Predicting the Number of Streams and Popularity of a Song on Spotify

Haneen AlHomoud

Introduction

Spotify is a digital music service that enables users to remotely source millions of songs on various record labels from laptops, smartphones, or other devices. It is one of the major music streaming platforms, with over 191 million active users, 40 million songs, and 2 billion playlists. A key feature of Spotify is the song recommendations to new users based on top music/tracks streams. Each music service provider is competing to give their listeners the joy of discovering the best music, and there are many algorithmic approaches. In this project, we want to predict the song's streams/popularity using the song's main features. This second project of the Bootcamp got me familiar with Machine Learning through web scraping techniques, APIs, and modeling methods such as Linear Regression

Data

To start collecting the data, I needed to get the top 200 music worldwide from 2016-2021. Thankfully, I used beautifulSoup, a Python library for pulling data out of HTML to scrape the Spotify Chart website *(https://spotifycharts.com/regional)* and got about one thousand an two hundred songs' basic information Spotify provide with their rank on each year info.
To get the song's features to build the model, I used Spotify's fantastic API service that allows the users to archive millions of songs' audio features in JSON format. You need to create an account at *(https://developer.spotify.com/)* Essentially, APIs work on a request-response basis.

Algorithms

1. Problem Understanding
2. Data Scraping
3. Dataset Exploration and Cleansing
    - Null Values

    - structural errors
    - Outliers
    - Duplicated rows
4. Exploratory Data Analysis (EDA)
5. Feature Selection
6. Modeling
7. Insights
8. Conclusion

Tools:

- Technologies: SQL, SQLite, Python, Jupiter Notebook.
- Libraries: NumPy, Pandas, Matplotlib, Seaborn, sqlalchemy, sklearn, spotipyre, requests, bs4, statsmodels.