Team: Capucine Philippine Leroux: cpl2146, Chris Berniel Cobashatse: cc4536, Narjes Al-Zahli: na2852, Vedant Gannu: vg2565, Xiangming Huang: xh2550

Background and context to the problem statement.

When a singer releases a song, it can be hard to predict if the song will be a hit. With streaming platforms, such as Spotify, releasing open source statistics on number of listenings, or feature extraction, we can start understanding better which features matter for a song to be a hit.

The main problem we want to address in this project is: How to predict whether a song will be a hit on Spotify? We chose Spotify because it has an open source web API and is the world's leading streaming platform with 489 million users and 205 million premium subscribers across 180 regions.

For this problem, we need to define what it means for a song to be a hit. Many different features can be considered to quantify the popularity of a song. Namely, relevant features could be whether the song was a hit in the long-run vs short-term (good longevity vs good start), the number of streams after the release date, the relative number of plays with respect to the artist's number of followers, famous chart records like the Billboard Hot 100, or music recording certifications (gold/platinum record).

In order to make a prediction, we need our popularity score to be consistent throughout our whole dataset. Besides this, we have to take into account the fact that popularity evolves over time: a song that was popular in the 50s might not be popular anymore.

The Spotify API gives open access to their own popularity score which is, according to their website, *"a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are."* As we don't have access to the number of streams per song, or the history of stream records, we decided to use the popularity score provided by the Spotify API to make our prediction. **Our goal is to predict the popularity score of a song in March 2023 based on the current popularity score of the Spotify API as of March 2023.**

Identification and description of the data set(s) you are planning on using along with their source

Our baseline is a dataset of songs available at this link: https://www.kaggle.com/datasets/yamaerenay/spotify-dataset-19212020-600k-tracks. The data was constructed using the Spotify web API (https://developer.spotify.com/discover/), it covers information about 600k+ songs and 1M+ artists between 1921 and 2020, and it recollects features such as: songs name, songs audio features (danceability, energy), songs release date, songs duration, songs popularity score, songs artist, artists name, artists genre, and artists number of followers. We want the popularity score to be as up-to-date as possible, so we will update the popularity using the Spotify API. We will complement this database as much as possible with all of the relevant features we can find on the API, such as acousticness, instrumentalness, key, liveness, loudness, speechness, tempo, time signature, and valence.

Additionally, we will use the songs' lyrics, using the Genius API (https://docs.genius.com/#/getting-started-h1).

Proposed ML techniques you are proposing on applying to solve the problem

- Random forest, XGBoost for regression

Reasoning: use regression techniques to predict the popularity score of a song.

- NLP LSTM on lyrics

Reasoning: perform sentiment analysis and extract other features from the lyrics