PROJECT - PROPOSAL & OUTLINE

Beat Analytics : Spotify Data Analysis and Song Popularity Prediction

Data Preparation & Analysis (CSP-571)

Project Group

Project Group Members:

Student Name	A number
Kasargod Kailash Chandra Shenoy	A20526053
Aditya Nayak	A20528097
Akshat Behera	A20516439
Veerendra Gopichand Karuturi	A20529571

Project Group Leader:

Student Name	A number
Kasargod Kailash Chandra Shenoy	A20526053

Table of Contents

Content	Page No.
Table of Contents	1
1. Project Proposal	2
1.1. Description	2
1.2. Questions	2
1.3. Proposed Methodology	2
1.4. Metrics	3
2. Project Outline	4
2.1. Literature Review	4
2.1.1. References	4
2.1.2. Supplemental Resources	5
2.2. Data	5
2.2.1. Overview	5
2.2.2. Feature Description	6
2.3. Data Processing and Pipeline	8
2.4. Data Stylized Facts	9
2.5. Model Selection	9
2.6. Software	10
2.6.1. R and RStudio	10
2.6.2. Libraries	10

1. Project Proposal

1.1. Description

The "Beat Analytics: Spotify Data Analysis and Song Popularity Prediction" project aims to explore and analyze Spotify music data to understand trends, patterns, and factors influencing song popularity. In order to extract insights from Spotify's enormous music archive, the project will make use of data analysis and machine learning. This will give musicians, record labels, and music lovers important information.

The R programming language will be used in our study for predictive modeling, data processing, and visualization. R is the best option for this project because of its vast libraries and statistical analysis features. To investigate the connections between various musical elements and song popularity, we will make use of a variety of data visualization approaches. Furthermore, we are going to create a prediction model to estimate song popularity based on certain attributes.

We want to use Spotify data analysis to find the formula behind hit songs, spot new trends in the music business, and give stakeholders useful information they can use to make wise decisions. In addition to improving our knowledge of the dynamics influencing music appeal, our study will show how data science is actually used in the music business.

1.2. Questions

The project seeks to answer the following questions:

What are the key features that influence the popularity of a song on Spotify?
How can we predict the future popularity of a song based on its features?
What trends and patterns can be observed in the top songs from 2010 to 2019?

1.3. Proposed Methodology

This section outlines the approach and methodology we plan to use for our analysis.

1. Data Collection: We will gather data from two primary sources:

- Top Spotify Songs from 2010-2019 by Year dataset from Kaggle/DataCamp.
- Ultimate Spotify Tracks Database from Kaggle.

2. Data Preprocessing: This step involves cleaning and preparing the data for analysis. We will:

- Handle missing values and outliers.
- Convert categorical variables into numerical formats if necessary.
- Normalize or scale the data if required.

3. Feature Engineering: We will extract and create new features from the existing data to better capture the underlying patterns. This may include:

- Aggregating song features by artist or genre.
- Creating time-based features like release month or season.

4. Exploratory Data Analysis (EDA): We will conduct an in-depth exploration of the data to uncover trends, patterns, and relationships. This will involve:

- Visualizing distributions of various features.
- Analyzing the correlation between different variables.
- Investigating how different features impact song popularity.

5. Model Development: We will develop a machine learning model to predict song popularity. The steps include:

- Splitting the data into training and testing sets.
- Selecting an appropriate algorithm (e.g., linear regression, decision tree, random forest, or gradient boosting).
- Training the model on the training data.
- Tuning hyperparameters to optimize model performance.

6. Model Evaluation: We will evaluate the model's performance using appropriate metrics such as R-squared, mean squared error (MSE), or mean absolute error (MAE).

7. Insights and Recommendations: Based on our analysis and model results, we will provide insights and recommendations to music industry stakeholders on how to improve song popularity.

1.4. Metrics

This section outlines the metrics we will use to measure the results of our analysis:

1. Popularity Score: The popularity score of a song on Spotify, which is a measure of how often the song is played and how recent those plays are.

2. Feature Correlation: The correlation between various audio features of a song (such as danceability, energy, acousticness, etc.) and its popularity score.

3. Model Performance Metrics:

- **R-squared (R^2):** Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
- **Mean Absolute Error (MAE):** Measures the average magnitude of the errors in a set of predictions, without considering their direction.
- **Root Mean Squared Error (RMSE):** Measures the square root of the average of the squares of the errors.

4. Feature Importance: Identifying which features are most influential in predicting the popularity of a song.

These metrics will help us evaluate the effectiveness of our analysis and the accuracy of our predictive model.

2. Project Outline

2.1. Literature Review

The "Beat Analytics: Spotify Data Analysis and Song Popularity Prediction" project's literature study examines previous research and studies on the subjects of machine learning applications in the music industry, song popularity prediction, and music data analysis.

2.1.1. References

1. "Are Hit Songs Becoming Less Musically Diverse?" by Andrew Thompson, Matt Daniels, Damián Gaume. This article explores the musical diversity of hit songs over time. *Link*: The Pudding

2. "Song Popularity Predictor" by Mohamed Nasreldin. This article discusses methods and techniques for predicting song popularity.
Link: Towards Data Science
(https://towardsdatascience.com/song-popularity-predictor-1ef69735e380)

3. Agarwal, S., Cherakkara, R., & Jhonnalagadda, M. (2018). Predicting Music Popularity with Machine Learning. IEEE Xplore, 1-6

4. Mauch, M., MacCallum, R. M., Levy, M., & Leroi, A. M. (2015). Measuring the Evolution of Contemporary Western Popular Music. PLOS ONE, 10(4), e0131732.

2.1.2. Supplemental Resources

1. Cher Lau's article on the 5 Steps of a Data Science Project Lifecycle provides a structured approach to handling data science projects.
Link: Towards Data Science

(https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492)

2. "R for Data Science" gives us a practical guide to data manipulation, visualization, and analysis using R, with real-world examples and case studies.
Link: https://r4ds.had.co.nz/

3. Spotify Engineering blogs offer insights into data analysis and engineering practices, which could be adapted to the Spotify data analysis project.
Link: Spotify Engineering Blog

4. Harvey, A. (2017). Music, Evolution, and the Harmony of Souls. Cambridge University Press.

2.2. Dataset and Sources:

2.2.1. Overview

a) Top Spotify Songs from 2010-2019 by Year

Description: This dataset comprises approximately 600 songs that were among the top songs of the year from 2010 to 2019, as measured by Billboard. It includes 13 features for exploration.

Source: Kaggle/DataCamp Dataset – Top Spotify Songs 10’-19

Data Origin: Extracted from organizeyourmusic.playlistmachinery.com

b) Ultimate Spotify Tracks Database

Description: This dataset provides comprehensive information on Spotify tracks, including various features such as acousticness, danceability, energy, etc., along with the popularity of the songs.

Source: Kaggle Dataset – Ultimate Spotify Dataset

Spotify Web API documentation on getting audio features: https://developer.spotify.com/documentation/web-api/reference/get-audio-features

2.2.2. Feature Description

Below is the feature description table for the 1st dataset - Top Spotify Songs from 2010-2019 by Year. It has 600 songs (i.e., Observations), along with 14 columns (i.e., Features) from which 13 can be used for the exploration and analysis.

Field Name	Type	Description
title	String	The title of the song.
artist	String	The artist or performer of the song.
top genre	String	The top genre classification of the song.
year	Integer	The year the song was released.
bpm	Integer	Beats per minute (tempo) of the song.
nrgy	Integer	Energy level of the song, likely subjective.
dnce	Integer	Danceability rating of the song.
dB	Integer	Loudness of the song in decibels.
live	Integer	Likelihood of the song being performed live.
val	Integer	Positivity or mood of the song.
dur	Integer	Duration of the song in seconds.
acous	Integer	Acousticness of the song.
spch	Integer	Presence of spoken words in the song.

pop	Integer	Popularity rating of the song, possibly subjective.

Following is the feature description table for the 2nd dataset - Ultimate Spotify Tracks Database. It has approximately 232, 725 tracks (i.e. Observations) spread across 26 genres of music. Where each genre approximately has 10,000 songs belonging to a particular genre. along with that 18 columns (i.e., Features) which can be used for the exploration and analysis.

Field Name	**Type**	**Description**
genre	String	The genre of the track.
artist_name	String	The name of the artist who performed the track.
track_name	String	The name or title of the track.
track_id	String	A unique identifier for the track.
popularity	Integer	The popularity score of the track.
acousticness	Float	A measure of the acoustic characteristics of the track.
danceability	Float	A measure of how suitable a track is for dancing.
duration_ms	Integer	The duration of the track in milliseconds.
energy	Float	A measure of the energy of the track.
instrumentalness	Float	A measure of the presence of instrumental sounds in the track.
key	Integer	The key the track is in.
liveness	Float	A measure of the presence of a live audience in the recording.
loudness	Float	The overall loudness of the track in decibels (dB).

mode	Integer	The modality of the track (major or minor).
speechiness	Float	A measure of the presence of spoken words in the track.
tempo	Float	The tempo of the track in beats per minute (BPM).
time_signature	Integer	The time signature of the track.
valence	Float	A measure of the musical positiveness conveyed by a track.

2.3. Data Processing and Pipeline

The data processing and pipeline for this project involve the following steps:

1. Data Collection: The 'Ultimate Spotify Tracks Database' and 'Top Spotify Songs from 2010-2019 by Year' datasets are gathered from Kaggle.

2. Data Integration: For analysis, the datasets are combined into a single data frame. In order to do this, the datasets must be combined using similar characteristics, such as song ID or artist name.

3. Data Cleaning: To guarantee the quality of the integrated data, it is cleaned. This covers dealing with missing values, eliminating duplication, and fixing inconsistent data.

4. Feature Engineering: To improve the analysis, new features are extracted from the current data. 'Year', 'Month', and 'Day' are extracted as distinct columns, for instance, by parsing the 'Date' field. Furthermore, musical attributes such as "acousticness," "danceability," and "energy" are normalized for comparison.

5. Data Aggregation: In order to facilitate analysis, the data is compiled based on pertinent attributes. For example, to investigate patterns over time or across categories, songs are classified by year, artist, or genre.

6. Data Visualization: Patterns can be found in the data by exploring and creating visualizations. Plotting the distribution of song elements, patterns in song popularity over time, and artist or genre comparisons are all examples of this.

7. Model Preparation: By dividing the data into training and testing sets, the data is ready for modelling. The goal variable is the song's popularity score, and features are chosen according to how relevant they are to song popularity.

8. Model Training and Evaluation: Test data is used to evaluate machine learning models, while training data is used to train them. R-squared, MAE, and RMSE are examples of metrics used to evaluate the models' performance.

The project attempts to accurately estimate song popularity and analyse Spotify data in an efficient manner by adhering to this data processing and pipeline.

2.4 Data Stylized Facts

In order to improve our analysis, our project's main usage of data stylization is the processing of attributes connected to music. Standardized elements like "Danceability," "Energy," "Acousticness," and "Valence" provide consistency between tunes.

For instance, the "Danceability" characteristic is normalized on a scale from 0 to 1, based on a mixture of musical factors, to determine whether a track is suitable for dancing. In a similar manner, "Energy" is calculated to indicate a track's activity and intensity on a scale of 0 to 1.

Then, using these processed features, patterns that influence a song's popularity are found and trends in musical tastes are examined. We can more effectively compare songs and derive insightful conclusions from our data by standardizing these properties.

2.5 Model Selection

After processing and styling data as explained in sections 2.3 and 2.4, we have the following predictors and response for our model:

- **Predictors:** Acousticness, Danceability, Energy, Valence, Tempo
- **Response:** Popularity Score

To estimate a song's popularity score, a linear regression model is applied to the training data using the aforementioned predictors and response. Metrics like R-squared, MAE, and RMSE are used to analyse the model's performance.

Our approach uses data visualization techniques in addition to linear regression to look for trends and patterns in the Spotify data. This gives us further insight into the variables affecting song popularity.

2.6 Software

The analysis and modeling for this project are conducted using R and RStudio, which are powerful tools for statistical computing and graphics.

2.6.1 R and RStudio

R is a free software environment and programming language used for statistical computing and graphics. For the development of statistical software and data analysis, statisticians and data scientists use it extensively. Numerous statistical and graphical methods are available in R, such as time-series analysis, classification, clustering, linear and nonlinear modeling, traditional statistical tests, and more.

The integrated development environment (IDE) for R is called RStudio. It offers an easy-to-use interface for managing packages, creating and running R code, and visualizing data. RStudio offers code editing, debugging, and project management tools to make working with R more productive.

R is the tool we utilize for data preprocessing, analysis, and modeling in our research. The platform used for coding and results visualization is called RStudio. R and RStudio work well together to provide a stable environment in which to analyze our data and create song popularity prediction models.

2.6.2 Libraries Used

Several R libraries are used in our research to help with data processing, analysis, and modeling. Among the important libraries utilized are:

tidyverse: A collection of R packages designed for data science, including ggplot2 for data visualization, dplyr for data manipulation, and readr for reading data.

lubridate: A package that makes it easier to work with dates and times in R, which is particularly useful for processing the "Date/Time" column in our dataset.

caret: A package for building predictive models and performing cross-validation, which is used for training and evaluating our linear regression model.

plotly: An interactive graphing library that allows us to create interactive plots and visualizations, enhancing the exploratory data analysis phase.

spotifyr: A wrapper for accessing the Spotify Web API from R, which can be used to gather additional data or features from Spotify if needed.

These libraries provide a range of functions and tools that streamline the data analysis process and enable us to build effective predictive models for song popularity.