**Predicting Spotify Popularity Trend**

Summer 2022

IST707 Final Project

John O’Rourke, Michael Nunn, Eugene Johnson

John O’Rourke, Michael Nunn, Eugene Johnson

[jtorourk@syr.edu](mailto:jtorourk@syr.edu), [mnunn@syr.edu](mailto:mnunn@syr.edu), [eujohnso@syr.edu](mailto:eujohnso@syr.edu)

Spotify Top 2000 Hits

Naive Bayes, Decision Tree, and SVM

07/20/2022

| Spotify Popularity Classification | |
| --- | --- |
| **Introduction** | Spotify has plenty of room for growth. American’s listen to 32 hours of music per week, which is significantly more than the 25 hours of listening per month of its average users.  While Spotify currently controls 31.7 percent of the podcast market. The purpose of this paper is to come up with a model that can predict the popularity of a song. By being able to predict the popularity of a song based on different characteristics we will be able to help Spotify utilize its capital better when it comes to negotiating royalties for artists. Giving Spotify a competitive advantage over other competitors in the streaming music industry.  We will start this project by cleaning data we obtained which is the top 2000 track from 2000 – 2019. Utilizing techniques that have been learned throughout this quarter. Next, we will analyze the data by looking at the correlations of the attributes that have been provided to see the independence of observations. We will also investigate the normal distribution based on the popularity of the songs.    In order to create models that can accurately predict what songs are popular we will create a nominal based on the values in the popularity attribute as listed below:  0-24: will be any value between 0 and 24  25-49: will be any value between 25 and 49  50-74: will be any value between 50 and 74  74+: will be any value 75 and above  Our goal for this project is to create prediction models that can be evaluated based on their accuracy. This will allow us to confidently recommend a solution that will provide Spotify with the ability to determine which songs will be popular and increase their profitability. |
| **Analysis** | data preparation and cleaning This dataset was retrieved from kaggle. It contains the audio statistics of the top 2000 tracks on Spotify from 2000 to 2019. The data contains 18 different attributes which includes artist, song, duration\_ms, explicit, year, popularity, genre, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. Looking at the basic analysis the following visualizations.  It is worth noting that this data is taken at a point in time. Meaning that while a song may appear on the top 100 hits for a given year, Spotify’s algorithm that determines a song's current popularity looks at total plays and how recent the plays were. This results in some songs returning a low popularity rating at the time the data was taken (April 2022) whereas when they were released, they may have had a much higher popularity score.  This dataset had 0 null values but due to the nature of how the data is structured, multiple duplicate entries. This allows for further understanding into how top hits are determined, while a song may have been created and released in 2012, it can make the top hits playlist for every following year if it retains its popularity.    To make classification of the popularity scores simpler to complete, the popularity scores were turned into factors of varying labels to ensure that all value ranges are covered. data exploration A number of different exploratory analyses were completed on the data to better understand the distribution of each song's attributes. Taking a look at the most and least popular songs allows for a better look into the thought process of how the “Top Hits” are determined.      The dataset’s popularity scores are right skewed with outliers near 0 on the scale. This further confirms our understanding of how Spotify handles popularity. This will affect the results of the models run as having popularity data set on a group of songs from 2001 and taking their popularity from 2022 is not necessarily an ideal condition for analysis.   models and methods Throughout this analysis, Naive Bayes, Decision Tree and SVM modeling techniques are utilized to properly classify popularity scores based on the numeric variables. analyses GOALS and Parameters When creating the Naive Bayes and SVM models, the dataset was split between training and testing at a set seed to ensure results were reproducible. A sample was taken to ensure that utilization of the full data set in training could be avoided. This allowed for an accurate representation of the data sets to be presented without bias.  These same samples were used within the model development for the support vector machines. Due to the overall size of the data, Sigmoid was not utilized for analysis. When attempting to run an SVM Model with the excluded kernel, it was determined to be an inefficient and impractical choice for this analysis.  All support vector machine models for Linear, Radial, and Polynomial kernels were run at tuned cost parameters being either 1 or 0.01 respectively. This allowed for insight into any possible variance in model accuracy.  Decision Tree modeling was staged at a maximum depth of 2 with a split of 100. This resulted in inefficient models due to hardware availability and overall resource usage. Analysis was not completed with Decision Tree after a full 24 hours of the model running. |
| **results** | technical results *Linear Regression*  The first model created was used to gather insights on the data. A linear regression model allows for knowledge on the dataset that impacts the decision for which variables are utilized in the classification models. Based on the results, the variables used for classification are danceability, loudness, duration in milliseconds, energy, and key. By looking at the correlation between each attribute, the results display that there is no real significance in the attributes either positive or negative toward the popularity.    After evaluation of the attributes, further testing was completed on the model for linear regression.    The result shows that based on the estimates, energy has the most impact on popularity then any of the other attributes at -7.99%, there is a standard error of 21.3 and a p-value of 0.03854.  *Decision Tree*  A successful Decision Tree model was not developed. Due to hardware restrictions and the complexity of the dataset, decision tree models being run at a Max Depth of 2 were taking over 24 hours to complete. This is inefficient and any results that would have been found with the model completed, would be immediately removed from any analysis.  *Naive Bayes*  Our first attempt at creating a Naive Bayes model was using the popularity label and factor as is. While this should have given us the best possible outcome it did not. The model we produced was only able to predict the popularity label with a 0.8% accuracy. In order to improve this factor we decided to change the value to a nominal and see which result would lead us to a more accurate model. With our first attempt we decided to use teens for the value which becomes a factor of 10.  This gave us a better accuracy than before at 14.3%. Looking for a more accurate model we decided to keep reducing the factor and see how that would improve our accuracy. The next attempt was to use a factor of 5. Once again this improved our accuracy to 28.75%.    Our final attempt to improve our accuracy was to use 4 factors and that gave us an accuracy rating of 54%.      Looking at the confusion matrix shows that our model had the most trouble predicting the values between 50% - 74%. If we were to reduce the number of factors to 3 we would probably get a more accurate model that would not be best for Spotify as they would need to be able to sort songs into different classifications that would show songs that are extremely popular versus songs that are moderately popular.  *Support Vector Machines*  We thought it would be best to tune our model to find out what the best cost would be for producing our SVM model. We tried ranges from 0.001 to 100 and we also used different kernels to see which would give us the best results. The attributes that we used to tune were the danceability, loudness, duration\_ms, energy, and key. First kernel we used was linear and there was no difference between any of the cost variables so we decided to use a cost of 0.001.    Second kernel we used was radial. Based on the result below you can see that the error was lowest with a cost of 1 so we used that cost for our SVM model.    Third kernel used was polynomial and looking at the evidence below you can see that the best cost was either 0.001 or 0.01 both produce the same error level and dispersion so we decided to use a cost of 0.001 for our SVM model.    We ran our SVM models for each different kernel and you can see that when we used the linear model the number of support vector was 822, when we look at the radial the number of support vectors was 932 and when looking at the polynomial the number of support vectors was 122. Since the polynomial provided the least amount of support this was the SVM model that we decided to use.        We choose to use the polynomial kernel because it used only 122 support vectors    The graph above shows how our SVM model using a polynomial kernel predicted our test data. But if you look at our table below you can see that this model did not do a great job of predicting the test data. Our model was only able to achieve an accuracy rate of 20.3%. This rate is far below that of the Naive Bayes model so we cannot suggest using this model to predict the popularity of the song for Spotify. We did not find any value in the Radial and Linear models, did not do any kind of prediction, it just put each value in the 50-74 label so we feel there is no value in these models. |
| **conclusions** | In conclusion we attempted to predict whether a song would be popular based on the song's danceability, the duration of the song, how loud the song is, the energy produced by the song and what key the song was produced in. This led us to creating models that used Decision Trees, Naive Bayes, and Support Vector Machines. Based on our research and data analysis we were able to determine that the best model for predicting songs popularity was Naive Bayes  While the accuracy is not what we were hoping for we think that the best way to improve upon our model is to first looking at an associate rule mining attempt to tell us more about the attribute that was provided in our data sample so that we can better determine which attribute would and their correlation would be best for determining our models.  Given the dataset with many attributes singling out the correct attributes to determine the popularity of the song was invalid. This analysis will provide a future perspective of how Spotify will determine the popularity of an artist. Given the dataset and Spotify’s determination of how an artist is determined as a popular hit is relatively overlooked in retrospect to the other attributes that is given.  In theory Spotify’s parameters set to determine a popular song are irrelevant as a popular song that is the top number one. For example the number one hit in 2005 might not be the number one hit in 2020. Since Spotify has taken an analysis of what is present. For future analysis a reformatted dataset of popular songs taken in present time will be more ideal.  With this being said, by creating a model determining how certain attributes of a song can positively or negatively affect its popularity, this leaves room for a possible gamification of the music streaming algorithms for giants in the space like Spotify or Apple Music. By allowing an artist to see into what attributes their next song must have to make it a top hit, they could effectively fine-tune their writing and performance to ensure they beat out other artists. This could have a negative impact on the music space as whole depending on how much this may catch on. |