Analysis of Spotify Top 2000 Tracks Data

Tamara Cooper, Ixchel Peralta-martinez, Carrie Andrews-Smith
STT 550, Fall 2022

Abstract. This report presents data analysis of the The Spotify - All Time Top 2000s Mega Dataset. Some data cleaning was performed to make a couple of the variables more meaningful and useful in the analyses. The cleaned data was applied to two models in which the Popularity variable was the response variable. The models were tested with various classification methods. The classification methods were then compared with cross validation procedures. We found that the Bagging and Random Forest classification methods produced the best results overall. However, the run times of these methods are significantly greater than other methods we used, so if the dataset were much larger the Bagging and Random Forest methods would not be optimal with the average home or business computers.

1. Introduction

Throughout history music has been an integral part of human lives. It's used to express our emotions, as a way to reflect into ourselves, as a way to relax or decompress, and as a way to make time pass when we're bored. It is also used to set the mood or add background at social events, in TV and movies, in plays, in offices, and shopping facilities. Overall, not a day goes by that most people do not hear music at some point in the day. The Spotify application has brought the world's music to many people since its launch in 2008. Since then more and more people sign up for an account and begin listening to their favorite songs and discovering new music. And, Spotify has been collecting audio data from these listeners. Can this data tell us anything about the most popular songs on Spotify's app?

The Spotify - All Time Top 2000s Mega Dataset is a collection of Spotify's top 2000 tracks for songs released from 1956 to 2019. This dataset contains information for 1,994 tracks including title, artist, genre, beats per minute (BPM), popularity, and nine other features. The data in this dataset was found on OpenML and Kaggle and was collected from the Spotify playlist - Top 2000s on PlaylistMachinery.

With this dataset I plan to answer the following questions. What audio features are predictors of songs' popularity? Did any one artist release more popular songs than the other artists? In order to answer these questions I will consider popularity as a response variable for regression and classification analysis. These are all supervised statistical learning problems. An unsupervised statistical learning problem that could be considered is whether or not there are multiple features that describe the most popular tracks. The data is summarized in Table 1.1 below.

Variable Name	Type	Description
Title	Categorical	Name of the track/song
Artist	Categorical	Name of the artist
Top Genre	Categorical	Genre of the track
Year	Quantitative	Release year of the track
Beats per Minute (BPM)	Quantitative	The tempo of the song
Energy	Quantitative	The energy of the song; the higher the value, the more energetic
Danceability	Quantitative	The higher the value, the easier it is to dance to this song
Loudness	Quantitative	The higher the value, the louder the song
Valence	Quantitative	The higher the value, the more positive mood for the song
Length	Quantitative	The duration of the song
Acoustic	Quantitative	The higher the value, the more acoustic the song
Speechiness	Quantitative	The higher the value, the more spoken words the song contains
Popularity	Quantitative	The higher the value the more popular the song

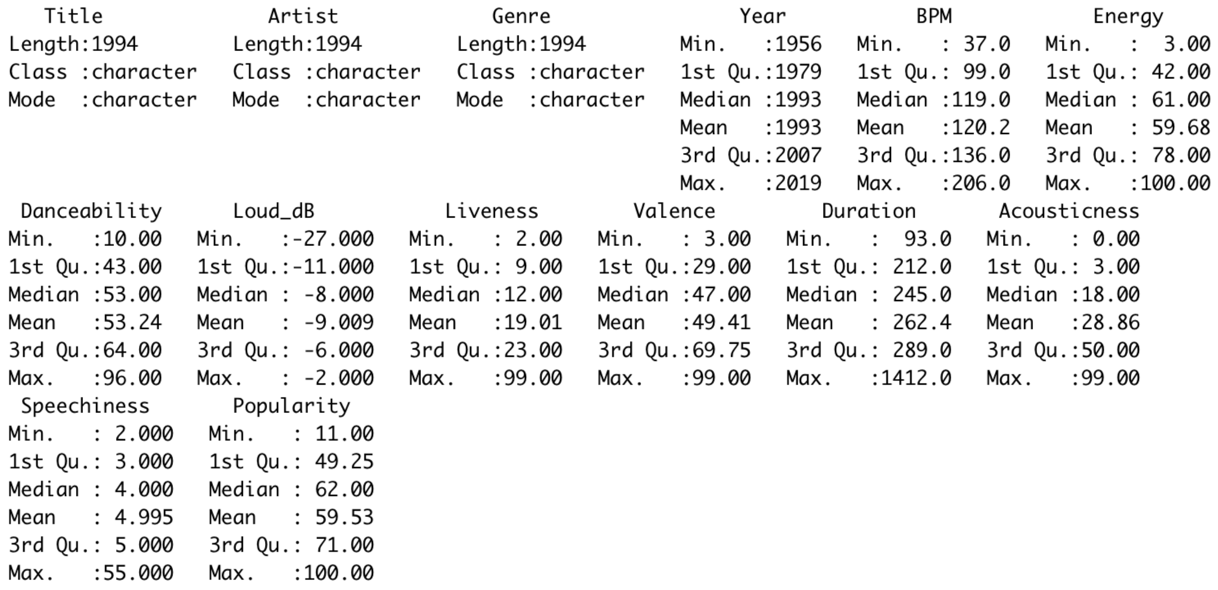
Table 1.1

2. Exploratory Data Analysis

The data in this project was found on OpenML and Kaggle, however we downloaded it from Kaggle since it was formatted in an easy to use data file on that site. This data was pulled from the Spotify playlist - Top 2000s on PlaylistMachinery. There are 1,994 observations and thirteen variables. The independent and dependent variables of interest all belong to this dataset and were not sourced from another dataset(s). We considered the Popularity variable as our response in order to see if the popularity of a song can be predicted by the other variables in this dataset.

2.1 Summary of data

The summary of all the data in figure 2.1.1 shows which variables are categorical versus quantitative. For the quantitative variables it will be easier to see the distributions through histograms and box-plots. However, it appears that Speechiness and Liveness are skewed to the right based on the means being very close to the 3rd quartiles.



```
    Title              Artist              Genre                 Year           BPM             Energy     
 Length:1994        Length:1994        Length:1994        Min.   :1956   Min.   : 37.0   Min.   :  3.00  
 Class :character   Class :character   Class :character   1st Qu.:1979   1st Qu.: 99.0   1st Qu.: 42.00  
 Mode  :character   Mode  :character   Mode  :character   Median :1993   Median :119.0   Median : 61.00  
                                                          Mean   :1993   Mean   :120.2   Mean   : 59.68  
                                                          3rd Qu.:2007   3rd Qu.:136.0   3rd Qu.: 78.00  
                                                          Max.   :2019   Max.   :206.0   Max.   :100.00  
  Danceability      Loud_dB           Liveness        Valence         Duration       Acousticness  
 Min.   :10.00   Min.   :-27.000   Min.   : 2.00   Min.   : 3.00   Min.   :  93.0   Min.   : 0.00  
 1st Qu.:43.00   1st Qu.:-11.000   1st Qu.: 9.00   1st Qu.:29.00   1st Qu.: 212.0   1st Qu.: 3.00  
 Median :53.00   Median : -8.000   Median :12.00   Median :47.00   Median : 245.0   Median :18.00  
 Mean   :53.24   Mean   : -9.009   Mean   :19.01   Mean   :49.41   Mean   : 262.4   Mean   :28.86  
 3rd Qu.:64.00   3rd Qu.: -6.000   3rd Qu.:23.00   3rd Qu.:69.75   3rd Qu.: 289.0   3rd Qu.:50.00  
 Max.   :96.00   Max.   : -2.000   Max.   :99.00   Max.   :99.00   Max.   :1412.0   Max.   :99.00  
  Speechiness       Popularity    
 Min.   : 2.000   Min.   : 11.00  
 1st Qu.: 3.000   1st Qu.: 49.25  
 Median : 4.000   Median : 62.00  
 Mean   : 4.995   Mean   : 59.53  
 3rd Qu.: 5.000   3rd Qu.: 71.00  
 Max.   :55.000   Max.   :100.00  
```

Figure 2.1.1

The histograms and box-plots of Popularity, BPM, and Energy in figures 2.1.2 and 2.1.3 show that the data for these features are close to normally distributed, there may be some outliers of concern in Popularity and BPM. The histogram of Year displays an almost uniform distribution, this could be interesting to look into more.

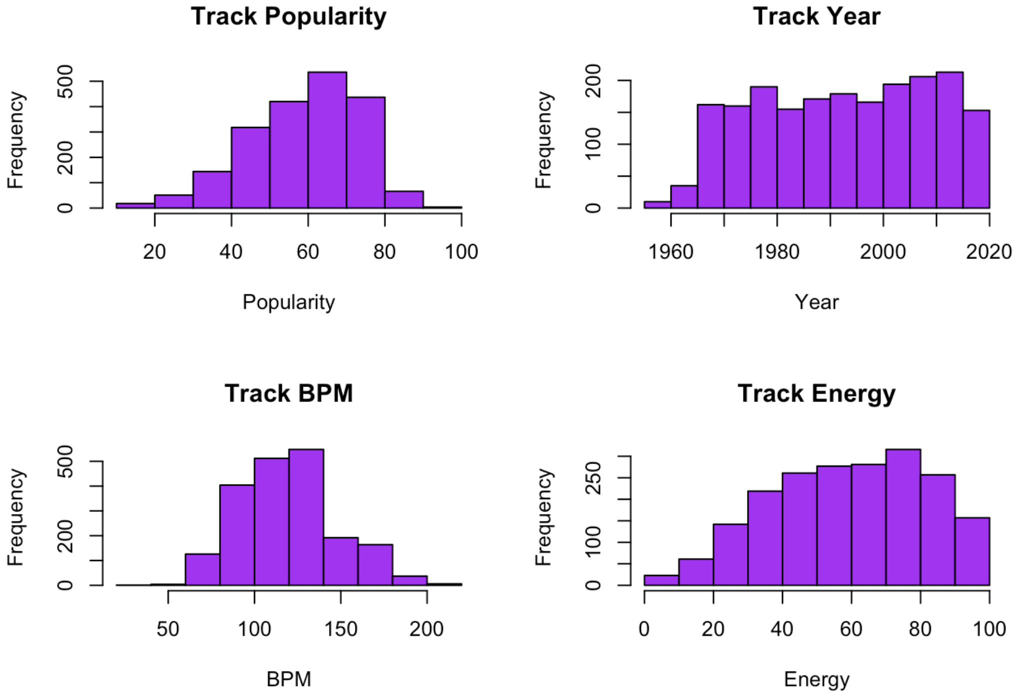


Figure 2.1.2

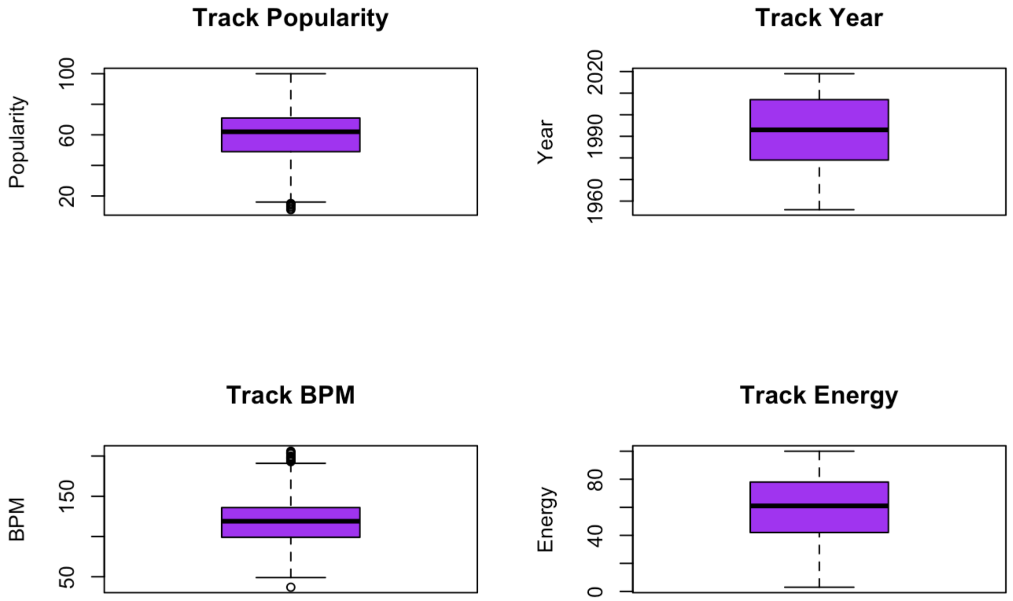


Figure 2.1.3

The histograms and box-plots in Figures 2.1.4 and 2.1.5 show that Danceability and Valence are close to normally distributed but Loudness is skewed left and Liveness is skewed right.

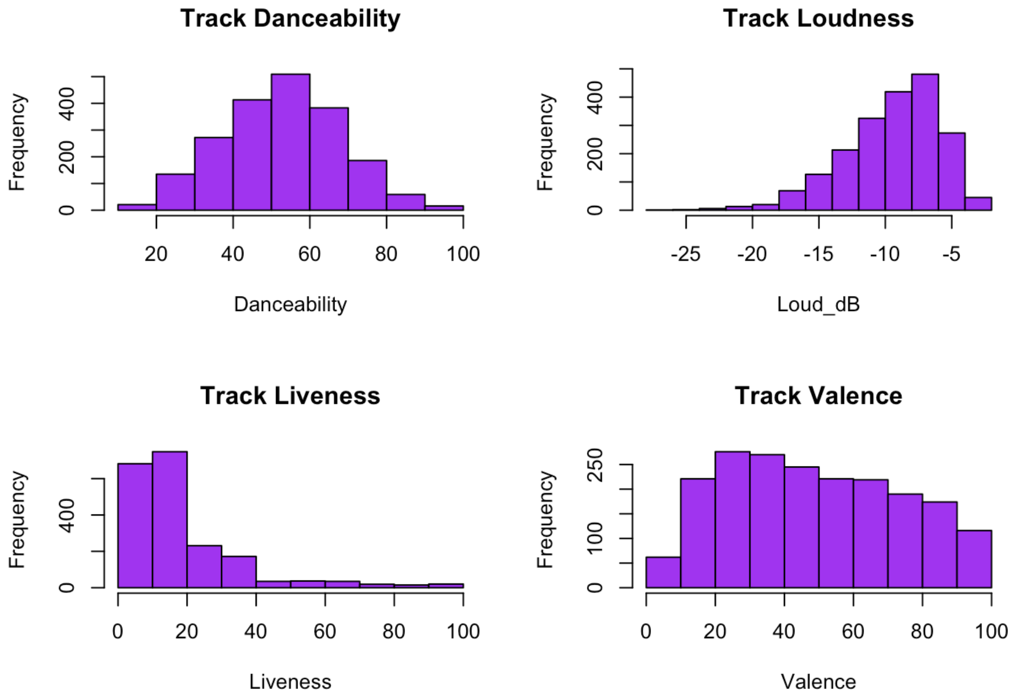


Figure 2.1.4

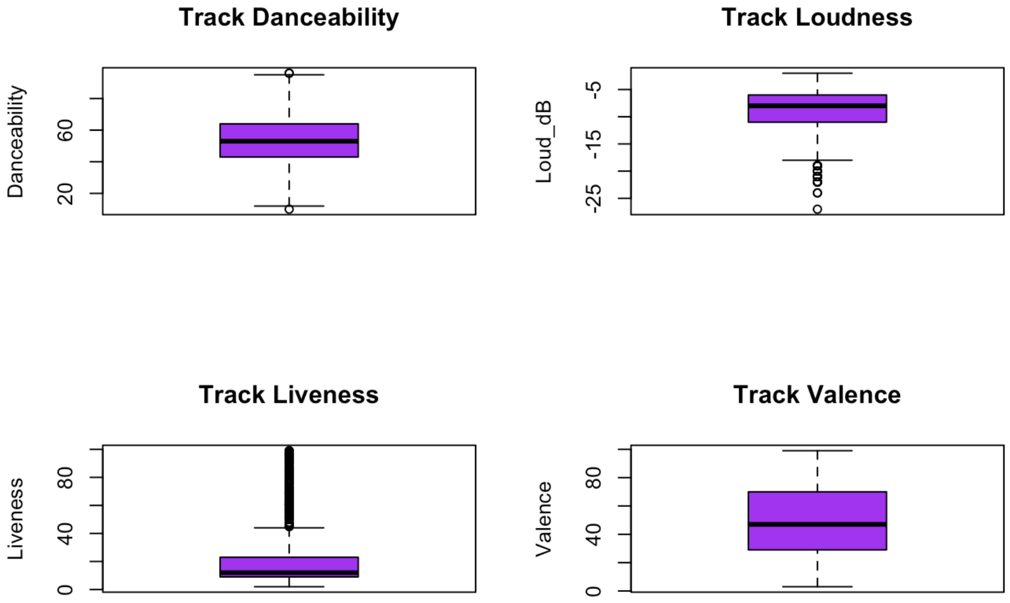


Figure 2.1.5

The histograms and box-plots in figures 2.1.6 and 2.1.7 show that the features Duration, Acousticness, and Speechiness are skewed right. Duration and Speechiness appear to have a lot of outliers that we may need to consider while conducting further analyses.

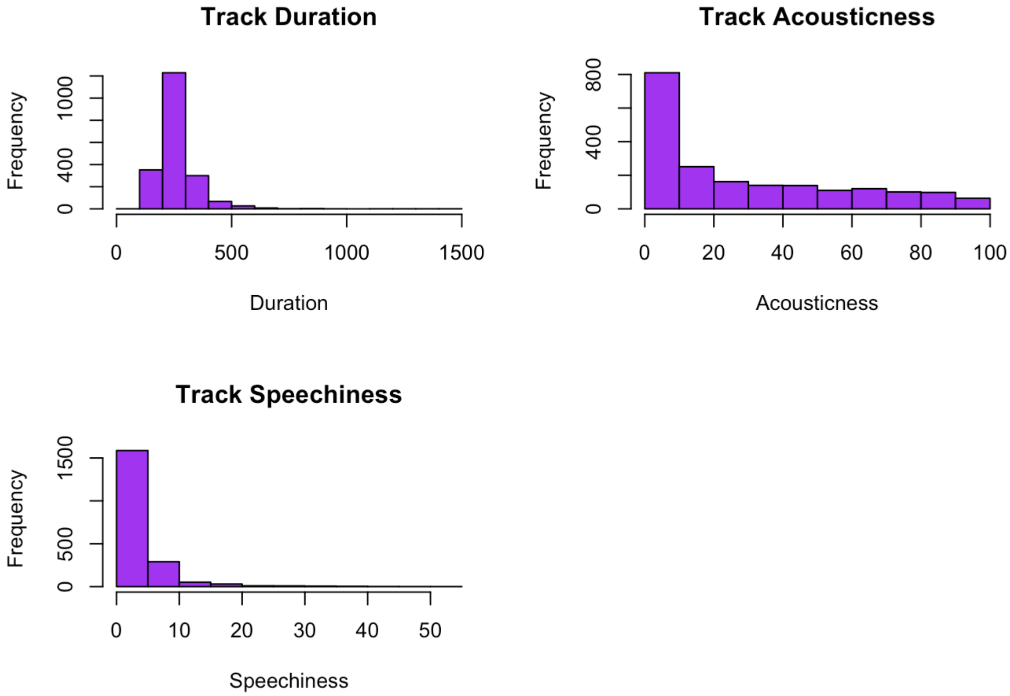


Figure 2.1.6

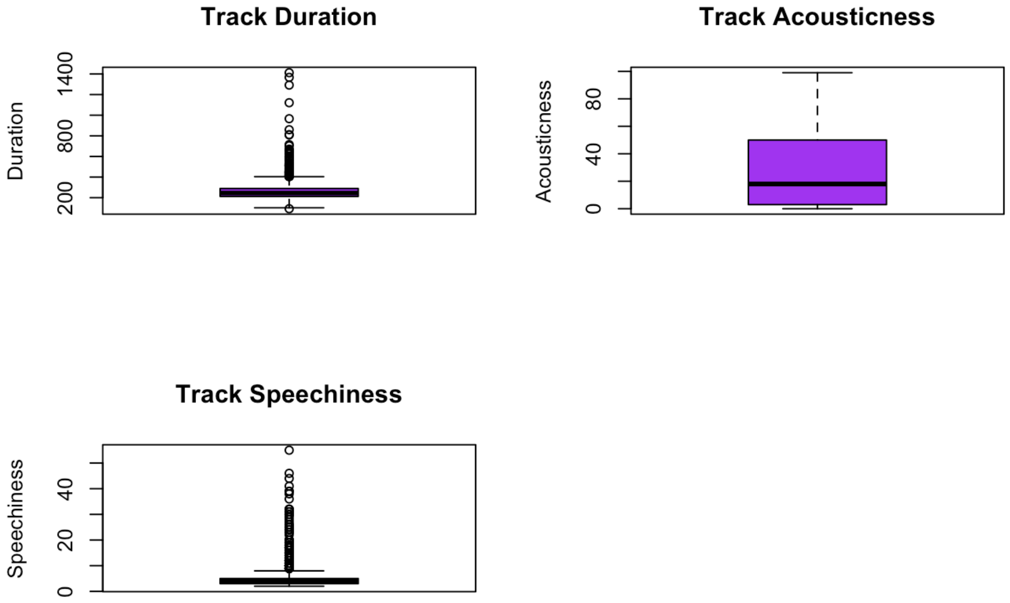


Figure 2.1.7

From the scatterplots in figure 2.1.8 it is clear that we cannot see much with this many features being plotted. Therefore, a heat-map may provide a more readable view.

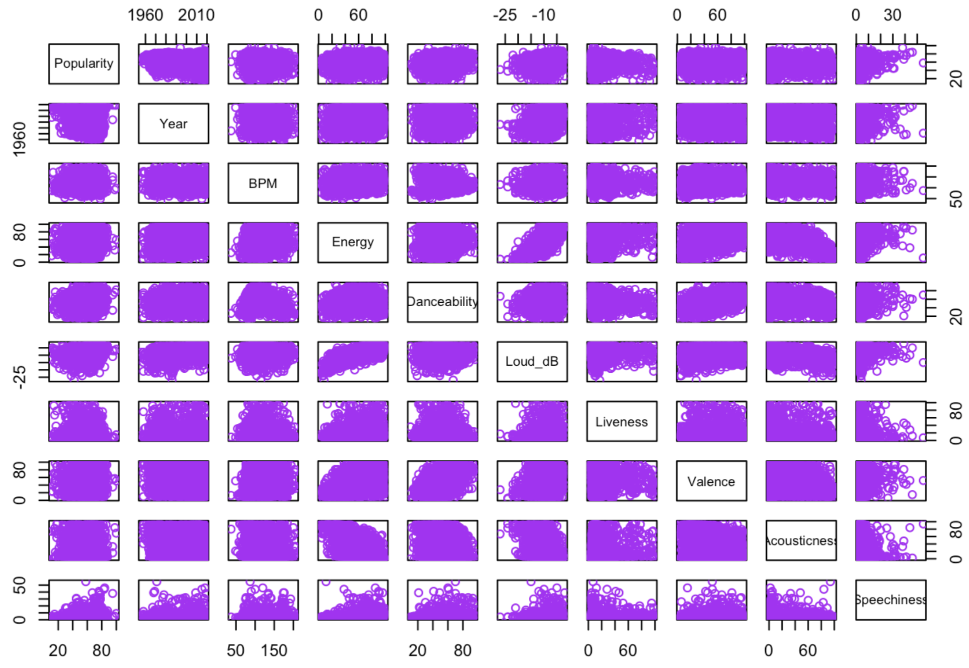


Figure 2.1.8

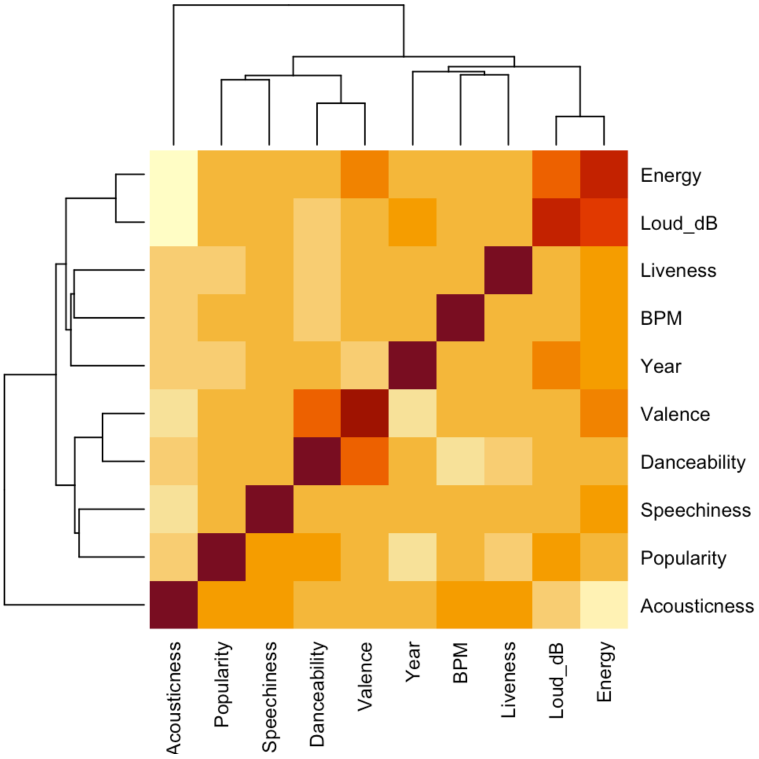
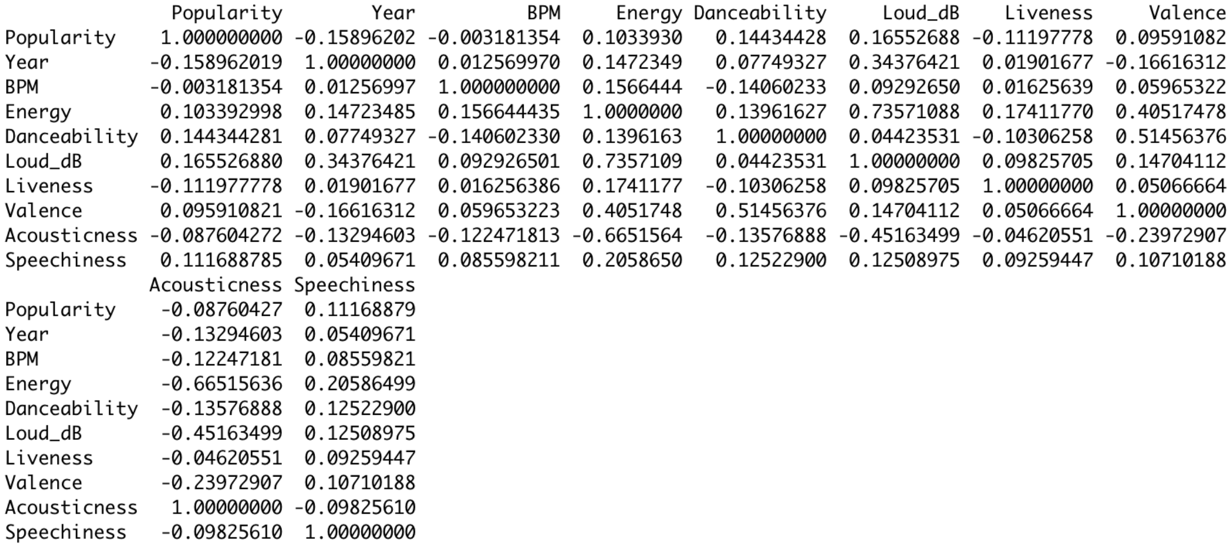


Figure 2.1.9

The heat-map in figure 2.1.9 shows that there is some positive correlation between Energy and Loudness, Energy and Valence, Danceability and Valence, and Year and Loudness. There also appears to be some negative correlation between Acousticness and Energy, Acousticness and Loudness, Danceability and BPM, Year and Valence, and Year and Popularity.

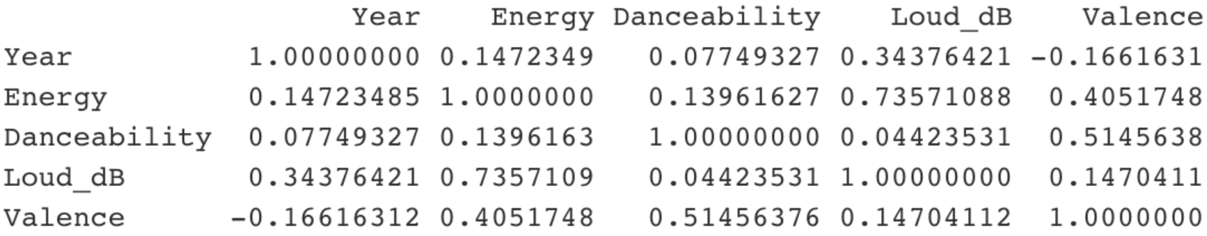
Taking a look at the correlation values provides a better understanding of the correlations between the variables. The correlation matrix in figure 2.1.10 is difficult to read but I wanted to view Popularity with all the other features first. We can see that Popularity is not strongly correlated with any of the other features. This leads me to believe that popularity must be associated with multiple features. Viewing smaller correlation matrices of the Energy, Danceability, Loudness, and Valence features can give more insight into the features that appeared to have some linear correlation in the heat-map.



```
             Popularity        Year          BPM    Energy Danceability     Loud_dB    Liveness     Valence
Popularity    1.000000000 -0.15896202 -0.003181354 0.1033930   0.14434428 0.16552688 -0.11197778  0.09591082
Year         -0.158962019  1.00000000  0.012569970 0.1472349   0.07749327 0.34376421  0.01901677 -0.16616312
BPM          -0.003181354  0.01256997  1.000000000 0.1566444  -0.14060233 0.09292650  0.01625639  0.05965322
Energy        0.103392998  0.14723485  0.156644435 1.0000000   0.13961627 0.73571088  0.17411770  0.40517478
Danceability  0.144344281  0.07749327 -0.140602330 0.1396163   1.00000000 0.04423531 -0.10306258  0.51456376
Loud_dB       0.165526880  0.34376421  0.092926501 0.7357109   0.04423531 1.00000000  0.09825705  0.14704112
Liveness     -0.111977778  0.01901677  0.016256386 0.1741177  -0.10306258 0.09825705  1.00000000  0.05066664
Valence       0.095910821 -0.16616312  0.059653223 0.4051748   0.51456376 0.14704112  0.05066664  1.00000000
Acousticness -0.087604272 -0.13294603 -0.122471813 -0.6651564  -0.13576888 -0.45163499 -0.04620551 -0.23972907
Speechiness   0.111688785  0.05409671  0.085598211 0.2058650   0.12522900 0.12508975  0.09259447  0.10710188
             Acousticness Speechiness
Popularity    -0.08760427  0.11168879
Year          -0.13294603  0.05409671
BPM           -0.12247181  0.08559821
Energy        -0.66515636  0.20586499
Danceability  -0.13576888  0.12522900
Loud_dB       -0.45163499  0.12508975
Liveness      -0.04620551  0.09259447
Valence       -0.23972907  0.10710188
Acousticness   1.00000000 -0.09825610
Speechiness   -0.09825610  1.00000000
```

Figure 2.1.10

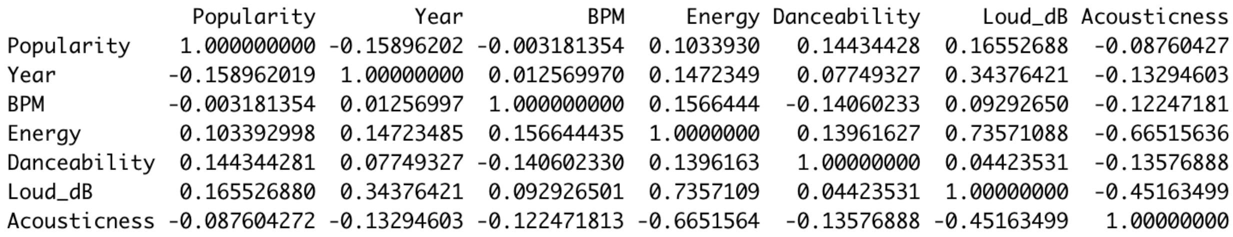
In the correlation matrix in figure 2.1.11 we see a fairly strong positive correlation between Loudness and Energy at 0.7357, there is also somewhat strong correlation between Energy and Valence at 0.4052 and Danceability and Valence at 0.5146. I will note that these correlations are strong compared to the other correlations between the features but 0.41 and 0.51 would not typically be considered strong linear correlations.



```
                    Year    Energy Danceability    Loud_dB    Valence
Year          1.00000000 0.1472349   0.07749327 0.34376421 -0.1661631
Energy        0.14723485 1.0000000   0.13961627 0.73571088  0.4051748
Danceability  0.07749327 0.1396163   1.00000000 0.04423531  0.5145638
Loud_dB       0.34376421 0.7357109   0.04423531 1.00000000  0.1470411
Valence      -0.16616312 0.4051748   0.51456376 0.14704112  1.0000000
```

Figue 2.1.11

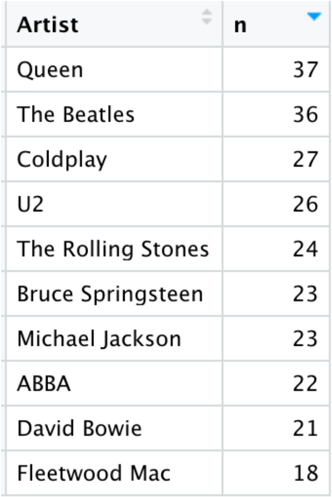
In figure 2.1.12 we see a fairly strong negative correlation between Acousticness and Energy at -0.6652, there is also a somewhat strong correlation between Acousticness and Loudness at -0.4516. Again, I will note that -0.45 is not typically a strong correlation but is strong here compared to the others.



```
               Popularity        Year          BPM    Energy Danceability     Loud_dB Acousticness
Popularity    1.000000000 -0.15896202 -0.003181354 0.1033930   0.14434428  0.16552688  -0.08760427
Year         -0.158962019  1.00000000  0.012569970 0.1472349   0.07749327  0.34376421  -0.13294603
BPM          -0.003181354  0.01256997  1.000000000 0.1566444  -0.14060233  0.09292650  -0.12247181
Energy        0.103392998  0.14723485  0.156644435 1.0000000   0.13961627  0.73571088  -0.66515636
Danceability  0.144344281  0.07749327 -0.140602330 0.1396163   1.00000000  0.04423531  -0.13576888
Loud_dB       0.165526880  0.34376421  0.092926501 0.7357109   0.04423531  1.00000000  -0.45163499
Acousticness -0.087604272 -0.13294603 -0.122471813 -0.6651564  -0.13576888 -0.45163499   1.00000000
```

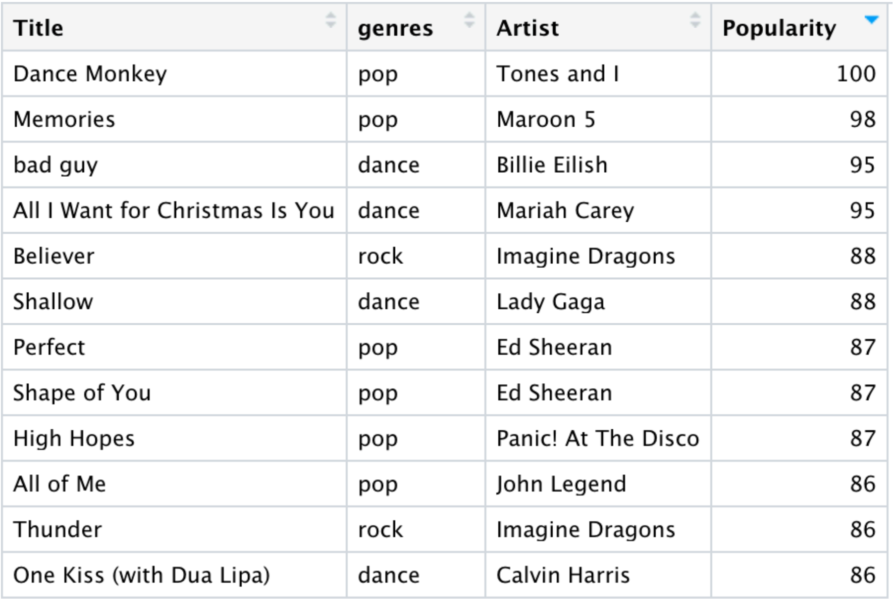
Figure 2.1.12

Below is a list of the artists that have eighteen or more songs in the top 2000 in table 2.1.1. And, in table 2.1.2 we see a list of the songs with highest popularity value. From tables 2.1.1 and 2.1.2 we see that none of the artists that have the most songs in the dataset are in the top twelve most popular songs.



Artist	n
Queen	37
The Beatles	36
Coldplay	27
U2	26
The Rolling Stones	24
Bruce Springsteen	23
Michael Jackson	23
ABBA	22
David Bowie	21
Fleetwood Mac	18

Table 2.1.1



Title	genres	Artist	Popularity
Dance Monkey	pop	Tones and I	100
Memories	pop	Maroon 5	98
bad guy	dance	Billie Eilish	95
All I Want for Christmas Is You	dance	Mariah Carey	95
Believer	rock	Imagine Dragons	88
Shallow	dance	Lady Gaga	88
Perfect	pop	Ed Sheeran	87
Shape of You	pop	Ed Sheeran	87
High Hopes	pop	Panic! At The Disco	87
All of Me	pop	John Legend	86
Thunder	rock	Imagine Dragons	86
One Kiss (with Dua Lipa)	dance	Calvin Harris	86

Table 2.1.2

As can be seen from the piechart in figure 2.1.13 there are many genres and many that appear be sub-genres of a larger genre.

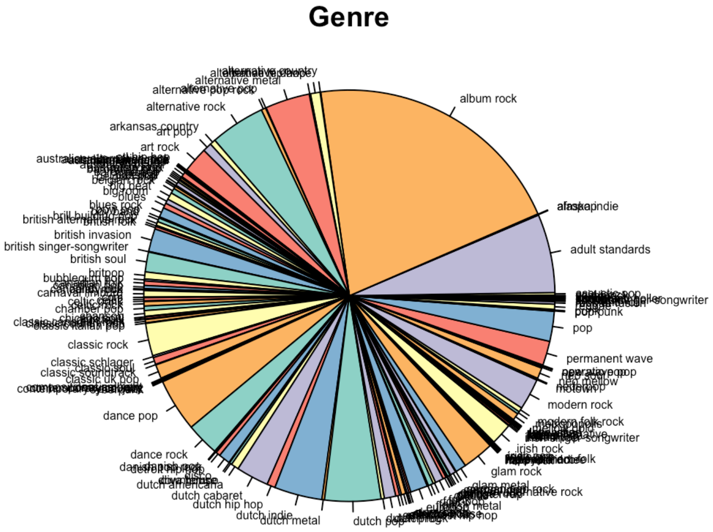


Figure 2.1.13

After condensing the genres I have found the following piechart. From figure 2.1.14, the piechart of condensed genres, we can see that Rock has the most songs in this data set with pop having the second most, followed by alternative.

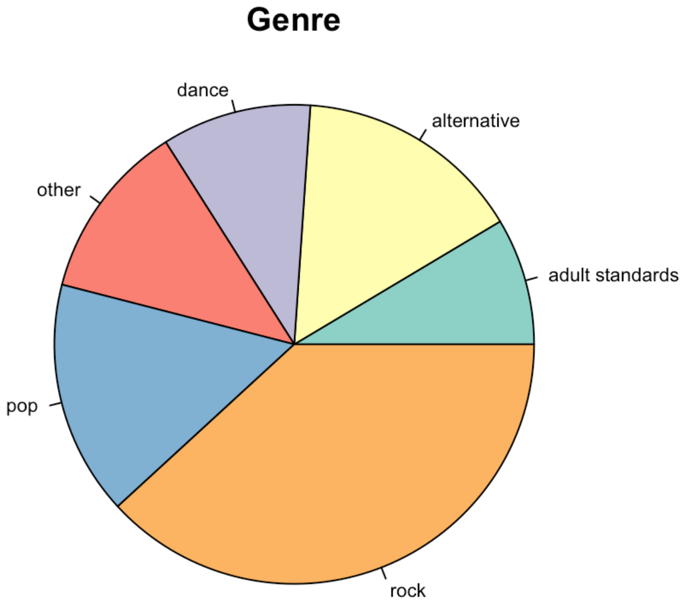
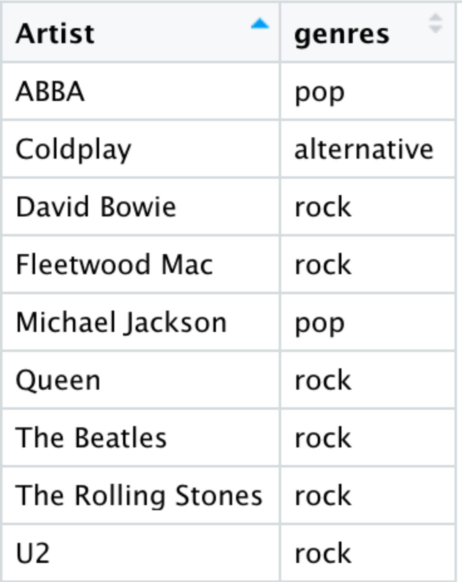


Figure 2.1.14

And finally, in table 2.1.3 we can see what genre each of the most popular artists belong. This table shows that among the most popular artists, rock is also the most popular genre followed by pop and alternative.



Artist	genres
ABBA	pop
Coldplay	alternative
David Bowie	rock
Fleetwood Mac	rock
Michael Jackson	pop
Queen	rock
The Beatles	rock
The Rolling Stones	rock
U2	rock

Table 2.1.3

3. Classification

Now we consider the Logistic Regression, Linear Discriminate Analysis, Quadratic Discriminate Analysis, and K Nearest Neighbors classification methods on this Spotify Top 2000 data. We intend to build predictive models that will describe Popularity as a categorical response variable. Here I have categorized Popularity in Low and High categories where Low Popularity values are equal to or below the median and High Popularity values are above the median. Recall that Popularity in this data set is rated with values ranging from 11 to 100, with higher values indicating that the song is more popular.

Since I found that Year, Danceability, and Loudness had the highest correlation to Popularity in the exploratory analysis, I will build a model using those as the explanatory variables. I have run the four classification methods on the model using the full data set and also using a 50% training and 50% testing data set.

3.1 Logistic Regression, Linear Discriminate Analysis (LDA), Quadratic Discriminate Analysis (QDA), and K Nearest Neighbors (KNN)

In section two with exploratory data analysis we found that none of the variables were highly correlated with Popularity, so we do not expect to find high accuracy rates with the proposed model applied to the four classification methods. In table 3.1.1 we see the accuracy, sensitivity, specificity, and run times for each of the methods using the full and 50% data sets. The model is defined as:

$$\text{Popularity} = \beta_0 + \beta_1\text{Year} + \beta_2\text{Danceability} + \beta_3\text{Loudness}.$$

Accuracy is defined as:

(Correct Low Popularity Predictions + Correct High Popularity Predictions)/(Total Popularity Predictions).

Sensitivity is defined as:

(Correct High Popularity Predictions)/(Total High Popularity Predictions)

Specificity is defined as:

(Correct Low Popularity Predictions)/(Total Low Popularity Predictions)

The accuracy, sensitivity, specificity, and run time results for the Logistic Regression, LDA, QDA, and KNN (where k = {1, 3, 5, 8, 10}) methods on the full and 50% data sets are shown below in table 3.1.1.

	Accuracy		Sensitivity		Specificity		Running Time (in seconds)	
	Full Data	50% Data	Full Data	50% Data	Full Data	50% Data	Full Data	50% Data
Logistic Regression	0.6083	0.5878	0.5571	0.6694	0.6554	0.5089	0.0217	0.032
LDA	0.6058	0.5727	0.555	0.7408	0.6526	0.4103	0.0644	0.070
QDA	0.6169	0.5777	0.6021	0.7633	0.6304	0.3984	0.0384	0.0227
KNN = 1	0.985	0.5256	0.9843	0.5429	0.9856	0.5089	0.0425	0.0216
KNN = 3	0.7593	0.5256	0.755	0.5408	0.7632	0.5108	0.0417	0.0202
KNN = 5	0.7101	0.5206	0.7037	0.5735	0.7161	0.4694	0.0422	0.0214
KNN = 8	0.6906	0.5236	0.6702	0.6224	0.7093	0.428	0.0428	0.0232
KNN = 10	0.6775	0.5366	0.6482	0.6449	0.7045	0.432	0.0438	0.0213

Table 3.1.1

Table 3.1.1 shows us that using only 50% of this data to train with produces predictions with about 52% - 59% accuracy for this model. However, we can also see that using the full data set is not much more accurate for most of the methods. Though, we may be able to find more accurate results by training on 70% of the data.

From this output we see that KNN = 1 on the Full data set produced the highest accuracy while Logistic Regression produced the highest accuracy on the 50% data set. We also see that KNN = 1 produced the best sensitivity and specificity on the full data set. QDA produced the best sensitivity on the 50% data set and Logistic Regression and KNN = 3 produced the best specificity on the 50% data set. We also can see that KNN = 3 had the lowest run time, but all of the methods except LDA had similar run times. We can see the accuracy, sensitivity, and specificity results more easily in the bar charts in figures 3.1.1, 3.1.2, and 3.1.3.

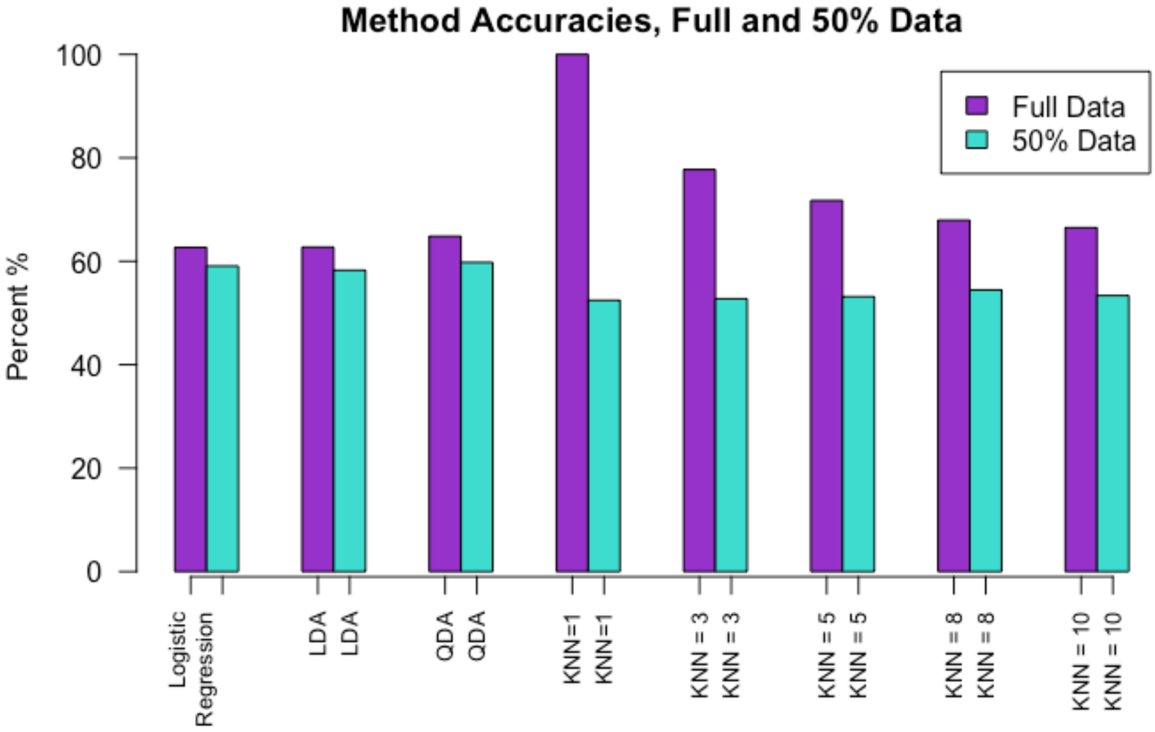


Figure 3.1.1

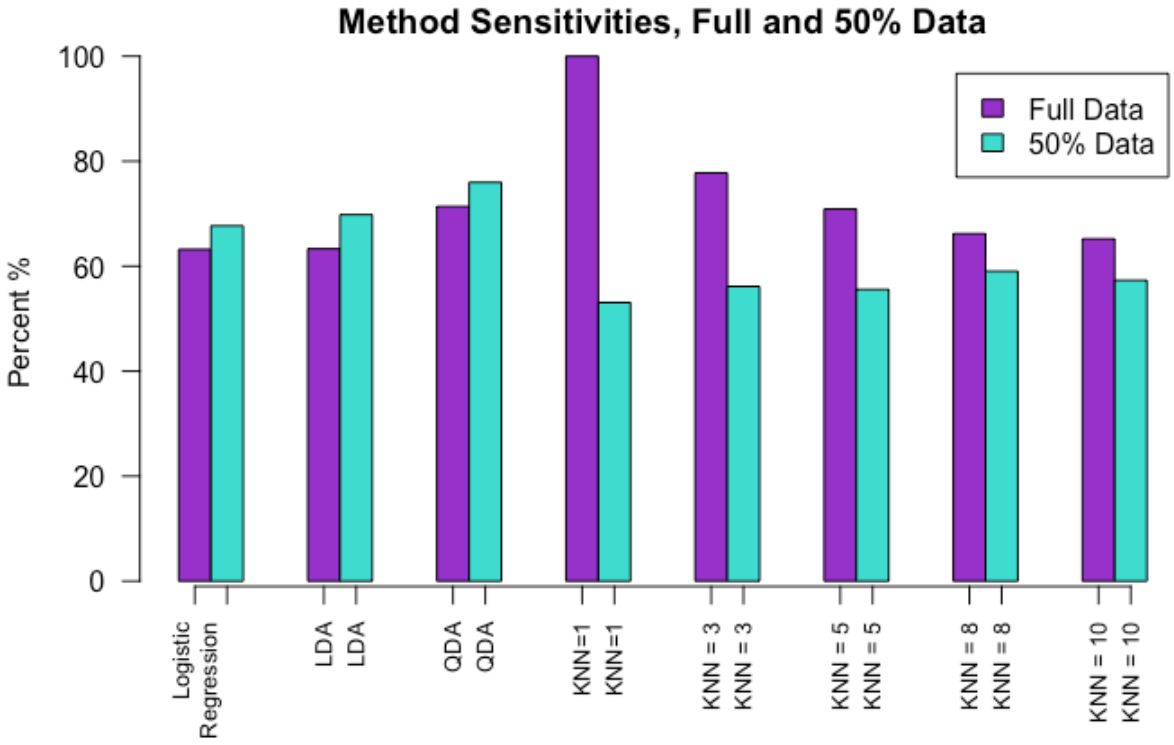


Figure 3.1.2

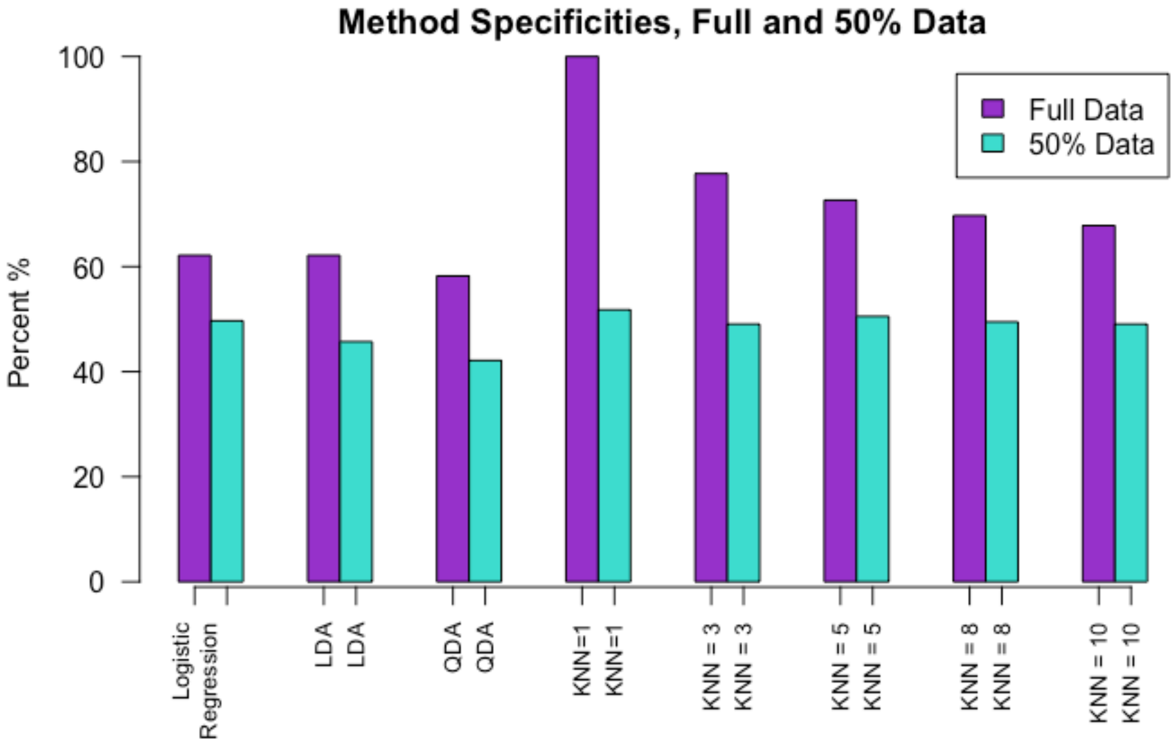


Figure 3.1.3

3.3 Conclusion

In conclusion it appears that the model presented is not the best predictor of Popularity. We might consider training on more data, however I suspect that this will not improve the accuracy by a significant amount. Using this model we find that KNN = 1 is the best classification method for the full data set and Logistic Regression is the best classification method for the 50% data set.

4. Resampling Methods

In order to obtain more information about the fitted model, we next applied some resampling methods. We considered two cross validation methods, the K-Fold and Leave One Out methods. These methods draw repeated samples from a training set and then refit the model on each sample. Once all the samples have been tested, we estimate the error rates and choose which method is best.

4.1 K-Fold Cross Validation Method

The first resampling method we applied is the K-Fold Cross Validation method with $K = 5$ (5-fold CV). The process of this method is described below:

1. Divide the data set into $K = 5$ different parts.
2. Remove the first part.
3. Fit the model on the remaining parts.
4. See how well the left out part predicts the response.
5. Repeat $K = 5$ different times by taking out a different part on each iteration.

Once all the iterations have completed we average the $K = 5$ different standard errors to obtain an estimated validation error rate for new observations. With this process we are also able to obtain estimated values for accuracy, sensitivity, and specificity on the classification methods for the proposed model.

Table 4.1.1 displays the overall accuracy, overall standard error, overall sensitivity, and overall specificity found with 5-fold CV for each of the four classification methods we considered in section three. From table 4.1.1 we see that the overall accuracy, sensitivity, and specificity found with the 5-fold cross validation method do not differ much from the values found for the full data set classification methods in section three. However, we see that the KNN values are closer to the values found with the 50% data set. We also see that the QDA classification method has the lowest overall standard error and the second fastest run time. From these values it appears that the QDA classification method is the best.

5-Fold Cross-Validation of Classification Methods

	Overall Accuracy	Overall Standard Error	Overall Sensitivity	Overall Specificity	Running Time (in seconds)
Logistic Regression	0.6063	0.3937	0.5581	0.6506	0.586
LDA	0.6058	0.3942	0.555	0.653	0.0582
QDA	0.6169	0.3831	0.6021	0.6304	0.0607
KNN	0.5893	0.4107	0.5571	0.6189	0.2074

Table 4.1.1

4.2 Leave One Out Cross Validation Method

The second resampling method we applied is the Leave One Out Cross Validation (LOO-CV) method. The process of this method is essentially the K-fold method with $K = n$, ($n =$ the number of observations in the data set) and is described below:

1. Divide the data set into a training set of size $n - 1$ and a validation set of size 1.
2. Fit the model on the training set.
3. Validate the model using the validation set and compute the standard error.
4. Repeat this process n times.

Once all the iterations are complete, we find the overall accuracy, standard error, sensitivity, and specificity. And, then we consider the classification method with the lowest overall standard error as the best method.

Table 4.1.2 shows the accuracies, standard errors, sensitivities, specificities, and run times of the proposed model for the four classification methods using LOO-CV. From table 4.1.2 we do not see any significantly different values for accuracy, standard error, sensitivity, or specificity across the classification methods. Again the QDA resampling method has the lowest standard error and second fastest run time. However, we can see that the run times are much higher with this method. These run times tell us that the LOO-CV method is more computationally intensive than the 5-fold method. Longer run times is a factor we must consider if we want to find the most efficient method. Since the standard error is not significantly different with the LOO-CV method than the 5-fold method for our model, we should opt for the more efficient 5-fold QDA method.

Leave One Out Cross-Validation of Classification Methods-1

	Overall Accuracy	Overall Standard Error	Overall Sensitivity	Overall Specificity	Running Time (in seconds)
Logistic Regression	0.6038	0.3962	0.5518	0.6516	17.83
LDA	0.6058	0.3942	0.555	0.6526	13.01
QDA	0.6169	0.3831	0.6021	0.6304	12.05
KNN	0.5928	0.4072	0.5665	0.6169	9.791

Table 4.1.2

4.3 Side by side comparison of 5-fold CV and LOOCV

To better help us select the best resampling method for our model across the classifications we look at a few visualizations. Below in figure 4.3.1 we see a bar chart display of the accuracies and standard errors found with the 5-fold CV and LOO-CV methods. Figure 4.3.1 shows us that the two resampling methods do not differ significantly across the classification methods for our model, and we see that the QDA method has the lowest standard error for both 5-fold and LOO-CV.

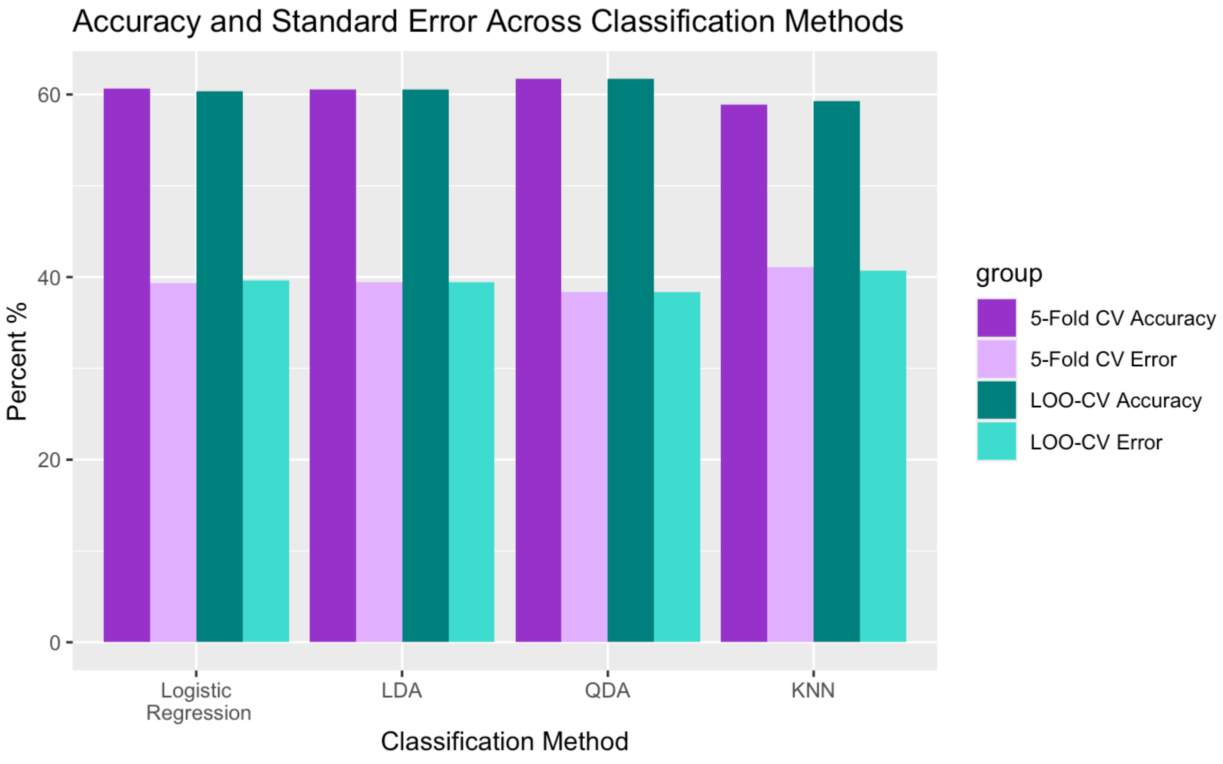


Figure 4.3.1

In figure 4.3.2 we see a bar chart of the sensitivities and specificities found with the 5-fold and LOO-CV methods. Again we do not see significant differences across the classification methods with these resampling methods. Figure 4.3.2 does, however, show us that these methods are correctly predicting the low popularity scores more often than they are correctly predicting the high popularity scores, especially for the Logistic Regression and LDA methods.

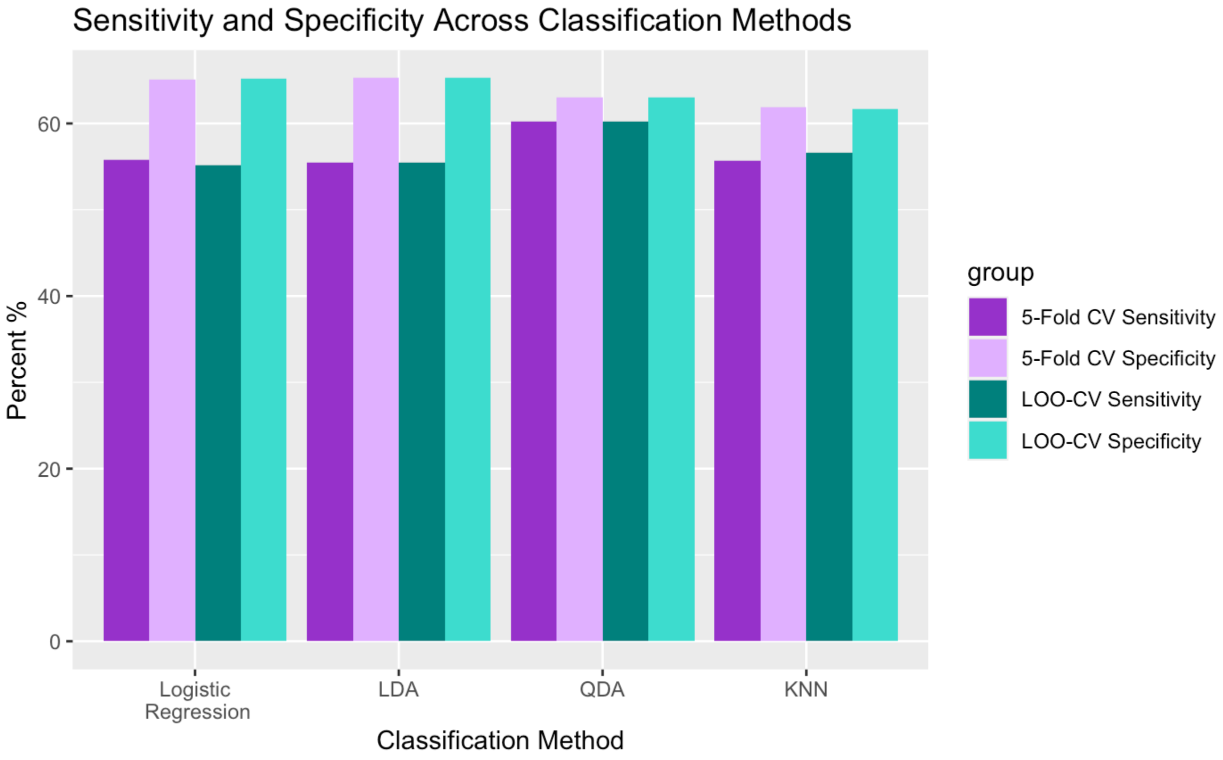


Figure 4.3.2

Finally, figure 4.3.3 displays side-by-side box plots of the accuracies for the 5-fold CV method across the four classifications. From the side-by-side box plots we see that the QDA classification method has the highest median accuracy and the smallest amount of accuracy variability of the four classifications.

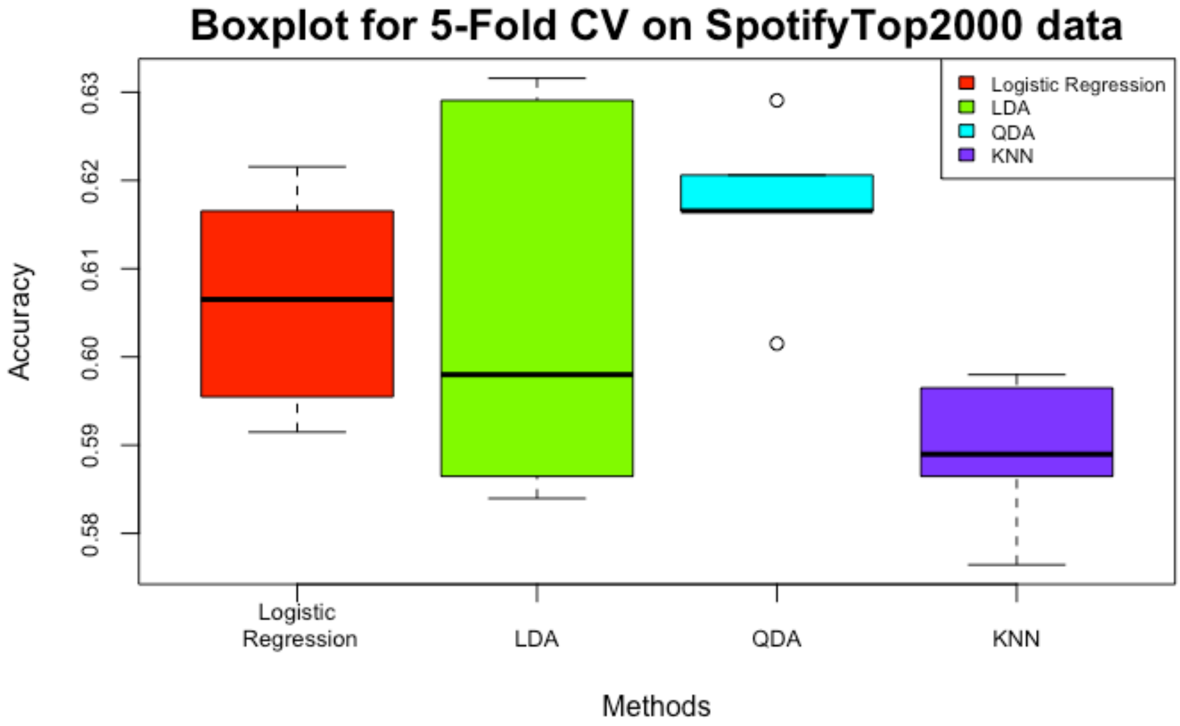


Figure 4.3.3

4.4 Conclusion

In conclusion we have found that for the proposed model the 5-fold CV and LOO-CV resampling methods perform close to the same for predicting popularity across the four classification methods, with the QDA method being the best. Since it is more efficient we should choose to use the 5-fold CV resampling method over the LOO-CV method. However, when we compare the 5-fold CV accuracies, sensitivities, and specificities to these values found with the full data set for logistic regression, LDA, and QDA we do not see any difference, but we see that the 5-fold CV run times are greater. Comparing the 50% data accuracies, sensitivities, and specificities the 5-fold CV resampling does perform better for predicting popularity when using a training set. Therefore, we would do best to choose the 5-fold CV resampling of the QDA classification method for our model.

5. Classification and Resampling with a More Complex Model and Advanced Classification Methods

In the above sections I performed classification methods on a model involving only three predictor variables. For this section I will use all the quantitative predictor variables including the trimmed down genres categorized numerically (as 1 - 6) in the model. This more complex model is defined as:

$$\text{Popularity} = \beta_0 + \beta_1\text{Year} + \beta_2\text{BPM} + \beta_3\text{Energy} + \beta_4\text{Danceability} + \beta_5\text{Loudness} + \beta_6\text{Liveness} \\ + \beta_7\text{Valence} + \beta_8\text{Duration} + \beta_9\text{Acousticness} + \beta_{10}\text{Speechiness} + \beta_{11}\text{Genre}$$

In this section I have also applied this model to the four classification methods in section three plus four more advanced classification methods. The methods under consideration in this section are Logistic Regression, LDA, QDA, KNN, Decision Tree, Bagging, Random Forest, and Boosting (for Bernoulli/Logistic Regression). The four more advanced methods are based on building trees and iterations to make predictions from a model. Bagging and Random Forest are similar, in fact Random Forest is a special case of Bagging, both of these methods use Bootstrapping to build many trees and then the average of the trees is taken for prediction. Random Forest builds on Bagging by randomly selecting m variables from each bootstrap sample to build the trees. And, in the Boosting method the trees are built sequentially from information gleaned from the previous tree.

5.1 Classification

With the above model I have performed the Logistic Regression, LDA, QDA, KNN (K = 1, 3, 5, 8, 10), Decision Tree, Bagging, Random Forest, and Boosting classification methods on a test set including 50% of the data. Table 5.1.1 shows the accuracies, error rates, sensitivities, specificities, and run times for the proposed model with all of these methods. From table 5.1.1 we see that the Bagging and Random Forest methods produce the most accurate results. However, these two methods have significantly longer run times. Since the SpotifyTop2000 dataset is not extremely large we can chose the Random Forest method with the greatest accuracy and lowest test error rate. But, if this were a dataset of the top ten thousand songs we would likely want to chose one of the less time intensive methods or we would need a more powerful computer. And since the Bagging method produces nearly the same accuracy, sensitivity, and specificity with a much lower run time we should chose this method instead of Random Forest. We can see a visual representation of these accuracies, sensitivities, and specificities below in figure 5.1.1.

	Accuracy	Error Rate	Sensitivity	Specificity	Running Time (in seconds)
Logistic Regression	0.5908	0.4092	0.6769	0.4969	0.0796
LDA	0.5827	0.4173	0.6981	0.4570	0.1454
QDA	0.5978	0.4022	0.7596	0.4214	0.0778
KNN = 1	0.5246	0.4754	0.5308	0.5178	0.0441
KNN = 3	0.5276	0.4724	0.5615	0.4906	0.0442
KNN = 5	0.5316	0.4684	0.5558	0.5052	0.0448
KNN = 8	0.5446	0.4554	0.5903	0.4948	0.0462
KNN = 10	0.5336	0.4664	0.5731	0.4906	0.04814
Decision Tree	0.6016	0.3984	0.5629	0.6408	0.0680
Bagging	0.7868	0.2132	0.8244	0.7487	0.3835
Random Forest	0.7888	0.2112	0.8184	0.7588	1.8167
Boosting	0.5163	0.4837	0.9980	0.0293	0.3006

Table 5.1.1

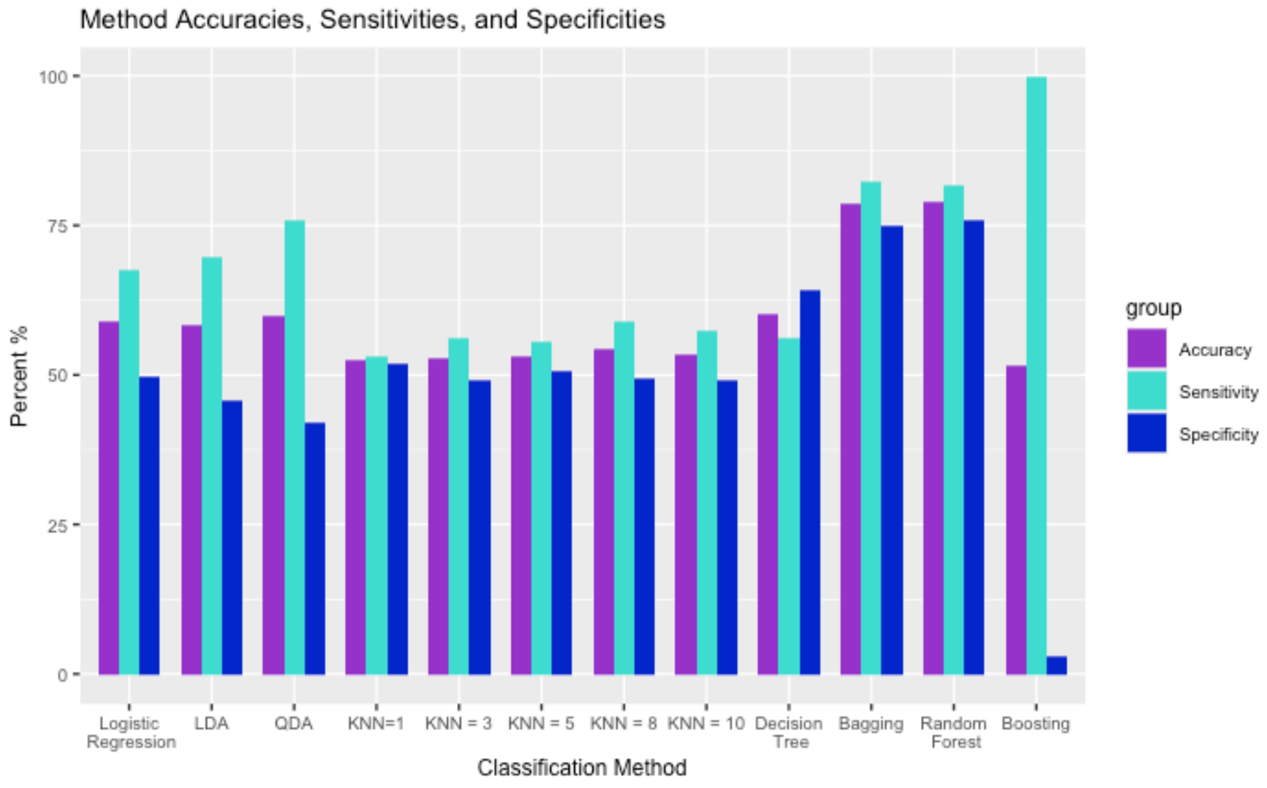


Figure 5.1.1

With the Random Forest and Boosting methods we are able to find the most important predictor variables. The Variable Importance Plot from the Random Forest method in figure 5.1.2 shows us that Year, Danceability, and Loudness (Loud_dB) and the most important predictor variables. This plot also shows us that we may want to consider the Genre variable as important predictor. Recall that in this model Genre has been condensed into six Genre categories.

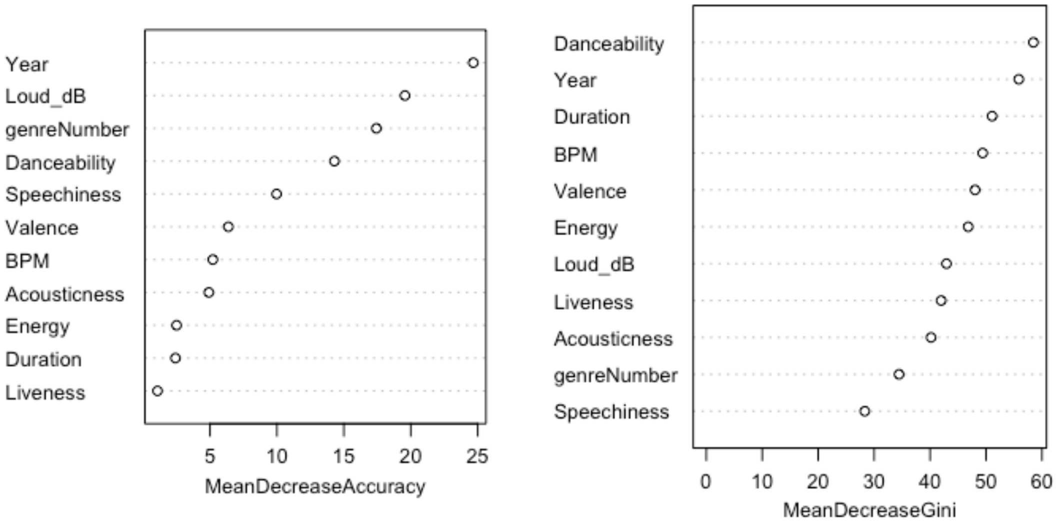
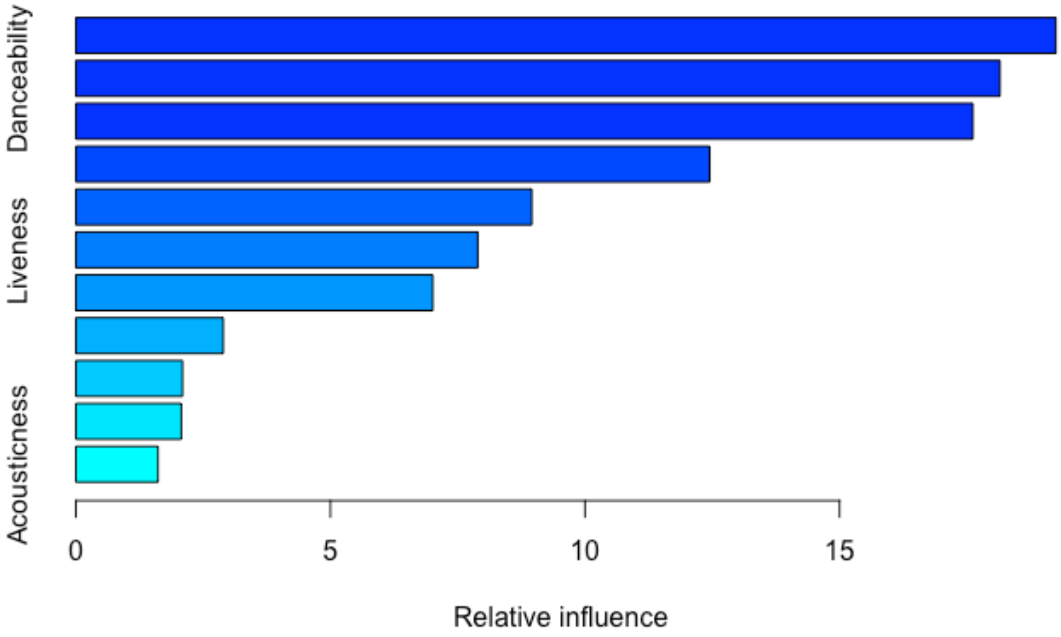


Figure 5.1.2

Additionally, table 5.1.2 and figure 5.1.3 display the variable importance determined by the Boosting method. Again we see that Year, Danceability, and Loudness are the most important predictors. And, we could consider Genre but it does not appear to be as important as Year, Danceability, or Loudness. These variable importance values and plots tell us which variables we might want to use in a simpler model.

Variable	Relative Influence	Variable	Relative Influence
Year	19.247870	Liveness	7.894001
Danceability	18.150101	Speechiness	7.010718
Loud_dB	17.618374	BPM	2.894884
Genre	12.449698	Valence	2.094760
Duration	8.954744	Energy	2.071149
Valence	2.094760	Acousticness	1.613701
Liveness	7.894001		

Table 5.1.2



5.1.3

5.2 Resampling

Finally, I applied the K-Fold Cross Validation method with K = 5 (5-fold CV) to the more complex model described in the introduction of section five. The accuracies, test error rates, sensitivities, specificities, and run times are listed below in table 5.2.1. From the accuracies and error rates we see that the Boosting method performed the best. The Boosting method does however have a longer run time than the comparably accurate QDA method. Yet, again we see that the QDA method's sensitivity and specificity are significantly different. This means that the QDA method is performing better at predicting the High Popularity songs than it is at predicting the Low Popularity songs. In that light, we may prefer the LDA method with an accuracy of 63%, sensitivity of 63%, and specificity of 62%. The LDA method has one of the lowest run times of 0.15 seconds as well.

The next best method is Random Forest. This method is predicting with an accuracy of about 64% and the sensitivity and specificity are nearly equal at 64%. Unfortunately, this method has a significantly longer run time of almost 11 seconds. So again we would not want to use the Random Forest method if our dataset were much larger. Therefore, for this dataset we can reasonably chose the Boosting method for good predictions and a fairly good run time.

	Accuracy	Standard Error	Sensitivity	Specificity	Running Time (in seconds)
Logistic Regression	0.6214	0.3786	0.6261	0.6165	0.1553
LDA	0.6274	0.3726	0.6331	0.6216	0.1511
QDA	0.6484	0.3516	0.7139	0.5822	0.1424
KNN = 10	0.5612	0.4388	0.5202	0.585	0.2195
Decision Tree	0.5958	0.4042	0.5025	0.6902	0.2131
Bagging	0.6199	0.3801	0.6171	0.6226	0.6457
Random Forest	0.6384	0.3616	0.6361	0.6408	10.8278
Boosting	0.6545	0.3455	0.6510	0.6579	2.8375

Table 5.2.1

Side-by-side box-plots of the accuracies across the classification methods for 5-Fold CV are presented in figure 5.2.1. And, a bar chart of the accuracies and test error rates across the classification methods for 5-Fold CV is presented in figure 5.2.2. These two plots provide visual representations of the findings presented in table 5.2.1. From the side-by-side box-plots it appears we should chose the QDA method since it has the best accuracy and appears to have the least variability. The bar chart seems to agree with the choice of the QDA method. Also, a bar chart of the sensitivities and specificities across the classification methods for 5-Fold CV is presented in figure 5.2.3. From this bar chart it appears that we would want to chose the Boosting method for the best predictions of High Popularity songs and Low Popularity songs. Although these visualization are helpful for quickly selecting the most accurate classification methods, we must consider run times and check with table 5.2.1 that we are choosing the best method for our dataset.

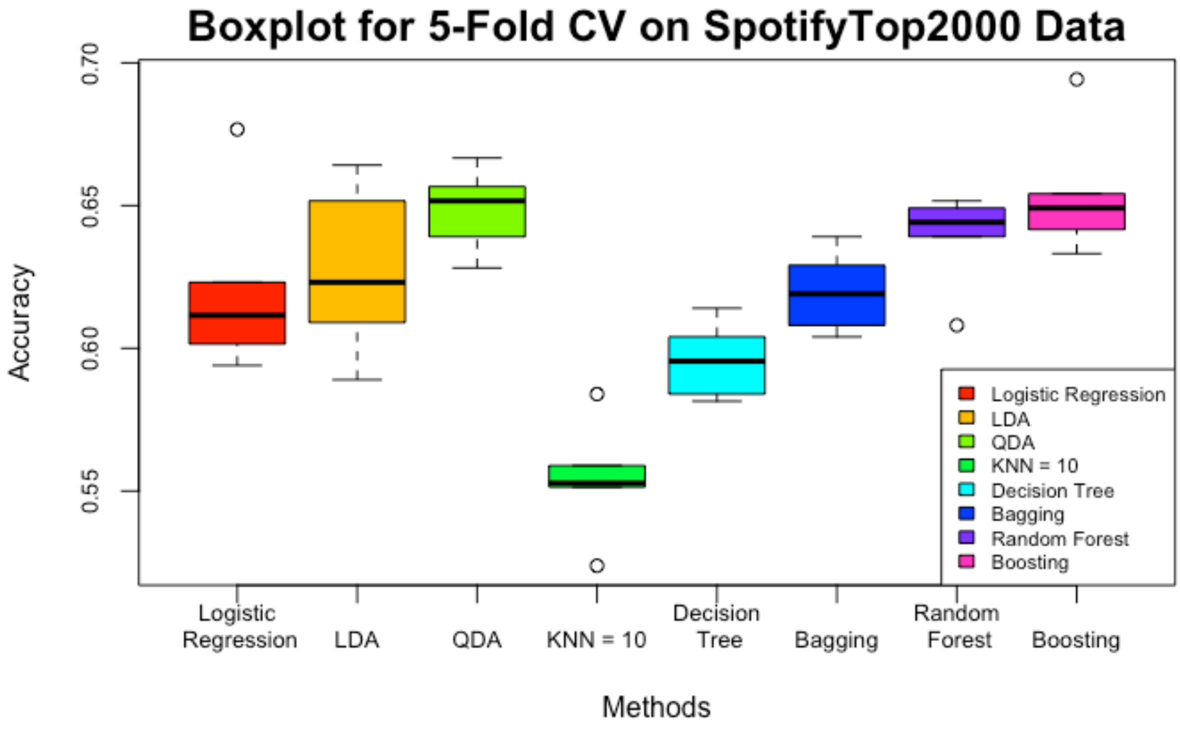


Figure 5.2.1

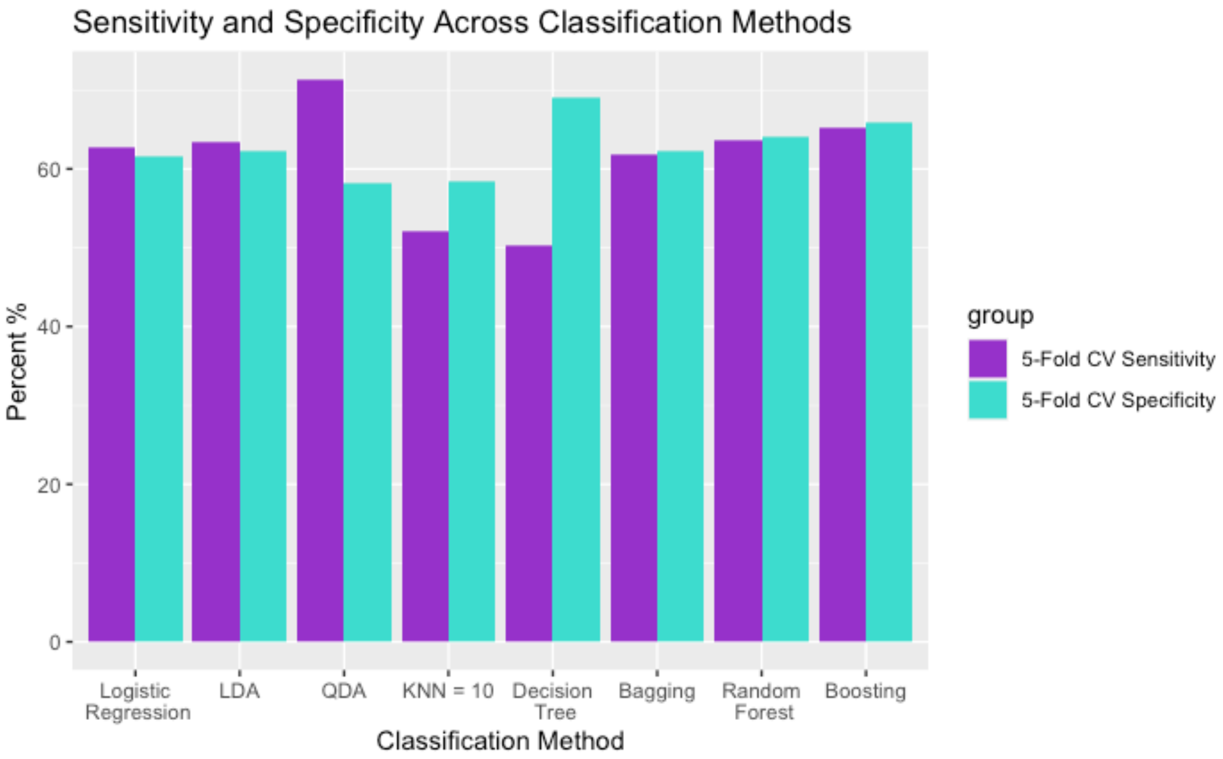


Figure 5.2.2

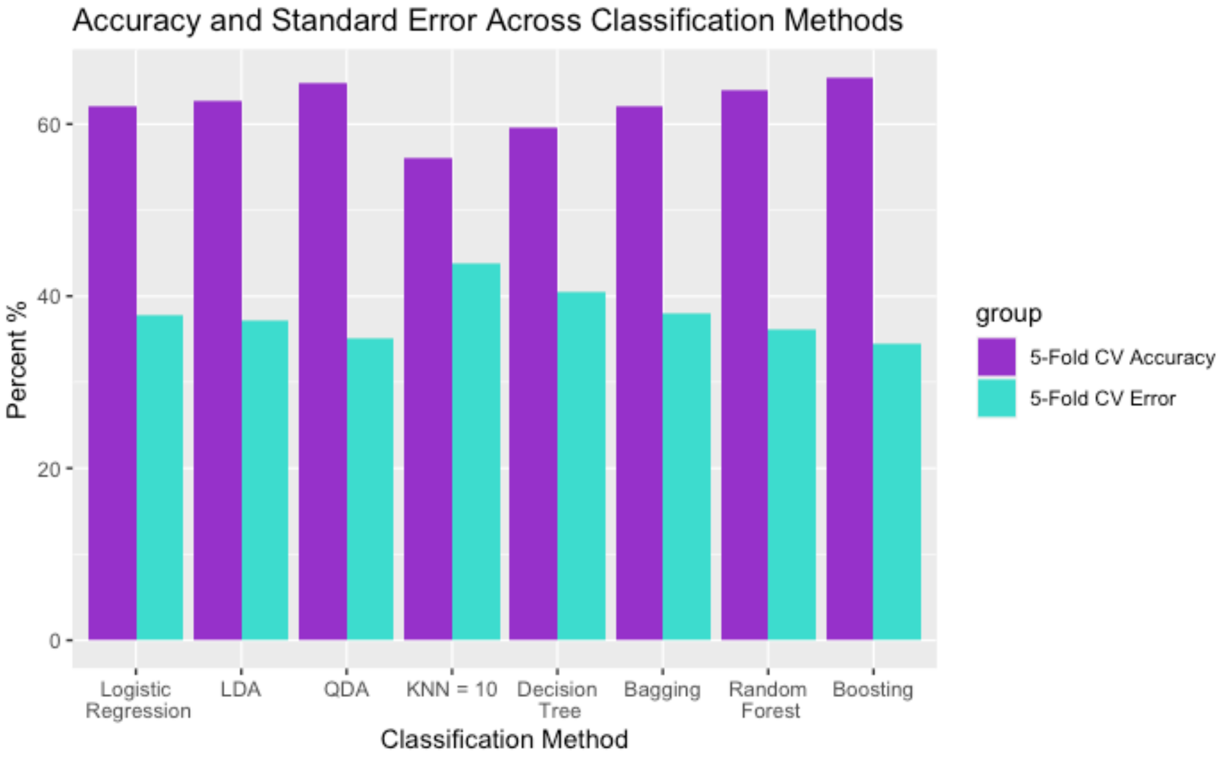


Figure 5.2.3

5.3 Conclusion

In conclusion when we consider the more complex model using all the numeric predictors and 50% of the data as a training set we found that the Bagging and Random Forest classification methods performed the best when it comes to accuracy, sensitivity, and specificity. However, we also found that both these methods were significantly more time intensive. Fortunately our dataset is not so large that the increased run time of these methods impacted our analysis in a considerable way. Yet, if our dataset was looking at Spotify's all time top ten thousand tracks rather than two thousand, we would likely find it difficult or too time consuming to use the Bagging or Random Forest methods. In that case, we found that either the QDA or LDA method is best for this more complex model. And when we ran 5-Fold Cross Validation of the eight methods using our complex model we found that the Boosting and QDA methods were the best. Additionally, we saw that the Boosting method had a higher run time but with our dataset this increased run time was not significantly impactful.

6. Final Conclusion

In closure we found that the simpler model with only Year, Loudness, and Danceability as predictor variables predicted song popularity with about 52% - 59%

accuracy using the Logistic Regression, LDA, QDA, and KNN classification methods, with Logistic Regression being the best method. We also found that the 5-Fold Cross Validation on the QDA method was the best for the simpler model. We then found that a more complex model using all the numeric variables and the Genre variable coded numerically in six categories as predictor variables predicted song popularity with about 52% - 79% accuracy using the Logistic Regression, LDA, QDA, KNN, Decision Tree, Bagging, Random Forest, and Boosting classification methods. We saw that the Bagging and Random Forest methods were the best but they both had significantly longer run times. And, with 5-Fold Cross Validation we found that the Boosting and QDA methods were best, with the QDA method having a lower run time. Through the Random Forest and Boosting methods we found that the Year, Loudness, Danceability, and possibly the Genre variables are most important in predicting song popularity. We also saw that our dataset was not so large that the greater run times of the advanced classification methods had a significant impact on our choice of best method.

So did we find answers to the proposed questions? We found that the audio features most important to predicting popularity are Year, Loudness, Danceability, and Genre. And we found that all the numeric predictors were reasonably good for predicting popularity using advanced classification methods. We discovered that the artist Queen had more songs in the dataset than any other artist. The Beatles were just behind Queen with one less song in the dataset. And, we saw that the genre Rock had the most songs in the dataset, suggesting that Rock is the most popular music genre on the Spotify application.

7. Source

- https://www.kaggle.com/datasets/iamsumat/spotify-top-2000s-mega-dataset

- “This data is extracted from the Spotify playlist - Top 2000s on PlaylistMachinery(@plamere) using Selenium with Python. More specifically, it was scraped from http://sortyourmusic.playlistmachinery.com/.”