Song Popularity Prediction

By: Marla Galván

May 2021

Introduction

Since the institutionalization of music to be a profitable business, there has been the quest to find the formula to create the sounds that will please the public. It is no surprise that there are many factors, external or interna, in which popularity is affected . However, this research will try to explain how the internal composition of the song could predict a formula which explains how the song features affect its popularity.

Dataset Information

Firstly, the data attributes were taken from "Spotify Dataset 1922-2021, ~600k Tracks" [1], these data set includes the following features were used for the implementation:

The Dependant Variable

1. Popularity: "The value will be between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity. Note that the popularity value may lag actual popularity by a few days: the value is not updated in real time"[2].

The Independent Variables

1. 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."[2]

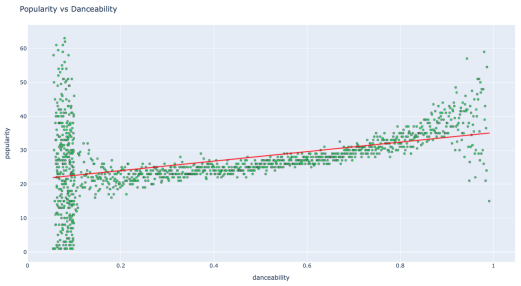


Fig 1.- Correlation between Popularity and danceability

Hypothesis: Danceability affects positively popularity

2. Energy: "measured 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. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy."[2]

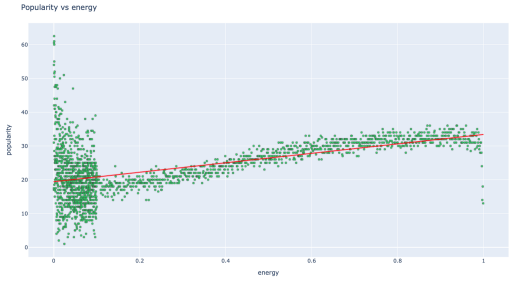


Fig 2.- Correlation between Popularity and Energy

Hypothesis: Energy affects positively popularity

3. instrumentalness: “Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.”[2]

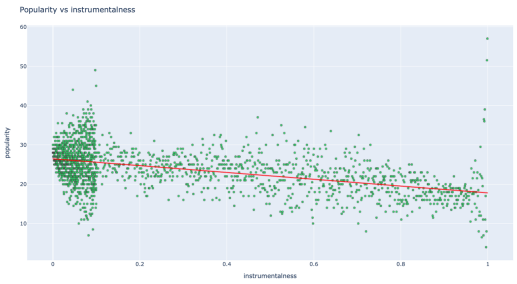


Fig 3.- Correlation between Popularity and instrumentalness

Hypothesis: instrumentalness affects negatively popularity

4. liveness: “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.”[2]

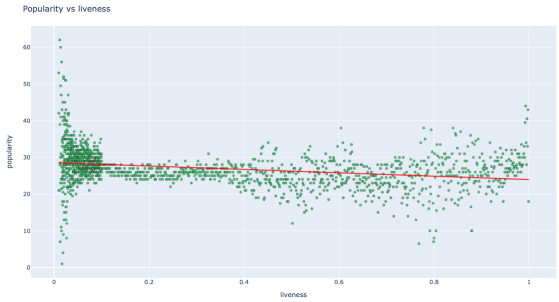


Fig 4.- Correlation between Popularity and liveness

Hypothesis: Liveness affects negatively popularity

5. 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. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.”[2]

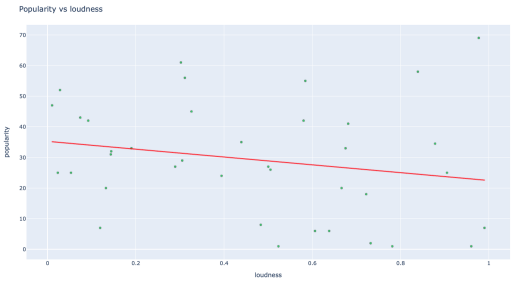


Fig 5.- Correlation between Popularity and loudness

Hypothesis: Energy affects negatively loudness

6. speechiness: “Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music.”[2]

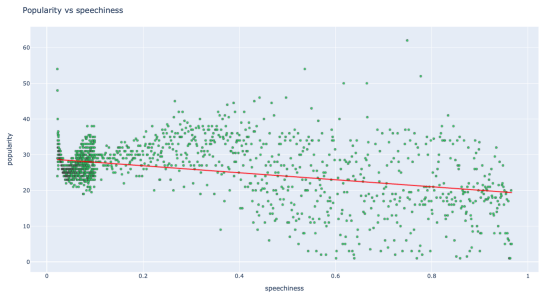


Fig 6.- Correlation between Popularity and speechiness

Hypothesis: speechiness affects negatively loudness

7. valence: "A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry)"[2]

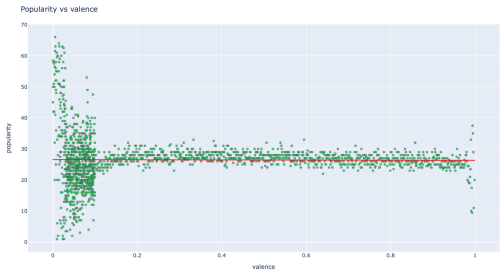


Fig 8.- Correlation between Popularity and valence

Hypothesis: valence affects positively loudness

Data Cleaning

The dataset provides additional variables, which were removed from this analysis since there was no correlation to popularity found.

Proposal Solution

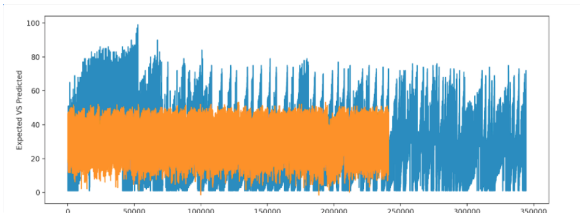
Since we want to know the relation that the dependant variables have over popularity, We are going to run an implementation on linear regression.The specified formula is the following:

popularity =

$\beta 0 + \beta 1 Danceability_i + \beta 2 Energy + \beta 3 Instrumentalness_i + \beta 4 Livenessi + \beta 5 loudnessi + \beta 6 speechinessi + \beta 7 valencei$

Analysis of the results

The percentage of training data was 30% since, the model fits better the expected popularity criteria, compare to other models where small bias is present.



The coefficients obtained were:

popularity = 12.27594118β0 + 1.24947653 β1 + 0.04968571β2 + -15.55597119β3 + -17.09652623β4 + -12.83364301β5 + -10.00844494β6 + -10.47530156β7

With the coefficients obtained we can check the proposed hypotheses:

- With the significant value r = 1.24947653 proves that danceability affects positively popularity
- With the significant value r = 0.04968571 proves that Energy affects positively popularity
- With the significant value r = -15.55597119 proves that instrumentalness affects negatively popularity
- With the significant value r = -17.09652623 proves that instrumentalness affects negatively popularity

- With the significant value r = -17.09652623 proves that Liveness affects negatively popularity
- With the significant value r = -12.83364301 proves that loudness affects negatively popularity
- With the significant value r = -10.00844494 proves that speechiness affects negatively popularity
- With the significant value r = --10.47530156 proves that valance affects negatively popularity

The obtained mean square error was 259.62, from which we can find that the model has significant errors on the test validation that can imply that model is underfitting.

The coefficient of the determination has a value of: -0.11 which implies that when the values in X decreases, Y increases.

Conclusion

In summary, we can see that the lower each of the dependent variables tends, the more popular the song becomes.

References

[1] https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

[2]https://www.kaggle.com/mindus/spotify-descriptive-and-exploratory-data-analysis/notebook