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](#_Toc38655996)

[2 Methodology and results 3](#_Toc38655997)

[2.1 Classification Trees 3](#_Toc38655998)

[2.2 Clustering 3](#_Toc38655999)

[3 Conclusion 4](#_Toc38656000)

[4 Bibliography 5](#_Toc38656001)

# 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).

A close up of a map

Description automatically generated

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:

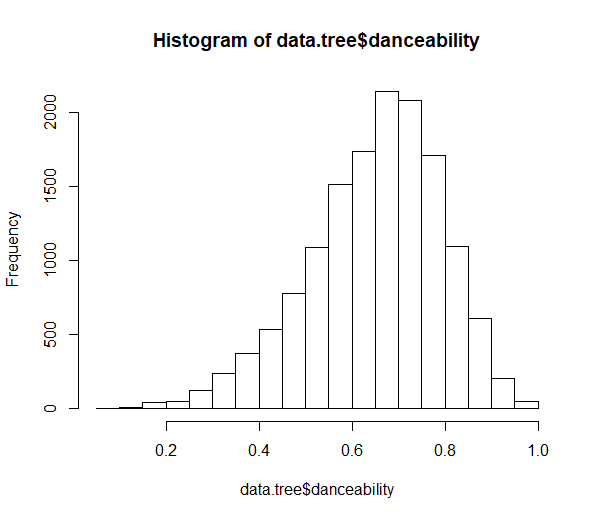
1. Danceability
2. Genre
3. Popularity
4. Gender

# 

# 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>