Carlos Gonzalez

CIS 3715

Dr. Hongchang

April 11, 2021

**Progress Report 2**

**What has been done?**

So far I completed my 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 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 the exploratory data analysis. This week was a bit more different then before. I decided this week to focus on the differentiating of the genres using different clustering models such as the K-means model. I had to check on my own if there was a specific number of k folds that would be the best in this case. I used the sum of squares method to see which number of folds would be the most beneficial and created a graph to show what number of folds would be appropriate for a dataset of this size.

**Chart, histogram

Description automatically generated**

I decided in this case to choose a k value of 25 since it has a decent SSE error compared to the others. I personally need to choose 25 because a k value of 200 would be too much for my personal laptop. I was able to decently differentiate the given genres. The genres in the dataset were very specific and be able to generalize them a bit more helps in the long run.

**Timeline

Description automatically generated**

I then decided to look at my project timeline that I had previously submitted and decided to add one more thing that will allow me to show my results in digestible way. What I mean is that I am learning how to use Tableau to make an interface that seems appealing to the reader and allows me to explain my data in way that wouldn’t be too confusing. Another thing that I started to do was to see the difference between Tableau and R’s library Shiny to properly show off the data and everything that I did. I am so far steering towards Tableau for time’s sake.

**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 done more research into Tableau before this, but that is not the source of worry because Tableau is more for aesthetics purposes.

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

I do have quite a bit left to go but I do have enough time to complete it. This week I’ll research more machine learning algorithms that will allow me to see if a song will be popular or not. As of right now, I do have some machine learning algorithms in mind such as Linear Regression/Logistic Regression, K Nearest Neighbor, and finally decision trees. I then plan on training the models and to see which one gives me the best results and choosing the correct algorithm. If I can plan everything correctly then I would like to combine use the results from my first part of my project for my second to see if that benefits the models or not. In the end, I do plan on using Tableau to make everything look nice and to be able to show off my results in a way that people can understand. This week may be a little more stressful because of exams and everything I have, but I will make most of my time and complete everything that I plan on doing. 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 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 to get a sense of 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. In the end, I do hope to use some of the work from here and submit it as a Kaggle task to see if others like it and to see how I could improve my models and just my thought process in general.