Title Page

- **Title:** Analysis of Factors Influencing Song Popularity on Music Streaming Platforms
- **Team Members:** Harsh Sahay, Srijan Gupta, Sahil Mehta
- **Team Name:** Skyfall21
- **Date of Submission:** 04/06/2024
- **GitHub Repository:** **Skyfall21's Project Github Repo**

Abstract

The objective of our project is to dissect and understand the myriad factors contributing to a song's popularity across key music streaming services like Spotify. Utilizing a comprehensive dataset from Spotify, we've embarked on a journey to unravel how song characteristics, artist clout, release timing, and ubiquity across different platforms intertwine to predict song popularity. Preliminary activities have included dataset acquisition and initial data cleaning, setting the stage for in-depth analysis and modeling.

Introduction

In the digital era, music streaming platforms have revolutionized how music is consumed and discovered. Our project, driven by the quest to decode the formula for song popularity on these platforms, leverages previous research and our dataset's rich attributes, poised to bridge gaps in understanding the multi-dimensional nature of song popularity.

Method

Dataset Overview

- **URL:** https://www.kaggle.com/datasets/arnavvvvv/spotify-music
- **Total Samples:** 953 songs, representing the most streamed tracks on Spotify, complete with attributes like track name, artist(s), release date, playlist inclusion, streaming statistics.

- **Preprocessing Required:** Yes, including normalization, missing value imputation, and categorical encoding.
- **Format and Labels:** JSON format; "popularity" label based on Spotify's index, ranging from 0 to 100.
- **Partitioning:** 70% training, 15% validation, 15% testing.

Progress on Methodology

- **Data Preprocessing:** Completed initial cleaning and standardization.
- **Exploratory Data Analysis (EDA):** Scheduled to commence next, aiming to identify patterns and relationships.
- **Feature Engineering and Predictive Modeling:** Post-EDA, we'll develop new features and apply machine learning models, including regression analysis, random forests, and gradient boosting machines.
- **Performance Evaluation:** Utilizing RMSE and MAE as primary metrics for model assessment.
- **Baseline Code Testing:** Utilizing open-source code from Elena Georgieva et al.'s study on "Feature Engineering for Predicting Billboard Hits" as a baseline to understand effective feature engineering techniques.

Preliminary Results

Our results are primarily from data preprocessing stages, where we have successfully normalized the dataset and identified potential variables for feature engineering.

Upcoming Tasks

- Conducting EDA to unearth insights and guide our feature engineering.
- Developing and refining our predictive models.
- Evaluating model performance and iterating based on findings.

References

1. Velmuruga, Dr.A. (2023). Machine Learning Approaches for Predicting Song Popularity: A Case Study in Music Analytics. INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT. 07. 1-11. 10.55041/IJSREM27361. This study explores various computational models to predict the popularity of songs. https://www.researchgate.net/publication/376387750_Machine_Learning_Approaches_for_Predicting_Song_Popularity_A_Case_Study_in_Music_Analytics/

2. Alison Salerno (2020): "Prediction of Spotify Song Popularity". This paper examines different machine learning techniques to predict the popularity of Spotify songs. Explore on IEEE Xplore. https://medium.com/analytics-vidhya/predicting-song-popularity-71bc3b067237/

3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019): "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". While primarily focused on natural language processing, BERT's methodologies can be adapted for analyzing textual data in song lyrics to predict popularity. https://aclanthology.org/N19-1423/

4. Elena Georgieva, Shiva Pentyala, Marco Giunta, Paolo Papotti, and K. Selçuk Candan (2018): "Feature Engineering for Predicting Billboard Hits". This study discusses feature engineering techniques crucial for predictive models in the context of Billboard hits, applicable to Spotify song popularity prediction. Find on ACM Digital Library. https://ccrma.stanford.edu/~egeorgie/documents/HitPredict_Final.pdf

Contributions

- **Harsh Sahay:** Spearheaded the project's conceptualization and dataset identification. Leading the data preprocessing phase.
- **Srijan Gupta:** Taking charge of the EDA phase and contributing significantly to model development.
- **Sahil Mehta:** Conducted a thorough literature review, aiding in the establishment of our methodological framework.