Analyzing and Predicting the Popularity of Songs on Spotify

Isha Vohra[1], Navya Khanna[1], Shresha Das[1], Yash Singhal[1]

[1] Georgia Institute of Technology

Abstract

The popularity of a song is of importance to artists and recording labels. Understanding the factors and features that influence the popularity of a song can help artists produce songs that resonate with the public. This would be beneficial to artists and record labels alike to help them get the recognition they deserve.

This paper utilizes statistical metrics and machine learning models to predict the popularity of a song on Spotify and understand the features that may affect its popularity. Linear Regression, Random Forest Regression, and Neural Networks are used to predict the popularity of the songs. Statistical methods such as Mean Squared Error, Mean Absolute Error, and R-Squared were used to measure the accuracy of our models. Lastly, we found the correlation between the features and popularity of the tracks.

Random Forest Regression best modeled the popularity of songs, while other models like Linear Regression and Neural Networks predicted the extent to which the features affected the popularity of songs, but were not as accurate. The study also found that genre is the most important feature in predicting the popularity of songs, with features such as acousticness and speechiness strongly contributed as well.

Analyzing and Predicting the Popularity of Songs on Spotify

The music industry has reached unprecedented levels of success and keeps growing exponentially every year. Out of the huge number of songs released every year, only a fraction of them manage to capture a large population's attention. A song's popularity means a lot to the artist and recording labels alike. Even though there are several factors contributing to the success of a song (marketing, videography, timing, etc.), the musical elements of a song contribute the most. Figuring out which elements are important will help local and unrecognized artists gain the platform they deserve. One way to do so is by leveraging Machine Learning. This paper discusses the extent to which machine learning can predict the popularity of a song based on its musical elements and identifies the features that contribute to a song's success.

There have been studies in the past that have utilized machine learning to predict a song's popularity. Some of these studies successfully used ML for prediction and classification, while others chose popular songs from Spotify and applied clustering and supervised learning algorithms on their datasets to analyze which features make a song popular. (Araujo et al., 2019; Beitawi et al., 2020; Yee & Raheem, 2022). Research conducted by Beitawi and Salehan uses K-means analysis on Spotify's top 100 songs from 2017 and 2018. The study shows that high danceability and low instrumentalness increases a song's popularity (Beitawi et al., 2020). Yee and Raheem in their study focused on social media features from YouTube videos and Spotify audio features and trained multiple ML models (Neural Networks, Support Vector Machine, Ordinal Regression, and Random Forest) to predict music popularity (Yee & Raheem, 2022). They found that instrumentalness and danceability were very important but a song's happiness factor (valence) doesn't matter a lot.

Another paper titled “SpotiPred” uses a dataset with ~200,000 tracks and uses Linear Regression, Random Forest, and K means to predict popularity of songs and to quantify the importance of every feature (Gulmatico et al., 2022). They came to the conclusion that acousticness matters the most, followed by loudness, valence, liveness, instrumentalness, and then danceability. This study gave a contrasting result to the previous studies conducted on select popular songs (Beitawi et al., 2020; Yee & Raheem, 2022) which stated that danceability was the most important element of every song and that valence (the happiness metric of a song) was not as important. But this study, which was conducted on a larger dataset using predictive models gave a different result. This study aligns with ours since we also use a large dataset and use predictive models. The main limitation of this model is that they only used numeric features and left out features like genre which could be encoded.

A popularity analysis conducted by UCLA analyzes the popularity of songs based on the danceability metric on spotify using the spotify API and the various metrics that they have available (Ochi et al., 2021). In this paper, instead of using traditional ML algorithms, they used the software SAP, which provides visualizations, to help them analyze and create the necessary visuals. The results were confusing and were not defined clearly as SAP does not work towards predicting popularity. After reading this study, we decided not to use SAP in our analysis.

One of the main limitations of a few of these studies was that the datasets used were very small and that they only contained English songs. The clustering analysis by Beitawi and Salehan (Beitawi et al., 2020) used 200 popular songs, Yee & Raheem (Yee & Raheem, 2022) considered only 1432 tracks. To cover this problem, we used a dataset that has 232,725 tracks and is a cumulation of popular and unpopular songs from different parts of the world to capture a variety of tracks.

Additionally, many of these studies gave conflicting answers. The type of dataset, the number of tracks selected, the features selected, the types of models used, etc. are some of the factors that affected the results of these studies. The contradictory result of the importance of valence is something that we decided to explore further (Beitawi et al., 2020; Gulmatico et al., 2022; Pareek et al., 2022). This also motivated us to include 14 features including genre, which has non numeric entries, by encoding its values.

To better understand how to structure the study and machine learning models to predict a real world metric, it becomes necessary to look at other fields where ML and Natural language Processing has been used to analyze the popularity of some data. We referenced a paper by Lee and Yee (Lee & Lee, 2014), where they use Natural Language Processing to identify top news stories based on their popularity in the blogosphere. Here, NLP has been used to process and identify which article may become popular on the blogosphere based on its many attributes. The passage discusses why it may be challenging to do this for news stories and how it becomes important to do so. They measure the popularity from the viewpoint of the content and the timeline in which it occurred, which may be important factors to consider when looking at the popularity. The authors propose several methods to estimate and develop language models for news stories and blog posts and evaluate how important they are. Moreover, they generate a temporal profile of the stories to evaluate their importance based on when they happened in time. They then move on to experimentally verify the effectiveness of their approaches. This helps us identify a structure that served as the backbone of our study.

While reading through previous work, we also noticed that some of them mainly used classifying algorithms to create a sort of rubric for artists (Pareek et al., 2022). Classifying algorithms in ML create multiple “bins”, where each bin in these studies represents the level of popularity (low, medium, high). We use prediction models (Linear Regression, random forest regression) to accurately predict popularity on a scale of 0 to 100, which is the scale present in our dataset curated by Spotify (Spotify, n.d.). To better understand feature reduction techniques and which one would fit our model the best, we referenced papers that discuss the relation between PCA feature reduction and decision tree classification (Nasution et al., 2018), and another paper on how feature selection affects the accuracy of music popularity classification using ML algorithms (Khan et al., 2022). We also looked at a paper that discusses MRMD2.0, a python tool for Machine Learning feature selection and reduction to compare and analyze different tools (He et al., 2020). In the study conducted by Khan (Khan et al., 2022), they used the same ML models once with feature reduction and once without. The result was that there was a 0.01% difference in accuracy between the two methods. The paper on PCA (Nasution et al., 2018) discusses how PCA can help increase accuracy for decision tree models. Based on these papers, we decided to use the PCA feature reduction technique to try to improve our prediction accuracy.

The above review and our interests have led us to the following questions:

1. Does a song's positivity/valence affect its popularity?
2. Can Machine Learning models be trained on current songs to predict a new song's popularity?

To address these questions, we trained our dataset on Linear Regression model, Neural Network, and Random Forest Regression to find the correlation between common musical features and song popularity. This also helped us predict popularity on a scale of 1 to 100. We used 3 ML models to gain a comprehensive understanding of our results and the strengths and weaknesses of every model. The rest of the paper follows the following order: in methods we detail how we collected the data, the types of models used, and how we ran them. In the results section, we discuss the performance of all the models based on our chosen metrics and provide visualizations of the same. Next comes the discussion section which relates our study's findings to past work in the same field. We also cover the strengths and limitations of our study in the context of our hypothesis and conclude with potential directions for future work.

Methods

In detailing our procedure, it is important to start by mentioning how this dataset was chosen. The research presented in this paper has made use of secondary data from Kaggle rather than using primary sampling and recruiting techniques since this is a Machine Learning study where using a dataset with a high usability score is sufficiently relevant. Specifically, the "Spotify Track DB" dataset from Kaggle was the dataset chosen (Hamidani, 2018).

This particular dataset was chosen because it contained the features that have been shown to be the best for the implemented models based on past research, and they had a high completeness score. We coded in the Python programming language and made use of sklearn's packages for our Random Forest, Linear Regression, and Neural Network models. The Kaggle dataset contains 232,725 tracks and covers 26 genres (~10,000 songs per genre). There are 18 features: genre, artist name, track name, track id, popularity, acousticness, danceability, duration, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time signature, and valence. Since the dataset was relatively clean, minimal cleaning was performed. Principal Component Analysis (PCA) was attempted to be implemented to reduce the dimensionality of the dataset, but this was not used since it led to a huge reduction in the explained variance. In other words, reducing one variable meant that the other variables explained the data less and were not representative of the full picture. Therefore, we used forward sequential feature selection with R^2 as the metric that this was based on.

For the measures of the independent variables of our study, we looked to the Spotify API's preset details on these to show how conceptual variables were operationalized for our study. For example, for the danceability variable, a value from 0.0-10.0 (where 10.0 is most danceable) was assigned as per how well a song was fit to dancing based on tempo, the strength of the beat, and regularity. In some cases, like that of the danceability independent variable, some of the features built off others that were evaluated in the study. This is exemplified by how tempo is one of the components that contributes to the danceability score. The dependent variable, which is track popularity, was measured as a value between 1 and 100 (where 100 is the most popular) as per the Spotify API as well. The score is calculated by an algorithm that is proprietary to Spotify. It is mainly based on the number of times a song has been played and the recency of listening.

Based on past studies, it was known that Linear Regression may not produce a high accuracy, but it was instrumental for feature selection. This is why the procedure for this study started with running this model. Random Forest and Neural Network models were also run from sklearn. The basic Neural Network model that was run was able to give more information about non-linearity that our other two models would not be able to show; although, the study did not move forward with this model because of its low accuracy relative to Random Forest. From here, since the Random Forest model had the highest level of success (highest R^2 value), this is the model that was used for the rest of the analysis of the study.

The design choice is supported by previous research which has successfully made use of similar methods in its analysis. One such study similarly used data from the Spotify API due to the generalizability of the data from this source (Araujo et al., 2020). The study then also made use of a model from sklearn, Support Vector Machine. In a study called Spotipred, in a manner that was similar to our research, cross validation techniques were employed (Gulmatico et al., 2022). The audio features that were evaluated in this study included tempo, acousticness, and danceability, and these were features that were considered in the study's analysis amongst others that haven't been studied in similar research.

The study has also been protected from confounds in its design through techniques like cross validation and bootstrapping of the models since these methods can help validate a model by showing accuracy and generalizability. Cross-validation involves resampling that divides the data into k equally sized subsets, training the model k-1 times, and then repeating this process for a total of k times while bootstrapping is a technique by which variability is determined by drawing randomly selected samples of the data with replacement. These techniques were used specifically in the Random Forest model. Another way that confounds were avoided was to make sure the data used was generalizable. Since Spotify is one of the most broadly used platforms for music streaming, the popularity scores and related metrics from this platform are representative of the population as a whole. Spotify data gives a relevant population demographic also since it has a user base of over 400 million users, and this large sample size can help provide reliable information about user preferences ("Spotify Web API", 2023).

Results

The results of our study indicated that genre was the most important feature in the prediction of song popularity, which was followed by acousticness, speechiness, and instrumentalness. The first null hypothesis was rejected which stated that there is no relationship between the positivity of a song and how popular it can be predicted to be. On the other hand, the second null hypothesis, which is that there is no relationship between the predictions on popularity that machine learning models give and their true popularity based on the number of streams, was rejected. This is because the results indicated that machine learning models like Random Forest Regression were able to accurately predict song popularity. This was based on the Random Forest implementation and is depicted in Figure 1 below. The results also indicated that Random Forest was the best model out of Linear Regression and Neural Networks to predict song popularity given the specific features of our dataset. For the statistical analysis, the model produced an R^2 score of 0.96, which is the coefficient of determination and was the regression metric we used to evaluate model success. It indicated the fraction of variation in the dependent variable (track popularity) which was explained by the independent variables in our regression model.

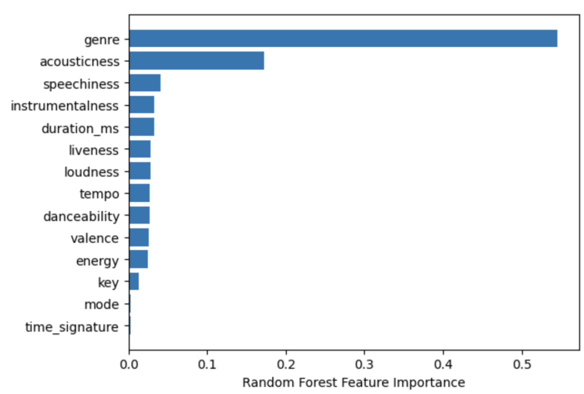

Figure 1: Random Forest Feature Importance

Metrics such as mean squared error and mean absolute error were used to measure the difference in the predicted values from the actual values to indicate model accuracy. Random Forest had the lowest error values with a mean squared error of 86.23 and a mean absolute error of 7.04. Linear Regression produced a score of 0.22 and errors of 256.68 and 12.88. The low score and high error values indicate that this model did not lead to successful prediction of song popularity. The Neural Network model produced a score of 0.74; however the values of its mean squared error and mean absolute error were 1961.19 and 40.69 respectively, which are very high. Therefore, both Linear Regression and Neural Networks displayed a poor performance relative to Random Forest.

Discussion

The purpose of our research paper was to investigate the extent to which machine learning can predict the songs that will become famous based on musical elements and more specifically, to predict which features are the most important determinants of the popularity of a song. Based on past research, this study focused on Random Forest, Linear Regression, and Neural Network models to predict song popularity given the features in the dataset to fill the gaps of current research. The rationale behind running three models was to compare and contrast the results of each to reason their strengths and limitations with respect to the research focus. This way the research question was considered holistically, addressing the limitations of each model. Python was used as the primary programming language and sklearn's packages were used for data preprocessing, feature selection, and model implementation.

Based on the results, genre was found to be the most important factor that contributed to song popularity by a large margin. This was followed by acousticness, speechiness, and instrumentalness. The past research that was encountered as part of the literature review attributed song popularity most strongly to high danceability and low instrumentalness (Beitawi et al., 2020), which was in contrast to the research findings.

The importance of song genre in the prediction of song popularity allows for categorization of songs based on their stylistic, technical, and thematic characteristics. These classifications of songs play a key role in guiding listener preferences and choices, ultimately influencing overall song popularity. Other important features that were found included acousticness, speechiness, and instrumentalness, each of which encapsulates different elements of musical characteristics that impact listener preferences and their likelihood of enjoying a given song.

Out of the three machine learning models that were implemented, Random Forest produced the best results. Given the specific features in the dataset, this model had the best prediction of song popularity with an R^2 score of 0.96. This value demonstrates if the model is a good fit and how good of a predictor it is for regression models, i.e., how far the predicted value is from the actual value, with the best possible value of 1.0. The efficacy of the Random Forest model was attributed to several factors. Firstly, the Random Forest model performs well with large datasets containing many features. This dataset has a high dimensionality with 232,725 tracks and 14 features, which was successfully handled by Random Forest. This is because Random Forest is an ensemble learning technique and involves the random selection of features in each decision tree, which reduces the chance of overfitting and variance. The flexibility of the algorithm in terms of being suited for both classification and regression problems allows it to handle the prediction of binary results such as if a song will be popular or not as well as the continuous popularity score values. This result was consistent with the results found in the study conducted by Pareek et al. (2022), which found that Random Forest had the most accurate predictions of song popularity among models like K-Nearest Neighbor and Linear Support Vector Classifier. This paper produced the given results with different metrics including precision, recall, and F1 score as well as by running different models, which further grounds the success of Random Forest as a good predictor of song popularity. This result was also supported by the results found in the paper by Gulmatico et al. (2022), which indicated an accuracy of 95.37% with Random Forest. The model also produced a mean squared error of 86.23 with Random Forest, which is indicative of the model's accuracy in predicting song success. This can be further reduced by accounting for variables such as social media features (Yee & Raheem, 2022) and the time frame that a song has been out (Araujo et al., 2019). The poor performance of linear regression (mean squared error - 256.68) in our study was attributed to the fact that the research problem was not linear and this model does not encapsulate the non-linearity in musical characteristics (Gulmatico et al., 2022). Neural Networks did not provide a good result since it was a single layer neural network and thus had limited ability to handle complex patterns and feature interactions.

The results failed to reject the null hypothesis, which states that there is no relationship between the positivity or valence of a song and its predicted popularity. Based on the analysis, this can be explained by the fact that valence is a highly subjective measure, the effect of which can be overpowered by features such as genre, acousticness, instrumentalness, etc. The second hypothesis's null hypothesis, which states that machine learning models can be used to predict the popularity of songs accurately, was rejected. While the linear regression and neural network models did not perform as expected, the random forest model produced results with a high accuracy and low error rate. By accounting for confounding variables such as song marketing and promotions, the artist name and popularity, and overall social media influence on general public perception, it is anticipated to produce more accurate results with different machine learning models. Moreover, for neural networks specifically, multiple layer models can be implemented to overcome the single layer limitations discussed above. Since this study did not consider these confounding variables in much detail, they could be a potential reason for the weaknesses in the design and results.

Despite its weaknesses, this study had well grounded strengths that provide a solid foundation for the accuracy of the results. First, a relatively clean and comprehensive dataset with 232,725 tracks, covering 26 genres was chosen which ensured that our results covered a wide variety of musical characteristics and were generalizable. Second, it was ensured that the data and thus our results were generalizable across song genres and other features like danceability, acousticness, valence, etc. by considering Spotify data, which is one of the most widely used music streaming platforms covering a large user base. Third, the results produced by the Random Forest model were verified by running other models to address the confounding variables and its potential limitations.

Based on the design and results, future work can explore focusing on specific musical genres and subgenres. Given that genre is one of the most important features for song popularity prediction, it is important to investigate how song popularity may differ across genres and what features specifically influence song success within subgenres. This can have important applications in developing more accurate genre specific models, which address a narrower problem space. Another area that future work can examine in greater detail is the incorporation of more diverse data since most research in this field so far has primarily focused on Spotify data. Addressing the confounds in this study by considering other data sources as well as characteristics such as listener demographics and virality trends on social media would thus help provide a more comprehensive and accurate prediction of song popularity. In conclusion, this research domain has widespread implications in the music industry in terms of the integration of technology with artist success, which is especially important for upcoming artists. Understanding listener preferences also has implications in consumer behavior and trends, which can have applications in interdisciplinary research. Thus, research in this domain is significant not only to the music industry but also other fields that harness the growing power of technology.

References

Araujo, C., Cristo, M., & Giusti, R. (2019). Predicting Music Popularity on Streaming Platforms. *Anais do XVII Simpósio Brasileiro de Computação Musical*, 141-148. 10.5753/sbcm.2019.10436

Beitawi, A. Z., Salehan, M., & Zhang, S. (2020). What Makes a Song Trend? Cluster Analysis of Musical Attributes for Spotify Top Trending Songs. *Journal of Marketing Development and Competitiveness*, *14*(3), 79-91.

Gulmatico, J. S., Susa, J. A. B., Malbog, M. A. F., Acoba, A., Nipas, M. D., & Mindoro, J. N. (2022). *SpotiPred: A Machine Learning Approach Prediction of Spotify Music Popularity by Audio Features*. IEEE. Retrieved April 10, 2023, from https://ieeexplore.ieee.org/document/9776765/authors

Hamidani, Z. (2018). *Spotify Tracks DB*. Kaggle. Retrieved April 1, 2023, from https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db

He, S., Guo, F., Zou, Q., & HuiDing. (2020, December). MRMD2.0: A Python Tool for Machine Learning with Feature Ranking and Reduction. *Current Bioinformatics*, *15*(10), 1213-1221. ProQuest. 10.2174/1574893615999200503030350

Khan, F., Tarimer, I., Hathal, S. A., Karadağ, B. C., Fayaz, M., Abdusalomov, A. B., & Young-Im, C. (2022). Effect of Feature Selection on the Accuracy of Music Popularity Classification Using Machine Learning Algorithms. *Electronics*, *11*, 3518. Proquest. 10.3390/electronics11213518

Lee, Y., & Lee, J.-H. (2014). Identifying top news stories based on their popularity in the blogosphere. *Information Retrieval*, *17*(4), 326–350. Springer. https://doi.org/10.1007/s10791-014-9241-z

Nasution, M. Z. F., Sitompul, O. S., & Ramli, M. (2018). PCA based feature reduction to improve the accuracy of decision tree c4.5 classification. *Journal of Physics*.

Ochi, V., Estrada, R., Gaji, T., Gadea, W., & Duong, E. (2021, August 5). *Spotify Danceability and Popularity Analysis using SAP*. Retrieved April 12, 2023, from https://arxiv.org/abs/2108.02370

Pareek, P., Shankar, P., Pathak, P., & Sakariya, N. (2022). Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify. *Journal of Development Economics and Management Research Studies*, *9*(11), 10-19. https://www.cdes.org.in/wp-content/uploads/2022/01/Predicting-Music-Popularity.pdf

Spotify. (n.d.). *Web API*. Spotify for Developers. Retrieved April 2, 2023, from https://developer.spotify.com/documentation/web-api

Yee, Y. K., & Raheem, M. (2022). Predicting Music Popularity Using Spotify and YouTube Features. *Indian Journal of Science and Technology*, *15*(36), 1786–1799. https://doi.org/10.17485/ijst/v15i36.2332

Appendix

Figures

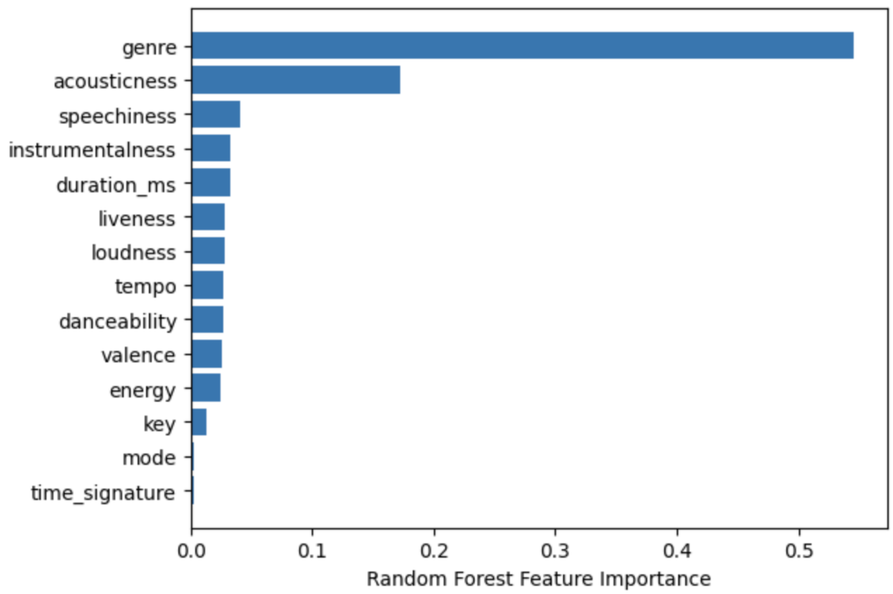

Figure 1: Random Forest Feature Importance

Dataset

The dataset for the study was taken from kaggle.

This is the explanation of every measure:

https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

Link: https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db