



Spotify hit songs combined dataset EDA and modelling

By: Abhiraaj Jadhav



Data Context


This is a dataset consisting of features for tracks fetched using Spotify's Web API. The tracks are labeled '1' or '0' ('Hit' or 'Flop') depending on some criterias of the author.

This dataset can be used to make a classification model that predicts whether a track would be a 'Hit' or not.



Data Attributes

- track: The Name of the track.
- artist: The Name of the Artist.
- uri: The resource identifier for the track.
- 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.
- 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. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
- key: The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C#/D?, 2 = D, and so on. If no key was detected, the value is -1.




- 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.

- mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.

- 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. Values below 0.33 most likely represent music and other non-speech-like tracks.

- 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. The distribution of values for this feature look like this:

- 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.

- 
- 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.
 - 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).
 - 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.
 - duration_ms: The duration of the track in milliseconds.
 - time_signature: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
 - chorus_hit: This is the author's best estimate of when the chorus would start for the track. It's the timestamp of the start of the third section of the track (in milliseconds). This feature was extracted from the data received by the API call for Audio Analysis of that particular track.



- sections: The number of sections the particular track has. This feature was extracted from the data recieved by the API call for Audio Analysis of that particular track.
- target: The target variable for the track. It can be either '0' or '1'. '1' implies that this song has featured in the weekly list (Issued by Billboards) of Hot-100 tracks in that decade at least once and is therefore a 'hit'. '0' Implies that the track is a 'flop'.



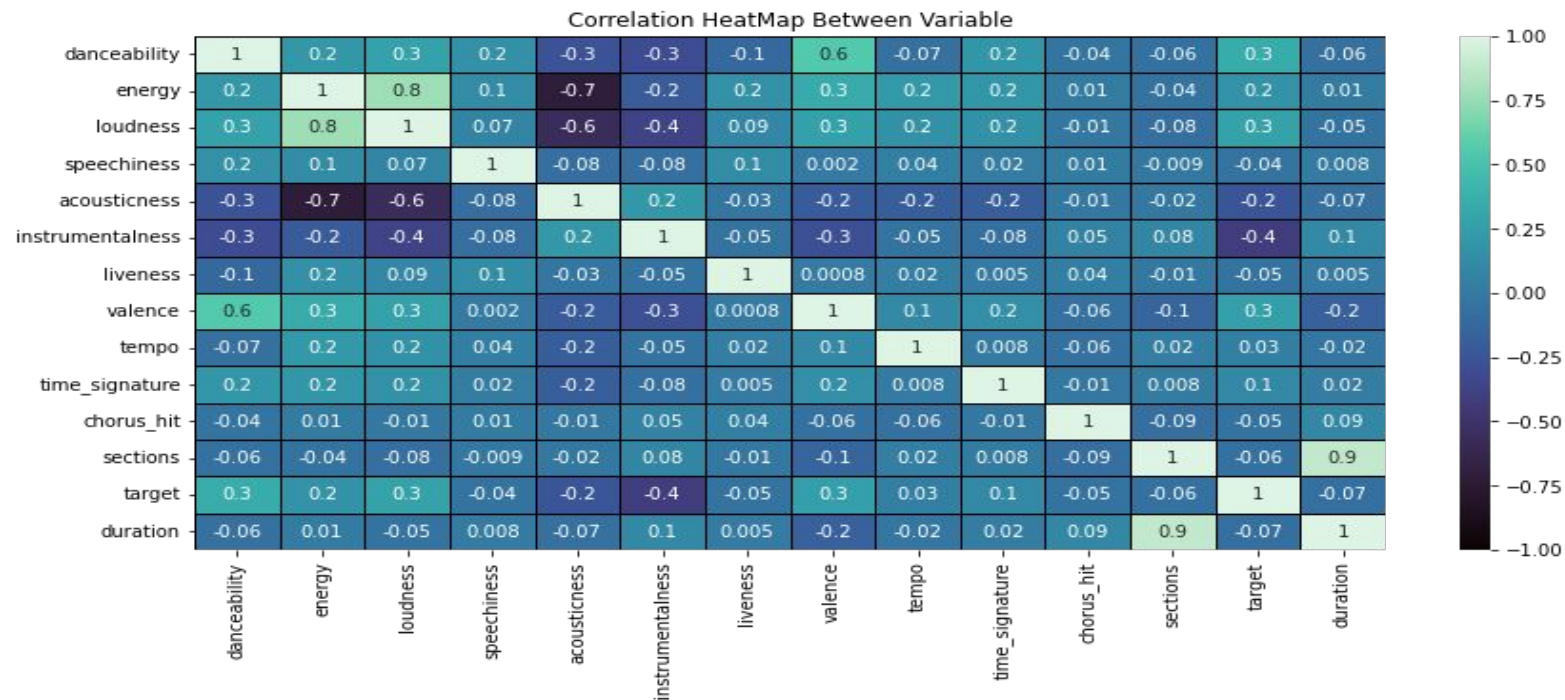
Exploratory Data Analysis

- `data.head()`: Reading the dataset to get to know more about the dataset.
- `info()`: Using info method to get to know about the data-types and the shapes of columns present in the dataset.
- `isnull().sum()`: To check for all the null values present in the dataset.
- `describe()`: To get the description of the data.
- Converted 'duration_ms' column that is time duration of songs in milliseconds to 'duration' column in seconds for better understanding of the column.



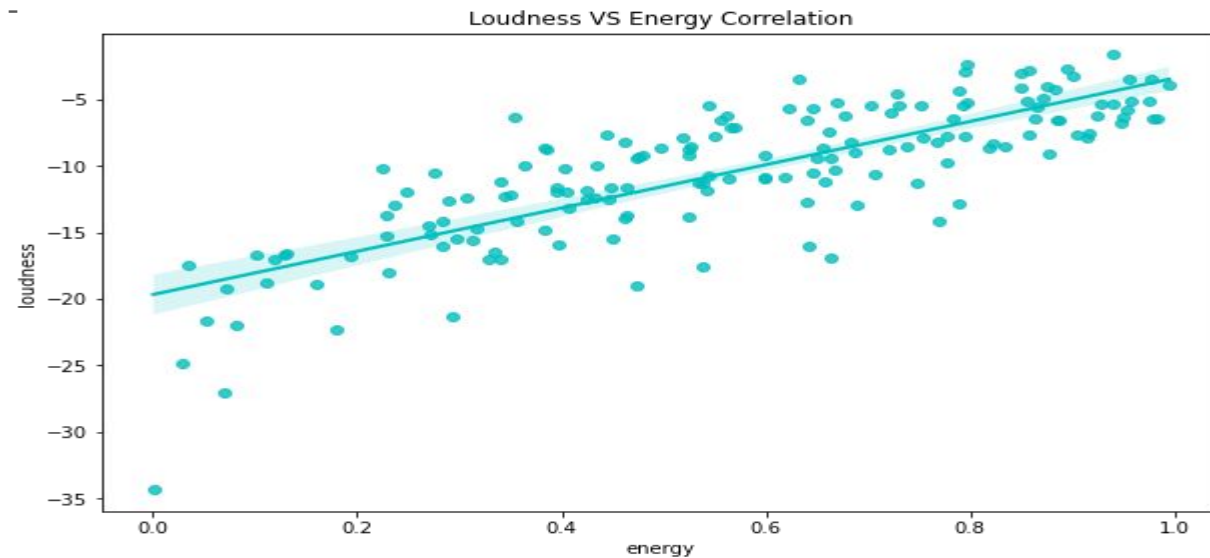
Data Visualisation

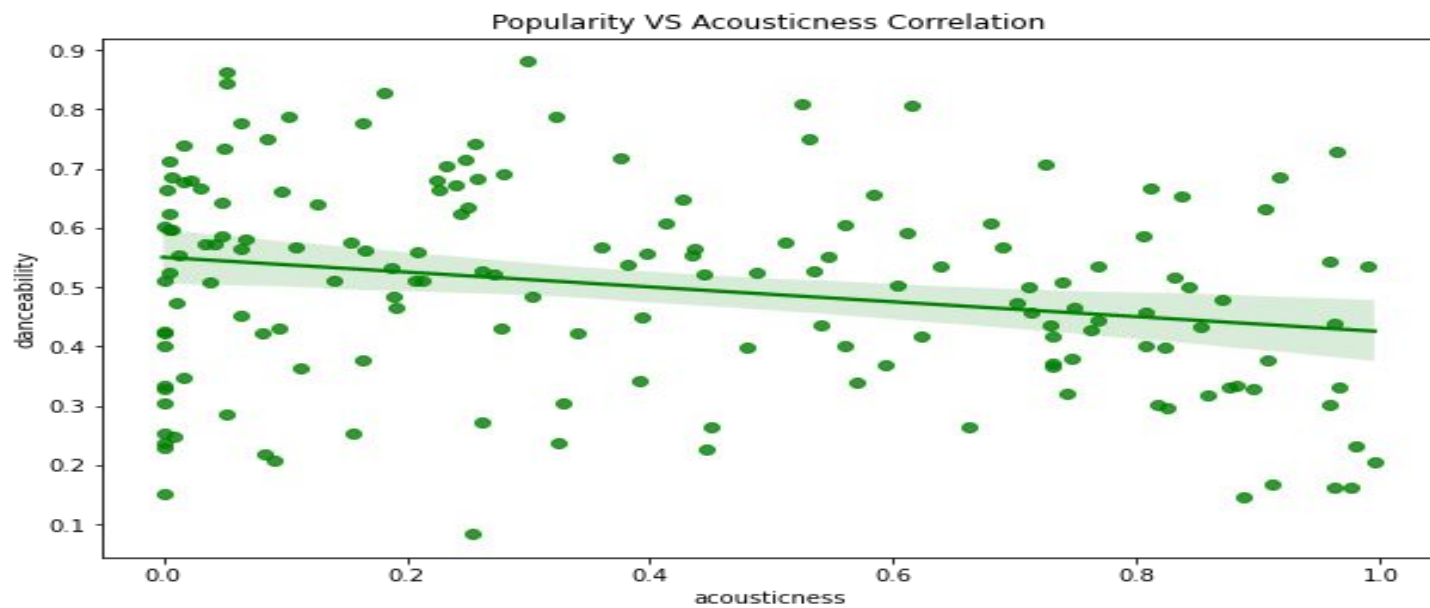
- Heatmap: This map describes the relationships between the different variables in form of colors and numbers:
- I have dropped the 'key' and 'mode' column due to the high cardinality.
-





- Regression Plot:







Modeling

Methods Used:

- Train Test Split Method: This method is used to prevent the model from overfitting and to accurately evaluate the model

Models Used:

- Logistic Regression
- K-Nearest Neighbors
- Decision Tree
- Random Forest
- Gradient Boosting



Results

Logistic Regression: 72.45%

K-Nearest Neighbors: 72.90%

Decision Tree: 70.28%

Random Forest: 78.17%

Gradient Boosting: 77.77%