Music Mood Classifier

Ankit Solanki
Data Science Capstone
Brainstation

Project Statement:

Classification of moods of the songs from Spotify's public playlists. From the metadata of the songs, classify the moods or the general sentiments of the tracks using unsupervised learning methods.

Background:

Music is typically categorized by genres (e.g. rock, jazz, blues). The way people consume music has changed drastically over last 15-20 years. With rise of digital streaming platforms and instantly downloaded contents, consumers are exposed to wide variety of choices and especially in music. Consumer patterns have moved towards more track level and playlist-oriented listening instead of particular artists or albums. With vast amount of choices for consumers, it gets difficult to keep track of the favorite music. The music descriptors like genres can help classify the music and distribute it in a manageable way. But Genres are usually not capable to evoke moods. If a place or a themed event needs a specific kinds of music that can represent brand image, it can only be done by mood descriptors.
Streaming services like Spotify does offer generalized mood playlists but can't provide personal level filtering of the music. Their Discover weekly playlists are somewhat similar but they rely heavily on the type of genres people listen to.
That's why a robust music mood classifier can be very useful for many situations.

Data Acquisition:

Although there many resources to get the music metadata online, I decided to build my own web scraper using python library called Spotipy and Spotify's web API support. The web scraper helped me fetched about 31000 song's info and their audio features.
The audio features in the dataset are the most important to define moods from their values. The important features include Acousticness, Loudness, Valence, Mode, Speechiness, Tempo etc. In total there are 19 columns and 31533 datapoints. The final dataframe is in the CSV format.

Data cleaning / Preprocessing:

In terms of cleaning the dataset, I was fortunate enough to get the incredibly detailed and fairly cleaned dataset from Spotify API. There were no null values in the dataset. For the preprocessing part – I first encoded explicit column with 1 and 0s. and then I have used MinMacScaler() to scale the rest of the audio features and numerical columns. Because it was going to be a unsupervised learning using cluster analysis, it's always recommended to scale the data.
In the EDA I found few audio features heavily correlated to each other which helped for feature selection process for cluster analysis. For example, Energy and loudness, energy and acousticness , acousticness and loudness explicit and speechiness have high correlations.
Ultimately, I ended up choosing those and a few more features wich were normally distributed - to define moods of the music. The features are acousticness, danceability, energy, loudness, speechiness and valence.

Cluster Ananlysis:

Because there are no target features, I have decided to use to Kmeans Clustering method with dimensionality reduction techniques like PCA to analyze and get distinct clusters that would be the moods of the tracks. After this process I was able to derive 4 clusters from the dataset. I have defined those clusters as 'cheerful', 'uplifting', 'chill, 'melancholy', and predicted and imputed those newly derived moods to their respective tracks. I have created a new CSV file with the mood labels as target features for future use.

Supervised Learning:

I now have the target feature to run machine learning models to and train them. For this part I have decided to go with three most robust classification models for the multiclass classification problem. The models I am using are Knearest neighbors, Support Vector Machines, and Random Forest Classifier. Surprisingly all three performed very well to the point that it seemed like the data was overfitting. But even after corss validation and hyperparameter tuning with grid-search, they all three gave very impressive test scores. After evaluating the models form their confusing matrixes , I came to the conclusion that Random Forest Classifier worked best for my dataset and had the least number of error. Therefore, I will be using Random Forests or Ensemble learning for the final production of the project.

Summary:

For the future work, I will be creating a landing page where anyone can upload the link of their spotify playlists and get the mood analysis report and recommend a new personalized playlist for the user.