The Anatomy of a Hit: Decoding Song Popularity with Machine Learning & SHAPing Insights

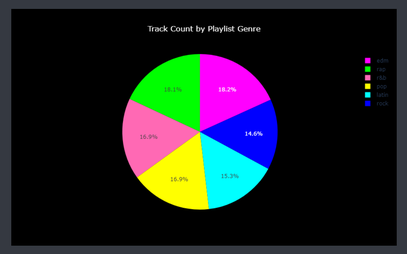
Anukeerthi Reddy Pothepalli
Institute of Insight, Robinson College of Business, Georgia State University

Introduction

- Music industry is massive and requires resources (time/effort/money)
- Oftentimes, number of listens on stream correlates with revenue
- Anatomy of a song, how does it correlate with Popularity?
- Does the features of the song define the genre of the song?

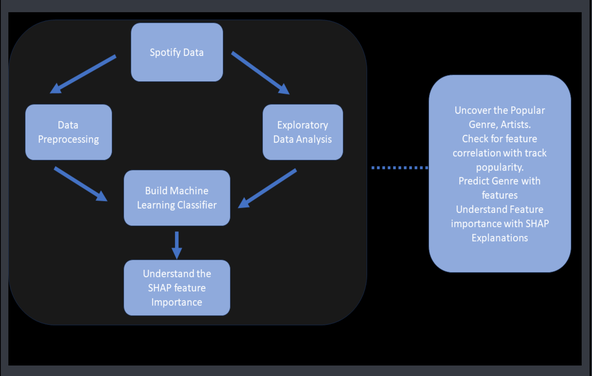
Dataset

- We have spotify data between 1957 and 2020.
- We have 32,833 Rows of Data across 23 features, 12 features concerning the sounf components of the song.
- First month of the year i.e. January has the maximum release
- The data is almost distributed across genres equally



Methodology

- Normalize the popularity with the features. find the important features corelating with Track Popularity



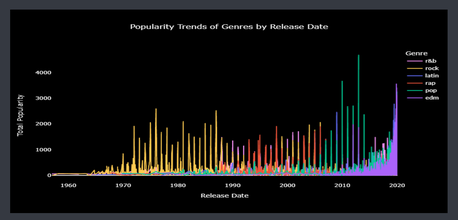
References

- DataSet
- SHAP Explainer
- Multiclass Classifier
- INFORMS Business Analytics Conference

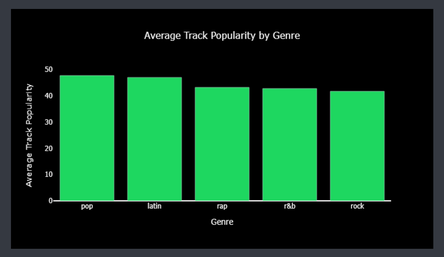
Pop, Latin, Rap Rule the Airwaves and Crack the Hit Code with Acoustics, Danceability, & Loudness of the song

Primary Analysis I

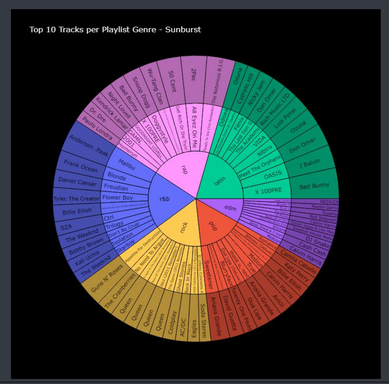
- Observing the trend in genre popularity reveals rock's dominance from the 1970s to the mid-1990s, followed by rap's surge in the mid-1990s to the mid-2000s.
- The mid-2000s marked a rise in popularity for pop, continuing through 2020. Notably, the recent years have witnessed a significant increase in the popularity of EDM genre artists.



- Observing the average track popularity by genre reveals that Pop, Latin, and Rock emerge as the most popular genres, capturing widespread appeal.
- Conversely, R&B and Rock stand out as the least popular genres, garnering comparatively lower levels of audience engagement

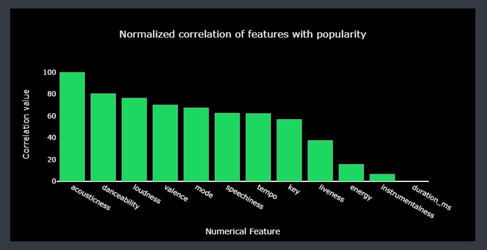


- Martin Garrix, Chainsmokers and David Guetta are the most popular Artists, while Trevor Daniel, Y2K, Don Toliver, Roddy Ricch have high mean popularity.
- If we examine the most popular tracks in each genre, songs by Queen, The Weeknd, Calvin Harris, Billie Eilish, David Guetta, and Ariana Grande stand out. Their tracks are loved across different genres and appear multiple times as the most popular.

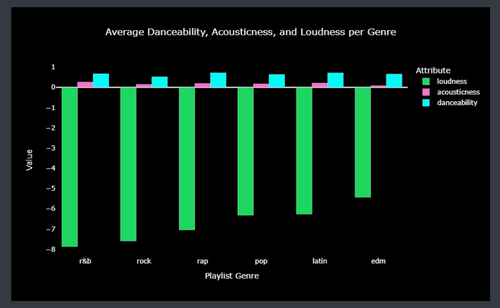


Primary Analysis 2

- An attempt to study the correlation between features and track popularity was conducted.
- While not finding a strong correlation initially, upon normalizing the features and comparing correlations, it became apparent that track popularity highly correlates with acousticness, danceability, and loudness of the song.
- Conversely, features such as duration, instrumentalness, and energy showed little to no correlation with track popularity.
- Additionally, a positive correlation was observed between energy and loudness of the song, while acousticness exhibited a negative correlation with both energy and loudness.



- Now that we've identified Acousticness, Danceability, and Loudness as crucial factors shaping track popularity, we delved deeper to analyze how these features vary across genres.
- Interestingly, only Loudness showed significant variation across genres, with popular genres like Pop, Latin, and Rap having a loudness value ranging between -6 and -7. In contrast, the average Acousticness and Danceability remained consistent across these popular genres.
- Furthermore, when examining energy levels across genres, EDM emerged with the highest average energy level, followed by Rock, Latin, Pop, Rap, and R&B.
- The song with the highest energy level is Rain Forest and Tropical Beach Sound in the latin genre with an energy level of 1.00
- latin genre is most suitable for dancing.
- Few rap songs are very less energetic, few of them have very high danceability and valence

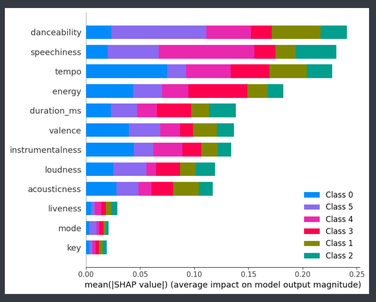


Random Forest

- Selected for its ability to classify Multiple Classes and its ability to manage a high number of features, ensuring accurate classification without overfitting in your diverse dataset.
- Achieved an Accuracy of 56% to classify into 6 different classes
- The model performed well to classify EDM, Rock and Rap Genres.
- The failed to correctly classify Pop and Latin and r&b. While classifying Pop, the model wrongly classified about 18.5% of Pop to edm and about 17.5% of pop to latin, does it mean that there is some similarity between these genres?

SHAP (SHapley Additive exPlanations) Values

- SHAP values are a way to explain the output of any machine learning model.
- Often the explanation between the Machine learning model is considered a black box, to an extent the Black box can be uncovered with SHAP values.
- The models considers Danceability, Speechiness and Tempo to be most be most influencing to classify the model while Key, mode and liveness are the least significant in the classification.
- Class Label Mapping: {0: 'pop', 1: 'rap', 2: 'rock', 3: 'latin', 4: 'r&b', 5: 'edm'}



Results

- Pop, Latin and rap are the most popular genres based on track popularity
- Latin music has experienced a exponential rise in popularity recently, with its share of popularity exploding by a staggering 280% in the past 3 years.
- Acousticness, danceability, and loudness exhibit a noteworthy correlation with track popularity, indicating their substantial impact on the overall appeal and reception of a song.
- The Random Forest model achieves a classification accuracy of approximately 57% across the six genres.