Analysis of the Popularity of a Song based on Spotify's All-Time Top 2000 Tracks

Ixchel Peralta-Martinez

Abstract: The dataset we analyzed consisted of 1994 values with information about a song. We focus on popularity as the response value with all the numeric variables as the predictors. Classification modeling will be used to analyze the relationship between a song's popularity with all the numeric variables. The response variable popularity will be a binary variable and categorized into zero and one based on the median. If the value of popularity is below the median it will receive a zero as the value and one otherwise. Classification methods that will be used include linear discriminant analysis, quadratic discriminant analysis, logistic regression, K-nearest neighbors, AdaBoost, bagging, boosting, decision tree and random forest. We will also use a five-fold cross-validation for the classification methods.

1. Introduction

This paper aims to research a song's popularity based on different predictors like beats per minute, energy, danceability, loudness, liveness, valence, duration of the song, acoustics, and the number of words spoken in the song and genre. Spotify audio statistics for the top 2000 tracks dataset contains data about songs ranging from the year 1956 to 2019 (Spotify---All-Time-Top-2000s-Mega-Dataset). The sample size for this data set is 1994 songs (Spotify---All-Time-Top-2000s-Mega-Dataset). This data set contains four categorical variables and ten numeric variables. The value given for a song's popularity is the response and a numeric variable (Spotify---All-Time-Top-2000s-Mega-Dataset). The predictors will be all the numeric variables except the genre. The

categorical variable we will use from the dataset is the different genres for tracks. Numerical variables for the dataset consist of the song's tempo measured with beats per minute, the level of energy with high-levels corresponding with more energetic songs, and the song's danceability with high-levels corresponding with easier to dance to songs, the loudness of a song with high-levels corresponding with louder songs, the valence with a more positive mood of a song scoring higher, duration of a song, level of an acoustic song higher value more acoustic the song, song word amount with a higher value containing more word spoken in song, and the popularity of a song(Spotify---All-Time-Top-2000s-Mega-Dataset).

We will use supervised statistical learning with classification modeling. The goal of the classification analysis will be to determine if any predictor variable has the most effect on popularity and look to see if different combinations of predictors lead to a popularity increase. We have one categorical variable, the genre of the song which we will use classification models.

2. Exploratory Data Analysis

2.1 Data Descriptions

When taking a closer look at genre values, many of the values are variations of other values. For example, a value in the genre is album rock, and a similar different value is rock-and-roll which can both be categorized under a larger category called rock. After counting all the unique values of the genre, the total was 149. The initial pie chart was not very useful, so some cleaning of the data was done to better categorize the genre values. Creating a unique list of words showed that many of the genres contained some variation of the word’s soul, jazz, hip hop, alternative, dance, indie, rock, and pop. After replacing genres that contained these words or were from a subset of these words, we got down to 47 categories, 6 were created as the main genres, and all other genres were put in a category called other.

2.2 Descriptive Statistics

	N	Mean	SD	Min	Q1	Median	Q3
Beats per minute	1994	120.22	28.03	37	99	119	136
Energy	1994	59.68	22.15	3	42	61	78
Danceability	1994	53.24	15.35	10	43	53	64
Loudness	1994	-9.01	3.65	-27	-11	-8	-6
Liveness	1994	19.01	16.73	2	9	12	23
Valence	1994	49.41	24.86	3	29	47	70
Duration	1994	262.44	93.6	93	212	245	289
Acousticness	1994	28.86	29.01	0	3	18	50
Speechiness	1994	4.99	4.4	2	3	4	5

Table 2.1. Five-Number summaries for the numerical variables

We start the analysis by looking at different predictors' five number summaries in Table 2.1. When looking at the variables beats per minute, valence, danceability, liveness, duration, acoustics, and speechiness, the mean is larger than the median, which means the distributions may be skewed to the right. Likewise, when looking at the variable's energy and loudness, the median is larger than the mean, which means the distributions may be skewed to the left. All variables have some skewness so a transformation may be needed.

We will check and examine the dot plots and boxplots for variables with more prominent differences between the mean and median.

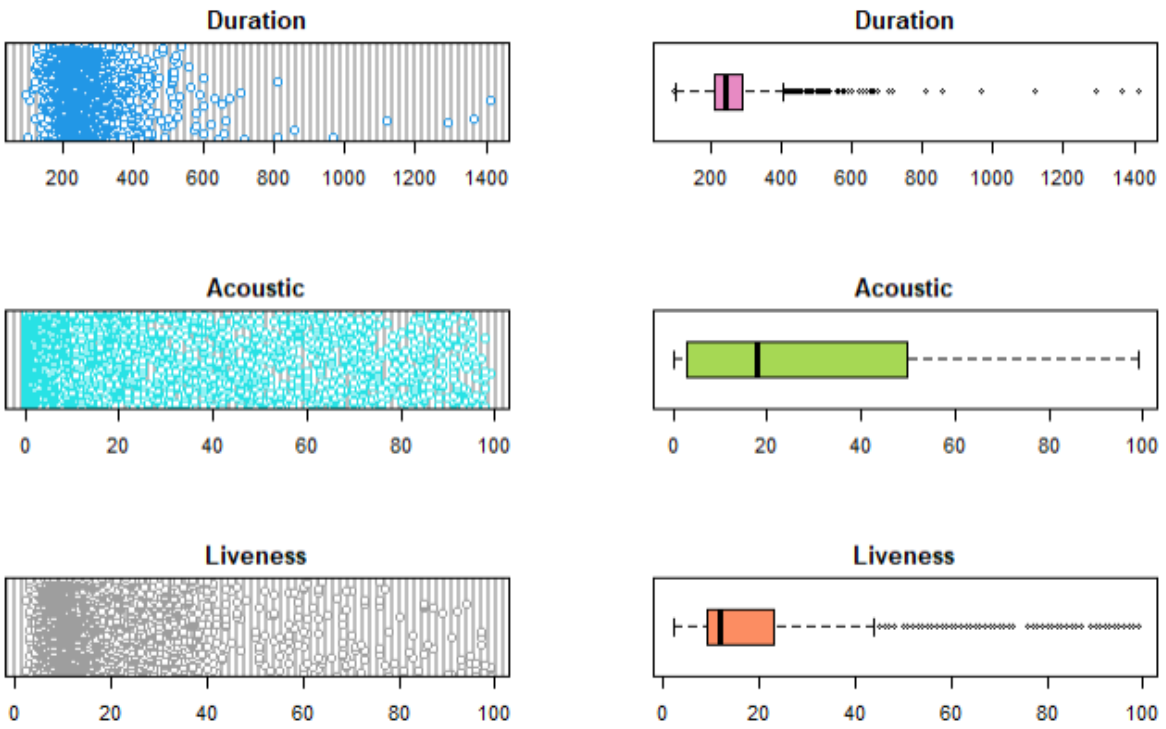

Figure 2.1 Dot plot and box plot for liveness, duration, and acousticness

Figure 2.1 dot plot for liveness shows a cluster from the values 0 to 20. Figure 2.1 box plot for liveness also shows that the median is less than 20. Again, the box plot shows that the distribution may be skewed to the right, with potential outliers occurring after the maximum. When looking the dot plot on Figure 2.1 for duration it shows a cluster from the values 200 to 400 with a couple of points outside the cluster. A similar finding is found when looking at the box plot. The box plot shows that the distribution may be normal, with potential outliers mainly occurring after the maximum. The dot plot on Figure 2.1 for an acoustic shows a cluster from the values 0 to 20 and even spread after. A similar finding is found when looking at the box plot. The box plot shows that the distribution may be skewed to the right, but there aren't any potential outliers. Most of the points for a song's liveness and duration are within a close range,

so the potential outliers may impact the regression model when having popularity as the response variable.

	Popularity	Beats per minute	Energy	Danceability	Loudness	Valence
Popularity	1					
Beats per minute	-0.00318	1				
Energy	0.103393	0.156644	1			
Danceability	0.144344	-0.1406	0.139616	1		
Loudness	0.165527	0.092927	0.735711	0.044235	1	
Valence	0.095911	0.059653	0.405175	0.514564	0.147041	1

Table 2.2. Correlation coefficient matrix for popularity, beats per minute, energy, dance ability loudness and valence

	Popularity	Duration	Acousticness	Speechiness	Liveness
Popularity	1				
Duration	-0.0654	1			
Acousticness	-0.0876	-0.10232	1		
Speechiness	0.111689	-0.02783	-0.09826	1	
Liveness	-0.11198	0.032499	-0.04621	0.092594	1

Table 2.3. Correlation coefficient matrix for popularity, duration, acousticness, speechiness and liveness

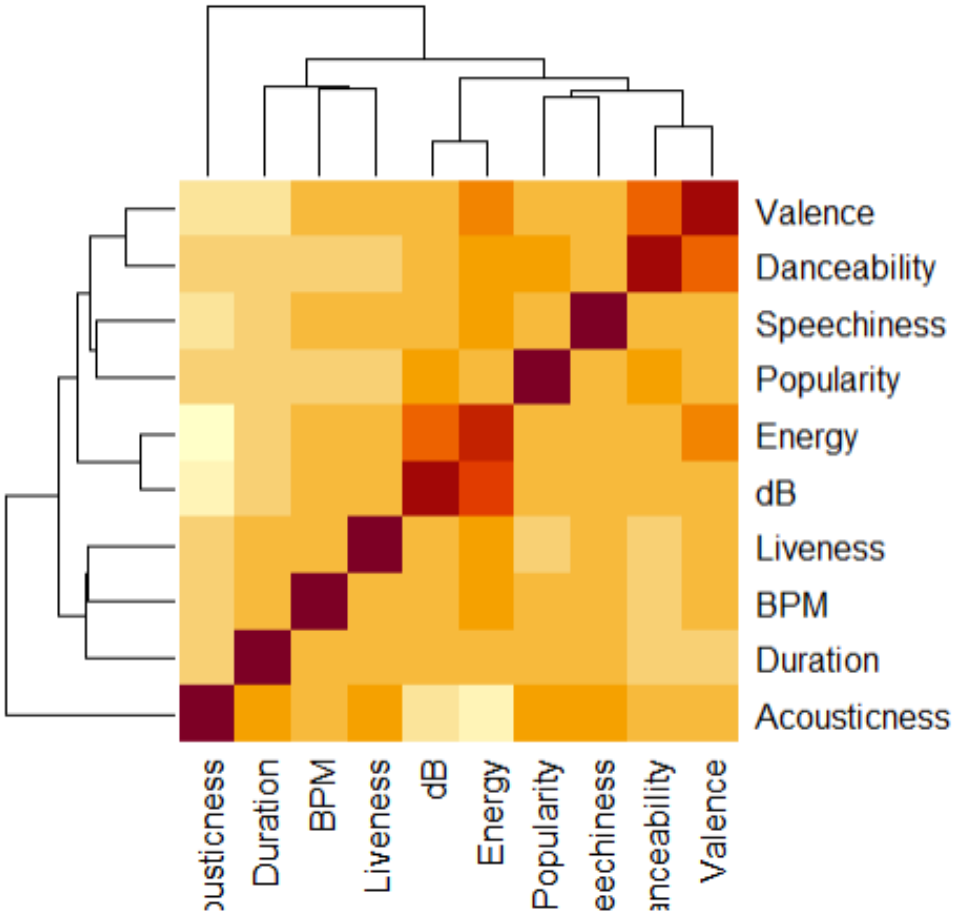

Figure 2.2 Heat map of numerical variables

Next, we will explore the correlation coefficients with popularity and the predictor variables. Table 2.2 and Table 2.3 show that most of the correlation coefficients are weak or have no relationship for all variables' positive and negative values. The most correlated among all pairs is the value for energy and loudness(dB), with r=0.736. Other significant correlation coefficients include valance and energy with r = 0.405, acousticness and energy r = -0.665, and valence and dance ability r = 0. 515.The popularity variable is the most correlated with danceability with an r = 0.14 and loudness r=0.166. The smallest correlation coefficient among the pairs is duration and speechiness, with a correlation coefficient of 0.023. Multicollinearity may be an issue when trying to fit a regression model because the correlation coefficient is

moderate to strong.

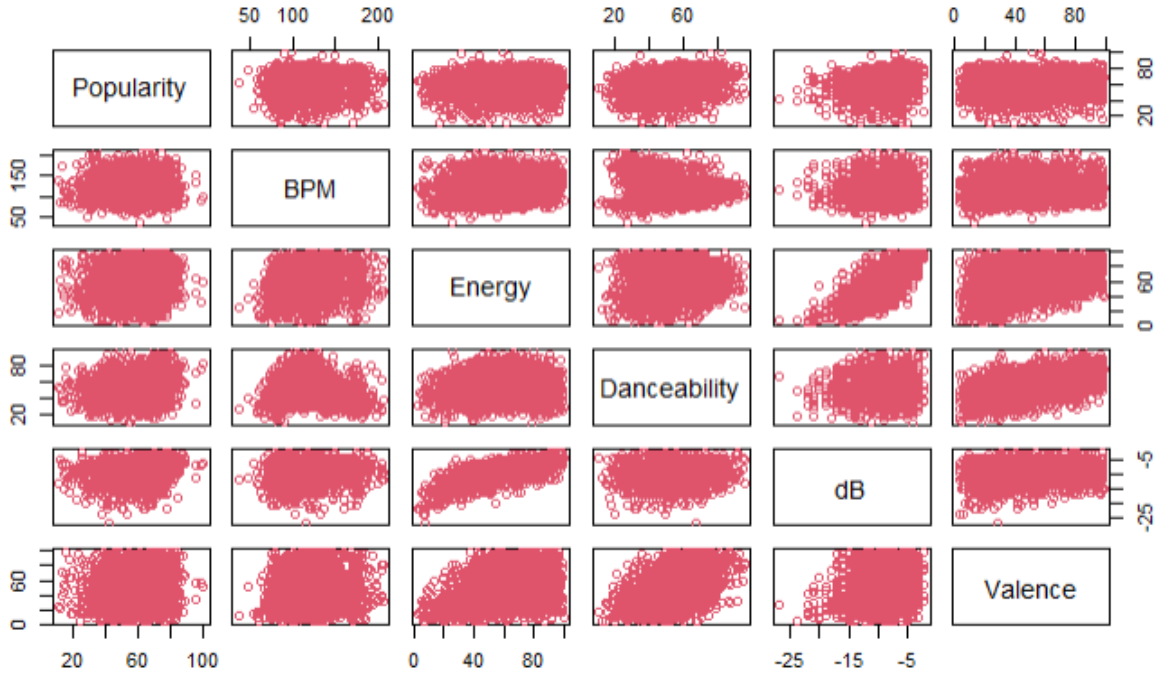

Figure 2.3 Scatter plot for numeric variables popularity, beats per minute, energy, danceability, loudness, and valence

Figure 2.3 shows the scatter plot for variable response popularity and predictor variables. When looking at the popularity and predictors variables, we don't see any variables with a significant linear relationship. However, the scatter plot does show that the predictor variables energy and loudness(dB) have a linear relationship.

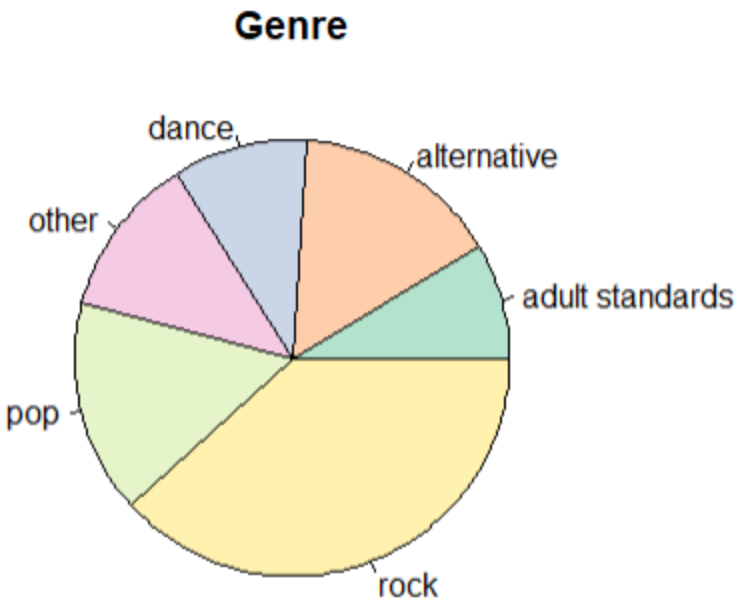

Figure 2.4 Pie chart for genres

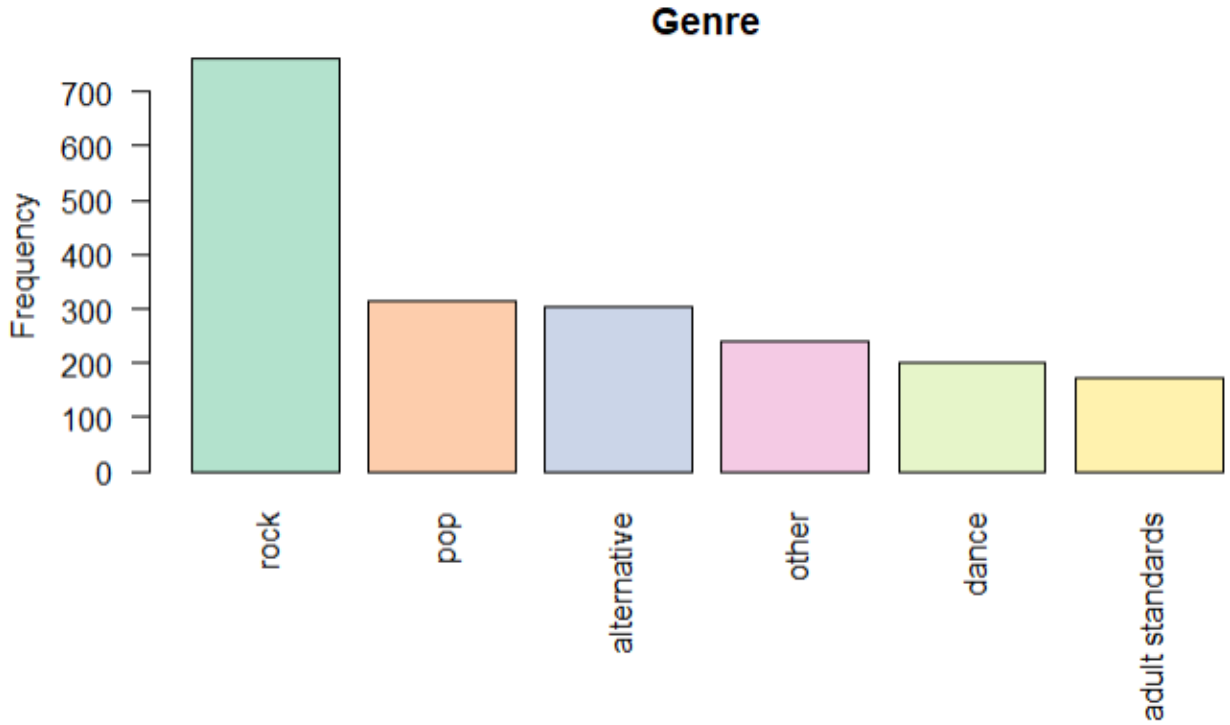

Figure 2.5 Pareto bar chart for genre

Now we will examine the categorical variable genre. After categorizing the genres into related categories, the pie chart Figure 2.4 shows almost half of the songs fall into the genres belonging to pop and rock songs. Some of the large slices of the chart include adult standards, dance, and alternative. The bar plot Figure 2.5 shows a similar finding to the pie chart. Again, the largest count is rock, with over 700 songs falling in the category. The second largest count is pop, with a count of close to 300 songs. The following largest counts are alternative and dances with counts of over 100 songs.

3. Classification

We introduce the categorical variable genre in our analysis and will examine the variance inflation factor to check if multicollinearity exists. From Table 2.2, we note that loudness and energy had a high correlation coefficient. Table 3.1 shows that none of the values exceed five, so there is no evidence of collinearity.

BPM	Energy	Danceability	Loudness	Liveness	Valence	Duration	Acousticness	Speechiness	Genre
1.075	4.102	1.511	2.432	1.073	1.852	1.110	1.871	1.083	1.048

Table 3.1. Variance inflation factor for numerical variables and genre

3.1 Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression and KNN

Next, we will categorize the numerical popularity variable into low and high, with a low value corresponding to less than the median of popularity and a high above the median of popularity. Then, we will use the numerical variables and genres to predict whether a song has a low or a high value for popularity. Finally, we will build all models with all 1994 observations as our training and test sets. Then we will use only half of our data set as our training set and the other half as our test set.

	Accuracy		Sensitivity		Specificity		Running Time (Seconds)	
	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data	Full Model	50% Data
LDA	0.588	0.567	0.603	0.589	0.572	0.499	0.046872s	0.027927s
QDA	0.581	0.544	0.681	0.7	0.503	0.4	0.049867s	0.036901s
KNN = 3	0.769	0.495	0.757	0.482	0.770	0.497	0.058931s	0.029441s
KNN = 5	0.718	0.507	0.692	0.504	0.725	0.499	0.050863s	0.022940s
KNN = 7	0.682	0.52	0.649	0.516	0.693	0.513	0.058835s	0.022935s
KNN = 10	0.663	0.525	0.646	0.518	0.668	0.521	0.050856s	0.022936s
Logistic Regression	0.588	0.49	0.603	0.512	0.572	0.454	0.028914s	0.036900s

Table 3.2. Full data and half of data accuracy, sensitivity, specificity and run time for linear discriminant analysis, quadratic discriminant analysis, logistic regression and KNN

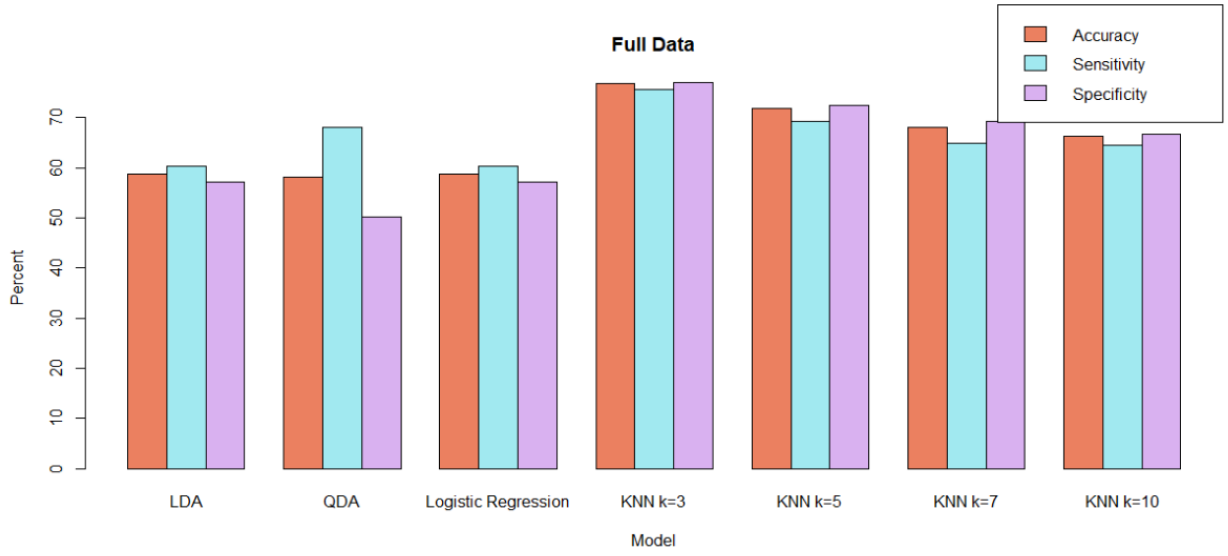

Figure 3.1 Bar plot for full data with the accuracy, sensitivity and specificity for linear discriminant analysis, quadratic discriminant analysis, logistic regression and KNN

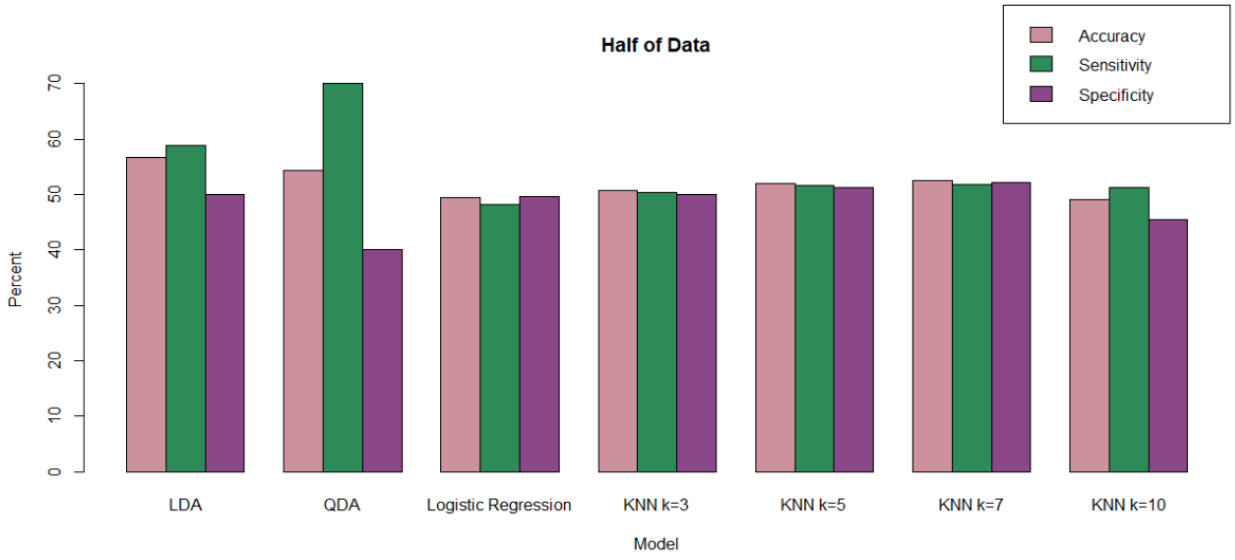

Figure 3.2 Bar plot for half of data with the accuracy, sensitivity and specificity for linear discriminant analysis, quadratic discriminant analysis, logistic regression and KNN

Table 3.2 shows the highest accuracy when using the complete data set is 76.9 percent when using the model KNN with k equal to 3. Table 3.2 also shows the highest accuracy when using half the data set is 56.7 percent when using the linear discriminant analysis. Table 3.2 also shows the highest sensitivity is 75.5 percent when using the complete data set and the model KNN with k equal to 3. The highest sensitivity is 70 percent when using half of the data set and quadratic discriminant analysis. The highest specificity is 77 percent when using the complete data set and the model KNN with k equal to 3. Finally, table 3.2 shows the highest specificity is 52.1 percent when using half of the data set and KNN with k equal to 3. The slowest run time is 0.058931s when using complete data set and KNN with k equal to 3. The slowest run time is 0.036901s when using half the data set and quadratic discriminant analysis.

Figure 3.1 shows that the best model for full data is KNN with k equal to 3, Although it is the slowest. Figure 3.2 shows that the best model for half the data is linear discriminant analysis because it has the highest accuracy and specificity.

3.2 5-Fold Cross Validation Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression and KNN

We continue our classification analysis with five-fold cross-validation using all of the numerical variables and genres to predict whether a song has a low or a high value for popularity. In the previous section, we analyzed the data with the complete data set and half the data set, but for five-fold cross-validation, we will split the data into five folds. Each fold will have a training and testing set.

	Accuracy	Standard Error	Sensitivity	Specificity	Running Time (Seconds)
LDA	0.569	0.012	0.592	0.546	0.059840s
QDA	0.565	0.008	0.647	0.496	0.069813s
KNN = 3	0.525	0.008	0.504	0.538	0.051863s
KNN = 5	0.529	0.009	0.500	0.549	0.049868s
KNN = 7	0.528	0.008	0.500	0.547	0.053853s
KNN = 10	0.528	0.008	0.498	0.550	0.056848s
Logistic Regression	0.569	0.012	0.587	0.549	0.070853s

Table 3.3. 5-Fold cross validation data accuracy, standard error, sensitivity, specificity and run time for linear discriminant analysis, quadratic discriminant analysis, logistic regression and KNN

Table 3.3 shows the highest accuracy using five-fold cross-validation LDA and logistic regression with 56.9 percent. Table 3.3 shows the highest sensitivity, 64.7 percent when using the five-fold cross-validation and the model QDA. The highest specificity is 55.0 percent with the model KNN and k=10. The standard error is similar for all the KNN models, with a standard error of 0.008 and 0.0009. Likewise, the standard error for LDA and logistic regression is the same, with a standard error of 0.012. The slowest run time is 0.070853s when using five-fold

cross-validation and logistic regression. The fastest run time is 0.049868s when using five-fold cross-validation, KNN, and k=10.

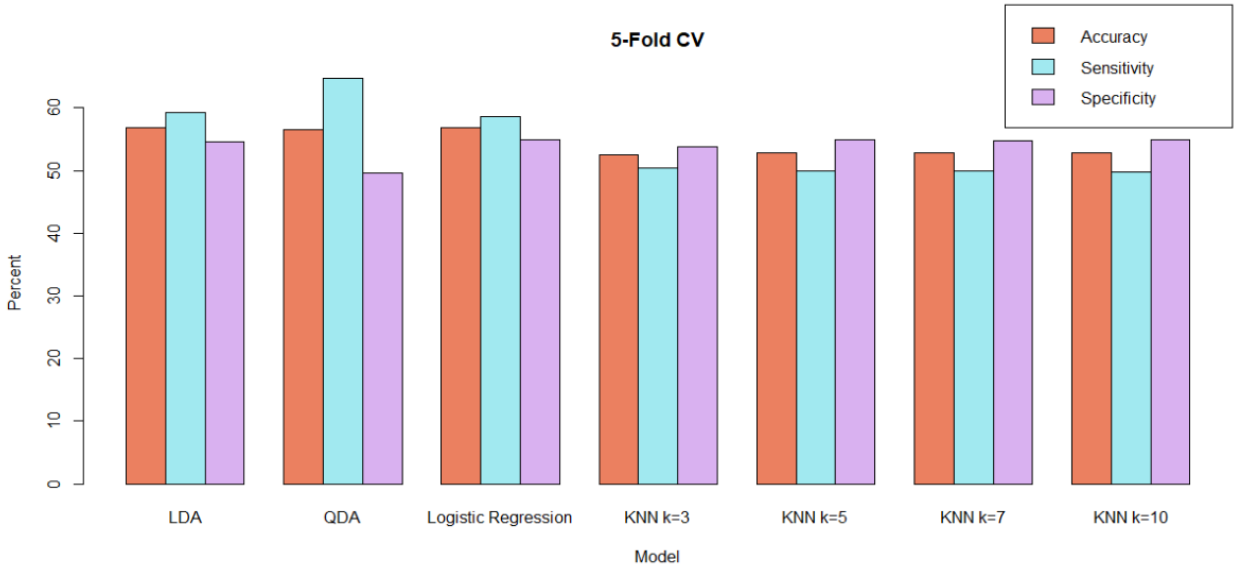

Figure 3.3 Bar plot with the accuracy, sensitivity and specificity for 5-Fold cross validation linear discriminant analysis, quadratic discriminant analysis, logistic regression and KNN

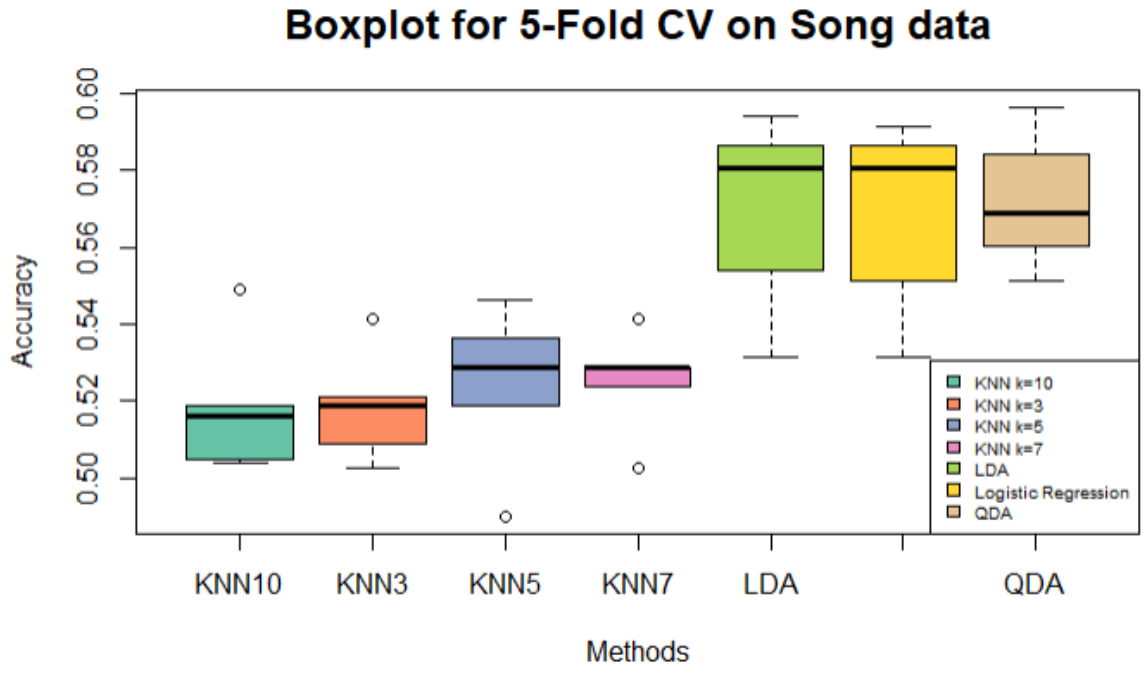

Figure 3.4 Boxplot with the accuracy for 5-Fold cross validation linear discriminant analysis, quadratic discriminant analysis, logistic regression and KNN

Figure 3.3 shows that the best model when considering all three accuracies, sensitivity and specificity is LDA and logistic regression. However, the slowest runtime is logistic regression. However, the overall best sensitivity is QDA. Figure 3.4 shows a similar finding to Figure 3.3 that the best accuracy is LDA and logistic regression

3.3 5-Fold Cross Validation Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, KNN, Decision Tree, Random Forest, Bagging, Boosting and AdaBoost

In this classification analysis with five-fold cross-validation, we will only use the numerical variables to predict whether a song has a low or high value for popularity. We will continue to use five-fold cross-validation. The data is divided into five folds. Each fold will have a training and testing set. In the previous section, we found that the best k for KNN is 5 so we will use k=5 for our KNN.

	Accuracy	Standard Error	Sensitivity	Specificity	Running Time (Seconds)
LDA	0.572	0.012	0.593	0.551	0.139411s
QDA	0.554	0.012	0.63	0.477	0.325419s
Logistic Regression	0.572	0.011	0.591	0.553	0.223734s
KNN k= 5	0.521	0.012	0.502	0.539	0.131046s
AdaBoost	0.576	0.010	0.563	0.588	2.994977s
Bagging	0.556	0.010	0.561	0.551	18.331620s
Boosting	0.576	0.012	0.566	0.586	2.735675s
Decision Tree	0.539	0.010	0.247	0.835	0.155894s
Random Forest	0.563	0.011	0.559	0.567	16.262940s

Table 3.4. 5-Fold cross validation data accuracy, standard error, sensitivity, specificity and run time for linear discriminant analysis, quadratic discriminant analysis, logistic regression, KNN, adaboost, bagging, boosting, decision tree and random forest

Table 3.4 shows the highest accuracy using five-fold cross-validation is boosting and AdaBoost with 57.6 percent. Table 3.4 shows the highest specificity is 83.5 percent with a decision tree and five-fold cross-validation. The second most significant is 58.8 percent and

model AdaBoost. The highest sensitivity is 63 percent with the model QDA. The standard error is similar for all the models, with a standard error ranging from 0.010 to 0.012. The slowest run time is 18.331620s when using five-fold cross-validation and bagging. The fastest run time is 0.131046s when using five-fold cross-validation and KNN with k=5. Although the decision tree has the highest specificity, it also has a very low sensitivity with only 24.7 percent.

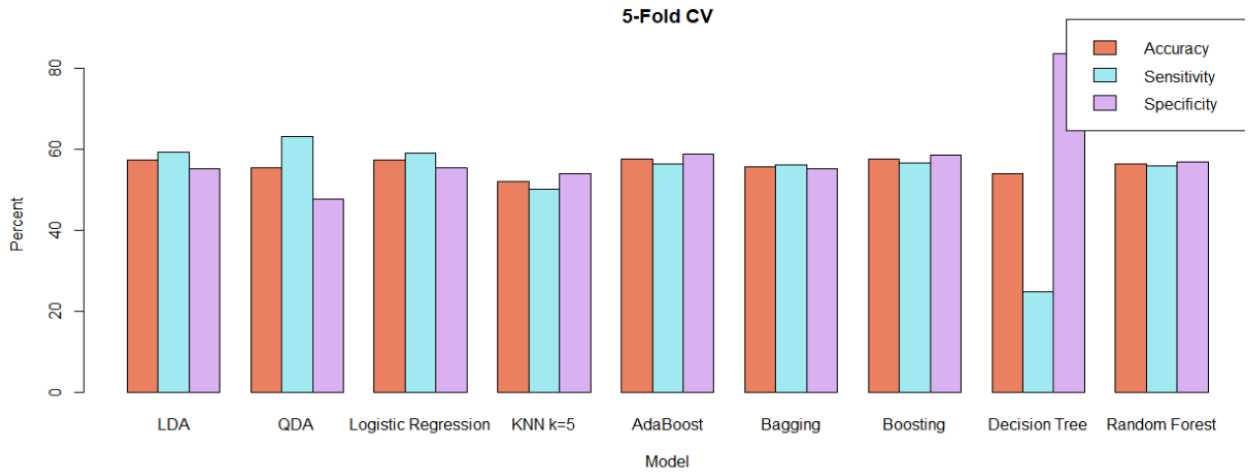

Figure 3.5 Bar plot with the accuracy, sensitivity and specificity for 5-Fold cross validation linear discriminant analysis, quadratic discriminant analysis, logistic regression, KNN, adaboost, bagging, boosting, decision tree and random forest

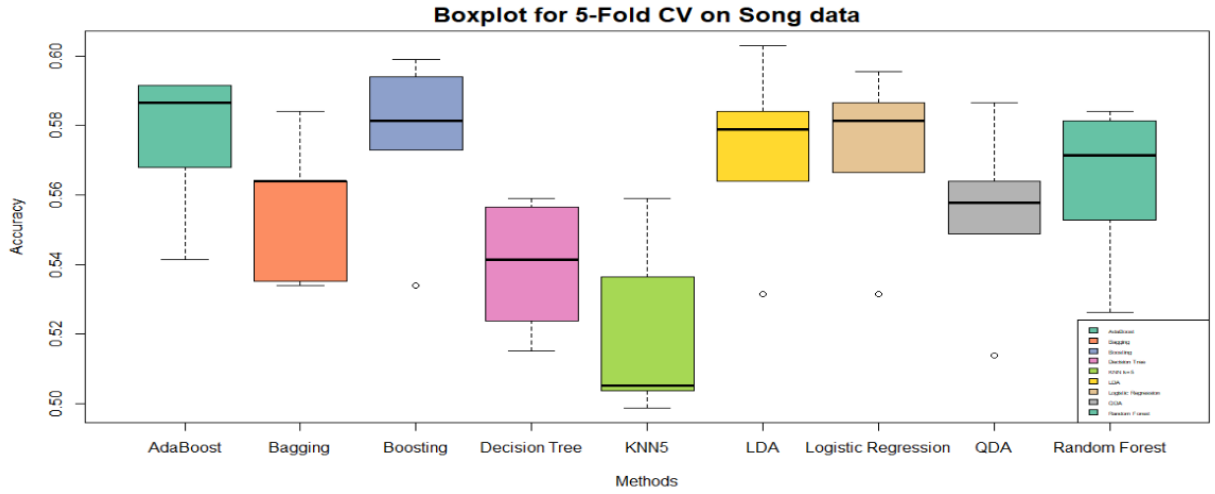

Figure 3.6 Boxplot with the accuracy for 5-Fold cross validation linear discriminant analysis, quadratic discriminant analysis, logistic regression, KNN, adaboost, bagging, boosting, decision tree and random forest

Figure 3.5 shows that the best model when considering all three accuracies, sensitivity and specificity is AdaBoost and boosting. Although, Figure 3.6 shows on average AdaBoost has slightly better accuracy. Figure 3.5 and Figure 3.6 shows that random forest is the second-best

model when considering all three accuracies, sensitivity, and specificity. However, the random forest has a long run time when compared to AdaBoost.

4. Conclusion

In the initial analysis of the data, we found that when using all of the numeric predictor variables and calculating the correlation coefficients, the most correlated variable with our response variable was danceability and loudness. Although, most predictor variables had a weak relationship with popularity. We continued our analysis, using classification models including linear discriminant analysis**,** quadratic discriminant analysis**,** Logistic Regression, and K- nearest neighbors. When splitting the data set into two halves and including the genre variable as a predictor, we found that the best model was linear discriminant analysis with an accuracy of 56.7 percent. Similarly, when using 5-fold cross-validation, the best model was also linear discriminant analysis for accuracy and specificity. The accuracy was 56.9 percent. In the final part of our analysis, we introduced five more models AdaBoost, bagging, boosting, decision tree, and random forest. When including these models, we found a slight improvement in accuracy, and the best model when predicting popularity is AdaBoost and boost. The accuracy is 57.6 percent. Overall, when predicting the popularity for the categories low and high and with all the numeric variables is AdaBoost and boosting.

Works Cited

"Spotify---All-Time-Top-2000s-Mega-Dataset." OpenML, https://www.openml.org/search?type=data&status=active&id=43386&sort=runs .