CISC 372 – Advanced Data Analytics
Project Proposal

Udbhav Balaji (19ub@queensu.ca)

Spotify Hit Predictor

Background & Motivation:
In the times of the internet, it is very easy for musicians and artists to be able to share their work with the rest of the world. However, this ease also makes it easy for any new artist to easily lose motivation if they aren't getting the numbers they need to succeed. In that sense, the Spotify Hit Predictor makes it easy for them to understand whether their song has the potential to be a hit (given the right promotions). It allows them to focus on the songs that really has that potential. Apart from this, music labels and record companies can also use this to allocate a promotion budget for each of the songs they release to get back the maximum return on investment.

Informal Description of the Project/Problem:
As of today, Spotify is the leading platform for artists to post their work and reach their audience. While there are a lot of established artists out there, there are thousands/millions of other artists who are looking for their first big break. Unfortunately, there isn't a tool that could help them realize the potential their songs might have. This is also a problem for music labels as they have no idea how a particular song will perform online, and hence, must take a big gamble on how much money they must spend promoting them.

General Direction of Solution:
This is a service, where the user can input the link/ID of the song and can find out whether the song has the potential to become a hit song. The way this would work would be, we use the link/ID to get the track features from the Spotify API endpoint and feed in these features to our ML model to predict whether the song could be a hit or not. Our machine learning model will use song data to see which songs have been a hit based on their track features.

Potential Datasets That Will Be Used:
For the ML model to be trained, we need data regarding the different songs, as well as whether they were hits or not. The datasets that I plan on using can be found here (as well as all relevant information regarding the different attributes). For the purposes of the project, I will only include the datasets for the decades 90s, 00s and 10s.

Measuring Performance of Results/Model:
Since we are looking at songs that have potential, an appropriate metric for evaluation would be the sensitivity of the data, since we want to put more emphasis on correctly classifying the hits as hits. There would be lesser costs if a hit was mis-classified as a flop song (but turns out to be a hit).