ShopifySongsClassification

May 16, 2024

1 Caltech AI & Machine Learning Bootcamp

Curse: CB-AIML-Core - Machine Learning

CB AIML JAN 2024 COHORT 2

Student: Eric Michel

2 Course-End Project 2: Shopify Songs Classification

2.1 Problem Scenario:

The customer always looks forward to specialized treatment, whether shopping on an e-commerce website or watching Netflix. The customer desires content that aligns with their preferences. To maintain customer engagement, companies must consistently provide the most relevant information.

Starting with Spotify, a Swedish audio streaming and media service provider, boasts over 456 million active monthly users, including more than 195 million paid subscribers as of September 2022. The company aims to create cohorts of different songs to enhance song recommendations. These cohorts will be based on various relevant features, ensuring that each group contains similar types of songs.

2.2 Problem Objective:

As a data scientist, you should perform exploratory data analysis and cluster analysis to create cohorts of songs. The goal is to understand better the various factors that create a cohort of songs.

2.3 Data Description

The dataset comprises information from Spotify's API regarding all albums by the Rolling Stones available on Spotify. It's crucial to highlight that each song possesses a unique ID.

	17 Distinct values	17 Distinct values
0	name	the name of the song
1	album	the name of the album
2	release_date	the day month and year the album was released
3	track number	the order the song appears on the album
4	id	the Spotify id for the song
5	uri	the Spotify uri for the song
6	acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the tr
7	danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements incl
8	energy	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically,
9	instrumentalness	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this con
10	liveness	Detects the presence of an audience in the recording. Higher liveness values represent an increased probab
11	loudness	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and ar
12	speechiness	detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk sh
13	tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the spe
14	valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valenc
15	popularity	the popularity of the song from 0 to 100
16	duration_ms	The duration of the track in milliseconds.

2.4 Summary: What I did in this project

1. Initial data inspection and data cleaning
2. Examine the data initially to identify duplicates, missing values, irrelevant entries, or outliers. Check for any instances of erroneous entries and rectify them as needed
3. Refine the data for further processing based on your findings
4. Perform exploratory data analysis and feature engineering
5. Utilize suitable visualizations to identify the two albums that should be recommended to anyone based on the number of popular songs in each album
6. Conduct exploratory data analysis to delve into various features of songs, aiming to identify patterns
7. Examine the relationship between a song's popularity and various factors, exploring how this correlation has evolved
8. Provide insights on the significance of dimensionality reduction techniques. Share your ideas and elucidate your observations
9. Perform cluster analysis
10. Identify the right number of clusters
11. Use appropriate clustering algorithms
12. Define each cluster based on the features

3 **Initial data inspection and data cleaning**

4 **Examine the data initially to identify duplicates, missing values, irrelevant entries, or outliers. Check for any instances of erroneous entries and rectify them as needed**

5 **Refine the data for further processing based on your findings**

6 **Perform exploratory data analysis and feature engineering**

7 **Utilize suitable visualizations to identify the two albums that should be recommended to anyone based on the number of popular songs in each album**

8 **Conduct exploratory data analysis to delve into various features of songs, aiming to identify patterns**

9 **Examine the relationship between a song's popularity and various factors, exploring how this correlation has evolved**

10 **Provide insights on the significance of dimensionality reduction techniques. Share your ideas and elucidate your observations**

11 **Perform cluster analysis**

12 **Identify the right number of clusters**

13 **Use appropriate clustering algorithms**

14 **Define each cluster based on the features**