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](#_Toc38642346)

[2 Methodology and results 2](#_Toc38642347)

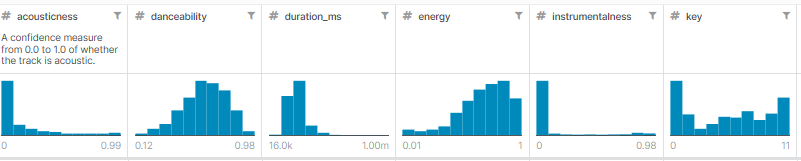
[3 Conclusion 3](#_Toc38642348)

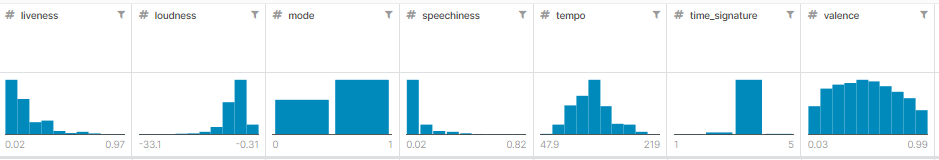
[4 Bibliography 4](#_Toc38642349)

# 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).





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 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

I used classification trees, so I want to show: ***Gini index*** or the ***Entropy***

Lost function for model training

Cost function for prediction

* Danceability
* Popularity
* Gender
* Genre

A screenshot of a video game

Description automatically generated

# 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>