Elementary Preprocessing and Data Cleaning:

Because there are string formatted features, I need to transform them into numerical values. For “key,” I dummy encoded the data because there is no inherent order in keys (within an octave, yes, but no information about octaves is given). For “mode,” I transformed it to binary variable “isMajor” because in mode, each song is either major or minor.

Before cleaning data, I examined the distributions of the features so that I can decide what to do with missing values. I found that “duration_ms” and “tempo” had 4939, and 4980 missing values, respectively (about 10% of the data). To preserve as much data as possible, I decided to impute for those missing values. Both duration and tempo are skewed (especially duration), so I used median to impute.

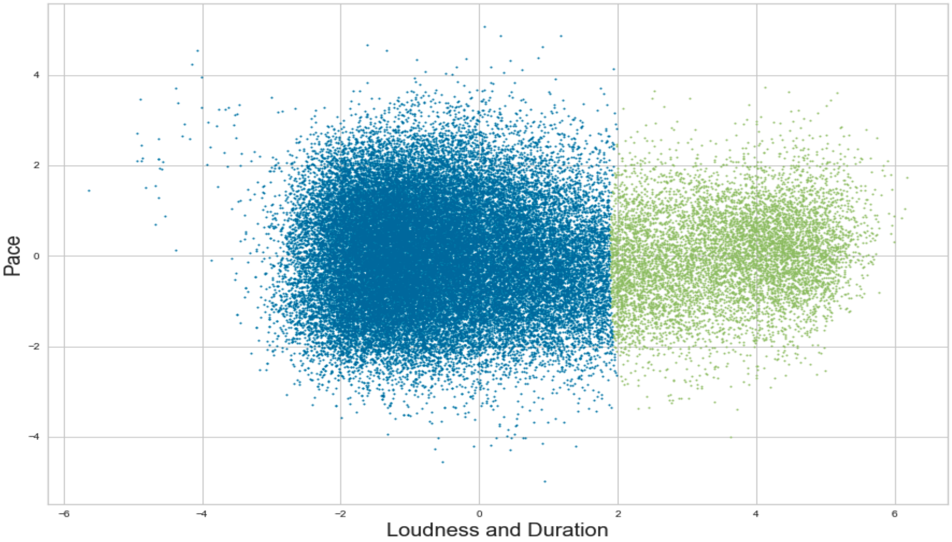
Data Transformation:

Because many features have serious skewness, namely acousticness, duration, instrumentalness, liveness, loudness and speechiness, they should be transformed. I used log transform for duration, speechiness, and liveness; sqrt transform for acousticness, cube root transform for instrumentalness, and loudness. I did those because I found those methods to be the best at moderating skewness.

PCA and Clustering:

Because PCA is meaningless for categorical data, I excluded them from PCA. Before PCA, I Z-Scored the rest of the data. To very briefly summarize the result, the first two PCs explain 45.36% of the variance, and the first 8 PCs explain over 90% of the variance. PC0 can be interpreted as loudness and duration. PC1 can be interpreted as pace (PC loadings see 3rd page). To cluster the 2 PCs, using the Silhouette method, I found that there are 2 clusters. Perhaps we expect 10 clusters, but because 2 PCs does not explain that much variance in data, in higher dimensions there might be 10 clusters, which unfortunately is impossible to visualize. I tried KMeans, but it failed to separate the clusters (see code):

Gaussian Mixture provided better clusters (on the right):



The clusters are separated fairly well. It seems like some genres have moderate loudness and duration, while others have higher. The pace for the clusters are fairly similar, although the green cluster has less variation.

To use PCA and clustering in order to enhance the models, I made 3 new datasets: one with 8 PCs combined with original categorical variables (DS1), one with original data combined with cluster probabilities generated by GM (DS2), and one with 8PCs combined with cluster probabilities (DS3). I choose to use cluster probabilities instead of labels because GM is soft clustering, meaning that there are points that can "belong" to both clusters, and the probabilities can be used as "weights" of clusters.

Baseline Model Performance:

First, I build a perceptron model to check the data's linearly separability. If the data is linearly separable, then the perceptron trained on the entire dataset should be able to predict the dataset with a high accuracy. However, this is not the case. The accuracy of perceptron was 0.33, even when I used the same data for training and testing. Therefore, non-linear models are more suitable for this classification task. The models I chose to experiment with are: random forest, decision tree, and adaboost. The micro-averaged AUCs: 0.923, 0.705, 0.8438.

Model Trained on New Datasets:

DS1 significantly decreased the model performances:. AUC dropped to: 0.893, 0.661, and 0.821. DS2 increased model AUCs:: 0.9249, 0.7087, 0.8447. DS3 decreased model performance, but it's better than DS1. AUCs: 0.921, 0.703, 0.844.

Because the aim is to increase AUC as much as possible, I will move forward with DS2 using Random Forest and conduct hyperparameter tuning. The reason why random forest performance worse with reduced dimensions is perhaps the algorithm itself performs cross validation, which can reduce overfitting, and reducing dimensions reduces the amount of information in data, inevitably.

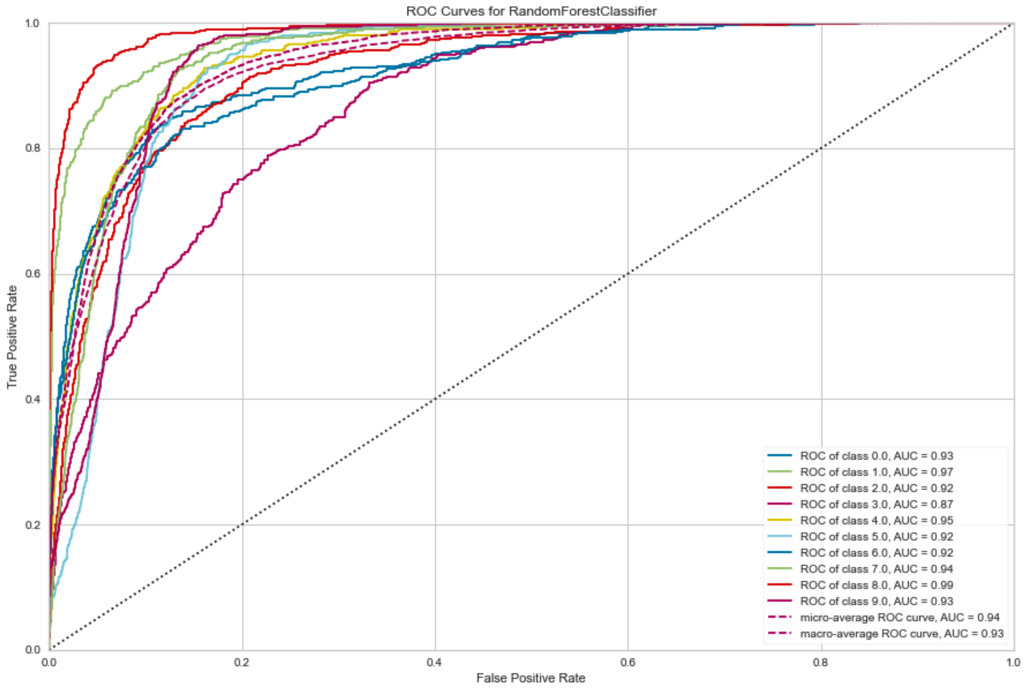
Hyperparameter Tuning for Random Forest:

Through many trials and errors, I found that the best hyperparameters are:

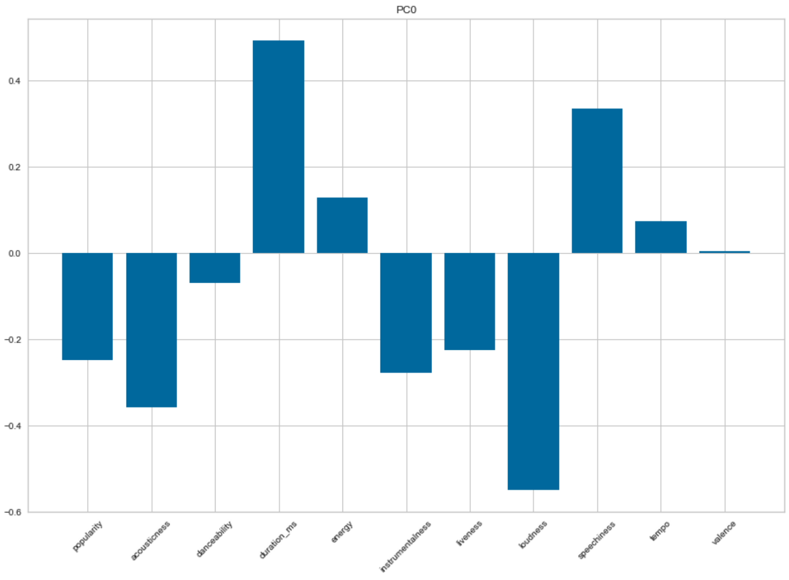
```
n_estimators = 200, max_depth=16, max_features= 0.4,n_jobs = -1,
min_samples_split = 14, criterion='gini'
```

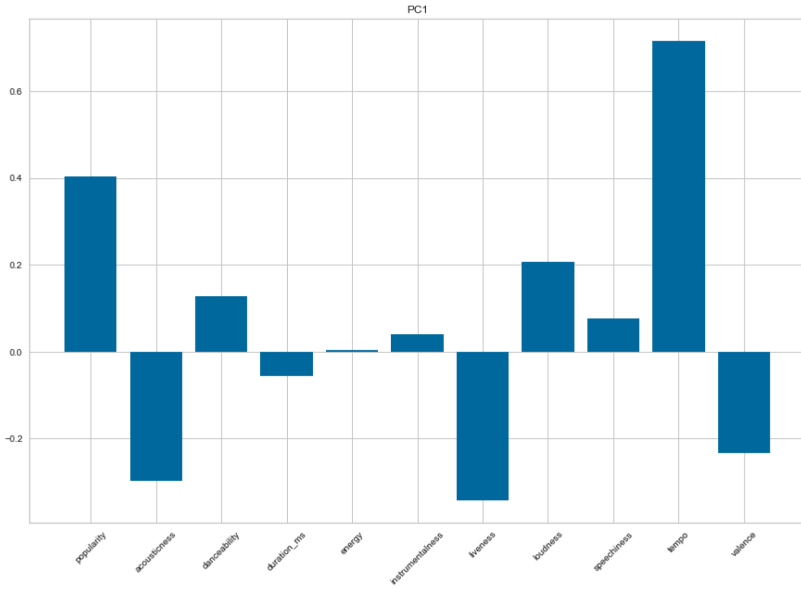
The AUC of this model is *0.9345*, an improvement compared to the baseline model AUC of 0.923.

The final model ROC curve:

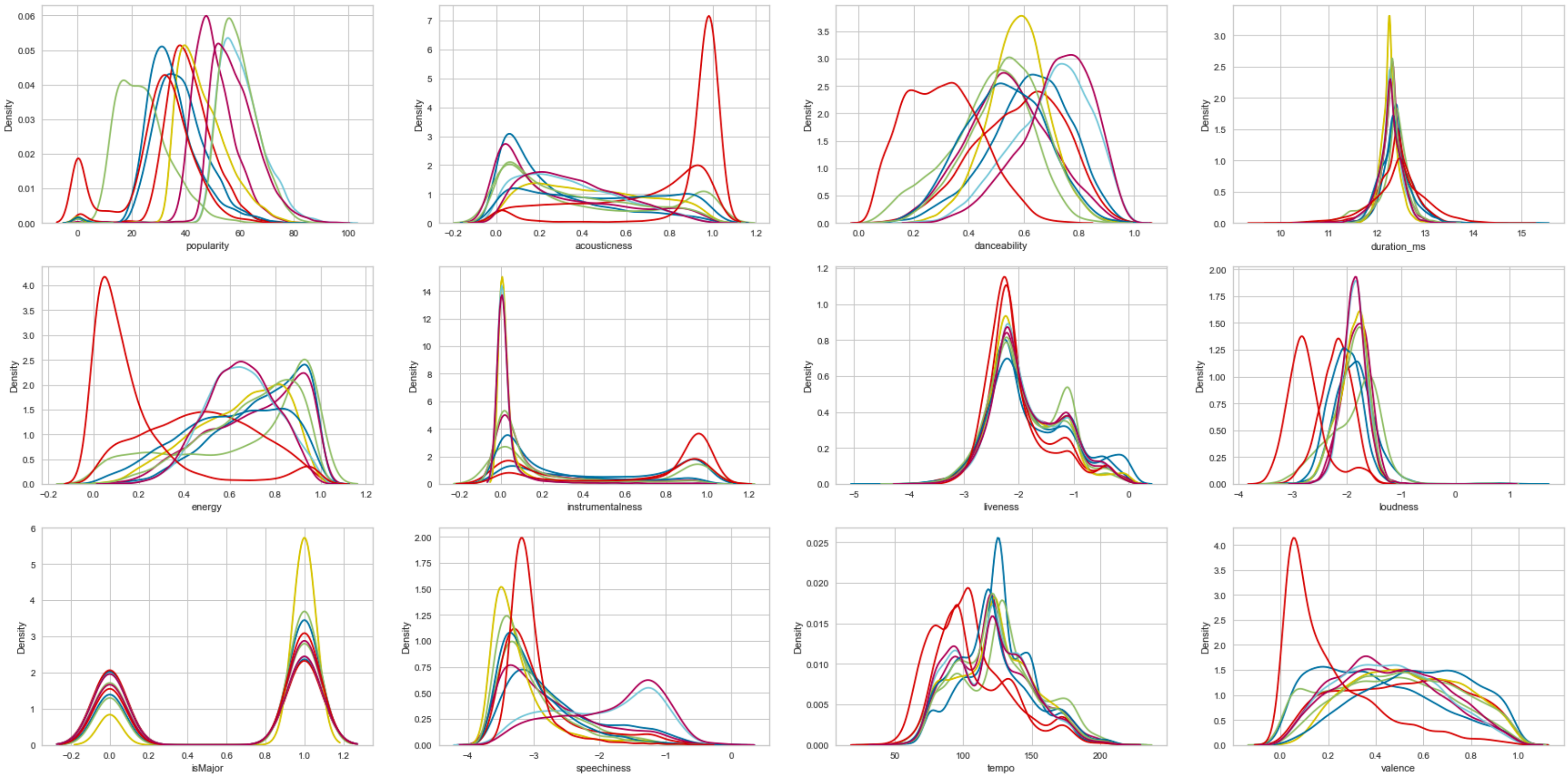


PC Loadings:





Extra Credit



Above is the distribution of the features of different genres. Distributions of duration and liveness for all genres are highly similar, while popularity of the genres are very different: different means and spread.

Perhaps the genre that differs from others the most is classicals (red line). It has the highest, on average, acounsticness, and lowest valence, energy, and speechiness.