**Song Popularity prediction using machine learning algorithms with insights from unsupervised learning**

**Authors:**

Abhinav Lakhani(0769062)

**Supervisor:**

Savita Seharawat

# 

# Abstract

In a digital age where independent artists are trying to break through the barriers installed by putative gatekeepers of the music industry, the golden key to becoming a successful artist is data. Without data, successful campaigns cannot be recreated and the unsuccessful ones cannot be learned from. Predicting song popularity is particularly important in keeping businesses competitive within a growing music industry. But what exactly makes a song popular? Starting with the Spotify song Dataset, a collection of audio features and metadata for approximately 40,000 songs, we evaluated different classification algorithms on their ability to predict popularity and determined the types of features that hold the most predictive power and also use unsupervised algorithms to get the insights from the dataset and use that insights to improve our classification algorithms.

The Spotify Popularity Index is a 0-to-100 score that ranks how popular an artist is relative to other artists on Spotify. As your numbers grow, you’ll get placed in more editorial playlists and increase your reach on algorithmic playlists and recommendations. The Index can be used to monitor and influence the progress of new releases. Each track has its own SPI calculated influencing the artist’s overall index. Yet, while the Popularity Index is majorly determined by recent stream count, other factors like save rate, the number of playlists, skip rate, and share rate can indirectly bump up or push down a song’s popularity index. This project uses similar dataset where the songs are labeled popular(1) or unpopular(0).

The chosen themes for this project will be Exploratory data analysis using unsupervised learning and using the knowledge gained from the unsupervised learning to build Classification/Clustering and Predictive Analysis.

We aim to answer:

* What are the different factors affecting song popularity.
* Are there any clusters/groups of songs identifying audience demographic like people who like rock music or jazz music, etc. so as to align with the artists’ theme.
* Build a classification algorithm to help with classifying whether a song would be popular.
* Make suggestions to the artists about their song: whether it would be popular or not and the attributes of their song that needs improvement.

Keywords:

Non Negative Matrix Factorization(NMF), PCA, Manifold Learning with t-SNE, Logistic Regression, SVM, Decision trees, Clustering, Random Forests, Cross Validation

# Related Work

The problem of predicting popularity is one that has been heavily researched. Salganik, Dodds, and Watts conducted an experimental study on popularity that focused heavily on the social influence of popularity. They found that the quality of a song only partially influences whether or not a song becomes popular, and that social influence plays an extremely large role [1]. Therefore, this project aims to use both acoustic features and metadata features to create a more accurate prediction model.

The work by Koenignstein, Shavitt, and Zilberman, which predicts billboard success based on peer-to-peer networks, potentially captures this social influence on song popularity. This group was extremely thorough with their work and used multiple regression and classification algorithms for their predictions [2].

Pachet and Roy investigated the problem of making predictions of song popularity and made the blunt claim that the popularity of a song cannot be learnt by using state-of-the-art machine learning [3]. In order to test the effectiveness of current machine learning algorithms, they tested the improvement of their classification models to a generic random classifier. Similarly to this work, Pachet and Roy consider both acoustic features and metadata; however, the study deals with an extremely large number of features (over 600) but does not mention any type of feature selection algorithm. As a result it is extremely likely that their model was subjected to overfitting. Pacet and Roy also considered features commonly used for music analysis which potentially could have affected the success of their results.

However, Ni et al have responded to the above definitive claim with more optimistic results on music popularity prediction, using a Shifting Perceptron algorithm to classify the top 5 hits from the top 30-40 hits (a slightly different problem from the aforementioned study) [4]. However, this study also uses more novel audio features which is a likely factor in their improved results

# Dataset and Features

This work uses dataset from recent Kaggle competition[5] which is an exhaustive collection of audio features and metadata for about 40,000 songs. The audio features include attributes about the music track itself, such as song duration, key, audio mode, time signature, The metadata uses more abstract features, such as danceability, energy, instrumentalness, liveness, etc. This Dataset is a subset which was derived from Spotify web api which provides access to user related data, like playlists and music that the user saves in the their Music library containing data regarding millions of songs and continuously being updated with new songs and changing user preference[6].

The dataset used for our analysis has 14 attributes within it, and are described below.

|  |  |  |
| --- | --- | --- |
| Attributes | Data Type | Description |
| song\_duration\_ms | float64 | The duration of the track in milliseconds. |
| acousticness | float64 | A confidence measure from 0.0 to 1.0 of whether the track is acoustic. |
| danceability | float64 | Describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. |
| energy | float64 | 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. |
| instrumentalness | float64 | 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". |
| key | category | The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = Câ™¯/Dâ™­, 2 = D, and so on.  levels: 12 |
| liveness | float64 | Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. |
| loudness | float64 | 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. |
| audio\_mode | category | 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.  levels: 2 |
| speechiness | float64 | This 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. |
| tempo | float64 | 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. |
| time\_signature | category | An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).  levels: 4 |
| audio\_valence | float64 | Describes 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). |

Table 1: Data Dictionary

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attributes** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| song\_duration\_ms | 35899 | 193165.8 | 45822.13 | 25658 | 166254.5 | 186660 | 215116 | 491671 |
| acousticness | 36008 | 0.276404 | 0.297928 | -0.01355 | 0.039618 | 0.140532 | 0.482499 | 1.065284 |
| danceability | 35974 | 0.570951 | 0.19001 | 0.043961 | 0.42476 | 0.608234 | 0.718464 | 0.957131 |
| energy | 36025 | 0.683932 | 0.212662 | -0.00168 | 0.539276 | 0.704453 | 0.870503 | 1.039741 |
| instrumentalness | 36015 | 0.036527 | 0.150024 | -0.0044 | 0.000941 | 0.001974 | 0.003225 | 1.075415 |
| liveness | 35914 | 0.198514 | 0.15167 | 0.027843 | 0.111796 | 0.135945 | 0.212842 | 1.065298 |
| loudness | 36043 | -7.4076 | 3.877198 | -32.1179 | -9.57814 | -6.34541 | -4.62071 | -0.87735 |
| speechiness | 40000 | 0.094107 | 0.083591 | 0.015065 | 0.0385 | 0.055881 | 0.118842 | 0.560748 |
| tempo | 40000 | 116.5628 | 26.16791 | 62.05578 | 96.99531 | 113.796 | 128.5174 | 219.1636 |
| audio\_valence | 40000 | 0.580645 | 0.237351 | 0.013398 | 0.398669 | 0.598827 | 0.759635 | 1.022558 |

Table 2: Numerical Statistics

|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes** | **key** | **audio\_mode** | **time\_signature** |
| count | 35935 | 40000 | 40000 |
| unique | 12 | 2 | 4 |
| top | 0 | 0 | 3 |
| freq | 5175 | 27154 | 23358 |

Table 3: Categorical Stats

# Methodology

Modeling various Supervised learning algorithm using train & validation set( cross validation)

Prediction

References

[1] Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. ”Experimental study of inequality and unpredictability in an artificial cultural market.” science 311.5762 (2006): 854-856.

[2] Koenigstein, Noam, Yuval Shavitt, and Noa Zilberman. ”Predicting billboard success using data-mining in p2p networks.” Multimedia, 2009. ISM’09. 11th IEEE International Symposium on. IEEE, 2009.

[3] Pachet, Franois, and Pierre Roy. ”Hit Song Science Is Not Yet a Science.” ISMIR. 2008.

[4] Ni, Yizhao, et al. ”Hit song science once again a science.” 4th International Workshop on Machine Learning and Music, Spain. 2011.

[5] “Song Popularity Prediction” *kaggle Dataset*. [https://www.kaggle.com/c/song-popularity-prediction](https://www.kaggle.com/c/song-popularity-prediction/code?competitionId=33890)

[6] [Web API | Spotify for Developers](Web%20API%20|%20Spotify%20for%20Developers) (<https://developer.spotify.com/documentation/web-api/>)

[7] GitHub Repository for this work: [abhinav3398/song-popularity: given features like acoustics, danceability, key, loudness, etc., predict the popularity of a song. (github.com)](https://github.com/abhinav3398/song-popularity)