Carlos Gonzalez

CIS 3715

Dr. Hongchang

April 4, 2021

Progess Report I

So far this week I was in the process of performing data preprocessing. This process was not that long. My first step was to look at the features of the data and check which are numerical or categorical. I then proceeded to define the features in order to allow the reader to understand. For example, liveness checks for an audience in the background to be able to tell people if the song that is in Spotify whether the song was recorded in front of a live audience. This numerical data ranges from 0 to 1. Now considering there was no data missing nor was anything messed up. I decided my next course of action would be to start with t he exploratory data analysis. I wanted to see if there was any correlation between the features so my I decided to create a heat map. A picture containing chart

Description automatically generated

As shown there is a correlation between certain features and how popular a song is. My next step was to look at the correlations more in depth. I decided to do some exploratory data based off the heat map. The first one I did was popularity and acousticness. I created a scatter plot along with a line of best fit.

Chart, scatter chart

Description automatically generated

This allows me to see that there is a level of acousticness that allows a song to be popular. This tends to be at 0.05 acousticness. I then did exploratory data analysis on the rest of the data. There is a slight chance that the data shows that people are looking for more upbeat music.

**What has not been done?**

I did complete what I had to and I’m glad I did. This dataset is quite larger than I expected but I am still confident in my ability to do this project. I do not have anything that I regret about this project. I could’ve have done some research about what machine learning would be the best and learn as much about the concepts behind it as I can. But that was not of the highest priority.

**What will be done the following week?**

I do have quite a bit left to go but I do have enough time to complete. This week I’ll be research more clustering algorithms to be able to complete my first task of being able to classify songs but genre based off the data given. I’ll look at different clustering algorithms such as the clustering algorithms caught in class and those not covered in class. I do hope to be able to find one that was not covered in class so I can take the opportunity to learn more about clustering. After that I’ll train the models and compare them. The week after that I plan on looking at different classifier model to be able to predict if a song will be popular. I’ll use classifiers such as logistic regression and others not talked about such as decision trees. I do plan on trying various models to get a good prediction and will compare each and everyone’s scores. After that I’ll take the week to prepare for the lightning talk and to choose the final models that I will use. For the last week I’ll compile all my results and make a final draft in which I use my chosen models but at the same time I will look also keep the other models to hand in to be able to show the progress throughout the project and how my thought process has changed over time. The final part will most likely be a zip file with everything in it. I do plan each model to its own Jupyter notebook so I do not have to worry about them getting messed up or not. I will keep in mind Principal Component Analysis to see if it will be beneficial to what I am trying to accomplish. I do plan on trying different methods to be able to learn more about numpy and pandas and any other important data science libraries needed. Also on the side I will be looking at various data science papers and projects in order to get a sense on how a proper data science paper or project should look like to be able how to figure out to display my findings in a great way.