BUS306 Final Project

Allyssa Utecht, Josh Rassin, and Mary Ardoin

Introduction

Knowledge of what makes a song popular is important for artists to maximize the profitability and the fame of their music. The dataset looks at release dates, playlist additions, streaming statistics, Apple Music presence, and audio features. These statistics are important for knowing the ideal release data to maximize profit, and taking advantage of music trends over time. Opportunistic release dates ensure that the music has the biggest impact and is most successful. Additionally, knowing where songs and albums are being added to playlists signals to music managers which platforms and companies are most used by listeners and customers, illustrating which are the most profitable. Playlists are an indicator of song and genre popularity. The streaming statistics of a song can be a helpful indicator of the success of marketing strategies for increasing song popularity. Additionally, streaming platforms like Apple Music and Spotify have different monetization models, so artists receive revenue differently; knowing which are most used by listeners of various genres is useful for artists to know where to stream their songs. In addition to maximizing the popularity and profit of their own music, artists may want to know statistics on playlists, platforms, and audio to understand their competition and establish benchmarks for comparison. This dataset and the conclusions that can subsequently be drawn provide a comprehensive overview of the music industry in 2023, which is useful for business and marketing decisions. This article illustrates how AI and ML can be used to predict hit songs, which is typically very hard to do. The ML model accurately predicted hit songs 82% of the time, illustrating the power of statistics and AI for influencing and understanding the music industry. Another study used random forests, an ML method, to understand how song trends have changed over time. Songs are becoming happier, more ‘party-like’, and less relaxed. The artist’s previous success is also related to a song’s success. This study illustrates how audio features and streaming statistics, two key variables in our dataset, greatly influence the success and profitability of a song. Music analytics provide a clear picture of listener engagement and which marketing initiatives, including platform and release date, are most successful. Listener preferences are key determinants of song popularity, which translates into profitability for the artist and producer. Understanding music characteristics and trends alongside AI and ML methods can help us predict what it takes to create a chart-topping song.

Data

The dataset was downloaded from Kaggle and comprehensively looks at all of the most streamed songs on Spotify in 2023. The columns are statistical descriptors, and the rows are the top songs streamed on Spotify in 2023. There are 953 rows, each containing a unique song that was one of 2023's hottest Spotify hits. In addition to the track name and artist name, 22 other variables contained information about the song and could be analyzed. Most of them were quantitative and considered things like how many times a song appeared in a Spotify playlist, its rank on a chart, the beats per minute, etc. Some of the major categorical variables were the mode of the song, the month it was released, and the key of the song; the descriptive statistics of these can be seen in the code. The average beats per minute was 122, the average times it appeared on Spotify playlists was 5,200, and the average 'energy level' was 64%. As for biases and limitations, while Spotify is one of the top music streaming platforms, it is not entirely indicative of the overall popularity and success of a song because music can be listened to on so many platforms and in so many ways. Spotify listeners are not entirely representative of the listening population, and certain genres are more successful on Spotify because of its listening demographic. Additionally, the rise of social media and platforms like TikTok have greatly skewed which songs become popular and how they become popular, so rankings in 2023 are not a great determinant of general trends in the music industry that determine success. While certain release dates, streaming platforms, and BPM and energy music styles may typically be more popular, the rise of social media may skew these trends. While the dataset is a great look at popular Spotify songs in 2023, it is not a comprehensive and holistic view of the whole music industry for that year.

Key Features (Columns):

- **track_name**: Name of the song
- **artist(s)_name**: Name of the artist(s) of the song
- **artist_count**: Number of artists contributing to the song
- **released_year**: Year when the song was released
- **released_month**: Month when the song was released
- **released_day**: Day of the month when the song was released
- **in_spotify_playlists**: Number of Spotify playlists the song is included in
- **in_spotify_charts**: Presence and rank of the song on Spotify charts
- **streams**: Total number of streams on Spotify
- **in_apple_playlists**: Number of Apple Music playlists the song is included in
- **in_apple_charts**: Presence and rank of the song on Apple Music charts
- **in_deezer_playlists**: Number of Deezer playlists the song is included in
- **in_deezer_charts**: Presence and rank of the song on Deezer charts
- **in_shazam_charts**: Presence and rank of the song on Shazam charts
- **bpm**: Beats per minute, a measure of song tempo
- **key**: Key of the song
- **mode**: Mode of the song (major or minor)
- **danceability_%**: Percentage indicating how suitable the song is for dancing

- **valence_%**: Positivity of the song's musical content
- **energy_%**: Perceived energy level of the song
- **acousticness_%**: Amount of acoustic sound in the song
- **instrumentalness_%**: Amount of instrumental content in the song
- **liveness_%**: Presence of live performance elements
- **speechiness_%**: Amount of spoken words in the song

Code

https://colab.research.google.com/drive/1lrgmWOiXtPJUtsV50BGXFkuJz-0DstvI?usp=sharing

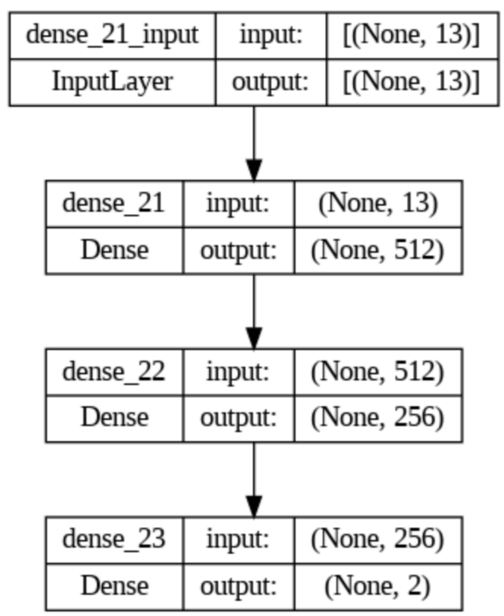
Discussion

One of the first things that was necessary to do was clean the data and select the variables we would be using. When cleaning the data, we created a 'top_100' column. This column was a categorical variable that measures the top 100 songs based on Spotify charts ('in_spotify_charts'), either having a value of 1 (top 100 songs) or 0. This variable was the y value in our models. For our x values, we wanted to choose variables that would impact the success of a song on Spotify, but not directly tell us it. We excluded song names, artists, information regarding other music streaming platforms, and direct measures of song success. The following are the X values selected in an attempt to classify whether a song is a top 100 song or not: 'artist_count', 'released_year', 'released_month', 'released_day', 'in_spotify_playlists', 'bpm', 'danceability_%', 'valence_%', 'energy_%', 'acousticness_%', 'instrumentalness_%', 'liveness_%', 'speechiness_%'.

After looking at all the models, you can see that each one pulls different conclusions from the data and reveals different trends. The first principal component analysis (PCA) model of 500 random data points simplified the training data set into a smaller set, with two principal components, to make it simpler to analyze but still accurate. We created a scree plot and then plotted the PCA into 2D space, to visualize the distribution after PCA, revealing potential patterns in the dataset. Most of the data we plotted is clustered on the left side, hugging the y-axis of the graph. Though there are no distinct patterns, there are a few outliers on the right side of the graph, with the number of points growing fewer as they spread away from the y-axis and large cluster. There looks to be a larger spread of values when it comes to non-top 100 songs (blue) along the y-axis, with red values seeming to stay between -40 and 20. Seemingly, there isn't any extreme distinction between the two classes.

The next model we ran was a decision tree classifier using the original training and testing data, not the PCA, as a rule-based approach to classify the data. Using this classifier, the training data had 100% precision and recall. The testing data had a high f1-score for non-charted songs (0.918) but a worse performance with Top 100 songs (0.300). However, the overall accuracy remained fairly high with a score of 0.853. We then used the 5-fold cross-validation method using the decision tree classifier model to verify the F1 and accuracy scores. The data was split into ten subsets that the model looped through, creating new training and testing datasets and producing results. The average F1 and accuracy scores paralleled the Decision Tree classifier, with the average F1 score being 0.267 and the average accuracy score being 0.871. Concluded from these tests, the model has difficulties with the precision of the top 100 songs but has a fairly high overall accuracy. One reason could be the class imbalance between non-top 100 songs and top 100 songs. Measures were taken to balance the class weights to help counteract this.

We then created a deep learning neural network model using Keras to analyze the top 100 songs, separating our selected X values and the 'Top 100' variable as the y value. Similar to our simple models, the class weight was balanced. We split the training and testing sets to check the model's performance on unseen data and then converted the y labels into categorical forms. The model takes dummy variables to create dense layers that dwindle into a final output. The following is a visualization representing the structure of our model…



The trained model then made predictions of the y-value based on the X-testing data. We then tested the accuracy of the predicted y values vs. the true y values. The result was a 0.895 accuracy score. This means that the model was fairly successful in predicting the correct y values based on the data.

Finally, we used OpenAI's API to create a model that uses our selected variables to determine if a song will be in the top 100. We utilized a few-shot classification model (with numerical features) and created a prompt text that should elicit a probability estimate (of a song being in the top 100) based on the variables/song features inputted (same as our other models). The prompt put into the GPT-3.5-turbo model produces multiple responses with randomized values in each variable and makes a prediction. The model's song prediction is compared to the real ranking of the song to evaluate the model performance. Upon testing the predictions against our testing data, the model performed extremely poorly. The R-squared score, similar to the F1 score was -1.539, a very poor fit for the data.

Overall, each model utilized different aspects of machine learning techniques that can be utilized for data analysis. The PCA model analyzed the variability of the data itself, the decision tree classifier to classify the data, and the 5-fold cross-validation to verify the scores of the decision tree. These did not predict anything about the data but gave insight into the strength and variability of certain features. We then used the Keras API of TensorFlow to construct a neural network, and OpenAI's GPT-3.5-turbo, for prediction. The Keras neural network was able to understand complex patterns in the data to make predictions, and GPT-4 is useful for understanding and generating text. Our results match their strengths in that our neural network performed well while our OpenAI model did not. In summary, the models all have different applicability and importance for data analysis. Each plays a different role in understanding, interpreting, and utilizing the Spotify dataset. There are multiple limitations we came across with our analysis. One drawback to the neural network and GPT-4 is the computational power required to run these models. Another large limitation is the data imbalance between the two classes. If we're looking at the Top 100 songs, that means the rest of the data is much larger, creating an imbalance. Throughout our analysis, we modified our class weights to reflect this imbalance. However, this still likely accounted for the low F1 scores with the Top 100 songs classification. Continuing on the issue with our data, our goal was to predict whether we could use song characteristics to determine whether it could be a Top 100 song. While we separated the top 100 songs from the data, the data still contained highly popular songs. Because all the songs were top-streamed songs within the year, it could make it harder for our models to accurately predict chart success (since most songs are nowhere near that level of success). Overall, our findings provide insight into the possibility of using AI and machine learning to manufacture songs that are statistically likely to "blow up". This application could be huge for music producers and could likely change the world of music.