Luca Ignatescu

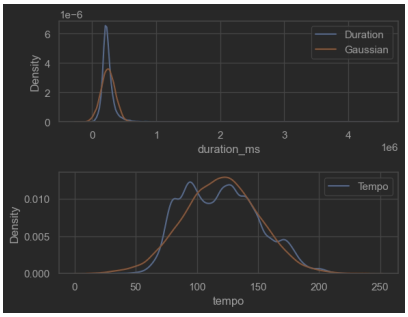
Capstone Project

Classification

Preprocessing

The first step performed was a preliminary selection of features. Firstly, instance_id was dropped, as it is just an arbitrary, Spotify specific feature. artist_name and track_name were also dropped, as using them in a meaningful way would require NLP techniques not covered in the class. Lastly, obtained_date was also removed because, although there might be correlations between obtaining period and genre, working with time series data was also not covered in the class.

The second concern to be addressed were the missing values. After analyzing the data, it was observed that the only rows with relevant missing values were duration and tempo. In total, 9440 rows were affected, roughly 20% of the dataset. Although a large percentage, I chose removal over imputation. This is because neither of these values are normally distributed. Below are the plots of the densities of these columns, on which gaussians with the same mean and variance are superimposed. Imputation risks a severe reduction in variance, particularly in the case of tempo, which has a very large standard deviation. As such, it was not used. In terms of duplicate values, only 8 in total were found, so they were kept, being statistically insignificant. There were also four empty rows in the middle of the datasheet, which were discarded.

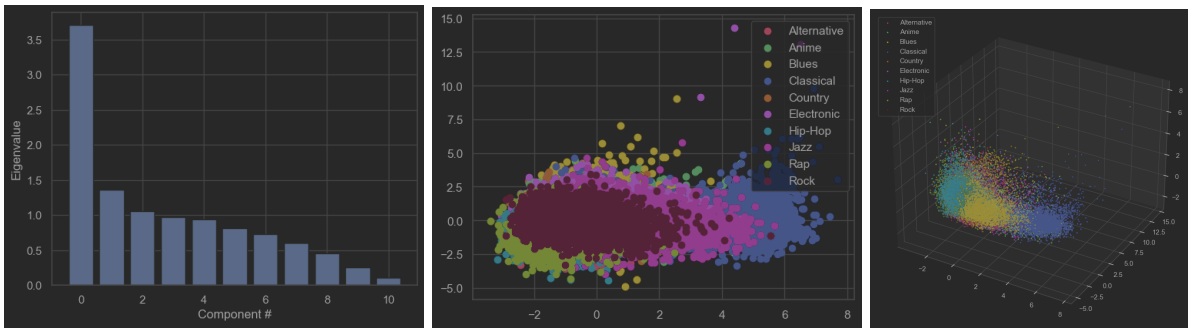


The next step was performing the specified split. Sampling was done without replacement, and whatever was sampled for testing was dropped from the training data, to prevent leakage. The upcoming visualizations are of the training set, as the testing set is meant to simulate real data and, as such, all mappings should be learned on the training set.

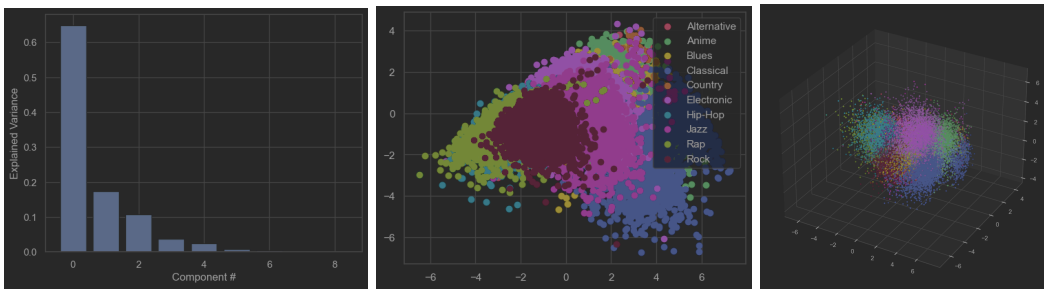
Dimensionality reduction was then performed. As the algorithms covered in class deal with continuous data, categorical variables were removed for this step, and then re-added to the model after the reduction. The chosen algorithms were PCA, for its well known reliability, as well as LDA, due to the fact that it's a technique conscious of class labels. Another reason is that both these algorithms learn a map between the higher and lower dimensional spaces, and, as such, can be used for prediction (as opposed to something like tSNE). Due to these choices, the data was z-scored before, as PCA requires centered data, and LDA is not affected by such rescalings, providing an additional reason for the removal of categorical variables before this process, as z-scoring categorical data is meaningless and possibly dangerous due to precision errors.

For PCA, the explained variances are plotted below**.** Considering the ratio of variances, I attempted both 2D and 3D PCA, as they capture 46% and 55% of the variance respectively. Considering the complex nature of music, this value is reasonable. The plots are also shown below. In the 2D case, the clusters overlap a lot, especially on the left side. As such, 3D PCA was attempted, with somewhat better results. Classical music is

now better separated (shown in blue), yet the rest of the clusters are still hard to differentiate, especially on the center-left, and a fuzzy mix of data points is present towards the middle of the picture.

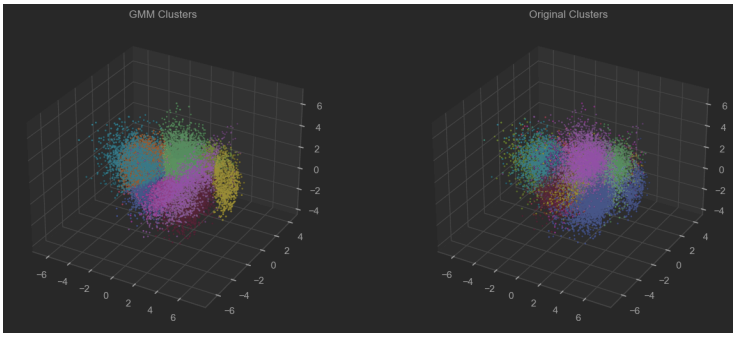


Considering this, LDA was also attempted. Below are the explained variances, as well as the 2D and 3D plots. In, 2D plot does not offer enough separation, but in 3D, the clusters are a lot better separated. Classical, Anime, Jazz, and Alternative are all very well distinguished. Furthermore, Jazz and Hip Hop overlap a lot and are separated from the rest, which is a promising observation considering the close relationship between the genres. When it comes to variance, the 3 components captured 92% of the variance. As such, LDA was chosen over PCA as the final dimensionality reduction technique, as it performs better at capturing the variance and separating the classes even though the data is not normally distributed (fact shown during the missing values discussion).



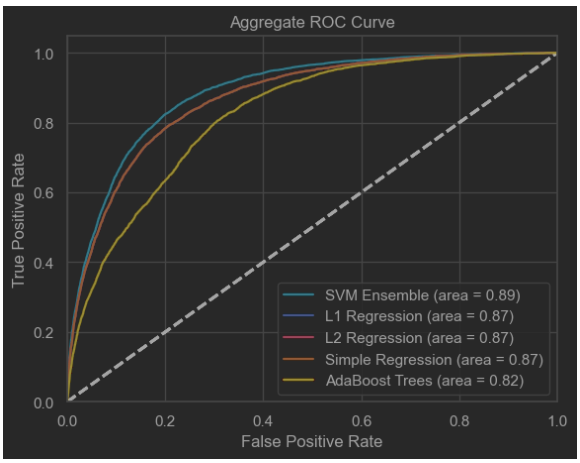
For the clustering method, Gaussian Mixture was chosen for two main reasons. Number one, the strong overlap between classes makes a hard clustering algorithm such as kMeans suboptimal. Furthermore, a visual inspection of the data shows some elliptical clusters, as opposed to fully spherical ones, further suggesting the use of something like Mixture Models, which can yield varied shapes. Lastly, GMM was chosen over DBSCAN due to the computational complexity of DBSCAN, the fact that we already know the number of clusters, and because outliers are not a severe problem, as can be seen visually.

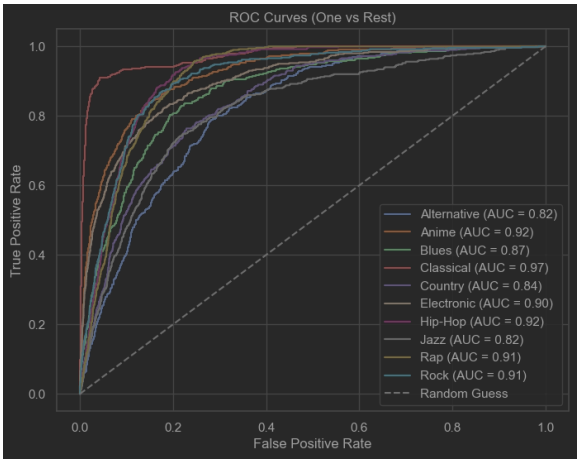
Below is the plot of the mixture model, compared with the real clusters (colors do not match because GMM is not aware of true labels). Visually, the model performed very well. It correctly identified part of the blue cluster (now red), as well as the superimposed rap and hip hop genres (in blue and orange in the left plot). It also correctly identified the fact that there are multiple overlapping clusters in the middle, such as the blue, purple and pink, although it couldn't tell that the original blue cluster had two parts. The yellow cluster also matches the original green cluster well. We can conclude that the clustering, and therefore reduction, was effective.



Model Fitting

A number of models were attempted. Firstly, simple logistic regression, with no cross validation, was attempted to establish a baseline performance, due to its rather simple nature. Next, L1 and L2 regularized logistic regressions were performed, with different regularization strengths, and 10-fold cross validation. Boosted trees were also used, specifically with AdaBoost and 5-fold cross validation, due to the nonlinear nature of some cluster separations as well as the presence of categorical data. Lastly, bagged SVM ensembles were attempted, with different numbers of estimators. These were preferred over simple SVM for training purposes, as SVM fitting is quadratic in the number of samples. The best performing models of each category are plotted below, with their aggregate ROC curves.





As shown in the plot, the best overall model was the SVM ensemble. Below are all the ROC plots for each class. The model performed really well on classical music, which was to be expected from the good separation of that cluster in the preprocessing step. Furthermore, the model was surprisingly adept at distinguishing Rap and Hip Hop, perhaps due to the random sampling involved in a bagging ensemble, which could also explain part of its superior performance over the other models. Furthemore, the ability of SVMs to deal with outliers might have also proven useful in such a large and dense data set. The model struggled the most with Alternative, Country and Jazz, perhaps because those classes were mostly clustered in the middle, as shown in the plot.

As such, it's likely that effective dimensionality reduction combined with an ability to deal with overlap and outliers are the most important factors when it comes to this classification, and a mix of bagging techniques with SVMs seems to tackle this well. The final AUC was **0.89.**