Exploring the predictive power of musical features via Spotify

Group 12: Ethan Hodys, Vincent Chiang

*Abstract*: Founded in 2006, Spotify’s primary business is providing an audio streaming platform, the "Spotify" platform, that provides DRM-restricted music, videos and podcasts from record labels and media companies (Wikipedia 2008). This paper is an exploratory survey of whether musical features and characteristics can be used predict how a song will be both received and classified by Spotify users. Using different methods for Classification Trees and Clustering in R, we make a case for the feasibility of prediction using a song’s meta data stored by Spotify.

**Contents**

[1 Introduction 2](#_Toc38799143)

[1.1 Histogram Plot 2](#_Toc38799144)

[1.2 Correlation Matrix 3](#_Toc38799145)

[2 Data Dictionary 4](#_Toc38799146)

[3 Methodology and results 7](#_Toc38799147)

[3.1 Classification Trees 7](#_Toc38799148)

[3.2 Clustering 9](#_Toc38799149)

[5. Conclusion 10](#_Toc38799150)

[6. Bibliography 11](#_Toc38799151)

# Introduction

Music can have a powerful effect upon people emotionally and has played an important role in since the first human society. Thus, being able to classify songs based on their musical features could prove to be a useful tool across many sectors. While this paper will in no way be an exhaustive survey of the topic it will provide an initial exploratory analysis to help guide future research

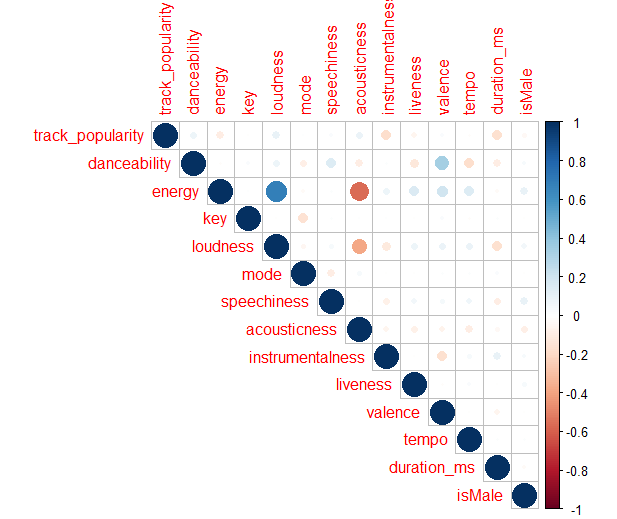
We have chosen to use the “Spotify Song Attributes” data set on Kaggle uploaded by George McIntire under a CC-BY License. Below is a list of the variables in the dataset with their distributions except for “track id” which is simply a unique identifier for a song (see appendix for details on the dataset).

## Histogram Plot

A close up of a map

Description automatically generated

## Correlation Matrix



We also added information on the biological sex of the performers using lists of popular names split between boy in girl. We import the data into MS Sql Server© to add the correct value for the new variable “isMale” which is a simple bit variable of 0/1 for female/male, respectively. From our survey of the data and personal interests we identified 4 response variables of interest to use for our analysis:

1. Danceability
2. Genre
3. Popularity
4. Gender

# 

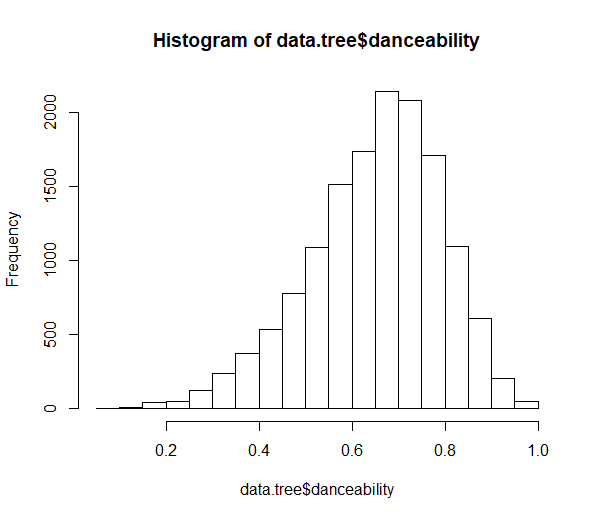
# Data Dictionary

| **variable** | **class** | **description** |
| --- | --- | --- |
| track\_id | character | Song unique ID |
| track\_name | character | Song Name |
| track\_artist | character | Song Artist |
| track\_popularity | double | Song Popularity (0-100) where higher is better |
| track\_album\_id | character | Album unique ID |
| track\_album\_name | character | Song album name |
| track\_album\_release\_date | character | Date when album released |
| playlist\_name | character | Name of playlist |
| playlist\_id | character | Playlist ID |
| playlist\_genre | character | Playlist genre |
| playlist\_subgenre | character | Playlist subgenre |
| danceability | double | 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 | double | 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 | double | 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 | double | 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 | double | 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 | double | 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 | double | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic. |
| instrumentalness | double | 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 | double | 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 | double | 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 | double | 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 | double | Duration of song in milliseconds |

# Methodology and results

This dataset is a more manageable collection of datapoints than the “FMA” data set archived on the UCI Machine Learning Repository (<http://archive.ics.uci.edu/>). The “FMA” data set is tagged with the ‘Classification’ and “Clustering’ meta tags. Since the “Spotify Song Attributes” is a much smaller subset of the information in “FMA” we split the analysis into the two different algorithms. For the Classification Decision Trees we selected the ‘*rpart*()’ and ‘*ctree*() functions in R. For the Clustering Analysis we chose *K-means*, *Hierarchical*, and *DBSCAN* for the clustering algorithms.

## Classification Trees



tree.pred H M

H 192 187

M 517 1975

.

[1] 0.7547893

n=9167 (2316 observations deleted due to missingness)

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 9167 2380 M (0.2596269 0.7403731)

2) valence>=0.6125 3250 1258 M (0.3870769 0.6129231)

4) tempo< 128.0345 2325 1048 M (0.4507527 0.5492473)

8) energy< 0.7885 1446 689 H (0.5235131 0.4764869)

16) tempo>=89.9755 1207 520 H (0.5691798 0.4308202) \*

17) tempo< 89.9755 239 70 M (0.2928870 0.7071130) \*

9) energy>=0.7885 879 291 M (0.3310580 0.6689420) \*

5) tempo>=128.0345 925 210 M (0.2270270 0.7729730) \*

3) valence< 0.6125 5917 1122 M (0.1896231 0.8103769) \*

tree.pred H M

H 322 411

M 869 1269

> tree.pruned

n=9155 (2328 observations deleted due to missingness)

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 9155 3658 M (0.3995631 0.6004369)

2) valence>=0.5965 3474 1519 H (0.5627519 0.4372481)

4) tempo< 128.0345 2489 939 H (0.6227401 0.3772599)

8) tempo>=89.9535 2151 738 H (0.6569038 0.3430962) \*

9) tempo< 89.9535 338 137 M (0.4053254 0.5946746) \*

5) tempo>=128.0345 985 405 M (0.4111675 0.5888325) \*

3) valence< 0.5965 5681 1703 M (0.2997712 0.7002288) \*

> sum(diag(confusion.matrix)) / sum(confusion.matrix) # the % accuracy on the test set.

[1] 0.5541623

tree.pred H M

H 0 0

M 379 2492

> tree.pruned

n=9167 (2316 observations deleted due to missingness)

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 9167 1287 M (0.1403949 0.8596051) \*

> sum(diag(confusion.matrix)) / sum(confusion.matrix) # the % accuracy on the test set.

[1] 0.8679902

## Clustering

# Conclusion

Start the section

# Bibliography

[Various. 2008.](Various. 2008. Wikipedia. Dec. Accessed 04 23, 2020. https://en.wikipedia.org/wiki/Spotify.) *[Wikipedia.](Various. 2008. Wikipedia. Dec. Accessed 04 23, 2020. https://en.wikipedia.org/wiki/Spotify.)* [Dec. Accessed 04 23, 2020. https://en.wikipedia.org/wiki/Spotify.](Various. 2008. Wikipedia. Dec. Accessed 04 23, 2020. https://en.wikipedia.org/wiki/Spotify.)

<https://rpubs.com/coleeagland/decisiontreesislr831>

<file:///C:/Users/Ethan%20Hodys/Documents/MastersWork/DataMining/Data-Mining-R-master/5.%20Tree%20models/5_Tree.html#regression-tree-boston-housing-data>