Machine Learning Capstone Project: Music Genre Classification

Overview

In this report, I will explain the preprocessing steps, dimensionality reduction steps, model building decisions, and the final results. The dataset used in this project contains 18 unique columns. The goal of this project is to predict the music genre from the rest of the data.

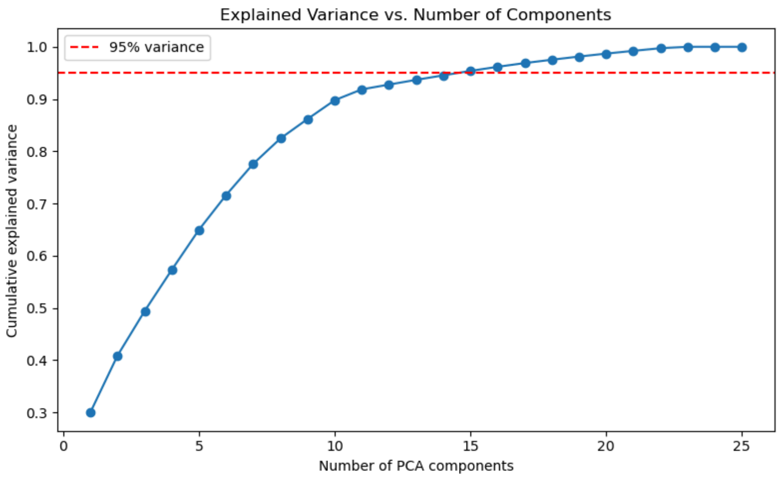
Preprocessing

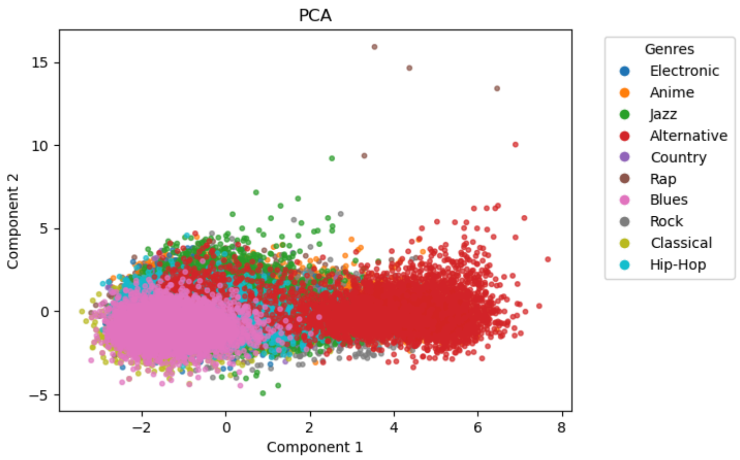
I first initialized the seed of the random number generator with my NYU N-number and then imported the dataset. I removed any null values which removed 5 fully null rows in the dataset that contained no relevant information on songs and was only a small portion of the dataset. Then for columns like 'tempo' containing '?' for some values and 'duration_ms' containing '-1', which don't make sense in the context of the columns, I decided to impute the missing values with the median value of their respective columns. I chose to impute the missing values rather than removing them as they consisted of about 10000 rows which is a big portion of the dataset and we needed to maintain the number of rows to ensure an even distribution of music genres for the train and test split in the future.

Following that, for the categorical columns such as the key and the mode, I one hot encoded those values in the dataset. There were also certain features such as 'instance_id', 'artist_name', 'track_name', and 'obtained_date' which were dropped from the dataset because they were not significant for this music genre classification. I also encoded the music genres to numerical labels to make it easier to identify results in the future. Finally, I normalized the numerical columns, ignoring the one hot encoded values, for the purposes of doing dimensionality reduction.

Dimensionality Reduction and Clustering

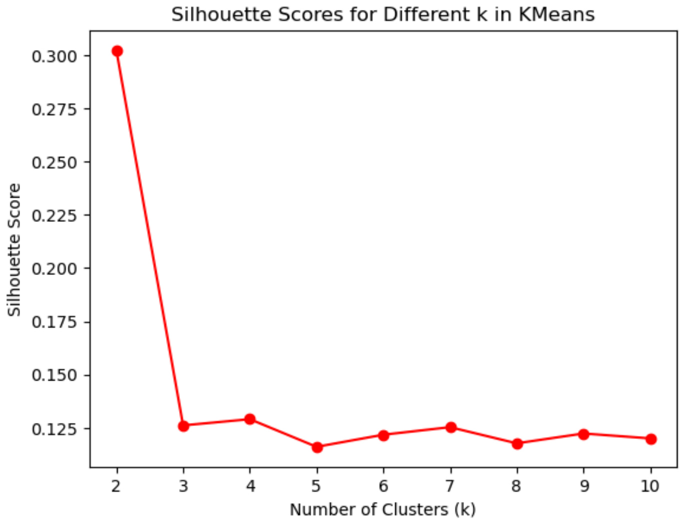
For this assignment I chose to perform PCA on my dataset to better understand our data. To figure out how many components I should use, I chose to find the number of factors that account for 95% of the variance. I found that we needed 15 components to do so and then plotted the scatterplot of the data using the first 2 principal components to visualize our data.

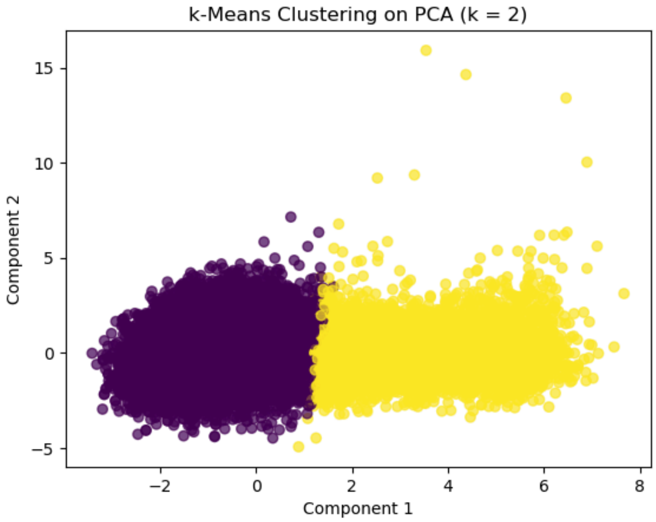




From the given scatterplot, we can see that most of the music genres are very close to one another, making it hard to tell clusters apart, and the only 2 real clusters that we can really make out are blues and alternative.

To further confirm this point, I chose to use kMeans to cluster the reduced dataset by first using the silhouette scores to determine the optimal number of clusters and then plotted the resulting scatterplot with the clusters.





We can see that the optimal number of clusters is 2 and is relatively clear with the scatterplot.

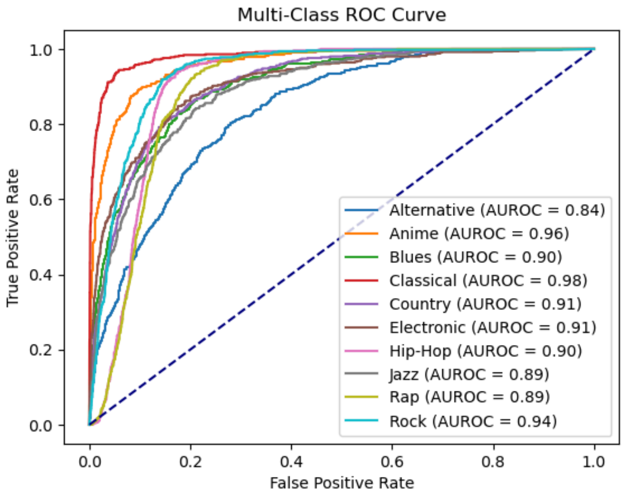
Building the Model

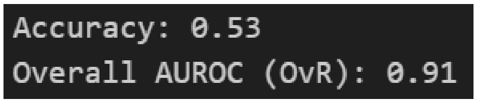
After reducing the dimensions of the dataset with PCA, I split the data into a training and testing set, ensuring that for each genre I use 500 randomly picked songs for the test set and the other 4500 songs for the training set by stratifying with the genres.

The classification model that I chose to use for this problem was a random forest classifier. Firstly, I performed hyperparameter tuning using GridSearchCV to determine the optimal hyperparameters for this model. Then I ran the model with the optimal hyperparameters on the data.

Results

Using the random forest model, I managed to get an accuracy of 53% and an overall AUROC (using a "one versus rest" approach since this is a multi-class classification) of 0.91.

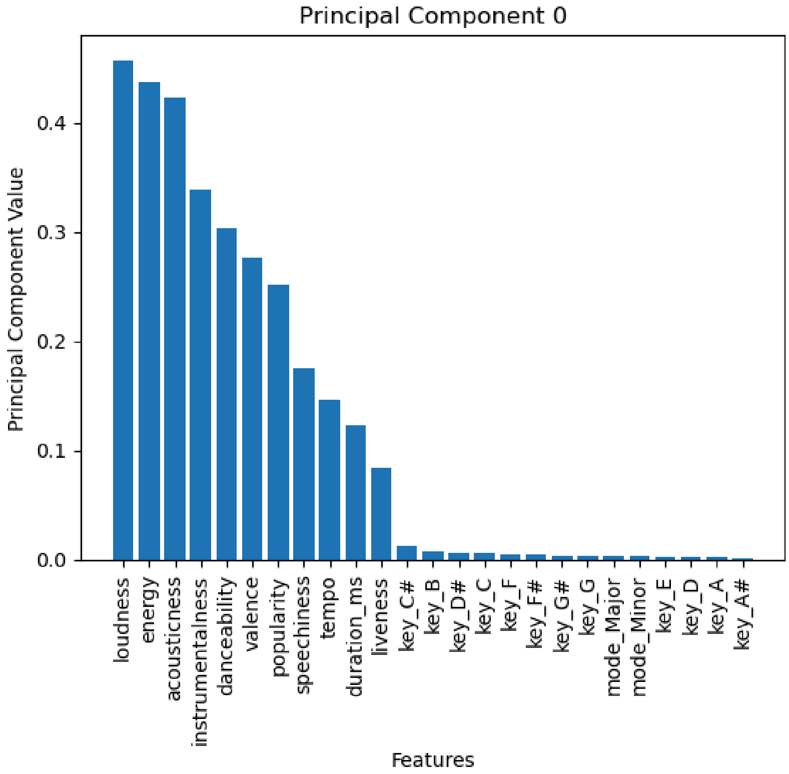




```
Accuracy: 0.53
Overall AUROC (OvR): 0.91
```

Throughout the course of this project, there were various challenges such as the missing data for some columns which was resolved through imputation. Similarly, removing unnecessary columns and one hot encoding the categorical columns seems to have made a positive impact on the performance of this model.

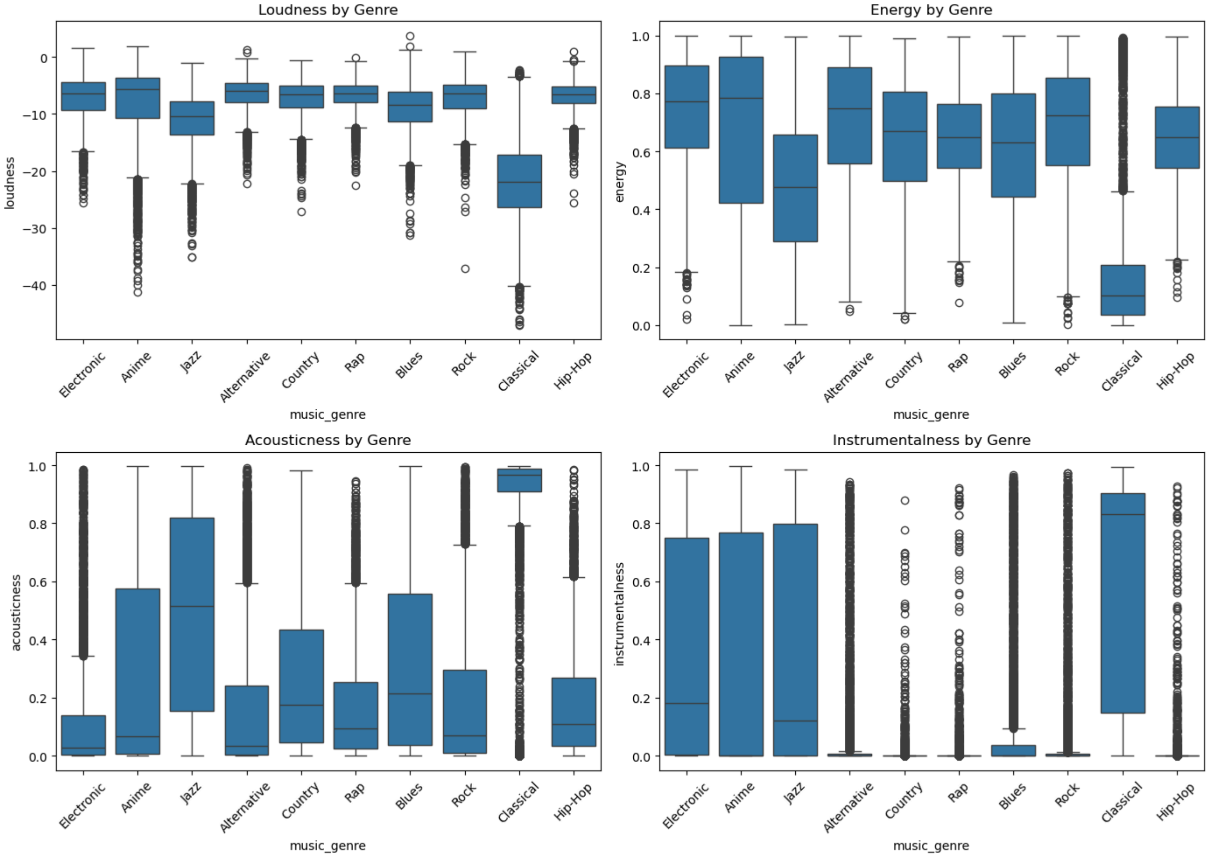
If we were to focus on the most important factors that underlies this classification success, we should look at the first principal component using PCA which has the greatest feature importance according to the model.



It's clear that the first principal component is mainly representing factors like 'loudness', 'energy', 'acousticness', etc. This indicates that these features have a significant impact on our classification of the music genres for each song.

Extra Credit

I chose to plot the 4 most important features of my classification model from above as boxplots for each genre to see if any specific features stand out for each genre. We can see that for things like classical music, the loudness and energy are generally lower than the other genres but it's acousticness is relatively higher. We can also see that the instrumentalness for electronic, anime, jazz, and classical genres are generally higher than the other genres as well.


Loudness by Genre
loudness
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop
music_genre
Energy by Genre
energy
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop
music_genre
Acousticness by Genre
acousticness
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop
music_genre
Instrumentalness by Genre
instrumentalness
Electronic
Anime
Jazz
Alternative
Country
Rap
Blues
Rock
Classical
Hip-Hop
music_genre