Final Project Proposal

Ashwini Marathe

Domain Background

Music has been an integral part of our culture all throughout human history. In 2012 alone, the U.S. music industry generated $15 billion. Of this $15 billion, the majority of the revenue is generated by popular, mainstream songs. Having a fundamental understanding of what makes a song popular has major implications to businesses that thrive on popular music, namely radio stations, record labels, and digital and physical music marketplaces. The ability to make accurate predictions of song popularity also has implications for customized music suggestions. Predicting popular songs can be applied to the problem of predicting preferred songs for a given population.

In this project I plan to work with Spotify data which has many music related attributes for a song. Spotify has an open developer API to download this data. Using this data I want to predict if a song will be popular or not based on the song and singer attributes. To decide if a song is hit or not I plan to threshold the song popularity variable.

Problem Statement

I would like to explore the impact of music related attributes as compared to the singer related attributes on popularity of songs. Previously two types of problems have been solved using the spotify data: song genre prediction and song popularity prediction. However, I would like to analyze the impact of singer attributes (age, gender, number of previous hits) in addition to the song attributes on popularity of songs.

Data

The spotify API provides 13 attributes for a song, a few of which are listed below:

- duration_ms: The duration of the track in milliseconds
- mode: Mode indicates the modality (major or minor) of a track
- acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic
- danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity
- energy: perceptual measure of intensity and activity
- instrumentalness: whether a track contains no vocals, and others

Spotify has data for millions of songs. I plan to work with data for around a million songs. However, the proportion of song gaining popularity is low and hence the data will be imbalanced. Different techniques might be required to overcome this problem.

Solution Statement:

Since it is a classification problem I plan to use logistic regression. This will be the baseline model. But due to imbalance in the data I want to compare the results of the baseline logistic regression with other tree based methods to evaluate which class of methods work better for imbalanced datasets.

Evaluation Metrics

I would like to know the accuracy of preductions. But since the dataset is imbalanced, maximizing the accuracy metric will not lend an optimal model. I will evaluate the models based on four metrics: * Accuracy * Precision * Recall * F1 score * AUC (are under the receiver operating curve)

Project design

The first phase of the project will be to collect data using the spotify web developer API. The API returns a maximum of 10,000 songs for a particular query (to avoid denial of service attacks). I plan to collect for the recent two years or so. Once the song idss are collected, I will also collect the artist information and the song features.

Once the data is collected, I will check if the data needs any cleaning or imputation. After data cleaning, I will visualize the different variables as comapred to the popularity of the song. I will threshold the song popularity to divide songs in two groups: hit and non-hit. I aim to predict if a song will be hit or not based its audio features and artist features.

Next step will be trying different models and tuning the hyper-parameters for the selected model.