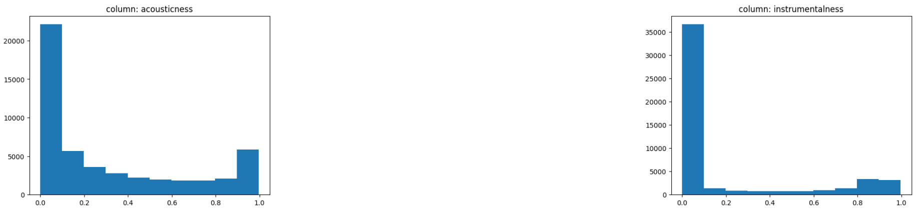
Snh362 Capstone Project Report (CS-UA 473)

snh362

May 2024

1 Data Cleaning and Prep

First what was necessary was a simple investigation of the data for abnormalities and strangeness. The following 2 plots were more interesting distributions of some of the features in the data. Besides this, there were some missing or nan values in tempo, some negative values in a couple features, and some nan row entries for some of the categorical variables. Since I was using a dataset with over 50,000 entries and the abnormalities only took up a small portion of that as well as the abnormalities not being disproportionate to any one group, I felt it was reasonable to remove any entries with them. I used the Standard Scaler to normalize the numeric features and OneHotEncoding to encode for artist name, key, and mode. We also use the LabelEncoder to encode the 10 different categories.



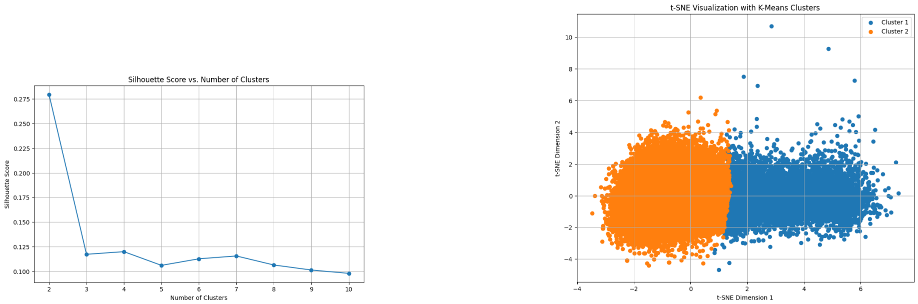
(a) Huge skew here in Acousticness but with a bump at the end

(b) Huge skew here in instrumentalness but with a bump at the end

Figure 1: Some more interestingly skewed data from the features

2 Data Clustering and Dimensionality Reduction

I reduced the over 6500 features due to the OneHotEncoding plus the standard numeric features to just 14 PCA features. Following this I plot the top 2 components from the PCA and find optimal clustering utilizing the silhouette scores, to then apply a k means clustering resulting the the plots below. The total variance explained after dimensionality reduction was approximately 81 percent and the top 2 components alone account for 35 percent of this. The spacing/clustering seems relatively reasonable given the PCA results, it does seem like there is maybe a ”more energetic vibe” on some music vs the ”less energetic vibe” on other music.



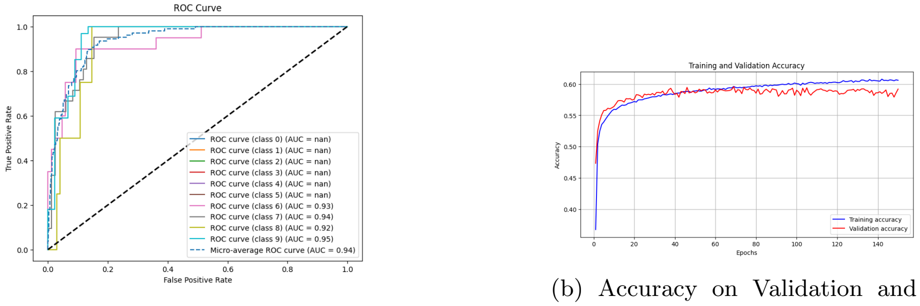
(a) Silhouette Scores vs Number Clusters

(b) t-SNE Visualization with K-Means Clusters in PCA space

Figure 2: Data Clustering plots

3 Modeling, AUC, ROC

Modeling was done on the post PCA/clustering data process and I use a neural network to classify the 10 classes based on the row by row info. The neural network architecture is simply a single layer neural network with 64 nodes followed by 32 nodes, utilizing SGD, relu activation, and a learning rate of 0.001. Training was done on 150 epochs since the model converges at that point. I also tried a XGB classifier but found the neural network to be more successful in classification. With the XGB classifier only achieving 51 percent accuracy compared to around 60 percent accuracy on the neural network.



(a) ROC curve after training

(b) Accuracy on Validation and Training sets

Figure 3: Model metrics plots