Tanuj Sistla

Capstone Project Report

Before we build our classification model, there are a number of cleaning procedures that are to be undertaken. Upon observation of the dataset, it can be noted that there are a number of categorical and non-numerical variables. Additionally, some variables have missing values that are to be handled.

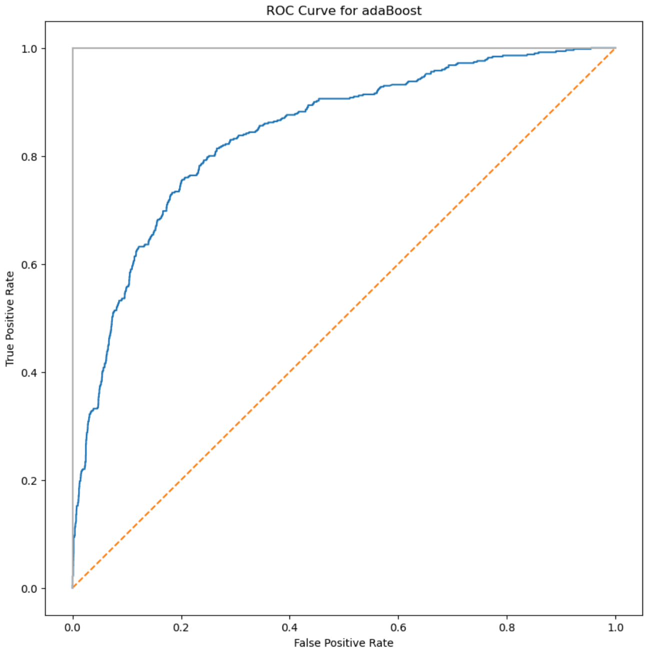
We begin by first saving the dataset within a pandas dataframe, and dropping all NA values. We also remove some variables that would not play a significant role in predicting music genre, such as track name, and date obtained. We also use one-hot encode the 'key' variable to make it numerical, more easily used for classification.

We then ensure that all values in each column fall within their required metrics. Upon handling missing values in the variables tempo and duration_ms, it was initially considered to simply drop the rows with missing values, but we would then not have the required number of rows to perform the train-test split specified in the spec sheet. So, for the song durations, we take the mean value of all song durations, and replace the missing values with this mean value. For the missing tempo values, we take the mean tempo of all songs within that specified genre, and replace the missing tempo value with this mean value specific to the track's genre. Finally, we map the music_genre variable to numbers from 0-9 (inclusive), and with this step, we are finished cleaning the data.

After performing the required train-test split, we use StandardScaler to scale our training and testing sets. Due to the large size of our dataset, we also must perform dimensionality reduction. Initially, I considered using PCA. Upon performing PCA, we find that the first 14 eigenvalues are above 1, which, under the Kaiser criterion, implies that the first 14 principal components are significant.

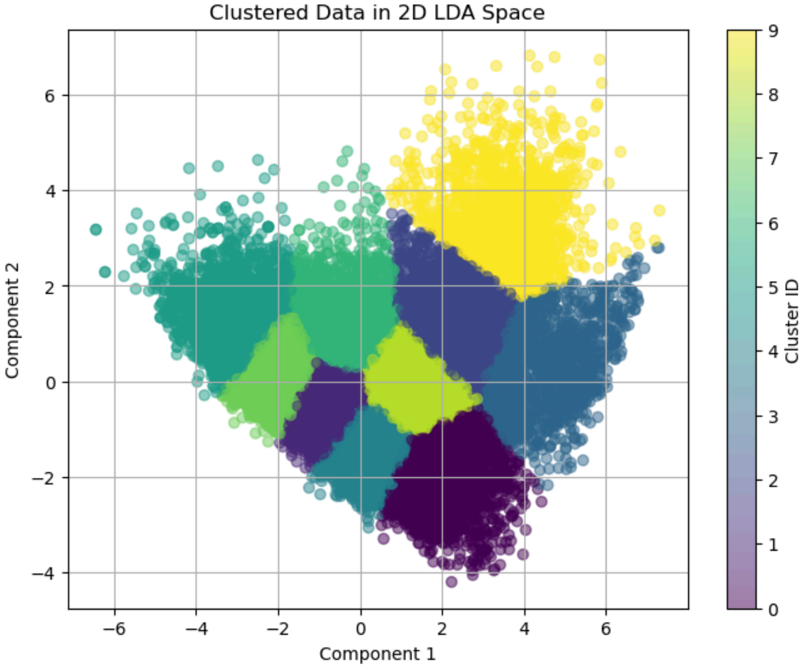
Given the presence of the class labels however, (since we are predicting music_genre), LDA presents itself as a better option than PCA. PCA did however tell us that a significant number of features can be important predictors, so we use the maximum possible number of components we can under Sklearn's LDA function less than or equal to 14, which is 9 (n_components cannot be larger than min(n_features, n_classes - 1)).

Finally, we build our classification models using adaBoost with single decision trees, and achieve an **ROC AUC score of 0.8758469111111111**. Below is the ROC AUC curve corresponding to it:



This ROC score is quite good, especially compared to the score received using PCA and adaboost, which was approximately **0.7398187111111112**. Note that it is possible to get a higher score using other classification models, such as random forest. For this project, we stick to adaBoost.

We then use the KMeans clustering algorithm to represent the data points in clusters of their genres, in the lower-dimensional space, i.e., we use our LDA applied dataset for clustering. Below is the plotted visualization of the clustering by the first two dimensions of our dataset reduced with LDA, with a specified number of 10 clusters.



It can be noted that the different clusters respective to their individual colors are of similar sizes to each other. While they are not exactly the same sizes, we can note with these results that the predictor variables provided in the given dataset are meaningful predictors of each track's music genre. This can also be noted by the ROC AUC score, and the significant components highlighted with PCA.