Applying the CRISP-DM Process on a Spotify Dataset

Sanjay Bhargav Kudupudi

September 27,2023

Abstract

This research paper explores the application of the Cross-Industry Standard Process for Data Mining (CRISP-DM) on a Spotify dataset spanning from 2010 to 2022. The primary objective is to identify the features influencing track popularity. The study reveals insights into track characteristics and their relationship with popularity, providing a foundation for further in-depth analyses and advanced modeling.

1 Introduction

The ever-evolving music industry relies increasingly on data analytics to understand listeners' preferences and trends. Spotify, as one of the leading music streaming platforms, offers a rich dataset for such analysis. This study uses the CRISP-DM process, a structured approach to data mining, to uncover insights from the Spotify dataset.

2 Methodology

2.1 CRISP-DM Overview

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It is a robust, structured approach that provides a detailed roadmap for data-driven problem-solving. The process consists of six phases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

2.2 Dataset

The dataset comprises tracks from Spotify playlists from 2010 to 2022. It contains features such as track name, artist, genres, and various musical attributes like danceability, energy, and valence.

3 Results

3.1 Data Understanding

Initial data exploration revealed information about track characteristics, artists, and genres. Most tracks in the dataset have high popularity scores, with rich musical attributes available for analysis.

3.2 Data Preparation

After handling missing values and outliers, the dataset was split into training and test sets, ensuring a representative sample for modeling.

3.3 Modeling

A basic linear regression model was employed to identify the features that influence track popularity. Although the model's performance was modest, it provided insights into the relative importance of various features.

4 Discussion

The analysis suggests that features like acousticness and danceability positively influence track popularity, while energy might have a negative impact. However, these findings should be interpreted cautiously, given the linear model's limitations. Further studies using advanced models or techniques might provide a clearer picture.

5 Conclusion

This study, though preliminary, underscores the potential of data analytics in understanding music trends and preferences. By leveraging structured methodologies like CRISP-DM, stakeholders in the music industry can derive actionable insights from vast datasets, paving the way for informed decision-making.

6 Acknowledgments

We express our gratitude to all contributors and experts who provided the dataset and tools essential for this research.

7 References

1. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

2. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

3. Spotify API Documentation. Available at: https://developer.spotify.com/documentation/

4. Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011, October). The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011).

5. Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.