Using Data to Find the Next Hit Album - Project Proposal - Jessica Wulzen

Problem statement:

Record labels want to release albums that will be successful, but trying to determine what acts will be well-received is subjective and imprecise. Music blogs and reviewers play a critical role in the promotion and discovery of new groups. By understanding trends in what reviewers respond positively to, record labels can make more informed decisions about what artists have the highest chance of being critically acclaimed.

To solve this problem, we will look at the scores assigned to albums by music reviewers, and use a variety of attributes from the songs to attempt to predict these scores. Steps will include collecting the data from multiple sources (more below), performing EDA to determine what musical attributes are most relevant, and attempting to develop a model to determine what albums will be better received by critics. Two aspects in particular will be given attention: first, to the music's genre, to help understand what traits are most popular in different types of music; and second, to the context of time, looking at if certain traits become overused and lose their popularity over time, or are cyclical and have a resurgence after a period of unpopularity.

Additionally, this data will allow us to answer a second question - what musical attributes are defining for each genre? Can we teach a computer to differentiate between electronic music, jazz, and rock and roll? This classification problem will serve as an opportunity to experiment with a variety of different machine learning algorithms.

Data:

A Kaggle dataset contains 18,000 reviews from Pitchfork.com, containing the review score, genre, and full review text. The data comes downloaded as a sqlite database, which can be read into Python using the sqlite3 package.

Spotify's API allows for access of musical attributes of all songs, including features such as key, time signature, and duration, along with other features determined by Spotify such as energy, acousticness, danceability, and many more. The package spotipy allows to access the spotify API through Python. Through this package, we can search for the albums contained in Pitchfork review dataset, find those albums in Spotify, access the song data for each song on the album, and then construct a variety of features to use in analysis based on the song data. This process will have challenges, as the program will have to match album names which may not be unique in each dataset, or may have differences with words such as 'remastered' in one dataset. Additionally, care will need to be given to handle different versions of albums, such as extended cuts, remixes, or live recordings vs the original studio recordings.

Along with the basic variables outlined above, opportunities exist to create many other variables for analysis. The full text of each review is available, making NLP possible as an avenue. Additionally, the Pichfork review score is at the album level, while the Spotify data comes at both the album and song level. Thus, additional variables can be created by looking at

the composition of song attributes in each album, such as looking at if the songs are similar to each other or if they vary on the musical elements we have access to.

Deliverables:

This project will result in several deliverables, including a written report, code, and visualizations depicting the findings and success of the algorithms developed.