SpotifyAnalyzer

December 6, 2024

1 Introduction

For my final project in CSC 335 (Machine Learning) at Adelphi University, I am building a machine learning model that will be able to predict a song's popularity based on it's attributes. These attributes are defined by Spotify in their documentation here: https://developer.spotify.com/documentation/web-api/reference/get-audio-features

The datasets I am using are from Kaggle: https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db and https://www.kaggle.com/datasets/dhruvildave/billboard-the-hot-100-songs?resource=download

2 Data Exploration

[1]:
```
# Imports
import numpy as np
import pandas as pd
```

[2]:
```
# Read in spotify data
spotify_data = pd.read_csv('SpotifyFeatures.csv', delimiter=",")
# Look at data
spotify_data
```

[2]:
```
        genre              artist_name                        track_name  \
0       Movie           Henri Salvador       C'est beau de faire un Show
1       Movie        Martin & les fées  Perdu d'avance (par Gad Elmaleh)
2       Movie          Joseph Williams    Don't Let Me Be Lonely Tonight
3       Movie           Henri Salvador    Dis-moi Monsieur Gordon Cooper
4       Movie             Fabien Nataf                         Ouverture
…       …                        …                                 …
232720   Soul                    Slave                    Son Of Slide
232721   Soul  Jr Thomas & The Volcanos                   Burning Fire
232722   Soul             Muddy Waters     (I'm Your) Hoochie Coochie Man
232723   Soul                   R.LUM.R                     With My Words
232724   Soul           Mint Condition     You Don't Have To Hurt No More

                      track_id  popularity  acousticness  danceability  \
0       0BRjO6ga9RKCKjfDqeFgWV           0       0.61100         0.389
1       0BjC1NfoEOOusryehmNudP           1       0.24600         0.590
```

```
2       0CoSDzoNIKCRs124s9uTVy          3      0.95200          0.663
3       0Gc6TVm52BwZD07Ki6tIvf          0      0.70300          0.240
4       0IuslXpMROHdEPvSl1fTQK          4      0.95000          0.331
…                        …         …          …           …
232720  2XGLdVl7lGeq8ksM6Al7jT         39      0.00384          0.687
232721  1qWZdkBl4UVPj9lK6HuuFM         38      0.03290          0.785
232722  2ziWXUmQLrXTiYjCg2fZ2t         47      0.90100          0.517
232723  6EFsue2YbIG4Qkq8Zr9Rir         44      0.26200          0.745
232724  34XO9RwPMKjbvRry54QzWn         35      0.09730          0.758

        duration_ms  energy  instrumentalness key  liveness  loudness   mode  \
0             99373   0.910          0.000000  C#    0.3460    -1.828  Major
1            137373   0.737          0.000000  F#    0.1510    -5.559  Minor
2            170267   0.131          0.000000   C    0.1030   -13.879  Minor
3            152427   0.326          0.000000  C#    0.0985   -12.178  Major
4             82625   0.225          0.123000   F    0.2020   -21.150  Major
…                 …      …                 …  ..        …         …     …
232720       326240   0.714          0.544000   D    0.0845   -10.626  Major
232721       282447   0.683          0.000880   E    0.2370    -6.944  Minor
232722       166960   0.419          0.000000   D    0.0945    -8.282  Major
232723       222442   0.704          0.000000   A    0.3330    -7.137  Major
232724       323027   0.470          0.000049  G#    0.0836    -6.708  Minor

        speechiness   tempo time_signature  valence
0            0.0525  166.969            4/4    0.814
1            0.0868  174.003            4/4    0.816
2            0.0362   99.488            5/4    0.368
3            0.0395  171.758            4/4    0.227
4            0.0456  140.576            4/4    0.390
…                 …       …          …        …
232720       0.0316  115.542            4/4    0.962
232721       0.0337  113.830            4/4    0.969
232722       0.1480   84.135            4/4    0.813
232723       0.1460  100.031            4/4    0.489
232724       0.0287  113.897            4/4    0.479

[232725 rows x 18 columns]
```

2.1 Attribute Explanation

The attributes in the dataset are, with a brief explanation sourced by the Spotify API…:

- 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."
- 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."

- Duration: “The duration of the track in milliseconds.”
- 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.”
- Instrumentalness: “Predicts whether a track contains no vocals. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.”
- Key: “The key the track is in. Integers map to pitches using standard Pitch Class notation.”
- 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.”
- Loudness: “The overall loudness of a track in decibels (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.”
- Speediness: “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.”
- Tempo: “The overall estimated tempo of a track in beats per minute (BPM).”
- Time Signature: “An estimated time signature. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).”
- 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).”
- Popularity: How popular a song is, from 1 - 100. The higher the number, the more popular a song is.

[3]:
```
# Read in chart data
chart_data = pd.read_csv("charts.csv", delimiter=',')
chart_data
```

[3]:
```
              date  rank                                 song  \
0       2021-11-06     1                          Easy On Me
1       2021-11-06     2                                Stay
2       2021-11-06     3                       Industry Baby
3       2021-11-06     4                          Fancy Like
4       2021-11-06     5                          Bad Habits
…                …     …                                   …
330082  1958-08-04    96                       Over And Over
330083  1958-08-04    97                     I Believe In You
330084  1958-08-04    98                      Little Serenade
330085  1958-08-04    99  I'll Get By (As Long As I Have You)
```

```
330086  1958-08-04   100                                 Judy

                              artist  last-week  peak-rank  weeks-on-board
0                              Adele        1.0          1               3
1       The Kid LAROI & Justin Bieber        2.0          1              16
2             Lil Nas X & Jack Harlow        3.0          1              14
3                        Walker Hayes        4.0          3              19
4                          Ed Sheeran        5.0          2              18
…                                  …          …          …               …
330082                Thurston Harris        NaN         96               1
330083               Robert & Johnny        NaN         97               1
330084             The Ames Brothers        NaN         98               1
330085                Billy Williams        NaN         99               1
330086                Frankie Vaughan        NaN        100               1

[330087 rows x 7 columns]
```

2.2 Data Set Creation

Let's combine these two datasets to get a cohesive grouping of information. We want to combine the data so that the track name/artist name are used as 'keys' in the combination. This will allow us to combine the song features with the chart data for analysis.

[5]:
```python
# Normalize column names for consistent matching
spotify_data['artist_name'] = spotify_data['artist_name'].str.lower()
spotify_data['track_name'] = spotify_data['track_name'].str.lower()

chart_data['artist'] = chart_data['artist'].str.lower()
chart_data['song'] = chart_data['song'].str.lower()
```

[6]:
```python
# Merge datasets on artist and track name
merged_data = pd.merge(
    spotify_data,
    chart_data,
    left_on=['artist_name', 'track_name'],
    right_on=['artist','song'],
    how='inner'
)
```

[7]:
```python
# Select relevant columns for the new dataset
song_data = merged_data[
    [
        'artist_name', 'track_name', 'rank', 'last-week', 'peak-rank',
        'weeks-on-board', 'date', 'popularity', 'acousticness',
        'danceability', 'duration_ms', 'energy', 'instrumentalness',
        'key', 'liveness', 'loudness', 'mode', 'speechiness',
        'tempo', 'time_signature', 'valence'
```

```
    ]
]

song_data
```

[7]:

```
           artist_name                    track_name  rank  last-week  \
0                usher            you make me wanna…    49       50.0
1                usher            you make me wanna…    50       48.0
2                usher            you make me wanna…    48       42.0
3                usher            you make me wanna…    42       39.0
4                usher            you make me wanna…    39       41.0
…                  …                           …    …        …
227145  mint condition  you don't have to hurt no more    32       34.0
227146  mint condition  you don't have to hurt no more    34       37.0
227147  mint condition  you don't have to hurt no more    37       42.0
227148  mint condition  you don't have to hurt no more    42       52.0
227149  mint condition  you don't have to hurt no more    52        NaN

        peak-rank  weeks-on-board        date  popularity  acousticness  \
0               2              47  1998-07-11          69        0.0359
1               2              46  1998-07-04          69        0.0359
2               2              45  1998-06-27          69        0.0359
3               2              44  1998-06-20          69        0.0359
4               2              43  1998-06-13          69        0.0359
…             …               …         …          …           …
227145         32               5  1997-04-26          35        0.0973
227146         34               4  1997-04-19          35        0.0973
227147         37               3  1997-04-12          35        0.0973
227148         42               2  1997-04-05          35        0.0973
227149         52               1  1997-03-29          35        0.0973

        danceability  …  energy  instrumentalness key liveness  loudness  \
0              0.761  …   0.639          0.000000   F   0.0945    -7.577
1              0.761  …   0.639          0.000000   F   0.0945    -7.577
2              0.761  …   0.639          0.000000   F   0.0945    -7.577
3              0.761  …   0.639          0.000000   F   0.0945    -7.577
4              0.761  …   0.639          0.000000   F   0.0945    -7.577
…                …  …     …               …  …      …       …
227145         0.758  …   0.470          0.000049  G#   0.0836    -6.708
227146         0.758  …   0.470          0.000049  G#   0.0836    -6.708
227147         0.758  …   0.470          0.000049  G#   0.0836    -6.708
227148         0.758  …   0.470          0.000049  G#   0.0836    -6.708
227149         0.758  …   0.470          0.000049  G#   0.0836    -6.708

         mode speechiness    tempo  time_signature valence
0       Minor      0.0539  164.088             4/4   0.922
1       Minor      0.0539  164.088             4/4   0.922
```

```
2        Minor      0.0539  164.088            4/4   0.922
3        Minor      0.0539  164.088            4/4   0.922
4        Minor      0.0539  164.088            4/4   0.922
…        …        …      …             …     …
227145   Minor      0.0287  113.897            4/4   0.479
227146   Minor      0.0287  113.897            4/4   0.479
227147   Minor      0.0287  113.897            4/4   0.479
227148   Minor      0.0287  113.897            4/4   0.479
227149   Minor      0.0287  113.897            4/4   0.479

[227150 rows x 21 columns]
```

Great! Now we have a complete dataset. With this, we can anaylyze how long a song will remain at the number one spot based on what type of song attributes it has. In other words, the question we want to answer is… *"What kind of musical attributes contribute to a song's longevity on the charts?"*