INF6027 Introduction to Data Science

Predicting Song Popularity Using Multimodal Data

Sumit Mishra (240175186)

Word Count:

University of Sheffield

**Motivation and Goal of the Investigation**

Predicting the song's popularity is a fascinating challenge that bridges the domain of data science and creative industries.  In the modern music industry, several platforms like Spotify and Apple Music rely on data-driven decisions to solve different use cases, such as personalized recommendations, predicting the success of an unreleased song, and more. In this context understanding what contributes to the success of a song is crucial since it can help streaming platforms deliver a better user experience but also empowers the artists and producers to make informed decisions and help them which song to buy copyright for and release on their platform which can lead to certain profit margins for instance, In 2022, Spotify reported over 456 million users globally, emphasizing the importance of accurate music recommendations and popularity predictions for improving user retention and satisfaction.

The goal of the research is to predict the popularity of the songs available in the MusicOSet Dataset which is a multimodal dataset containing diverse features such as song, album and artist characteristics and hence using these features to train different machine learning algorithms to compare the algorithms and the dataset has lyrical dataset allowing the extraction of advanced features like sentiment polarity, subjectivity. This allows for an analysis of how lyrical attributes impact the predictive models, providing a comprehensive understanding of the factors influencing song popularity.

**Literature Review**

The prediction of song popularity has been a compelling area of research due to its applications in music recommendation systems, marketing strategies, and understanding listener preferences. Several studies have utilized machine learning techniques to address this challenge. For example, the study *"Predicting Song Popularity Using Spotify Web API"* analyzed over 160,000 songs and employed models like Random Forest and XGBoost to predict popularity using features such as tempo, loudness, and genre, achieving high predictive accuracy [1]. Similarly, a Stanford University project titled *"Song Popularity Prediction"* applied regression models, neural networks, and support vector machines, emphasizing the importance of metadata and audio features in predicting song success [2]. These studies highlight the role of multimodal data in achieving accurate predictions, aligning with this research's focus on integrating diverse song, artist, and album characteristics.

Beyond structural and audio features, the role of lyrical content has gained attention for its impact on a song’s success. The paper *"Quantitative Sentiment Analysis of Lyrics in Popular Music"* analyzed Billboard Hot 100 songs over decades, revealing correlations between emotional content and popularity trends [3]. Another study, *"Hit Song Prediction: Feature Engineering with Metadata and Lyrics,"* demonstrated the influence of sentiment polarity and repetition metrics on model performance, with Random Forest and Logistic Regression models outperforming others [4]. Furthermore, the paper *"Beyond Beats: A Recipe to Song Popularity?"* explored multimodal approaches combining lyrical, audio, and artist data, showing that integrating diverse features significantly enhances prediction accuracy [5]. This body of work supports the approach in this study, which incorporates lyrical features like sentiment polarity alongside traditional song attributes to analyze their collective impact on song popularity.

**Research Aim**

The aim of this study is to develop and evaluate predictive models for forecasting song popularity using multimodal data, including song features, artist attributes, and album characteristics. Initially, a baseline model will be trained using these core features to establish foundational insights. Subsequently, feature engineering will be performed to extract lyrical features, such as sentiment polarity and linguistic attributes, which will be incorporated to assess their impact on model performance. This study also involves a comparative analysis of multiple machine learning models, including hyperparameter tuning, to determine which approach provides the most accurate and interpretable predictions of song popularity. By examining the influence of various features and comparing model performances, the research seeks to identify the most significant factors contributing to song popularity while highlighting the strengths and limitations of each modeling approach.

**Research Questions**

1. What combination of song, artist, and album features contributes most for accurate predictions of song popularity for different machine learning algorithms?
2. How does the inclusion of lyrical features, such as sentiment polarity and other linguistic attributes, enhance the predictive accuracy of machine learning models for song popularity?
3. Which predictive model demonstrates the best performance in forecasting song popularity using all available features after tuning for this specific task, and what are the associated strengths and weaknesses of each approach?

**References for Literature review**

1. Yashraj Kakkad, *"Predicting Song Popularity Using Spotify Web API,"* [GitHub Repository](https://github.com/yashrajkakkad/song-popularity-prediction).
2. Stanford University, *"Song Popularity Prediction,"* [CS229 Project](https://cs229.stanford.edu/proj2015/140_report.pdf).
3. Alberto Acerbi, *"Quantitative Sentiment Analysis of Lyrics in Popular Music,"* [UC Press](https://online.ucpress.edu/jpms/article/30/4/161/106385/Quantitative-Sentiment-Analysis-of-Lyrics-in).
4. Paper, *"Hit Song Prediction: Feature Engineering with Metadata and Lyrics,"* [arXiv](https://arxiv.org/abs/2301.13507).
5. Paper, *"Beyond Beats: A Recipe to Song Popularity?"* [arXiv](https://arxiv.org/pdf/2403.12079).