CS334 Project Proposal

Predicting the Popularity of a Song

Group Members: Jonathan Kim, Anna Mola, Mohammed Alsalamah

1. Description of the problem and the data.
    a. The Spotify Hit Predictor Dataset
    b. The problem: Art is highly subjective, and it can often be hard to determine whether a given song could be a hit. As such, it may be useful for music producers and labels to algorithmically determine the potential sales of a track. The goal of this project would be to provide a more objective metric that can be used to predict whether a given song will be a hit or not.
    c. What does your dataset look like?
        i. The dataset is titled "The Spotify Hit Predictor Dataset (1960-2019)". It is a database of over 40,000 hit songs from the last 6 decades, each uniquely identified by 19 features. A song is determined as a "hit" if it landed within the Billboard's weekly top 100 during its lifespan. The data is split up into 6 .csv files, with each csv representing the hit songs of a given decade. This data can be used in conjunction with a machine learning model to predict whether or not a new song would be a hit.
2. Description of what you plan to do. For example, what are the machine learning methods you plan to apply or improve on. What are the experiments you plan to run and how do you plan to evaluate the algorithms?
    a. We're going to assess the dataset with a wide variety of models. After doing some research and looking at other evaluations of this dataset, we've observed that the random forest machine learning model seemed to perform well. We want to focus on this model as it is most promising with training accuracy and we can run additional tests to determine which parameters provide the most accurate results. For example, what should be the value for the maximum depth of the random forest tree or minimum number of samples required to be at a leaf node. We may use F-1 as a scoring metric to measure and find the most optimal values for the hyperparameters mentioned above. In addition to the random forest model, it would be useful to analyze other models we have learned about in class like K-Nearest Neighbors, Decision Trees, Linear and Logistic Regression, and possibly Neural Networks.
    b. Our plan is to compare the test accuracy of all the models to determine which model and set of parameters for that model can provide the most accurate results.
    c. From there we can select the most accurate model and possibly use song data from outside the dataset to make a prediction about that song's popularity.
3. Short list of references to showcase what has been done to date.

- https://techxplore.com/news/2019-09-spotify-songs.html
    - This article covers a pair of students at the University of San Francisco who used a similar dataset to predict billboard hits. They compared four different learning models, which, sorted for performance, were Logistic Regression, Neural Networks, SVM, and Random Forest. Their performance was admiral, with their models achieving up to an 88 percent accuracy using a test dataset.
- https://www.kaggle.com/gcdatkin/spotify-hit-prediction
    - This user on Kaggle evaluated the same dataset we are using by assessing the data using 8 different models and comparing them against each other.
- https://arxiv.org/pdf/2007.03137.pdf
    - Research paper demonstrating six different algorithms' (Logistic Regression, Decision Tree, Random Forest Classifier, Gradient Boosting-XGBoost, Neural Networks) performance on predicting whether a song will be a billboard hit.
- https://www.diva-portal.org/smash/get/diva2:1214146/FULLTEXT01.pdf
    - Research paper with an overview of work done in predicting songs' popularity based on audio-only features. They determined that they couldn't predict a song's popularity just based on audio features alone.