DS 740 – Data Mining and Machine Learning

Final Project Executive Summary

“A strange art – music – the most poetic and precise of all the arts, vague as a dream and precise as algebra,” a quote by Guy de Maupassant, a 19th-century French author, attempts to explain the unique nature of music with a single sentence. It is this premise that gives music such a universal power. Music knows no language or creed, nor does it see prejudice or cast judgement.

Music has been a prominent part of many people’s lives for as long as history has been recorded. Online, music streaming services allow listeners to have access to virtually all recorded music at any time, in any location. Though hard copies of music, in the form of records, CD’s, and cassettes are still used and collected, digital music has taken the globe by storm in the digital age. Convenience, cost, and mobility are major contributing factors to why this may be so. Companies like YouTube, Apple, Pandora, and Spotify have long since capitalized on people’s love for music. Music streaming companies constantly look for ways to improve their services and provide a better, customized user experience to their subscribers.

My investigation is focused on the ability to predict a song’s genre based on a series of attributes that are thought to define a song. Country and Rock music share many attributes and being able to accurately identify one over the other is highly important when providing the listener with music that matches their preferences. Though my analysis focuses only on *two* of the available genres, it can be expanded to evaluate other genres. The data scientists at Spotify (in this case) that create music selection algorithms will be particularly interested in the answer to this question. Being able to tweak and improve their predictive methods to drive a highly accurate, personalized recommendations is highly attractive for Spotify listeners. The differentiation of closely related genres can be particularly beneficial, which is the focus of this study.

Logistic Regression for classification as well as Random Forest were used to answer the question above. Logistic Regression will be used as a simple yet effective means of classification that generally performs better than KNN when the response is dichotomous. Random Forest will be used to focus on accuracy when considering that there might be predictors that are more informative than others with respect to the response variable. Both methods are well suited for classification and take different approaches on determining the result based on the predictor variables which will allow me to accurately identify if a song is “rock” or “country” based on the information provided.

A large data set containing information about 50,000 songs was used in this analysis. A subset of the data (10,000 rows) was used to maintain reasonable run times as well as focus on two specific genres (Rock and Country). 13 variables were included in the analysis, which was distilled down from 17 total variables and the response variable (music genre). The variables that were removed were identifying variables that were specific to each song and irrelevant for the analysis. 2 of the remaining 13 variables were categorical in nature, being Key and Mode.

Approximately 1000 values throughout the entire dataset were missing or removed and therefore imputed using the MICE package in R. Amongst those removed were negative values of song duration, which is erroneous.

Histograms for the numerical variables were created to explore the distributions of the data. Instrumentalness and Liveness were transformed logarithmically based on being right skewed. Acousticness, Energy, and Speechiness underwent Box Cox transformation based on their distributions. The Loudness variable underwent Inverse Hyperbolic Sine transformation due to its negative values/left skew.

The original values were replaced in the data set by the transformed variables.

Prior to running the full, double cross-validation process, I evaluated various tuning parameters for the Random Forest analysis. I found that the best values for ‘mtry’ were between 2-5 as represented by the following graph displaying accuracy.

Chart, line chart

Description automatically generated

This process was done external to the double CV steps to improve overall run time. No tuning parameters for Logistic Regression were evaluated.

Based on the results of the double CV, the supervised learning method, Random Forest, proved to be the most accurate method for predicting music genre with an ‘mtry’ value of 4. This method provided an accuracy of 82.19%.

Popularity (on a scale of 0-100) proved to be the most influential variable in this case, by a significant margin.

Chart

Description automatically generated

As popularity increased, so did the likelihood of being a rock song. This makes sense because rock music is generally more popular than country music, at least in terms of record sales. Base on the Boxplot below, there is a clear difference between popularity of rock and country songs.

Chart, box and whisker chart

Description automatically generated

Instrumentalness was the second most influential variable. Instrumentalness is a measure of the vocal content of a song. As the value for instrumentalness increases, so does the likelihood that the track contains no vocal content. That being said, rock songs generally contain more instrumental music, and less vocals compared to country songs which can be used as an identifier. Based on the Boxplot, there is a fair amount of overlap between instrumentalness between genres, but there is a tendency for rock songs to show greater instrumentalness and vice versa for country music.

Chart, box and whisker chart

Description automatically generated

Based on the model performance, accurately predicting music genre ≈82% of the time, the model is sufficiently accurate to be used on new data, especially given the context of the data. This model is used for predicting music, not diagnosing a disease, so 82% accuracy is perfectly fine.

There are some suggestions on how to improve model performance. Many of the numerical variables are calculated on a scale of 0 to 1, indicating more or less of a particular characteristic. Turning those variables into categorical variables based on their value could provide slightly more clear results. Additionally, exploring more types of transformation for the numerical variables could also improve the performance of the logistic regression model.

Applying the Random Forest model to other subsets of the data, other genres (rap vs pop), could also help identify which variables are particularly important between other genres. Tracking which variables are most pertinent to each genre of music could be highly valuable when tuning the model for predicting target genres.

Lastly, applying the model to the *entire* dataset would provide a holistic view of how all the genres compare to one another. All these subsequent analyses can help paint a clearer picture of what truly causes one genre to differ from one another, and which characteristic are shared between genres.

Because music is a complex art, there can sometimes be difficulty in categorizing a song into a genre. This study does a great job at quantifying, with the precision of algebra, an art, that can be as vague as a dream.