Johnny Koponen

CS4375.001

Machine Learning Project

Music is a very popular form of culture that has emerged in the 20th century. Similar to most things, music can become boring when it gets played over and over again. What if users were able to find the next up and coming popular song right when it releases?

This dilemma is particularly interesting to me because I am a very big Spotify user. I tend to listen to music for at least 30% of the day. I often hear similar songs and they sometimes become repetitive, forcing me to try and find new releases.

I initially approached this idea by thinking a Binary Classification utilizing genre, duration, artists, and song expression would be able to create efficient machine learning models. I was planning on creating my own “score” for each instance of data by assigning numeric values to specific genres, artists, and other categorical data. After receiving feedback on my project proposal, my perspective on how to approach this project changed. At first, I had a dataset that had only the popular songs of the recent years. There were no songs that were deemed as unpopular which would’ve made training my model very offset. Luckily, Professor Iyer was able to recommend a new dataset to me. [Link](https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-01-21/readme.md). This dataset had a track popularity score as well as instrumentalness, valence, tempo, duration, genre, and other features. Using this new data, I would now be able to train my model properly and try to predict the track\_popualrity score that was a part of the dataset.

Through my Exploration phase of the data, I noticed multiple things. Most of the songs in the dataset were roughly 3 minutes long. The original track\_popularity values were a bit skewed to the left, values seemed to hover around 50-60. The overall tone that was most apparent for the songs in the dataset was a C, and most songs were near the 120 bpm tempo range which is considered pretty upbeat. During the exploration phase, I also noticed some columns weren’t needed. Song name, artist, track\_album\_id, track\_album\_name, playlist\_id, and many others were unnecessary. I decided to drop those columns and run my algorithm with most of the features being danceability, energy, key, loudness, mode, etc. With these I tried to predict the track\_popularity. After running it through my models, I noticed that the accuracies were very low <10% and the r^2 values were very irregular. I believe this was occurring because the track\_popularity was a very specific value between 0 and 100.

To combat this issue, I decided it would be best that I added a new column to the dataset to help my models. I decided to add a new ‘Target’ column under a condition where if the track\_popularity was greater than or equal to 50, then Target would be equal to 1. Otherwise, the target would equal 0. I then made my models try and predict the target value based on the danceability, energy, key, loudness, and mode. Ultimately, my models saw an increase in their accuracy ~60% and their r^2 values were better.

Though my accuracy still only reached up to 65%, I learned a lot during the project. Data exploration really helped me identify features that are necessary when creating machine learning models. Some columns are very unique and won’t help us train our models if we include them. It is very important choosing the correct columns that are most likely to have a correlation to what is trying to be predicted. Different youtube videos, towards data science, and articles over pandas really helped. Personally, I enjoyed creating this final project more than anything else this semester. Seeing my program work on something I am very passionate about was very rewarding.