Spotify Analytics and Predictive Model for Hit Songs

Shun-Ting (Trista) Wang
stw@umd.edu
University of Maryland College Park

I. Introduction

Spotify is a music streaming service that allows users to listen to millions of songs, podcasts, and other audio content from a wide range of artists and record labels. It was founded in 2006 in Sweden and launched to the public in 2008. Today, it is one of the most popular music streaming services in the world, with over 345 million active users as of December 2021.

Spotify has played a significant role in the music industry by providing a convenient and easy-to-use platform for music listeners. It has become an important part of the modern music landscape, and continues to evolve and adapt to meet the changing needs of its users and the industry.

What's more, the company also provides analytics tools that allow artists, record labels, and other industry professionals to track the performance of their music on the platform. "Spotify Analytics" is a feature that provides detailed data and insights on the performance of an artist's music on the platform. It allows users to see how many streams their tracks have received, where their listeners are located, and other key metrics.

Spotify Analytics was introduced in 2015 as part of the company's efforts to provide more transparent and comprehensive data to artists and industry professionals. Since then, the feature has undergone several updates and improvements, including the addition of new metrics and the ability to track the performance of individual tracks, albums, and playlists.

It has become an essential tool for musicians, record labels, and other industry professionals who want to understand how their music is performing on the platform and reach new listeners. It has also helped to increase transparency and accountability within the music industry, allowing artists to better understand how their music is being consumed and shared.

II. The stakeholders and desired business outcome

Musicians, record labels, music industry professionals, music producers, and the Spotify users are the potential beneficiaries for this topic. Musicians and bands are the primary creators of the music that is streamed on Spotify, and they stand to benefit from understanding the performance of their music on the platform. Spotify Analytics provides data and insights that can help artists understand how their music is being consumed and shared, and can help them identify opportunities to reach new listeners.

Record labels invest in the production and promotion of music, and they need to track the performance of their artists' music to understand the return on their investment. Spotify Analytics can help record labels understand the popularity and reach of their artists' music and make informed decisions about future investments.

Music producers can also better understand which tracks are popular with listeners, identify trends in the music industry, and make informed decisions about their music releases and promotion.

Music industry professionals in the music industry, such as managers, agents, and marketers, can also benefit from Spotify Analytics. By understanding the performance of music on the platform, they can better support and promote the work of their clients and identify opportunities for success.

Ultimately, the intended beneficiaries of Spotify Analytics and hit song prediction are the users of the platform. By providing data and insights to artists and industry professionals, Spotify can help users discover new music that they will enjoy and increase the overall quality and variety of content on the platform.

III. Description of the types of analytics that are used

This dataset provides audio information for around 600,000 Spotify tracks. The data is divided into around 20 columns, each of which describes the track and its characteristics. Spotify's API offers a wide range of audio features that developers can analyze the following characteristics of tracks:

1. duration_ms: The duration of the track in milliseconds.
2. key: The estimated overall key of the track. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
3. mode: Mode indicates the modality (major or minor) of a track. Major is represented by 1 and minor is 0.
4. time_signature: An estimated overall time signature of a track.
5. acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.

6. danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.
7. energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.
8. instrumentalness: Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
9. loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Values typically range between -60 and 0 db.
10. speechiness: Speechiness detects the presence of spoken words in a track.
11. audio_valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
12. tempo: The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
13. song_popularity: Song popularity score as of April 2021 on a normalized scale [0-100] where 100 is the most popular.
14. liveness: Detects the presence of an audience in the recording.

In this analysis, I compared the audio characteristics of the top 1000 songs on Spotify to the characteristics of the entire dataset and found that the distributions were similar, but the two groups had some notable differences. The top songs were more likely to be released in recent years, have shorter durations, be more explicit and danceable, and have higher energy, loudness, and slower tempos. However, some of the top artists, including Justin Bieber, Billie Eilish, and Juice WRLD, had higher acousticness than the average of the top 1000 songs, which is unusual because most acoustic songs are not as danceable. This unique combination of high acousticness and high danceability may set these artists apart from others on the platform. Alternatively, it could be that these artists are able to create highly-acoustic songs that still become hits due to their emotional appeal, lyrics, or background story.

I also examined the correlations between different audio characteristics and popularity and found that danceability had the highest correlation with popularity, followed by acousticness, explicitness, and speechiness. Valence and danceability were also highly correlated. These findings suggest that while certain characteristics, such as high energy and danceability, may increase the likelihood of a song's popularity, other factors such as emotional appeal and lyrics may also play a role. Overall, we can conclude that it is not solely the audio attributes of a song that determine its popularity, but rather a combination of factors that may include the artist, the song's background story, and its emotional appeal.

I plotted a graph to analyze the change in audio characteristics of songs on Spotify over the years and discovered that characteristics such as acousticness and instrumentalness have decreased, while energy, loudness, and danceability have increased. As of 2021, the order of these features is loudness, danceability, energy, and valence, with valence remaining relatively stable around 0.6, indicating that most songs are generally happy.

IV. Predictive model for popular song

In the preprocessing stage, I split 20% of the data as testing data, which will be used to evaluate the accuracy of our machine learning model. I also created a highly popular column for songs with a popularity score greater than 50, and used this feature as the target for predicting song popularity. To build our model, I created a pipeline to normalize the data and apply machine learning techniques to train and test the model. This made the model accurately predict the popularity of songs based on their audio characteristics and other relevant features.

I used logistic regression, random forest, and XGBoost as the methods to predict the hit songs. The logistic regression model gave the result of 67.52% accuracy which was not a good number. The random forest model has 87.47% accuracy which was the best result I have. XGBoost had 70.64% accuracy which was a moderate result. Although I have the accuracy of 84% from the predictive model, the precision of predicting the popular song is only 53%. This number could be improved if I have more popular song's data.

A random forest model is a type of machine learning algorithm that is used for classification and regression tasks. It is an ensemble model, meaning that it is made up of multiple decision trees working together to make predictions.

A random forest model could be used to predict the popularity of a song based on its audio characteristics and other relevant features. To do this, the model would be trained on a dataset of songs with known popularity scores, using the audio characteristics and other features as input variables. Once the model has been trained, it can be used to make predictions about the popularity of new songs by inputting their audio characteristics and other relevant features into the model.

One advantage of using a random forest model for this task is that it is able to handle large amounts of data and can handle data that has a large number of features (such as audio characteristics). It is also relatively robust to overfitting, meaning that it can generalize well to new data. Additionally, the fact that it is an ensemble model means that it can make more accurate predictions than a single decision tree.

The importance of audio features for the popularity of songs in the general dataset differs from that of the top 1000 songs. For the general dataset, the top features related to popularity are year, loudness, energy, explicitness, and danceability. However, for the top songs, the most important features are year, danceability, acousticness, explicitness, and speechiness. This may be because the recent popular genre of music has been hip-hop, which tends to be more speechy and explicit. In terms of the feature importance for popularity, as determined by a random forest model, the order is loudness, acousticness, duration, valence, and speechiness.

There may be differences in the feature importance as determined by a random forest model and a correlation matrix because these two techniques measure feature importance in different ways. The random forest model calculates feature importance by looking at how much each feature contributes to the accuracy of the model's predictions, while the correlation matrix calculates feature importance by measuring the strength of the linear relationship between each feature and the target variable.

Additionally, the results of a random forest model may be affected by the specific parameters used for training the model, such as the number of trees in the forest or the depth of the trees. On the other hand, the results of a correlation matrix are not affected by these parameters and are simply a measure of the strength of the linear relationship between the features and the target variable.

IV. Two examples of using analytics and prediction for this topic

Here are two examples of using analytics for Spotify analysis and a predictive model for hit song prediction. First of all, analyzing audio features to determine the characteristics of hit songs: One way to use analytics in a Spotify analysis is to analyze the audio features of popular songs to determine what characteristics they have in common. For example, you could use a machine learning model to analyze features such as loudness, tempo, and genre, and identify patterns or trends in the data that are correlated with song popularity. Lastly, using a predictive model to forecast the success of new songs: Another example of using analytics in a Spotify analysis is to build a predictive model that can forecast the success of new songs. This could involve training a machine learning model on a dataset of past hits and using that model to predict the likelihood of a new song becoming popular. The model could be trained on features such as the artist's popularity, the genre of the song, and the audio features of the track.

V. Conclusion and future work

I used audio features to predict the popularity of songs in a dataset containing approximately 600,000 songs. I defined songs with a popularity score greater than 50 as the target for prediction and used a random forest model, which had the highest accuracy of 87.47%. While the overall accuracy of the predictive model was 84%, the precision for predicting popular songs was only 53%. One way to improve the way that analytics methods are used for Spotify analytics is to incorporate a wider range of data sources into the analysis. While audio features are certainly important in understanding the characteristics of popular songs, there are many other factors that can influence a song's success, such as the artist's popularity, the marketing efforts behind the song, and the lyrics of the song. By incorporating data on these additional factors into the analysis, it may be possible to get a more complete picture of what drives song popularity and to make more accurate predictions. Another way to improve the use of analytics for Spotify analytics is to focus on using more advanced machine learning techniques. While traditional statistical methods can be useful for understanding patterns and trends in the data, more advanced techniques such as deep learning or natural language processing can potentially provide even more insights. These techniques can be used to analyze complex relationships in the data and to make more accurate predictions. Finally, it may be helpful to involve a diverse team of experts in the analytics process, including data scientists, music industry professionals, and marketing experts. This can help to ensure that all relevant perspectives are considered and that the analytics efforts are aligned with the goals and needs of the business.

There are many potential areas for future work in Spotify analytics and model prediction for hit songs on the platform. First of all, we can incorporate more data sources into the analysis, such as social media data, marketing data, or data on the lyrics and themes of the songs. This could help to provide a more complete picture of what drives song popularity and enable more accurate predictions. Another area of focus could be on developing more advanced machine learning techniques for predicting hit songs on Spotify. For example, techniques such as deep learning or natural language processing could be used to analyze complex relationships in the data and make more accurate predictions. Another potential direction is to focus on analyzing trends in specific genres or markets, such as hip-hop or electronic dance music. This could involve developing specialized predictive models or analyzing the characteristics of popular songs within a specific genre or market. Still another area of focus could be on improving the accuracy of predictions made by machine learning models. This could involve developing new methods for evaluating the performance of the models, or fine-tuning the model parameters to optimize their predictive power. Last but not least, we can apply the desired song to three predictive models and use a voting system to decide if it is likely to be a hit or not. This could help music professionals to make decisions.

VI. References

1. Spotify Song Popularity Analysis. https://rpubs.com/mary18/919897
2. Rutger Nijkamp (2018) Prediction of product success: explaining song popularity by audio features from Spotify data.
3. Al-Beitawi et al. (2020) What Makes a Song Trend? Cluster Analysis of Musical Attributes for Spotify Top Trending Songs

4. Spotify Popularity Prediction-ML Practice
https://www.kaggle.com/code/pelinsoylu/spotify-popularity-prediction-ml-practice/notebook
5. Taylor Fogarty (2019) Predicting the Future (of Music)
https://towardsdatascience.com/predicting-the-future-of-music-c2ca274aea9f
6. Alex Hsieh et al. (2019) Music attributes and its effect on popularity
https://ahsieh53632.github.io/music-attributes-and-popularity/