." With our logistic regression, we initially observed high accuracy (test R-squared ~ 0.91) due to a significant class imbalance, as only 8.3% of songs contained "love" in the title. After adjusting our regression with class weighting, model performance dropped to more realistic levels (Test R-squared ~ 0.31), reflecting the necessity of optimising models to effectively predict improbable events. We experimented with various feature sets and found that simpler models using a few carefully selected variables—such as track duration, danceability, and explicit content—outperformed more complex models. A reduced logistic regression model achieved a higher test R-squared of ~0.41, suggesting that overfitting and multicollinearity were issues with the full model.

We then implemented a random forest regressor to account for variable interactions our logistic regression missed, so that we could better assess variable importance. This model identified liveness, speechiness, loudness, tempo, and track duration as the most influential predictors, suggesting that songs with "love" in the title share subtle acoustic and structural characteristics. Exploratory data analysis, including pairplots and single-variable regressions, supported these findings. Interestingly, Spotify's proprietary popularity metric outperformed our custom Billboard-based metric in predictive power. Overall, this study demonstrates that while no single feature strongly predicts the presence of "love" in a title, a combination of audio and

popularity features can offer meaningful insights. Future work could explore ensemble models or natural language processing of lyrics to enhance predictive accuracy further.

Introduction

The Billboard Hot 100 ranks the United States' most popular songs, providing a summary of the current pop music zeitgeist. The chart allows researchers and music enthusiasts alike to explore how mainstream music has changed since the Billboard's inception in the mid-20th century : it has expanded to encompass genres from country to electronic dance music, in addition to reflecting consumption through streaming, radio airplay, and digital sales (Trust, 2022). Though the chart gives a sense of what artists and songs resonated with the public during certain time periods, it provides little insight into the actual composition of the music. Spotify and other streaming services, however, provide vast registries of music data—including information on Billboard Hot 100 hits—and we could use the information they provide to investigate what features exist in pop music while also using the Billboard to study their success.

Songs about love and relationships have been relevant to the Billboard for decades, so we found ourselves particularly curious about compositional and popularity-based metrics of romantic music. Using Spotify's vast API and the Billboard's detailed record of weekly charts, we conducted a research project in which we explored which metrics were predictors of songs with "love" in their titles. Our analysis provides insight into how lyrical themes correspond with measurable audio features and commercial performance.

Despite the ubiquity of love songs, less than ten percent of Billboard hits actually mentioned love in their title, so our first models had unrealistically high accuracy when used on our entire dataset. Therefore, most of our regressions weighted underrepresented data so that the model would not predict "false" on every occasion. Our first model used all variables and was direly overfit, so we regressed on a handful of variables with the highest R-squared, and then finally constructed regressions on only one variable at a time. A histogram was used to visualize individual regression coefficients, and we determined that our most important variables were track duration, liveness, and acousticness. We then ran our random forest model because the logistic regressions did not account for variable interactions; though it confirmed liveness was one of the most impactful variables, it also found loudness and speechiness were strong predictors. The likelihood a song was performed live appears to be the strongest indicator of the presence of "love" in the title, though the models disagreed on other potential indicators.

Challenges of our project involved choosing a project direction and resolving problematic results from our models; and limitations were introduced by the influence of subjective variables and the limited scope of the project. Our research could be expanded to account for music outside of the Billboard Hot 100, or its focus could shift towards predicting compositional attributes instead of title data.

Data

Our study used two datasets posted on Kaggle, but the data itself was derived directly from Spotify's API and from the Billboard magazine's Hot 100 rankings. While we used the Hot 100 chart data to identify hit songs and determine their success, the Spotify data enabled us to assess aspects of the music's composition. Most song attribute variables measured objective values, including tempo (in beats per minute), duration (in milliseconds), and loudness (in decibels); however, a handful were more subjective, such as "valence" (a calculated measure of the song's happiness) danceability, and "liveness" (probability a song was performed live), and were represented with values from zero to one. Meanwhile, Billboard-derived variables included week-by-week data on songs' rank and the week they charted, reflecting a timeframe from the beginning of the chart's inception to 2022.

Though purging, imputing, or changing values in our datasets was not necessary, we still manipulated our data for use in our study. To gauge Billboard success, we used an algorithm that assigned a song of a given rank points for each week on the chart—100 for a song of rank 1, 99 for a song of rank 2, and so on—and totaled the points for each song to calculate a popularity score. The column containing the scores was then concatenated to the Spotify API dataset so that we could use them for our analysis. To isolate relevant data, we also created a new .csv file containing *only* information corresponding to songs with "love" in their titles.

Ideally, logistic regression should account for variables' influence over one another and their interactions' effect on predictions. As part of our exploratory data analysis, we printed scatterplots of all metrics of song composition and popularity—such as spotify_track_popularity, acousticness, and tempo—and investigated relationships between our features. None of our plots displayed easily discernible correlations (linear or not), so we decided to use logistic regression without any transformations reflecting variable interaction.

Methods

Our group used supervised learning—specifically logistic regression—to analyze our dataset. Because the data was already labeled and the target variable was already known, our regression models used binary classification to predict whether or not a song had the word "love" in the title. We used various song characteristics from our dataset such as valence, tempo, and liveness to make predictions. To measure a regression's success, we looked at train and test accuracies so that we could determine performance on both their input data and new data. Our initial goal was to first model single features, then build more complicated models with different combinations of single features with high accuracy scores. We expected the training and testing accuracies to vary greatly depending on the combination of features used.

After logistic regression, we used a random forest, another supervised learning method, to determine which features were most important in the model's decision making based on mean decrease in impurity. Decision trees make splits to best reduce impurity (i.e. the variability in the data it sorts in a given category), and a higher value means a feature contributed more to making splits. We created a bar graph to visualize how influential our variables were in predicting whether a song title contains "love" according to our model.

Results

When we trained our model on all variables, we found that the R-squared values were the same regardless of what variables were selected; additionally, they were improbably high at around 0.917 for our training data and 0.913 for our testing data. Our target variable was highly imbalanced, as the vast majority of songs in our data did not contain "love" in their titles. Therefore, we added a parameter called "class_weight" that allowed us to adjust how the model weighted the data; by setting class_weight to "balanced," we adjusted each weight to be inversely proportional to its frequency in the data (GeeksforGeeks, 2024). The balanced model put greater emphasis on our underrepresented class—i.e. our target class—so our R-squared values dropped to ~0.3079 for our training data and ~0.3077 for our test data, which appeared far more realistic for regression that was likely overfit.

To discern which variables had the greatest influence on our data, we printed their R-squared coefficients to see which had the largest values. All coefficients were below 0.1, and our overall R-squared was ~0.1770 for our training data and ~0.1817 for our test data, so we concluded that the model was too overfit to draw reasonable conclusions about the effects of each feature. Running the three variables with the highest coefficients did not return a very high R-squared output, either: our resultant values for the whole model were only ~0.4244 and ~0.4204 for training and test data respectively. Because we wanted to study the impact of individual variables, we created a dataframe that would store the train and test R-squared for a logistic regression on each variable at a time; then, we displayed the values on a bar graph. The predictors with the highest R-squared values were track duration (train ~0.9175, test ~0.9137), liveness (train ~0.6430, test ~0.6359), and acousticness (train ~0.6163, test ~0.6025). However, logistic regression typically uses multiple interacting variables as predictors, so the regression on individual features may not have fully reflected their influence on the data.

To compensate for what logistic regression may have missed, we ran a random forest regressor on our dataset to see which variables best predicted whether a long had love in it. We created another bar graph—this time displaying feature importance—to identify the most impactful variables. According to our model, liveness, speechiness, and loudness had the greatest influence, with feature importance scores of ~0.0957, ~0.0955, and ~0.0955 respectively. Both models recognized liveness as a strong predictor, but their conclusions otherwise diverged.

Conclusion

Through unsupervised learning, our study investigated possible predictors of songs with "love" in their titles. After experimenting with a logistic regression model, we identified liveness, speechiness, and loudness as the three most influential factors in determining if a song had the word "love" in the title. Our random forest model likewise suggested liveness was a strong predictor of a song's title; but unlike our regression, it identified acousticness and speechiness as the other two most important features. Not all love songs reflect our most predictive variables, but some of the most famous love songs—such as *Shape of You* by Ed

Sheeran and *All of Me* by John Legend—are consistent with the attributes greatly influential in our models.

One of the greatest challenges of the project was establishing its direction. At first, our aim was to predict Billboard success of songs based on compositional characteristics, but we found a wealth of existing studies that asked research questions almost identical to ours and were concerned about the originality of our study. We also wanted to account for lyrics of songs on the chart, but time constraints and the difficulty of acquiring the information made our goal unfeasible. After discussing how to determine songs' subject matter without lyric data, we settled on using song titles; therefore, we narrowed the scope of our project only to songs with "love" in their titles and decided our new research question was original enough to proceed with making Billboard success predictions. Upon examining our available data and tools, we decided that we were best equipped to *find* predictors of a characteristic of the music, hence the current topic of our study.

At different points of the study, we were also concerned that our calculations reflected imprecision in our models. When we first ran the logistic regression, we noticed our R-squared values were improbably high; however, we concluded that so few songs in our dataset contained "love" in their title that our unadjusted regression would always be highly accurate, and we recalibrated our model. During later stages of the project, the different results of our logistic regression and random forest made us wary of the regression's lack of consideration for variable interaction. Regardless, contradictions in our models' predictions may be difficult to avoid because the algorithms we chose draw conclusions about data differently: while logistic regression searches for correlations between variables, random forest sorts data based on its characteristics. The models' incongruous conclusions should be used as inspiration for further research, as we could investigate why each model prioritises certain features and foster a stronger understanding of how different algorithms work.

Performing analysis on music based on its characteristics also poses difficulties due to the subjective nature of some of our variables. For example, both our popularity metrics reflected different definitions of musical success: while the songs' popularity points we calculated accounted for songs' position and time on the Billboard, popularity scores from Spotify's API gauged total numbers of plays on the streaming service. Additionally, variables such as "valence" cannot be quantified in practice; a song that sounds happy to one person may seem melancholic to someone else. Our subjective variables do not invalidate the results of our study, as our models were consistent in identifying "liveness" as a strong predictor, implying that our conclusions were not necessarily obscured by the presence of imprecise features.

While our analysis certainly contains many limitations, we provided sufficient validity to our processes through thorough variable exploration, comparing different analyzation methods, and ensuring we did not make too broad of a statement regarding the correlations that we found. We identified highly correlated variables for popular songs that had the word "love" in the title, and we did not attempt to generalize our results to all love songs, as our project scope was restricted to Billboard data.

The study offers diverse opportunities for further research, from utilizing new datasets to focusing on variables besides song titles. We could attempt to predict "love" in the lyrics or even just choruses, as some love songs do not have titles that contain the word. This extension presents its own challenges: though correlations we discover between song subject matter and compositional characteristics would be more accurate, preparing and analysing our data would require intensive text parsing. We could more easily expand our analysis to popular music from other countries and to charts for specific genres; the scope of our project would be reflective of music beyond the most mainstream hits of the United States. Another extension of our analysis could involve investigating whether characteristics of songs predict the presence of others; for example, songs with high speechiness may have longer track durations. We would not be dealing with lyric or title data, so text parsing would not be necessary, which could make analysis simpler and more realistic to carry out.

References

GeeksforGeeks. (2024, August 7). *How Does the class_weight Parameter in Scikit-Learn Work?* https://www.geeksforgeeks.org/how-does-the-classweight-parameter-in-scikit-learn-work/.

IBM. (2021, March 12). *Supervised vs. Unsupervised Learning: What's the Difference?* https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning.

TheDevastator. (2023). *Hot 100 Audio Features.csv* [Data set]. Kaggle. https://www.kaggle.com/datasets/thedevastator/billboard-hot-100-audio-features

TheDevastator. (2023). *Hot Stuff.csv* [Data set]. Kaggle. https://www.kaggle.com/datasets/thedevastator/billboard-hot-100-audio-features

Trust, G. (2022, August 4). *64 Fun Facts From the Billboard Hot 100's First 64 Years: From Ricky Nelson to Lizzo & More.* https://www.billboard.com/music/chart-beat/hot-100-64-years-fun-facts-ricky-nelson-lizzo-1235122282/