Finding Billboards 100 Predictores using Supervised Models and Unsupervised Models

Smith College Department of Statistical and Data Science, Nicole Sanchez Flores, Debora Camacho, Ari Cross, Vanessa Pliego, Kimberly By Goytia

Introduction

In the United States music industry, the Billboard Hot 100 is the primary music chart used to rank songs by popularity. The Billboard is a ranked-list that identifies the top 100 most-played songs nationally based on a combination of sales, online streaming, and radio airplay. The Billboard uses a confidential formula that implements these factors to produce the weekly chart. Additionally, paid streams are weighted higher than free streams. Since the Billboard is regarded highly within the music industry, an artist might be interested in exploring whether their song can rank a high position on the Billboard or not. There are numerous qualities about a song that can contribute to its popularity, but what qualities are the most relevant? Perhaps a song's genre, how energetic it is, danceability, etc. With these qualities identified, artists and labels might better understand how their upcoming release will be received by the public in terms of their popularity ranking on the Billboard, using a predictive model. To better understand music trends within the Billboard Hot 100 that can be used for model building, data will be sourced from the "Billboard Top Songs" dataset by Samay Ashar housed on Kaggle. This dataset contains 5,000 songs blending Spotify charts data with synthetic entries and is being actively updated as of March 19, 2025.

Research Question

What aspects (i.e. genre, danceability, energy, tik tok virality, etc.) of a song influence a high position on the Billboard Top 100? Based on these specific traits, will a song land a top 10 position on the Billboard Top 100?

Methods

Hierarchical Clustering

- Created a preliminary complete Hierarchical Clustering model with 4 clusters to evaluate the relevance of the features in our dataset but to also capture any trends without a pre-specified label through unsupervised learning.
  - We find that variables **"streams", "genrepop",** and **"peak position" cluster** which suggests that popular songs with high streams and peak positions are **usually pop songs**.
  - Additionally, **"weeks on chart"** and **"energy"** cluster which indicates higher energy songs might stay on chargers longer than less energetic songs.

Logistic Regression

- This supervised learning approach was used to explore the probability of a song appearing on Billboard Top 10 based on danceability. We found largely that **none of these predictors yielded statistically significant predictions** to whether a song would appear on BB100 or not in both single predictor models as well as models encompassing all predictions.

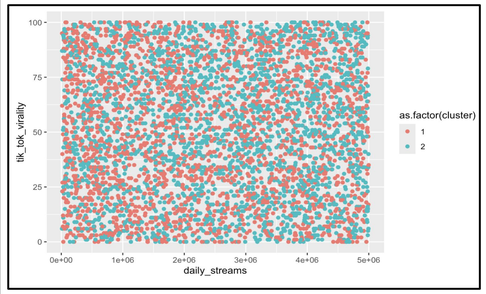
K-Means

- An unsupervised approach that uncovers patterns by grouping similar data points.
  - **Variables Used:** streams, daily_streams, weeks_on_chart, lyrics_sentiment, tik_tok_virality, danceability, acousticness, energy
  - **Clusters:** In order to decide what the optimal quantity of clusters would be we utilized silhouette cluster method within the **fviz_nbclust()** function. This method resulted in an optimal cluster amount of **2**.

Methods

K-Means cont.

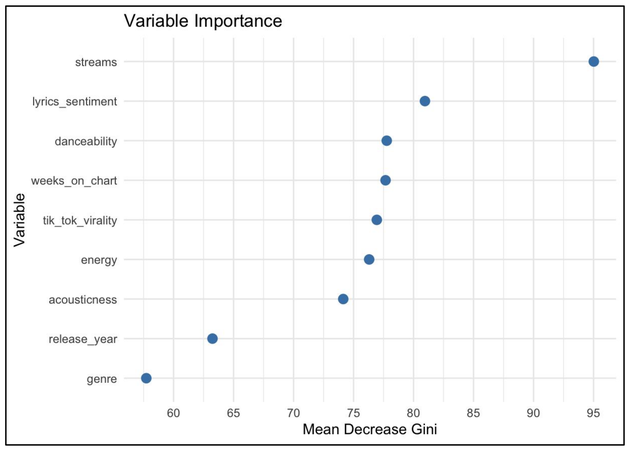
- The plot below displays the relationship between **tik_tok_virality** and **daily_streams.** There isn't a clear relationship between the variables, which was also observed with the other variable pairs. This may be due to a lack of distinct clustering, suggesting that the data points are relatively similar across groups. The k-means clustering also showed a high overlap between the optimal cluster models of 2 which indicates lack of quantifiable difference between our chosen variables.



Random Forest:

- Made a random forest model that predicts if the song will be in top 10 using variables such as danceability and acousticness, using an 80/20 split for training/testing.
- **Most Important Variables:** Streams, Lyrics Sentiment, Danceability, Weeks In Chart, and TikTok Virality
- **Least Important Variables:** Release Year, and Genre

Results



From the plot and what is shown above, we can see that streams is by far the most important variable since it has a **MeanDecreaseGini of 95%**. The next most important variable is lyric_sentiment since it has a **MeanDecreaseGini of ~82%**. Danceability and the number of weeks that the song is on the chart are the next most important variables with ~ **77% MeanDecreaseGini**. Then tiktok virality is next with ~**76% MeanDecreaseGini**.

Results

- From our assessment of two unsupervised and two supervised models we determined that the ideal modeling technique for our data set would be **random forest**, with a 91% accuracy rate.
- With random forest we found that the top variables that influences a song to be on the Billboards are streams, lyrics_sentiments, and danceability and weeks_on_chart. We also found that the model used achieved a 91% accuracy. From the graph above, you can see that the mean decreases for each variable.

Discussion

Limitations

- The data found in the Billboard Hot 100 dataset contains a mix of synthetic and real world music information. Due to the manufactured information present in a few songs it presents issues with accurately predicting the impact of our predictors on Billboard chart longevity and peak position.
- Some results seem to contradict each other from one model to another. An example of this is that the Random Forest model implies that genre pop is not an important feature, but in hierarchical clustering it seems that it is.

Future Directions:

- Explore unsupervised learning in depths. Though we had some findings using these techniques, perhaps an in-depth exploration can yield a more fruitful result
- Consider using a real world based data set with little to no synthetic information to create an accurate model for predicting the association between our predictors and a song's resulting Billboard Hot 100 peak position.

Implications

- Though streams is one important feature there are other factors that play a role in a song being on the Billboards 100. Songs that generally do well tend to be high energy songs which tend to be "pop." However, it is not the only factor in this model "streams" seem to the most common factor throughout all of the models.