Cultural Influence of Music as Propaganda

Noah Cillo and Jeremy Schlegel

Department of Systems Engineering, United States Military Academy, West Point, New York 10996

Corresponding author's Email: mr.noahcillo35@gmail.com

Author Note: Noah Cillo is a cadet at the United States Military Academy at West Point. He is a Systems Engineering Major with a Minor in Applied Statistics. This research was conducted by Noah Cillo as part of an independent undergraduate research project. SGM Jeremy Schlegel, an instructor at West Point and affiliate of the West Point Music Research Center, provided academic guidance and advisory support during the development of this study. The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of Defense.

Abstract: This project seeks to provide a baseline of qualitative and quantitative variables of musical forms of communication found within populations and are likely to have the most significant impact on their narratives. Using system-based skills and approaches, the project will provide Sandia National Laboratories with a preliminary understanding of their continued research. A statistical model will be developed to identify and predict the popularity of contemporary music.

Keywords: Aspects of Music, Qualitative and Quantitative Analysis, Random Forest Model, Predicting Song Popularity

1. Introduction

Music has provided a unique influence on culture spanning back centuries. From Yankee Doodle in the Revolutionary War to the jazz and blues of the Harlem Renaissance, music has consistently passed down tradition and promoted social movements (Hodge, 2023). As much as music was used by the people, it was also censored and manipulated by governments. Nazi Germany banned jazz music because of its popularity in Jewish and African communities (Street, 2003). Both the Soviet Union and the Nazi regime exploited their country's folk music for the promotion of their agendas (Street, 2003). To identify the political use of music in communities, there first needs to be a way to identify and predict the popularity of music itself.

This literature review and model analysis seeks to provide a baseline for the analysis of musical forms of communication. The functional objective of this project is to identify aspects of music that may influence a population to support an ideological movement. The research will provide the most significant qualitative and quantitative variables of popular music used to influence a population.

Sandia National Laboratories is a contracted research and development center for the U.S. Department of Energy's National Nuclear Security Administration. The findings will be able to provide Sandia National Laboratories with background information to build on the Dynamic Multi-Scale Assessment Tool for Integrate Cognitive-behavioral Actions, also known as DYMATICA (Sandia, n.d.). The DYMATICA model is a tool used to answer why a country or population acts a certain way, with applications in identifying popular music and propaganda in foreign countries (Sandia, n.d.). It is used to anticipate how the population would respond to changes in the government, economic, or socio-cultural landscape.

2. Background

The purpose of this research is to assist in identifying propaganda in music and its effects on foreign populations. The scope of the project is centered around the popularity of music. The model is not focused on propaganda identification, as Sandia already has this modeled with DYMATICA.

The methods used are both qualitative and quantitative analysis. A tool used to assist this method is constructing a functional hierarchy. A functional hierarchy takes a single functional objective and breaks it down into quantitative value measures. The functional hierarchy, shown below in *Figure 1*, was centered around the core functions and further broken down into four sub-objectives. The scope of this research was focused on producing popular music and methods of distributing music. Each function is made up of four value measures, which can be quantified and analyzed. The method of modeling will use linear regression and random forest models in R Studio to create an accurate predictive model around the response variable. The paper will conclude with recommendations for further research, along with the intended application of our findings by Sandia National Laboratory.

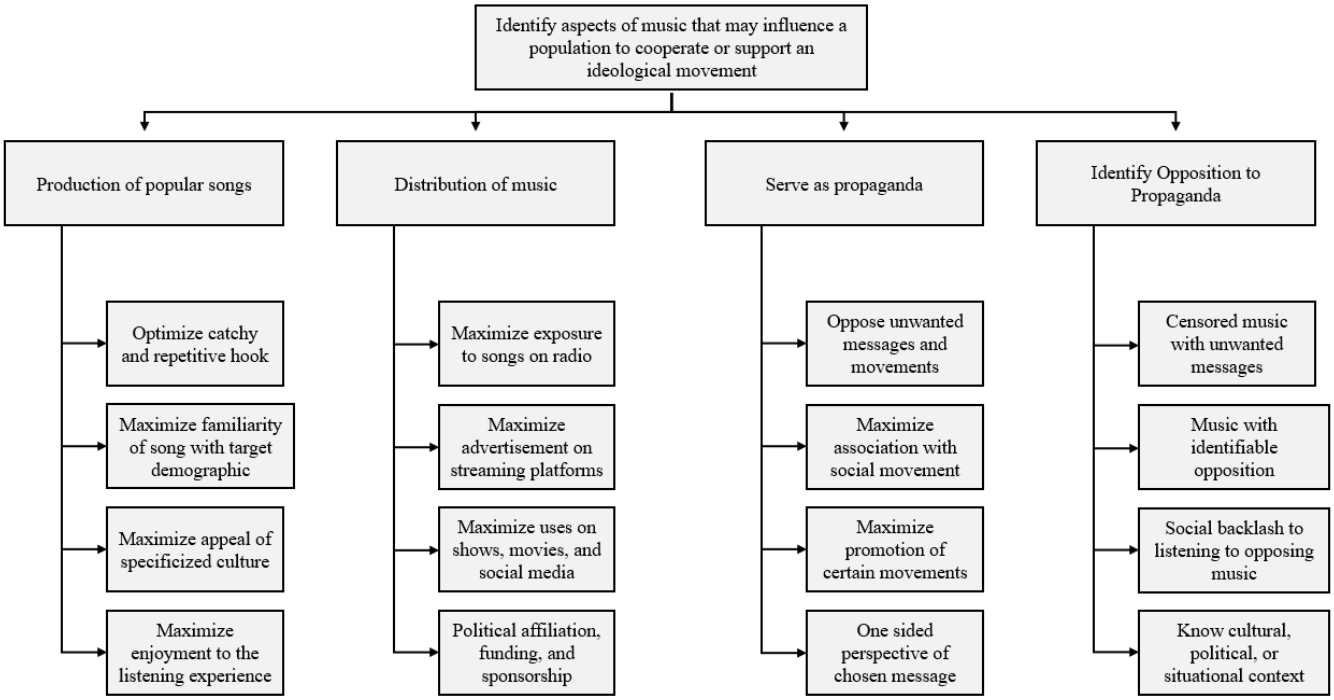

Figure 1. Functional Hierarchy

3. Literature Review

The literature review will focus on the production of popular songs and the impacts of their distribution methods. This will include ways to quantify and analyze music, along with studies found that provide analysis relevant to the scope of this research. Genre and culture are major factors influencing a song's popularity within specific demographics. A song's exposure, advertisement, and publicity have a strong relation with its popularity and national recognition. Each variable will be analyzed and provided with background information relevant to the modeling process.

3.1 Aspects of Popular Music

Music's popularity can be influenced by the melody, dynamics, or catchiness of the hook (Rothstein, 2021). Though it is difficult to make statistical analyses on music, extensive datasets have been made to analyze the characteristics and genres of music.

3.1.1 Quantifying Data

The initial findings focused on quantifying music. One extraction method was the Music Information Retrieval (MIR) toolbox in MatLab. The *MIRtoolbox* can convert audio samples into data, such as timbre, tonality, rhythm, and form (Lartillot and Toiviainen, n.d.). Parameters can be used to analyze the data and determine trends of certain songs and their respective lyrics.

A premade data set that can be used is the Million Song Dataset. This dataset has over 40 variables for each song (Bertin-Mahieux et al., 2011). These variables include objective variables such as artist and song name, length, and release date. Other variables may be more subjective, such as danceability and energy. This data set was not used as it did not include a variable for the song's contemporary popularity. The quantifying of music is important because it allows statistical analysis to be done objectively and build a model that can be applied to other studies.

3.1.2 Aspects of Music

One of the most important variables corresponding to popularity is the hook. Repeated riffs, known as hooks, contribute to the listening factor of a song (Rothstein, 2021). The catchiness creates a cognitive 'itch', getting the song stuck in the listener's head and increasing the probability of the listener replaying the song (Anuta, 2006). The best way to induce a cognitive itch is to increase familiarity with the song, matching the genre of music with the targeted audience.

A study done at Dartmouth College viewed the brain's response to auditory imagery (Kraemer et al., 2005). Cognitive itching occurs when the brain is imagining a tune or phrase. The study's findings concluded that when patients listen to songs, their brain subconsciously fills in the lyrics (Kraemer et al., 2005). The auditory association cortex, shown in *Figure 2*, showed significantly more activity in excerpts of songs that had no lyrics while coupled with familiar genres (Kraemer et al., 2005). The cognitive stimulation of catchy songs and familiarity show the importance of song recognition and repetition. This supports the notion that genre familiarity is closely related to popularity within demographics.

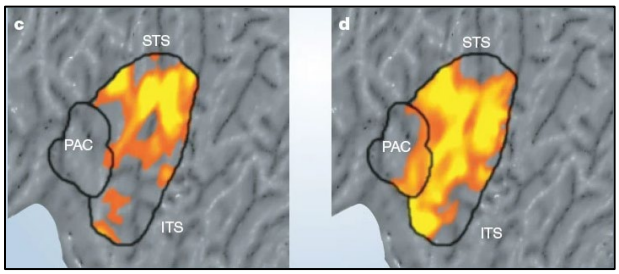

Figure 2. Contrasting stimulation: Songs with lyrics (left image) and songs with embedded gaps in lyrics (right image), (Kraemer et al., 2005)

3.1.3 Genre and Demographics

The genre of contemporary music is one of the biggest factors in a song's popularity. A study done by Coleman Insights Media Research showed the demographic balance of genre popularity (Milkman, 2021). The study took the top 100 popular songs within different demographics and comprised what percent of their top 100 were in certain genres. This study concludes that there are preferred genres in different demographics (Milkman, 2021). The results showed that 70% of the music in the Top 100 among African American consumers was Hip-Hop/RnB, whereas both Hispanic and African American consumers only had 2% of their music consumption comprised of country music. Hip-Hop/RnB was also shown to make up 43% of the music for consumers ages 12-34, but only 11% ages 35-54. The importance of this is the effectiveness of targeting a demographic. When creating music for a social movement or political cause, it is essential to know the demographic of the intended audience.

3.2 Impact of Distribution

Much of the qualities of music had no significant importance in making the music popular. Genre and familiarity were limited by demographics, and the catchiness of the hook was too broad. Other factors this paper investigates are less of the actual music, and more of its marketing. Many factors contributing to a song's popularity include exposure, availability, and advertisement.

3.2.1 Advertisement and Distribution

One of the most important elements of popular music is its distribution. Music is business; distribution is a key element in any business. The difference between local recognition and nationwide popularity is how accessible the music is (Rothstein, 2021). Most means of distribution come from external agencies known as record labels. Record labels play multiple roles in the promotion of music: aggregators and distributors. Aggregators ensure the exposure of music through its availability. They ensure the content is available on every digital and streaming platform (Lulibo, 2022).

Traditionally, distributors handled getting music released nationwide at music stores by vinyl and CDs. With most of the music being streamed, distributors mainly focus on ads and playlisting (Lulibo, 2022). The biggest influence record labels have on music is promotion, paid advertisement, and social media campaigns. Labels leverage their industry connections, gain radio promotions, and assert recognition by 'tastemakers' (Lulibo, 2022). Tastemakers are responsible for curating playlists on streaming platforms such as Spotify, Pandora, and Apple Music (Lulibo, 2022). These playlists get a lot of web traffic and increase song exposure.

3.2.2 Song Exposure

Song exposure can also take the form of music videos. Music videos keep audiences talking about an artist and further increase their popularity. A large part of the success of Michael Jackson and Madonna was their showmanship, highlighted in music videos and performances on MTV (Munsalud, 2018). Music videos can also change the perception of songs, such as further developing a socio-political message in the video.

3.2.3 Trends, Memes, and Social Media

Modern media is centered around user-specific content. Users determine who they follow, what they listen to, and when they listen to it. This challenges traditional media and changes industry strategies. With the preferences and demands of users much more important, one of the biggest factors of popularity is word of mouth. Word of mouth takes the form of online posts, texts, and videos. This affects the popularity of music through online buzz. Traditional advertisement requires long and expensive campaigns across multiple mediums (Grome, 2023). Cheaper and faster ways to promote attention are through online posts and viral trends on platforms such as YouTube, Instagram, and TikTok (Grome, 2023).

4. Methodology

The scope of the following models will be to analyze the function "produce popular music" from the functional hierarchy, *Figure 1*. The objective of this model is to determine what are the most significant aspects of contemporary music that make songs popular and determine if we can accurately predict the popularity of a song given its quantitative components.

Three statistics used to evaluate the models are the out-of-sample R^2, mean absolute error, and the root-mean-squared error, shown in *Equation Set 1*. The out-of-sample-R^2 (OSR2) is the percent of variability which is explained by the model. An OSR2 value closer to 1 explains more variability than one which is closer to 0. The mean absolute error (MAE) is the average error of the predicted values from the actual values, measured by the sum of errors divided by the sample size. The root-mean-square error (RMSE) involves squaring the error, then taking the average of errors and square rooting the average. The RMSE is different from the MAE because it is more sensitive to outliers, giving it a different sensitivity statistic to analyze the effectiveness of the models.

$$OSR^2 = 1 - \frac{\sum_{i=1}^{n}|x_i - \hat{x}_i|}{\sum_{i=1}^{n}|x_i - \bar{x}|} \; ; \; MAE = \frac{\sum_{i=1}^{n}|x_i - \hat{x}_i|}{n} \; ; \; RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{n}} \quad (1)$$

4.1 Modeling

The first model used to determine the significance of certain variables was linear regression. Variable visualization was utilized to see any significant relationships between variables and popularity. Lastly, a random forest model was used to determine if we could effectively predict the popularity of a song using 70% of the data set to train the model and test it on the remaining 30% for effectiveness. Both data analysis models are completed in R Studio.

4.1.1 Spotify Tracks Dataset

The dataset used was the Spotify Tracks Dataset (Pandya, 2022). This dataset includes 125 genres with 1,000 songs in each genre for an equal distribution of music. Each song has a title, artist, genre, and 14 quantitative variables about the song. The response variable analyzed is "popularity". The popularity of the song is calculated by an algorithm based mostly on the total number of streams and the streams' recentness.

A few issues were initially identified with the dataset. First, there were too many observations. With 125 genres and 1,000 songs each, 125,000 observations hindered the computing process. Additionally, songs were in multiple genres, making many songs counted and analyzed multiple times. The data was cleaned by only keeping ten genres. This was decided because most of the genres were niche sub-genres of already listed predominant genres. For example, alternative rock, grunge, hard rock, hardcore, grindcore, j-rock, punk rock, punk, and rockabilly were all taken out of the data set and were represented by the overarching genre of rock-n-roll. This left 10,000 observations through the encompassing ten genres.

Through further analysis, more cleaning had to be done due to the data collection method. There was a significant number of observations that were not given data. These observations had little or no popularity and had some variance of error, such as copyright infringement songs, unreleased music, and other incomplete observances. This was mediated by eliminating all songs with 0-10 popularity score, leaving 6,165 observations over the ten genres to complete the model.

4.1.2 Variable Visualization

Initially, plots were made on the significant variables to determine any correlation between popularity and confounding variables. As seen in *Figure 4*, the graph depicts every observation's danceability score and popularity score. The blue trend line shows the relationship with an overall positive correlation. However, there is a significant amount of variation away from the line on either side. Every other variable was plotted and no single variable could be an accurate predictor of popularity.

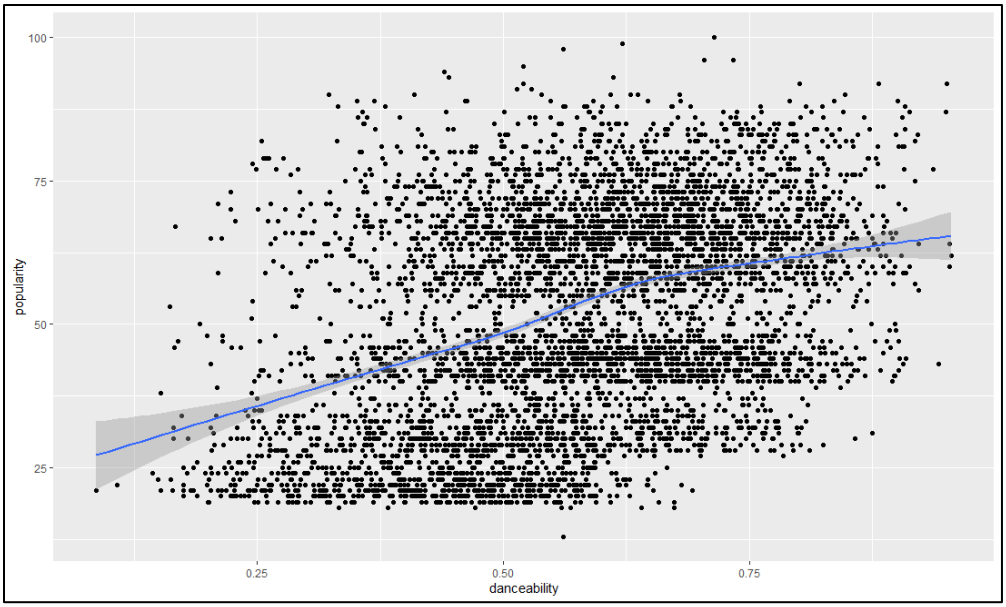

Figure 4. Scatterplot of danceability and population with regression line

From previous literature reviews, we know how significant genre can be in determining popularity (Anuta, 2006). With little luck finding a correlation between the continuous variables and popularity, we plotted the overall popularity of each genre to determine if there were any trends. *Figure 5* shows that some genres tend to be more popular than others, but most genres have a significant number of outliers and variations in the data.

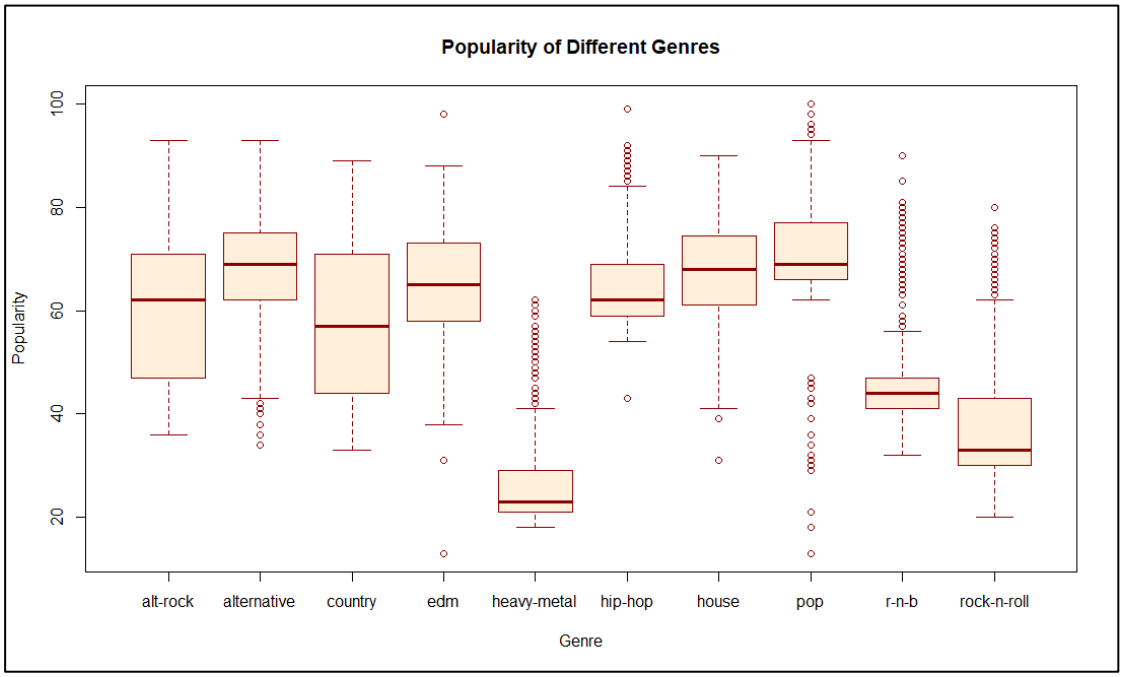

Figure 5. Boxplot of genres and their assessed popularity, showing each quartile and outliers for each genre

4.1.3 Linear Regression

The first model we used was linear regression. The first step in either model was to use a pairs plot to show the relationships between all the variables. High correlation and collinearity are problematic for the model and would need to be removed. However, no variables contained above a 90% collinearity, so no variables were removed.

The next step was to partition the data. The data was split into two data sets, a training set and testing set. The training set was used to build the regression model and the testing set was used to evaluate the effectiveness of the model. This allows us to determine if this model would be useful in application. After completing the model and applying to the test set, the OSR2, MAE, and RMSE were calculated to determine the effectiveness of the mode.

4.1.4 Random Forest

After completing the linear regression model, we decided to model a random forest model as well. A random forest model allows us to handle many variables and assess variable importance (Grömping, 2009). Random forest models use an ensemble of random trees to predict based on the forest's average (Grömping, 2009). The data was split into a training set of 70% and a testing set of 30% of the observations. We then built the random forest model of 500 trees and a node size of 5. The model was used to predict the test set and then evaluated for effectiveness.

5. Results

Each of the models were calculated for error and OSR2 score, along with a plot showing how accurate the model was at predicting the popularity of the test data sets. *Table 2* shows the test statistics for each model.

5.1 Linear Regression Model and Random Forest Model Results

The linear regression model gave us the coefficient results, *Table 1*, and provided us with an overall inaccurate model. The coefficient table shows that all the variables besides tempo should be included, due to its low significance. The linear regression model is a simple formula to apply to an unknown observation. When given the statistics of the observations, it uses the values to provide a predicted popularity. The variance of the prediction error is too high to gain any use from the model. The model had an error between 16 and 19, while only being able to explain 20% of the variance.

Table 1. Coefficient Table from Linear Regression Model

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   57.609952   2.673103  21.552  < 2e-16 ***
danceability  48.321769   1.849950  26.121  < 2e-16 ***
energy       -20.391476   2.110255  -9.663  < 2e-16 ***
acousticness  -4.244985   1.147245  -3.700 0.000218 ***
liveness      -7.778776   1.225716  -6.346 2.40e-10 ***
speechiness   15.421193   3.288480   4.689 2.81e-06 ***
tempo          0.010040   0.008292   1.211 0.226050
loudness       1.542353   0.110393  13.971  < 2e-16 ***
valence      -14.559908   1.363642 -10.677  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.03 on 4991 degrees of freedom
Multiple R-squared:  0.2315,    Adjusted R-squared:  0.2303
F-statistic:   188 on 8 and 4991 DF,  p-value: < 2.2e-16
```

Table 2. Statistic Table with given model and stat

		Model	
		Linear Regression	Random Forest
Statistic	OSR2	0.1975	0.7073
	MAE	15.8293	10.1133
	RMSE	18.8827	7.5738

The random forest model provided much better results. Since all variables are not significantly correlated, the random forest model used all variables included in the dataset to build the model. With an MAE of 10 and RMSE of 7.5, there are much fewer outliers in this model and will usually predict a song's popularity within a margin of 10 popularity score. This model's independent variables explain 70% of all variability and is much more accurate at predicting popularity. The visualizations of the model's effectiveness are seen in *Figure 6* and *Figure 7*, showing the random forest model is more accurate. Although the random forest model can give a general idea if a song will be popular or not, both models do not provide enough accuracy to be effective in a research capacity.

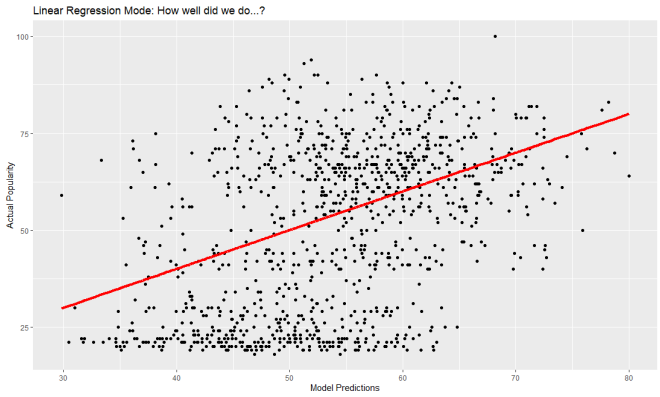

Figure 6. Linear Regression predictive model results

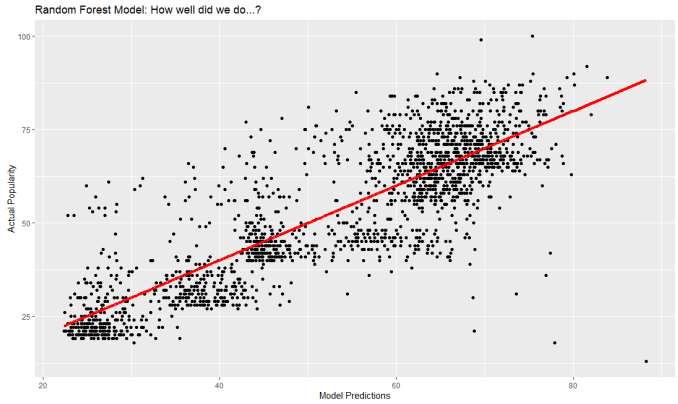

Figure 7. Random Forest predictive model results

5.2 Application

This model can be applied to predict the effectiveness of a song before its release. If Sandia wanted to create a song with the most popularity potential to distribute a particular message, they would invest their money in maximizing certain characteristics that are shown in this model. This can also be applied to record labels when promoting songs. Before investing thousands of dollars in the promotion of a new song or album, the record label would want to use a similar model to predict the song's popularity potential and ensure a satisfying payoff from their investment.

6. Conclusion

The aspects which make music popular allow the same music to be influential across a population. To understand how music can be used to support a political, social, or civil movement, it must first be understood how music becomes popular among the people of a country. Some factors such as genre, danceability, and a catchy hook contribute to making a song repeatedly played. When referring to a targeted demographic, music of familiarity increases the relistening value of the music.

The random forest model on the Spotify Tracks Dataset provides a predictive model that can estimate the popularity of a song given its characteristics within an average of 10% popularity. The random forest independently predicts the response variable from decision trees built on the unknown observation's independent variables. The model then aggregates all the decision trees and reports the average response variable. This may be effective for quick and initial estimations but does not provide enough accuracy for other models and research that require accurate predictions.

6.1 Future Research

Since we determined the aspects of music do not provide enough data to determine if a song is popular, we will want to investigate data on popular songs and their distribution and advertising. Future data analysis will involve similar models with data on money spent on advertisement, distribution, record label status, name recognition, and other similar variables. Adding these variables will add more explanation for probability and produce a more accurate predictive model.

Another area we want to focus on is how music is popular within different demographics. A limitation of our dataset was it was nationwide popularity. However, if popularity within demographics such as age, race, and sex, were included, we might be able to provide a better predictive model for popular music.

With the analysis of what makes music popular and how to make music popular, Sandia National Laboratories and other research teams can apply these findings when investigating the role of propaganda in music. Identifying popular music allows researchers to view possible narratives in a chosen message. When a single side of an issue is in popular music, it is likely the other side may be suppressed. Listening to certain music may cause a social backlash. It is most important to identify the cultural and political context of the situation. Future models will lead to the application of propaganda identification in foreign countries.

7. References

Anuta, J. (2006, June). Probing the Question: What Makes a Song Catchy? [Review of Probing the Question: What Makes a Song Catchy?]. The Pennsylvania State University. https://www.psu.edu/news/research/story/probing-question-what-makes-song-catchy/

Balenzano, K., Feder, E., & Yuan, D. (2020). A Visual Analysis of Top Spotify Song Data [Review of A Visual Analysis of Top Spotify Song Data]. Carnegie Mellon University. https://www.stat.cmu.edu/capstoneresearch/spring2020/315files/team3.html

Bertin-Mahieux, T., Ellis, D., Whitman, B., and Lamere, P. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

Dewan, S., & Ramaprasad, J. (2014). Social Media, Traditional Media, and Music Sales. MIS Quarterly, 38(1), 101–122. https://www.jstor.org/stable/26554870

Grome, Justin. Utilizing Social Media in the Music Industry [Review of Utilizing Social Media in the Music Industry]. Forbes. https://www.forbes.com/sites/forbesbusinesscouncil/2023/09/22/utilizing-social-media-in-the-music-industry/?sh=19344b161ac3

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. The American Statistician, 63(4), 308–319. http://www.jstor.org/stable/25652309

Hodge, M. (2023, September 10). The Pitches of Protests: How Music Makes Movements [Review of The Pitches of Protests: How Music Makes Movements]. Harvard Political Review. https://harvardpolitics.com/pitches-of-protest/

Kraemer, D., Macrae, C., Green, A. et al. Sound of silence activates auditory cortex. Nature 434, 158 (2005). https://doi.org/10.1038/434158a

Lulibo. (2022, February). *The Core Functions of a Record Label in 2022*. Retrieved from Stereo Fox: https://www.stereofox.com/resources/the-core-functions-of-a-record-label-in-2022/#

Milkman, S. (2021, May 11). *Slicing Contemporary Music Tastes by Demographics and Consumption*. Retrieved from Coleman Insight: https://colemaninsights.com/coleman-insights-blog/slicing-contemporary-music-tastes-by-demographics-and-consumption

Munsalud, J. (2018, December 13). *The Impact of Music Videos on Pop Culture*. Retrieved from New University.

Pandya, Maharshi. (2022). Spotify Tracks Dataset [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/4372070

Rothstein, A. (2021, July 12). What are the Elements of Popular Music? [Review of What are the Elements of Popular Music?]. Institute of Production & Recording: College of Creative Arts. https://www.ipr.edu/blogs/audio-production/what-are-the-elements-of-popular-music/

Street, J. (2003). “Fight the Power”: The Politics of Music and the Music of Politics. Government and Opposition, 38(1), 113–130. http://www.jstor.org/stable/44483019