Unce Shahid
Pascal Wallisch
Principles of Data Science
12/19/23

Capstone Project: Analysis of 52,000 Songs

Preprocessing:

- Row-Wise Removal of Duplicate Tracks
    - This is an instance of pseudoreplication. Some of the data points are not independent, and so they were removed. For example, songNumbers *8052, 8053, 8055, 8056, 8057, 8061, 8062, 8063, 8064, and 8065* all have the same exact characteristics, with the only difference being that they belong to different albums. The method of removal used is row-wise exclusion of replicants with repeated track names.
- Row-Wise Removal of Track Remixes
    - Some tracks have multiple versions present within the dataset. The different versions share similar characteristics to each other in some aspects, but also differ in other aspects. Ultimately I decided to exclude multiple versions because these data points are only partially-independent from each other. The method of removal was including the first version of a track and excluding the rest.
    - The sample size was reduced from 52,000 down to 38,512 after cleaning.
- Dimension Reduction
    - For the questions where dimension reduction is beneficial, I normalized the data via z-scoring and performed a principal components analysis, making use of the kaiser criterion to cut off eigenvalues below 1.

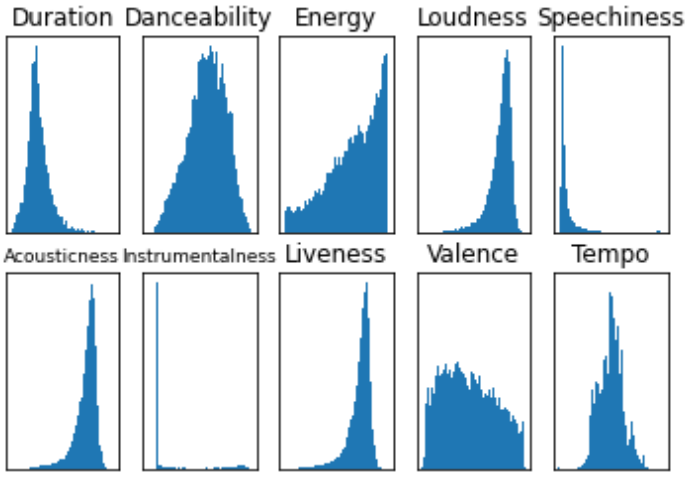


Figure 1*: Distributions of the 10 characteristics being analyzed.*

Question 1: Consider the 10 song features duration, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence and tempo. Are any of these features reasonably distributed normally? If so, which one?

Duration has a right-skewed distribution. Duration has a highly concentrated center, with one horizontally-lengthy tail (similar to a cauchy tail), and one short, compact tail (similar to a gaussian tail). The right tail is a lot longer, so this is a right-skewed distribution. A normal distribution has little to no skew, so duration is not normally distributed.

Loudness, acousticness, and liveness all have left-skewed distributions. These 3 distributions are similar to the duration distribution, except instead of being right-skewed, they are left-skewed. Their left tails are a lot longer than their right tails, so they are not normally distributed.

Speechiness has a distribution much like the exponential distribution. Speechiness has a very tall peak at a value ~ .04, with a descent towards a value of ~.40 that is very reminiscent of an exponential distribution. Speechiness also acclimates very quickly from the start of 0 to .04, making it appear even more like an exponential distribution. Speechiness does not have a normal distribution. Interestingly, the distribution has close to 0 data points after a value of ~.40, but a resurgence starts to appear at around a value of ~.90.

Instrumentalness does not have a normal distribution. There is an extremely tall peak at a value of 0, meaning most songs are not instrumental at all. After the peak, there's a small spread of instrumentalness ranging all the way from the minimum (0) to the maximum (1). Instrumentalness does not have a normal distribution.

Energy has a positively-sloping monotonic distribution. Songs are more likely to have higher energy, than lower energy when compared at any energy level. The distribution is not monotonic for every specific level of energy, but that can be accounted for by variance. Energy does not have a normal distribution.

Valence is somewhat uniformly distributed. Valence is the most uniformly distributed out of any of these 10 distributions. Except for the small tails, the highest difference between any two levels of valence would be around double. What this means is that songs make use of all measures of valence. Valence is not normally distributed.

Tempo does not have a gaussian distribution. There are many valleys in the tempo distribution. These valleys likely mean that when songs are being produced, people might experiment more with gaps (.5 to .6 instead of .55) in tempo. The tails are similar to that of a normal distribution, however the middle is not distributed like how a normal distribution would be. Tempo does not have a normal distribution.

Danceability is the closest to having a normal distribution. The danceability distribution is almost symmetrical. The left half of the distribution is pretty similar to the right half, however, it still is not a normal distribution.

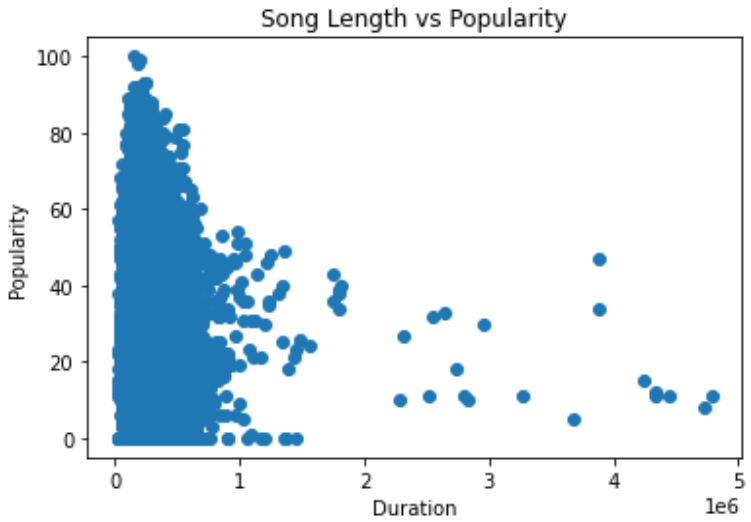


Figure 2: *Scatterplot of duration vs popularity.*

Question 2: Is there a relationship between song length and popularity of a song? If so, is the relationship positive or negative?

Song length and popularity have no correlation with each other. Just from the scatterplot alone, it is evident that there is no correlation between song length and popularity as there is no interpretable slope. I calculated the Pearson correlation coefficient between duration and popularity and obtained the value -0.097. This value is very close to 0, so there is a large possibility that there is no correlation at all. However, looking at the general trend of the scatterplot, it is likely that there is just a very weak negative correlation.

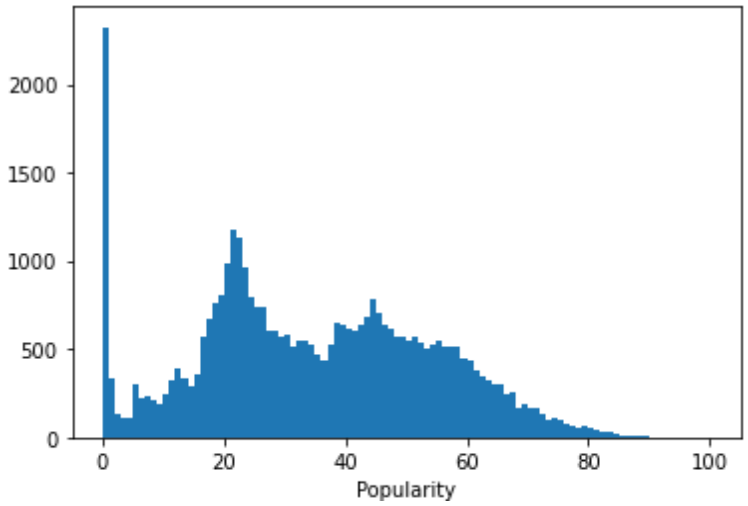


Figure 3: *The distribution of popularity.*

Question 3: Are explicitly rated songs more popular than songs that are not explicit?
Explicit songs are indeed rated as more popular compared to non-explicit songs (with statistical significance). The first decision I had to make was determining which type of significance test to use. Above is the distribution of popularity, where you can see there are a very large amount of replicants with 0 popularity. Since 0 is an extreme value, mean-based significance tests would not be recommended for use. Therefore I chose to use the Mann-Whitney U test, a nonparametric significance test. The Mann-Whitney U test provided a p-value of 4.23×10^{-21}, which is statistically significant at even the 99% level. There is a large size gap between the explicit sample size (4,375) and the non-explicit sample size (34,137), however I performed a power calculation via gPower that resulted in a power of 1. Therefore, I can say with confidence that explicit songs are rated more popular than non-explicit songs.

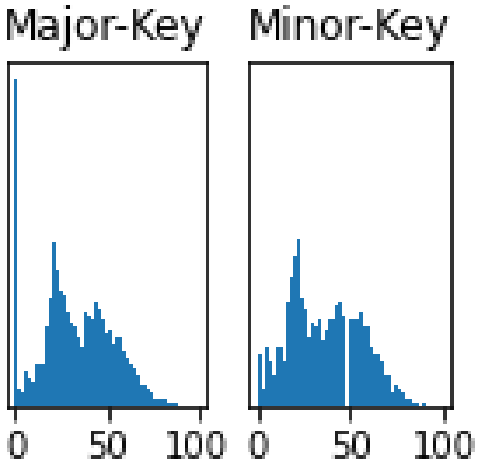


Figure 4: *Distributions of popularity for major-key songs and minor-key songs.*

Question 4: Are songs in major key more popular than songs in minor key?
Songs in minor-key have higher popularity than songs in major-key. Notably, from Figure 4, major key has so many more songs with a popularity of 0. The major-key sample size is 24,043 and the minor-key sample size is 14,469. As both sample sizes are large, power for the following significance test is high. Just like in question 3, I ran a Mann-Whitney U significance test and obtained a p-value of .028. This p-value is statistically significant at the 95% level, therefore minor-key songs have higher popularity than major-key songs.

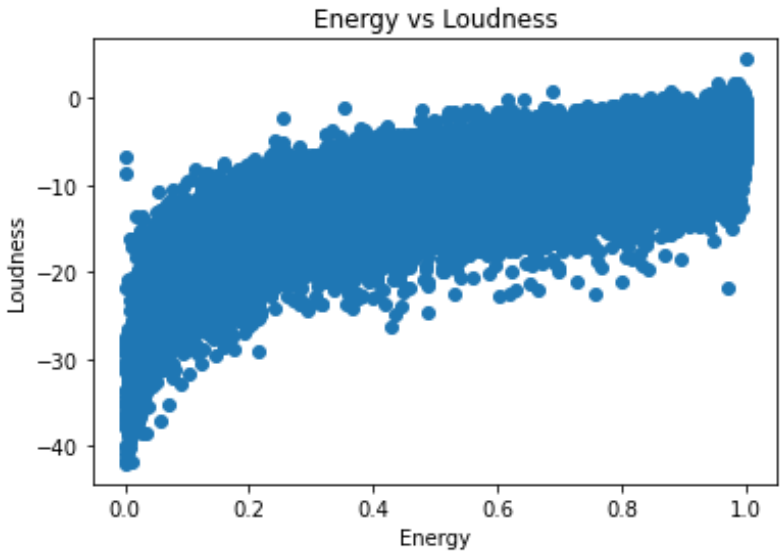


Figure 5: *Scatterplot of energy vs loudness.*

Question 5: Energy is believed to largely reflect the "loudness" of a song. Can you substantiate (or refute) that this is the case?
Energy does largely reflect the loudness of a song. The scatterplot above shows a positive trend, where higher energy is associated with greater loudness. To find out how well energy reflects loudness, I calculated the Pearson correlation coefficient between energy and loudness, which was .78. A coefficient of .78 is pretty large, so there definitely is a sizable positive association between energy and loudness.

popularity	1
instrumentalness	0.04514
energy	0.00955
duration	0.00939
speechiness	0.00586
liveness	0.00534
danceability	0.00519
loudness	0.00466
acousticness	0.00297
tempo	0.0007
valence	2e-05

Figure 6A: *R^2 values for each predictor.*

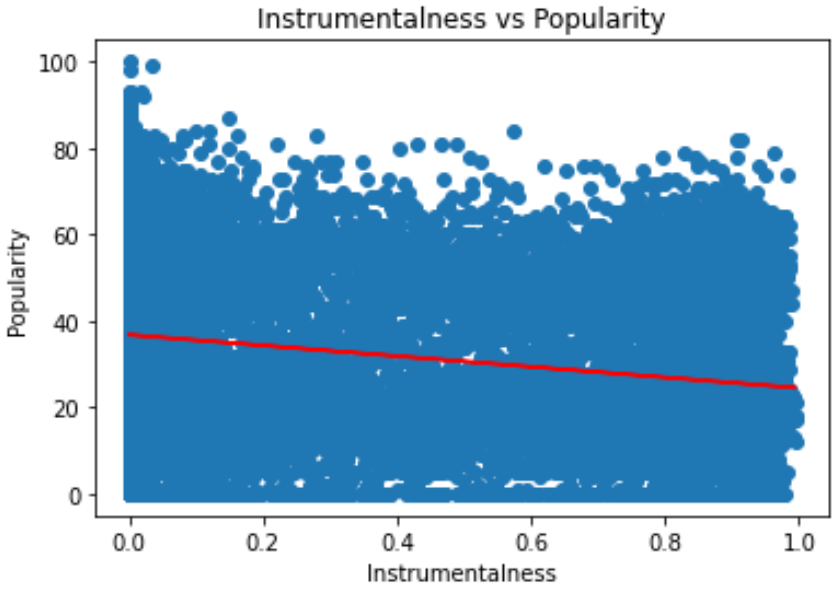
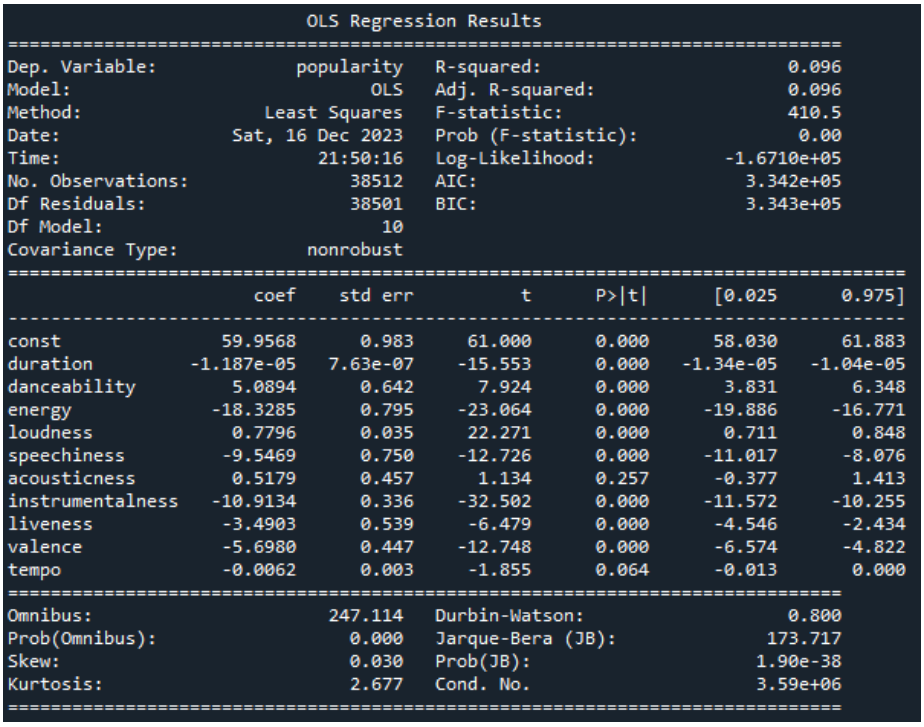


Figure 6B*: Scatterplot with a red line predictor.*

Question 6: Which of the 10 song features in question 1 predicts popularity best? How good is this model?

Out of all 10 of these features, instrumentalism is the best predictor. The coefficient of determination (COD, also known as R^2) is a commonly used measure to determine how good of a predictor a model is. The COD is a measurement of how much variance in the outcome is explained by the predictors. Figure 6A shows the COD values for each of the 10 predictors, where instrumentalism has the highest value of R^2. This would make instrumentalness the best predictor out of these 10 features. However, the R^2 value of the model is just .045, which is really really low. As seen in Figure 6B, the predicting capabilities of instrumentalness is very bad, despite being the best predictor of the bunch.



OLS Regression Results

Dep. Variable:	popularity	R-squared:	0.096
Model:	OLS	Adj. R-squared:	0.096
Method:	Least Squares	F-statistic:	410.5
Date:	Sat, 16 Dec 2023	Prob (F-statistic):	0.00
Time:	21:50:16	Log-Likelihood:	-1.6710e+05
No. Observations:	38512	AIC:	3.342e+05
Df Residuals:	38501	BIC:	3.343e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	59.9568	0.983	61.000	0.000	58.030	61.883
duration	-1.187e-05	7.63e-07	-15.553	0.000	-1.34e-05	-1.04e-05
danceability	5.0894	0.642	7.924	0.000	3.831	6.348
energy	-18.3285	0.795	-23.064	0.000	-19.886	-16.771
loudness	0.7796	0.035	22.271	0.000	0.711	0.848
speechiness	-9.5469	0.750	-12.726	0.000	-11.017	-8.076
acousticness	0.5179	0.457	1.134	0.257	-0.377	1.413
instrumentalness	-10.9134	0.336	-32.502	0.000	-11.572	-10.255
liveness	-3.4903	0.539	-6.479	0.000	-4.546	-2.434
valence	-5.6980	0.447	-12.748	0.000	-6.574	-4.822
tempo	-0.0062	0.003	-1.855	0.064	-0.013	0.000

Omnibus:	247.114	Durbin-Watson:	0.800
Prob(Omnibus):	0.000	Jarque-Bera (JB):	173.717
Skew:	0.030	Prob(JB):	1.90e-38
Kurtosis:	2.677	Cond. No.	3.59e+06

Figure 7: *Results of the 10 factor model.*

Question 7: Building a model that uses *all* of the song features in question 1, how well can you predict popularity? How much (if at all) is this model improved compared to the model in question 6. How do you account for this?

I built a model with all 10 features as predictors via multiple regression. The R^2 of the model is .096, which is higher than the R^2 for the model in question 6 (R^2 of .045). This new model using all 10 features is better than the previous model of 1 feature by a factor of 2x. However, despite having double the R^2, I don't think this model is any good at predicting the popularity of songs. A COD of .096 means my model can only explain 9.6% of the variance in population, which is incredibly poor for a model with 10 features. There are definitely overfitting issues with this model (more features = higher R^2), so it's still nowhere close to being a good predictor of popularity. The higher COD can be accounted for by collinearity issues.

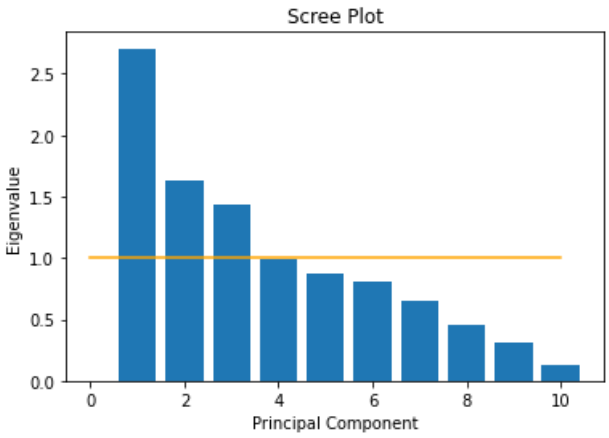


Figure 8A: *Scree plot with orange kaiser criterion line.*

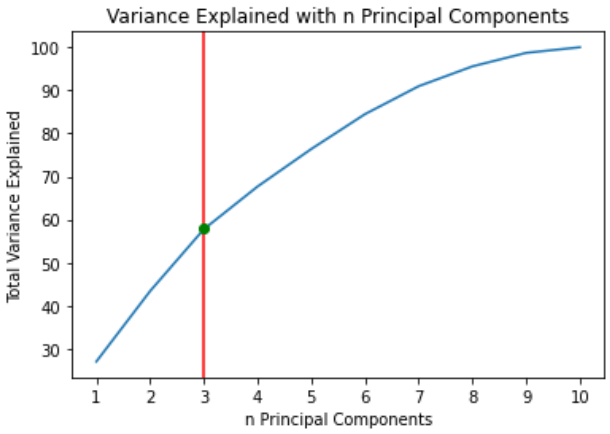


Figure 8B: *Variance explained with n components plot.*

Question 8: When considering the 10 song features above, how many meaningful principal components can you extract? What proportion of the variance do these principal components account for? Using the principal components, how many clusters can you identify?

I found there to be 3 meaningful principal components. First, I created a dataframe with the 10 features being analyzed and then z-scored them. Following that, I ran a principal components analysis, extracted the eigenvalues and eigenvectors, and rotated the data into the new orthogonal vector space. Using the results from the PCA, I created a Scree plot and a variance explained plot. I chose to use the Kaiser criterion, which has a cutoff at an eigenvalue of 1. This criterion follows the belief that each meaningful principal component should justify their inclusion. That is, they must add more value than the cost of inclusion (an eigenvalue >1). The Kaiser criterion gave me 3 total meaningfully principal components, which altogether explains 57.8% total variance. Performing loading analysis for the principal components, I found that there are 4 clusters. The first cluster is made up of energy, loudness, and acousticness. The second cluster consists of danceability, valence, and instrumentalness. The third cluster is formed from speechiness and liveness. Lastly, the fourth cluster is made up of duration and tempo. The fourth cluster falls just short below the Kaiser criterion with an eigenvalue of .991.

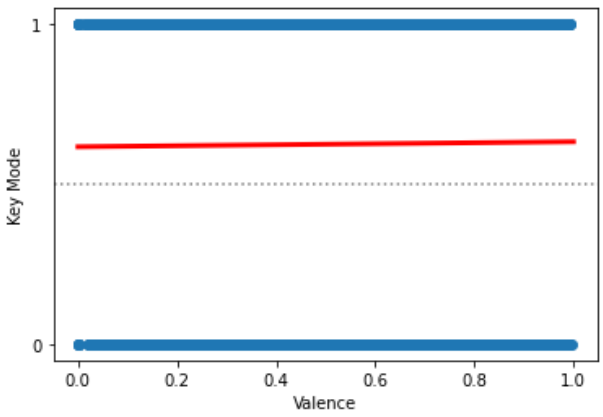


Figure 9A: *Graph of logistic regression model with valence.*

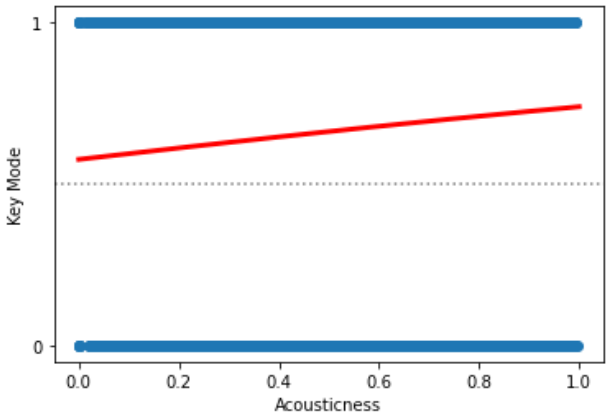


Figure 9B: *Graph of logistic regression model with acousticness.*

Question 9: Can you predict whether a song is in major or minor key from valence? If so, how good is this prediction? If not, is there a better predictor?

Valence is a terrible predictor of the key a song will be in. The blue dots in Figure 9A represent the scatter plot of valence. Every level of valence has songs in both minor-key and major-key, so there isn't a good way to differentiate between keys by using valence as a predictor. The logistic regression model just resulted in giving a flat 62.8% chance that a song with any valence will end up being in major-key, which just means there are more major-key songs than minor-key songs. This is not a useful model. I ran logistic regression for all 15 numeric features, they were all not that good. Every model suffered from the same issue of having both modes of key at each level, so it was difficult to distinguish between which key to predict. However, I found acousticness to be the best predictor of key. As seen in Figure 9B, acousticness actually has a positive slope, so you can glean some information that was not available in most other predictors. When acousticness is at a higher level (there are more acoustics in the song), then the song is more likely to be in major-key.

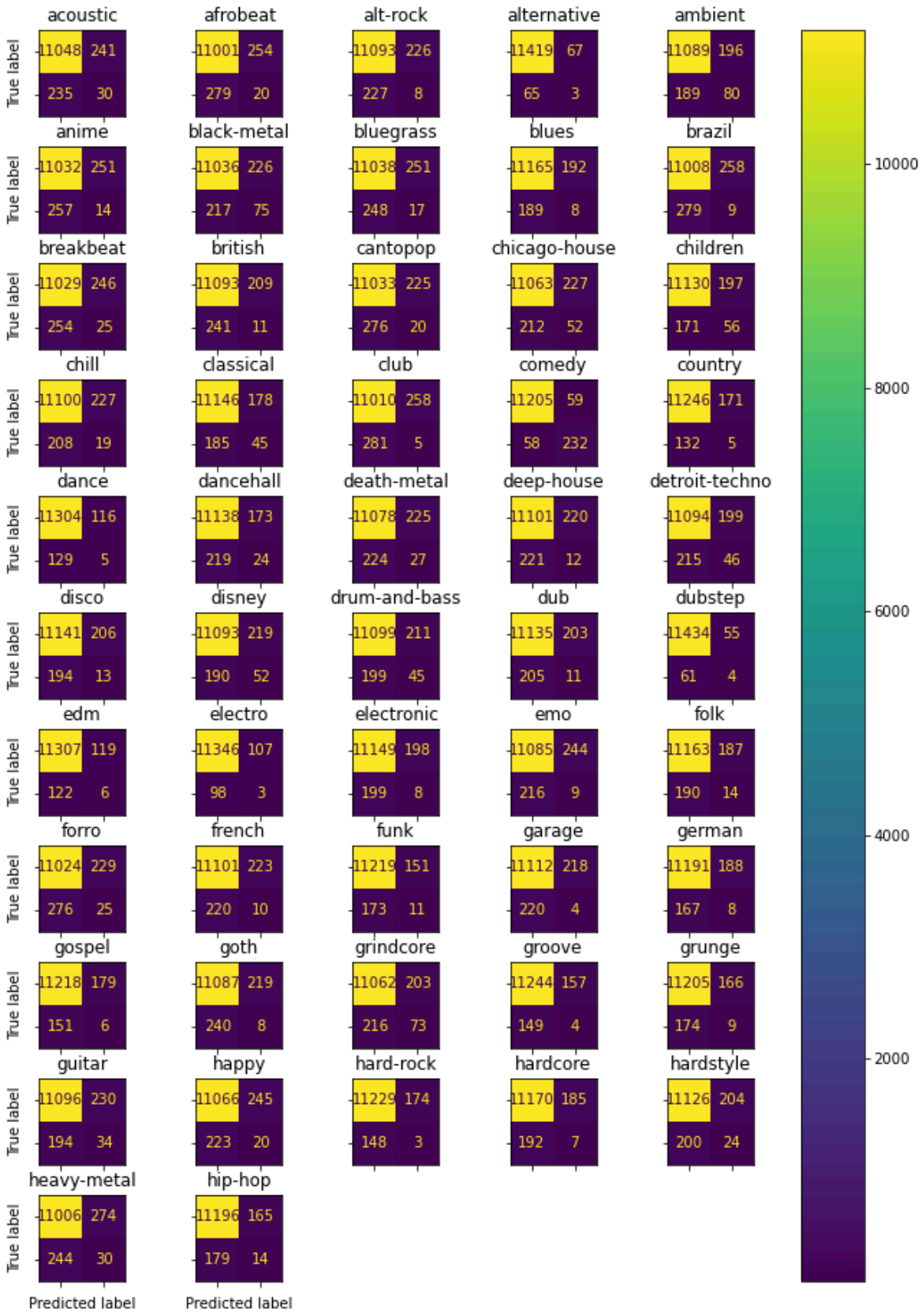


Figure 10: *A multilabel confusion matrix for each of the 52 genre labels.*

Question 10: Can you predict the genre, either from the 10 song features from question 1 directly or the principal components you extracted in question 8?

My model can not predict genre labels. I converted genre titles from strings into numeric labels and created a multi-label classification tree based on the gini index, with the 3 meaningful principal components collected earlier as predictors. Additionally, I split the data into 70% training and 30% testing sets. Figure 10 above shows the confusion matrix for each genre label, obtained via sklearn's multilabel confusion matrix function. The reason my model is bad is because it performs poorly at every metric. The accuracy, specificity, and sensitivity of the model are all .11, which is really low. The core issue here is that there are too many genre labels to predict and not enough predictors holding independent information to help classify the genres. The model is trying to predict 52 labels from just 3 predictors afterall. If the model was classifying labels in a more-binary manner, then it would likely perform a lot better, but it's difficult to reduce genres to a binary form and still retain useful information.