Requiem for a stream: Analyzing Spotify Song Attributes as Indicators of Popularity using Machine Learning Methods

Ethan Kallett, Sabrina Peltier, Sabhya Raju, Rohin Shivdasani

12/05/2021

Contents

1 Introduction **2**

2 Literature Review **3**

3 Data Review **4**

4 Methods and Findings **5**
4.1 Clustering . . . 5
4.1.1 Standardize Data . . . 5
4.1.2 Selecting K . . . 5
4.1.3 Clusters and Popularity . . . 6
4.1.4 Naming Clusters . . . 6
4.1.5 Visual Analysis . . . 7
4.1.6 Conclusion . . . 8
4.2 Continuous Response . . . 8
4.2.1 Overview . . . 10
4.2.1.1 Ordinary least squares regression . . . 10
4.2.1.2 Best subset selection . . . 10
4.2.1.3 Ridge regression . . . 10
4.2.1.4 Lasso regression . . . 10
4.2.1.5 Random Forest . . . 10
4.2.2 Findings . . . 10
4.2.2.1 Test MSE . . . 10
4.2.2.2 Coefficients from regressions . . . 11
4.3 Discrete Response . . . 12
4.3.1 Overview . . . 12
4.3.1.1 Logistic regresion . . . 12
4.3.1.2 Decision tree . . . 12
4.3.1.3 Random forest . . . 13
4.3.1.4 Boosted tree . . . 13
4.3.2 Complications with Discrete Response . . . 13
4.3.3 Findings . . . 14
4.3.3.1 Model fit . . . 14
4.3.3.2 Model interpretation . . . 15

5 Conclusion **17**

6 Individual Contributions **18**

1 Introduction

As the music industry has moved onto the internet, streaming platforms like Spotify have increased access to a diverse array of songs and allowed many songs to reach new heights of popularity. With the end of the year approaching, many of us are encountering our connections on social media sharing their Spotify "Wrapped" which is a review of the user's listening during the year. With more music released annually than any time in history, music consumers are empowered to search, discover, and follow disparate songs, artists, playlists. With such a broad ability for consumers to sample songs, we are interested in studying how songs' intrinsic feature qualities and attributes dictate (or not) their popularity.

Media in other forms like but videos across platforms such as Facebook, TikTok, and Youtube, has evolved around the consideration that creators have a greater understanding of the social media sites' algorithm for promoting and recommending content to viewers. That is, content creators have a deep understanding of the specific video attributes that they believe effect popularity; with Mr. Beast on Youtube as perhaps the most notable example. This has resulted in creators and producers fine tuning their content for attributes such as the video length, thumbnail images, title, upload consistency, and more. However, the same "popularity hacking" mentality is less prevalent and pervasive across the music industry.

As streaming platforms like Spotify make available more specific data and the relationship between song attributes and popularity is better understood over time, we believe we will see artists adopt a similar "hacking" ethos in how they create songs by premeditating some of the qualities analyzed in this paper, such as energy or speechiness, into their song in ways they wouldn't otherwise. This leads us to our research question: What makes a song popular?

There is limited research on the relationship between song audio features and popularity as measured by

stream count which makes this study important and crucial. Record companies, Sportify, and artists can utilize the findings from this research to better understand and predict how songs climb to the top of charts.

We approach this question with a multi-faceted analysis. First, we analyze popularity as a discrete variable – whether a song ranked on the Billboard Hot 100 chart, achieved a Grammy award or received over a million streams over 2 weeks. Then, we analyze popularity through stream counts as continuous, specifically using ordinary least squares regression and random forest, as well as selection and shrinkage of methods LASSO and Ridge regression selection methods to elucidate the ability to song attributes as predictor variables of streaming volumes on Spotify. We also deploy unsupervised learning classification methods to investigate the broad popularity outcomes associated with different music genres.

Our findings indicate that song attributes consistently show they are strong predictors of song popularity by as defined by presence on the Billboard Hot 100 ranking, but less reliably for predicting a Grammy Award or predicting the number of streams achieved on Spotify. Moreover, the presence of heterogeneity of the treatment effect, or in other words certain features being beneficial in certain genres and less so in others, really challenge the significance of the findings of this study and necessitate a fair amount of future research to be conducted.

2 Literature Review

The paradigm of popularity in music and songs has been the subject of extensive academic interest across academic disciplines, including but not limited to data science. The particular question of focus in this paper, which is to investigate the determinants of song popularity using song attributes, features, and qualities, has been a widely studied question, as have other considerations surrounding song popularity and its determinants. The specific research field on this topic is called Hit Song Science (HSS).

Previous research most directly similar to the analysis we conduct in our paper utilize standard measurements of song features and attributes to compare their impact on popularity across songs across songs. This research relies most often and heavily on the same song data we use in our paper such as speechiness, acousticness, valence, duration, acousticness, etc. measured by the Echo Nest Corp. until they were acquired by Spotify in 2014. Spotify now provides this data via their API. A 2013 study by Ceulemans and Detry titled Does Music Matter in Pop Music? The Impact of Musical Characteristics on Commercial Success and Critics' Ratings examined the impact of song attributes on popularity using data from Billboard's Hot 100 rankings for popularity and song feature data from Echo Nest Corp. across a sample of 514 songs. In addition to the song attribute data, (Ceulemans and Detry 2014) custom added data about the song artist, specifically artist gender, artist nationality, and whether the artist is associated with a major label. Both the peak position achieved by a song on Billboard's Hot 100 chart as well as "survival" – the number of days a song remains on Billboard's Hot 100 chart were used as the measures of popularity. The paper found that the mode of a song and critics' rankings of the song impacted the survival of a song in charts. The paper also discussed that for a song to be commercially successful, a song must be promoted and in line with current trends.

Askin and Mauskapf's 2017 paper titled What Makes Popular Culture Popular? Product Features and Optimal Differentiation in Music instead studies the impact of artist reputation and previous success on peak position and survival on the Billboard Hot 100 Chart. The paper utilizes independent variables such as if the artist was associated with a major record label, the number of previous charting songs by an artist, and a custom criterion called "typicality" – or similarity to past successful songs from the genre, that is built using the Spotify / Echo Nest Corp. song attributes such as acousticness, valence, tempo, key, and genre. (Askin and Mauskapf 2017) find that a song's position on charts is driven by artist familiarity, genre affiliation, institutional support, and perceived proximity to its peers. The paper also found that songs that were differentiated (i.e. had a low typicality) and that did not sound too much like previous and contemporaneous productions are more likely to reach a higher position on the chart meaning

A paper by (Suh 2019) titled International Music Preferences: An Analysis of the Determinants of Song Popularity on Spotify for the U.S., Norway, Taiwan, Ecuador, and Costa Rica investigates popularity by analyzing the effect of song attributes on popularity but with a focus on their different impacts across countries. The paper examines three main various audio and artist features of songs, including whether or not

a track features a guest artist, a happiness index based on the height of valence levels, and an energy level based on abrasiveness and loudness, the latter two of which are constructed from the same Spotify-provided song attributes examined in our paper. The paper runs a 10 separate ordinary least squares regressions, using two measures of song popularity from the Spotify's Top 200 chart in each of the following 5 countries: the U.S., Norway, Taiwan, Ecuador, and Costa Rica. Popularity is measured via two dependent variables using data from Spotify's daily Top 200 chart in each of these countries: (1) a track's peak position on the chart and (2) the number of days it survives on a country's Spotify Top 200 chart. The paper concludes that in most all countries, the presence of a featured artist on a track increased both peak position and survival on the Top 200 chart. However, louder and more abrasive songs demonstrated significantly shorter chart lifespans in three of five countries: the U.S., Norway, and Taiwan. Meanwhile, happier songs achieved both higher peak chart positions and longer chart lifespans in two of the five countries: Norway and Taiwan.

In a paper titled Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market, (Matthew J. Salganik and Watts 2006) construct an artificial market to study social influence as a determinant of song success as measured by downloads. (Matthew J. Salganik and Watts 2006) found that the number of downloads affects both the degree of popularity achieved by a song with new listeners and the reported level of enjoyment listening to the song for those with prior song download number knowledge. (Matthew J. Salganik and Watts 2006) asserts that high downloads have a signaling effect and create a superstar effect, as music listeners with access to this information assume that highly downloaded songs are more worthy of being listened to. Thus, the paper concluded that in addition to personal preference and song quality, consumer song choice in the music industry is a function of the number of downloads attached to a song.

A paper by (Nijkamp 2018) called Prediction of product success: explaining song popularity by audio features from Spotify data, researched the relationship between song audio features like key and tempo from the Spotify database and song popularity measured by the number of streams a song has. The research used a novel attribute-approach to study whether these audio features could account for the number of streams. The paper finds that audio features can explain a higher number of streams only moderately.

3 Data Review

This project principally relied on four different datasets sourced from Kaggle. The first Spotify Charts contained daily observations from 2017 through the end of 2019 of the most popular 200 songs in 50 distinct countries/regions around the world. It was absolutely essential helping us calculate the streams achieve song had achieved.

The second major data set was The Spotify Hit Predictor Dataset (1960-2019). This dataset contained information about all the song attributes we examined. The song features we examined from this dataset are: `danceability` (a 0-1 scale of how suitable for dancing), `energy` (a 0-1 scale of the song's energy intensity), `key` (mapping of keys to integers, interpreted as factor), `loudness` (overall loudness in decibels), `mode` (major or minor key, interpreted as factor), `speechiness` (0-1 scale assessing the presence of spoken words), `acousticness` (0-1 scale of the confidence that the track is acoustic), `instrumentalness` (0-1 scale measuring the lack of spoken words), `duration_ms` (duration of the song in milliseconds), `liveness` (0-1 scale assessing the presence of an audience in the recording), `valence` (0-1 scale assessing the musical positiveness conveyed by the track), `tempo` (measured in beats per minute), `time_signature` (beats per bar, interpreted as factor), `chorus_hit` (the author's best estimate of when the chorus begins, measured in milliseconds), and `sections` (number of sections in the track).

The third major dataset was Data on Songs from Billboard 1999-2019. This contained the response variable of whether the song had been featured in the Billboard Hot 100 or not. This data set also contained the artist metadata `Followers`, `NumAlbums`, `YearFirstAlbum` (which we converted into `years_since_1st_album`), `Gender` (converted to `is_male`) and `Group.Solo` (converted to `is_group`).

Finally, we used the dataset The Grammy Awards which contained the list of all the Grammy awards given since its inception in 1958.

Table 1: Summary of Key Features

Features	Source
streams	Spotify Charts
danceability	The Spotify Hit Predictor Dataset
energy	The Spotify Hit Predictor Dataset
key	The Spotify Hit Predictor Dataset
loudness	The Spotify Hit Predictor Dataset
speechiness	The Spotify Hit Predictor Dataset
acousticness	The Spotify Hit Predictor Dataset
instrumentalness	The Spotify Hit Predictor Dataset
liveness	The Spotify Hit Predictor Dataset
valence	The Spotify Hit Predictor Dataset
tempo	The Spotify Hit Predictor Dataset
duration_ms	The Spotify Hit Predictor Dataset
time_signature	The Spotify Hit Predictor Dataset
chorus_hit	The Spotify Hit Predictor Dataset
sections	The Spotify Hit Predictor Dataset
Followers	Billboard 1999-201
NumAlbums	Billboard 1999-201
years_since_1st_album	Billboard 1999-201
is_male	Billboard 1999-201
is_group	Billboard 1999-201
billboard	Billboard 1999-201
grammy	The Grammy Awards

In summary, the key features used in this study can be seen in Table 1.

4 Methods and Findings

4.1 Clustering

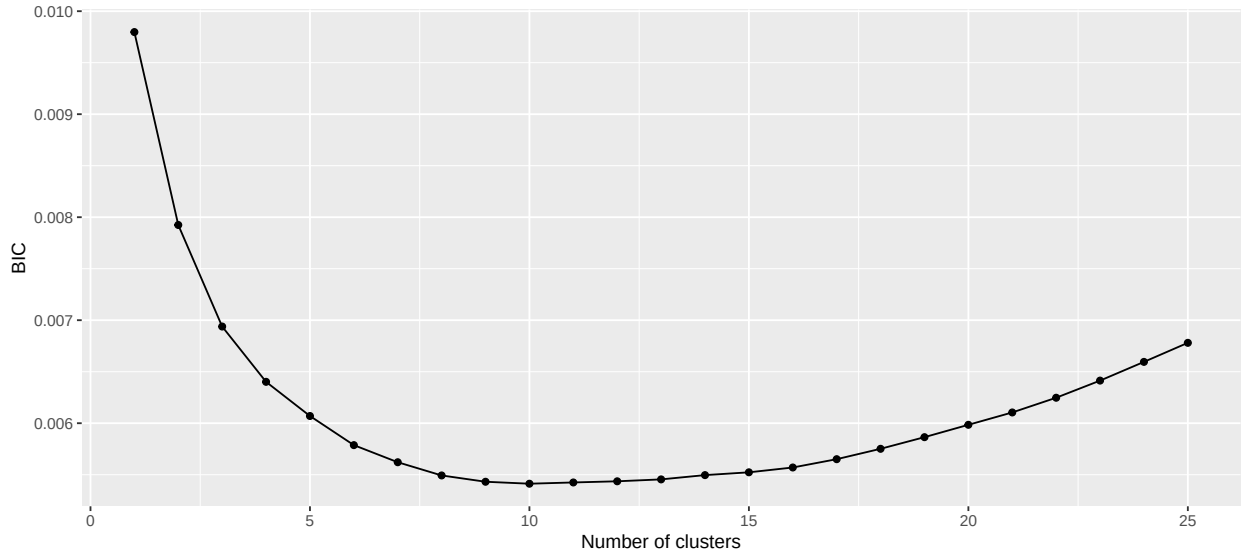
4.1.1 Standardize Data

4.1.2 Selecting K

We will begin our analysis by applying clustering to our data to investigate how popularity outcomes differ with music genres. To select the number of clusters we used BIC, where

$$\text{BIC} = \text{RSS} + S\frac{\ln n}{n}$$

RSS, the residual sum squared errors, measures the error which is the distance an observation is from its cluster center. Looking at a plot of cluster numbers and BIC, 10 clusters were selected. This value was chosen based on what appeared to be the elbow of the graph and reasonable to analyze.



4.1.3 Clusters and Popularity

Once the clusters were determined, the variables "track" and "artist" were added back to the scaled data. The "scaled_clustering" data included the track name, artist, scaled attributes, and the cluster assigned to each song. With this data set, we were able to combine it with the larger aggregated Spotify data set which includes the streams for each song. With this we were able to group all songs by their cluster and find the average stream per song in a given cluster. This tibble shows there is a clear discrepancy in popularity between clusters. The most popular (cluster #1) having an average of 138,970,246 streams per song and the least popular cluster (10) having an average of 10,460,082 streams per song. This indicates the similar qualities of songs can contribute to the amount of streams it will receive. Looking at the number of songs in each cluster, there is a range from around 500 to around 10,000. Unsurprisingly, the cluster with the highest average stream per song has the third most songs in it's cluster (6,193), indicating artists are creating more songs with attributes that are correlated with more streams. Also following this pattern is cluster #9 which has the largest number of songs and has the second highest average number of streams per song. However, the least streamed cluster falls in the middle of the ranking with 3,358 songs in the cluster, meaning there are decent amount of songs with less streams.

Table 2: Custer Popularity, Average Streams per Song

Cluster	Streams
4	35658083
7	33752641
1	30608114
2	28226636
8	26586400
9	25993999
3	18510218
5	17459569
10	13325158

4.1.4 Naming Clusters

Next, I examined a sample of songs selected for each category to see how they could be qualitatively described. From each cluster I looked at the first 25 songs to gain insight into the types of songs. From this I was able to title the different clusters.

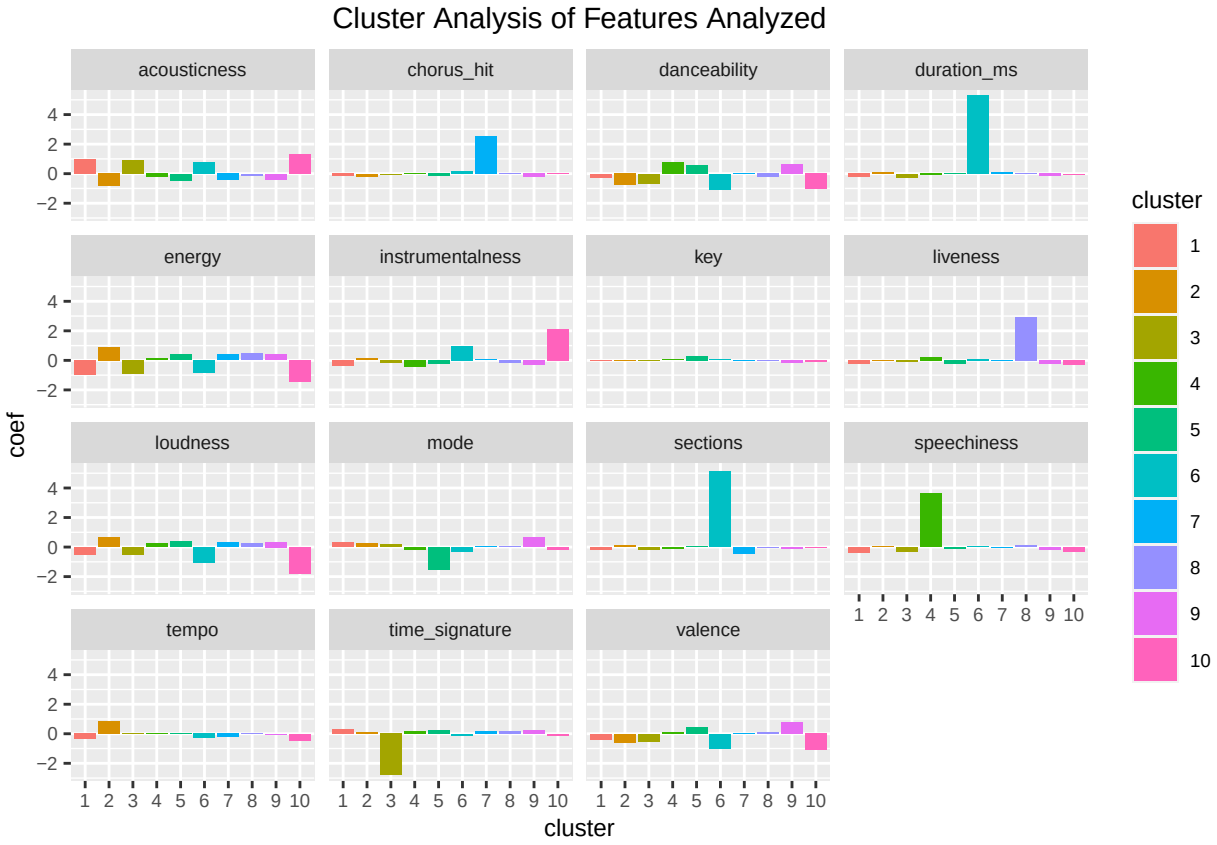
Table 3: Subjective Naming of Clusters

Cluster	Style/Type/'Vibe'	Observations per Cluster
1	Slow, Soul, Breakup	6193
2	Dark, Epic, Video-Game	5499
3	Slow Country, Movie Montage	2545
4	Trendy, Hip-Hop, Club	1715
5	Pop Rap, Dance, Less Well Known	6522
6	Orchestra	551
7	Harsh Rap, Intense, Metal	2154
8	Indie Rap, Alternative	2335
9	Popular, Radio, Head-Nod (Not Dance)	10234
10	Background Music	3358

A challenge presented itself when working on this section because, despite there being massive amounts of data, there is still an artistic quality to music. It was easy for me to feel how two songs were similar, but putting it into words was a bit difficult. Table 3 summarises my conclusions. Looking at the words used to label the clusters and popularity, something that surprised me is the "breakup" style cluster (#1) had the most streams per song. However, it does make sense since the artists in this cluster, such as Cole Swindle, One Direction, Carrie Underwood, and Sam Smith, are extremely well known. Unsurprisingly, cluster #9, which I described as popular, radio music, was second on the list of average streams per song. Another observation was the lack of streams per song for cluster #10 which could only be described as background music. Another interesting aspect of this section to note is that a measure of popularity was not used to cluster these songs, however there were observable trends and well known songs were clustered together. For example, cluster #5 ranked low in terms of average streams per song and the artists in the cluster are not mainstream. This indicates there are measurable attributes of a song that can contribute to it's success or failure.

4.1.5 Visual Analysis

Finally, we are created a graphic to explore the differences in measured attributes across clusters. To do this, we are grouped the songs by cluster and found the mean value for each song component for each cluster using the standardized data.



The duration for cluster 6 is much higher than any of the other clusters, which makes sense since classical music tends to be long. A similar observation can be made for the variable "sections" where cluster 6 is significantly higher than other clusters.

Energy is similar across clusters, however we observe a slight dip in clusters 1, 3, 6, and 10. This matches the descriptions given (soul, slow country, orchestra, background music).

Another interesting observation is the time signature graph. All the clusters have similar, near zero, values for the average time signature. However, cluster #3's average is significantly lower. When classifying cluster #3 I found it difficult to pinpoint what the exact "vibe" was and how this group differed from others. The data indicates the "feeling" cluster #3 was different than others was most likely due to the low time signature.

4.1.6 Conclusion

K-means clustering was able to group similar songs together based on their quanatifable attributes. This method is applicable to real life because Spotify offers multiple daily playlists unique to each user (Daily Mix 1-5) and each playlist tends to be a different "vibe", not necessarily genre. This style of machine learning could be used to create playlists for users. Spotify could also use this style of unsupervised learning to market new songs by finding their cluster and suggesting the song to people who heavily listen to that cluster. Spotify is highly specific and goes beyond traditional genres, such as "emo", to create more precise genres, such as "midwest emo". Since Spotify has a massive number of songs available, many more clusters could be added to find hyper-specific groups of similar songs for users. This style of machine learning has the potential for users to be matched with large of songs they enjoy based on qualities it is difficult to distinguish by just listening.

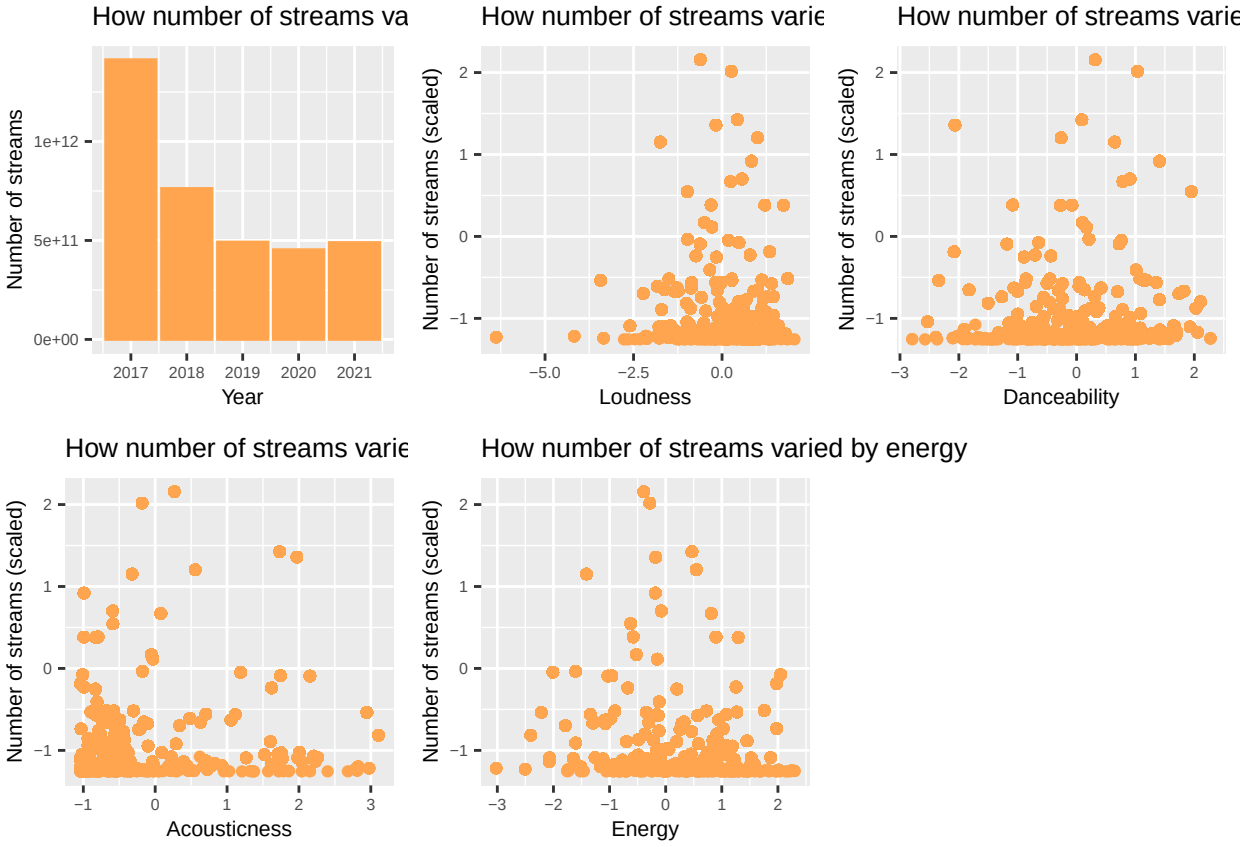
4.2 Continuous Response

Before we begin, we will complete a very basic data exploration.

Table 4: Summary Statistics of Features

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
sum_streams	26949	133958526.755	106459997.413	147537	51700652	205325705	363547095
danceability	26949	0.651	0.142	0.255	0.56	0.759	0.974
energy	26949	0.587	0.17	0.073	0.489	0.726	0.978
loudness	26949	-6.983	2.404	-22.32	-8.465	-5.069	-2.063
mode	26949	0.706	0.455	0	0	1	1
speechiness	26949	0.127	0.12	0.025	0.04	0.185	0.461
acousticness	26949	0.241	0.231	0	0.058	0.352	0.957
instrumentalness	26949	0.003	0.024	0	0	0	0.837
liveness	26949	0.144	0.098	0.037	0.09	0.164	0.979
valence	26949	0.452	0.204	0.05	0.288	0.599	0.978
tempo	26949	124.179	30.591	48.718	99.933	147.073	191.9
duration_ms	26949	220085.041	43636.128	107147	201159	240760	484147
chorus_hit	26949	42.209	18.166	16.73	29.659	50.233	114.411
sections	26949	9.908	2.438	3	8	11	17
year	26949	2018.279	1.469	2017	2017	2019	2021
month	26949	6.084	3.464	1	3	9	12
day	26949	3.951	1.998	1	2	6	7

Now, we will look at the distributions of a key features in relation to number of streams.



4.2.1 Overview

For this part of the study, we attempted to study song popularity as a continuous variable. We proxyed for popularity of a song by looking at its number of streams. The number of streams of a song is a straightforward and intuitive metric to studying how popular a song since it indicates how well liked a song is for users on Spotify. A limitation of using this metric is that it does not provide insight on how many of the streams were repeat streams vs new streams over a time period.

We used the same datasets as the previous analysis with popularity as a discrete variable. Specifically, we used Spotify Charts and The Spotify Hit Predictor Dataset (1960-2019) — the two datasets that gave us information on individual songs and its streams. We did not use the datasets with information from Billboard Hot 100 and Grammies as they did not provide relevant data for our analysis, which is primarily focused on song characteristics and its popularity. The independent variables we were interested in for this section included: rank, danceability, energy, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, chorus_hit, sections, year, month, day.

We ran the following predictive models: ordinary least squares regression, best subset selection, ridge regression, lasso regression, and random forest.

4.2.1.1 Ordinary least squares regression this regression model fits the data with a regression line that minimizes the sum of the squares of the residuals. We used this regression model to look at the linear relationship between the number of streams and the independent variables.

4.2.1.2 Best subset selection involves identifying a subset of the regressors that are believed to be related to number of streams. We use AIC/BIC to determine the optimal subset size.

4.2.1.3 Ridge regression is used when data suffers from multi-collinearity, which was a very real concern in our data. We used cross volidation to find the optimal lambda for our regression.

4.2.1.4 Lasso regression uses shrinkage to predict linear relationship in data. Again, we used cross validation to find the optimal lambda for our regression.

4.2.1.5 Random Forest this classification method that uses decision trees to predict. We leveraged cross validation to find the tuning parameter mtry.

4.2.2 Findings

Table 5: Model Comparison for Continuous Response

Model	Test MSE
OLS	0.71
Best Subset	0.75
Ridge	0.71
LASSO	0.71
Random Forest	0.00

4.2.2.1 Test MSE As we can see from the above values, the test MSE is lowest for Random Forest. We were surprised to see that the test MSE for Random Forest was 0. This likely because the percentage of variance explained in the model was 99.99, which is incredibly high for a random forest model. Apart from Random Forest, the Lasso regression model produced the lowest test MSE and had a very similar test MSE value to Ridge regression. Both Ridge and Lasso regression marginally improved on the OLS regression's regression fit suggesting that regularization did not make a significant difference in our problem. The low test

MSE of OLS regression suggests that a linear relationship between number of streams and the independent variables well explains the relationship between the two in our analysis. The Best Subset model also produced a low test MSE, which makes it a very comparable model to the rest of the regression models. Overall, apart from the tree regression method Random Forest, the rest of the regressions had very similar (low) test MSEs which suggest that their fit to the data is high. That is, linear regression models are able to explain the relationship well.

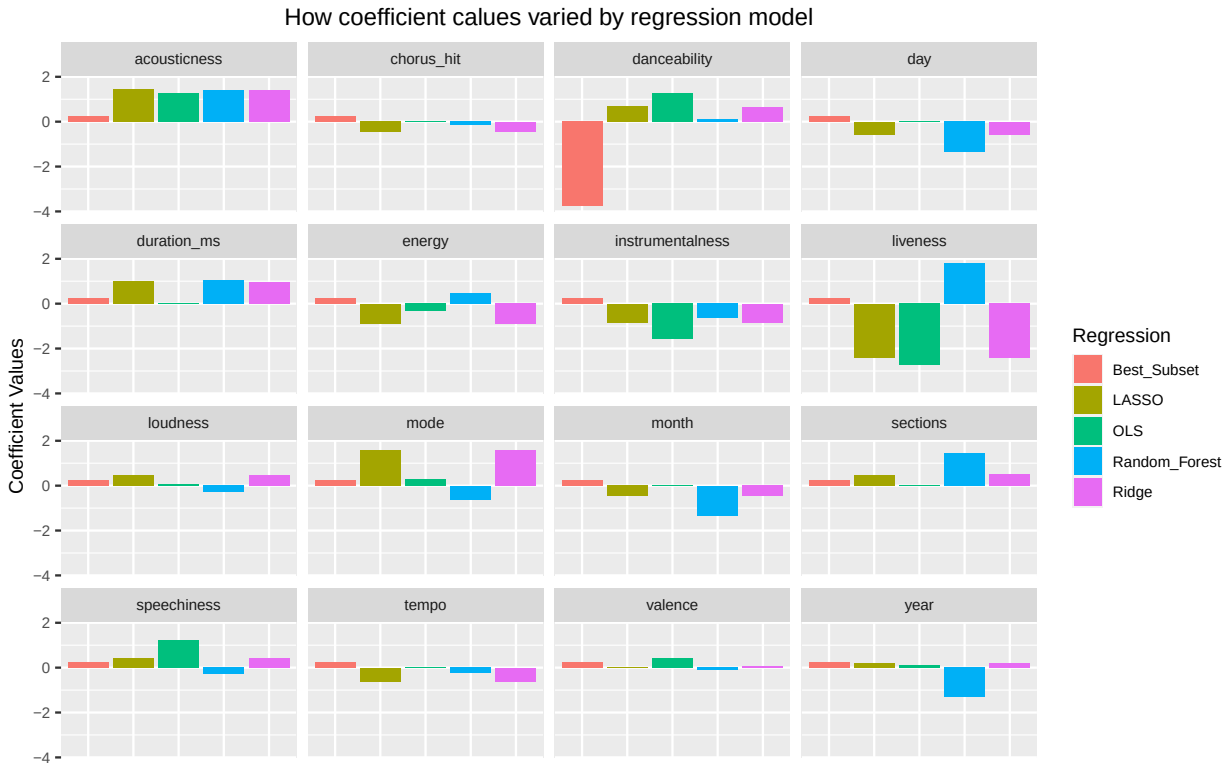


Figure 1: Comparison of Model Coefficients

4.2.2.2 Coefficients from regressions From looking at the above plots, we can immediately tell that there is quite a bit of variability in the coefficient values from each prediction model. We see that acousticness is an important feature in all the models with a relatively high positive coefficient value. Further, it is interesting to see that this coefficient values are similar across the various models we used. This suggests that the higher the confidence of a song being acoustic (measured by acousticness feature), the higher the chance the song is popular. Another interesting finding from this analysis is that danceability's coefficient value is varied across models, in fact, four of the five models indicate a negative relationship between danceability and number of streams. Only, random forest includes a positive coefficient value for danceability. The coefficient values themselves are higher in absolute terms than other coefficients suggesting the importance of a song being danceable in its popularity. Since random forest had the most accurate predictions and its model included a large positive coefficient value for danceability we can say with confidence that this feature is a key indicator. Unsurprisngly, we also see low coefficient values for day, month, and year suggesting that the date of data capture is not revealing of a song's stream count. Loudness and sections are two other coefficients that had low values across our regression models. Interestingly, our random forest model and linear models disagreed in regards to the sign substantially and value frequently of the coefficients. Nearly 12 of the 16 coefficients had a different sign and significant difference in value when comparing the random forest model to the linear models.

Overall, we see that the variables with the greatest net positive influence are `acousticness`, `duration_ms`,

`sections` and the variables with the greatest net negative influence are `instrumentalness`, `liveness` and `tempo`. In contrast to the result from (Nijkamp 2018), our findings suggest that audio features explain the number of streams sizably, especially our random forest model.

4.3 Discrete Response

4.3.1 Overview

This part of the study attempted to assess song popularity as a binary value taking either 0 (unpopular) or 1 (popular). This binary value was calculated using a variety of metrics.

The first criterion took the value of 1 for a song that won a Grammy award (any award for any category in any year since 1960). We understood that if we selected just one kind of Grammy award, our results might be biased towards the features that made a song popular in that category, rather than creating a robust model that is generalizable to the overall features which make a song popular.

The second criterion is whether or not the song was featured in the Billboard Hot 100. This criteria was equal to 1 for any song that was featured on the listing for any amount of time from any year between 1999-2019.

Finally, the third criterion we considered looked at the streaming history of a song on Spotify. We first calculated the number of days that a song achieved more than 1 million streams globally. Then, looking at the distribution of days that this song had over 1 million streams globally, we decided that a fair threshold for our purposes would be 14 days, i.e. the song popularity would take the value 1 if there were 14 days in which the song achieved over 1 million listens on Spotify and 0 otherwise. These days did not have to be consecutive and the 1 million streams could be distributed among any number of countries/regions.

We believed that the three criteria we were considering were different enough that it did not make any sense to aggregate them into one single metric for song popularity. Rather, at every step of the way, we ran each model on each of the separate criteria and interpreted results. Furthermore, we understood the importance that artist metadata (such as number of followers on Spotify, length of musical history, gender, etc.) could have on song popularity. For that reason, for each formulation and each response variable, we created two models, one for the artist metadata and the song and one that simply looked at the song-level variables. All told, this meant that six different models were ran for each formulation, which gave us the ability to look at any given model's findings with scrutiny.

Table 6: Description of six models run

Models without Metadata	Models with Metadata
Grammy as Response	Grammy as Response
Billboard as Response	Billboard as Response
Streams as Response	Streams as Response

Table 6 summarises the different models we calculated for every different model formulation. For each grouping of data to build these models, a subset was withheld as validation data.

We then used four different formulations to calculate our predictive models: logistic regression, decision tree, random forest and boosted tree.

4.3.1.1 Logistic regresion As expected, we computed six OLS logistic regressions. We decided to run this OLS logistic regression to give ourselves a baseline and understand how much improvement on the standard error we were getting through the different tree methods we were trying.

4.3.1.2 Decision tree As trees are one of the most readable regression models, we decided to make decision trees for each model. To avoid overfitting, these trees were then pruned to an optimal size, with the parameter penalizing tree size chosen through cross validation.

4.3.1.3 Random forest Random forests were grown for each model. Again, the tuning parameter *mtry* was found through cross validation.

4.3.1.4 Boosted tree For the boosted tree models, the parameters number of trees M, learning rate λ and tree size $|T|$ were chosen through cross validation. During parameter tuning, the learning rate λ was held fixed at 0.01 and different tree sizes $|T|$ 1 through 6 were tried. With each combination of λ and tree size $|T|$, the optimal number of trees M was chosen. For a few models which, from the cross-validation results, seemed to benefit from the complex interactions of many different variables, the values 3 through 8 were tried for optimal tree size.

4.3.2 Complications with Discrete Response

The first complications that arose from this discrete methodology was the massive class imbalance that was presented in some of the data sets.

Table 7: Class imbalance in the final data

Num Songs in Data	Award Criterion	Metadata?	Num Awards in Data	% in Positive Class
41106	Grammy	Not Incl	104	0.25
41106	Billboard	Not Incl	20553	50.00
5020	Grammy	Incl	51	1.02
5020	Billboard	Incl	4847	96.55
190	Streams	Incl	66	34.74
495	Streams	Not Incl	209	42.22

The class imbalance was particularly a problem for the Grammy data set, since so few Grammy awards matched songs in the other data set (not to mention that few Grammy awards have been given in total). Using a song's presence on the Billboard Hot 100 as a criterion was not necessarily a criterion until joining the song with the artist metadata, which reduced the negative class considerably. It is intuitive, though, that the songs that are less famous would have less information about their artists. In addition to the class imbalance, we can see a lack of observations in the Spotify data in general. This is especially accentuated after joining the Spotify stream data with the artist metadata. However, because of the criterion chosen, class imbalance is less of an issue.

In order to best deal with these issues, the training data set downsamples from the majority class by a substantial margin. Moreover, it samples from the minority class twice. This procedure was done so that the a proportion (at least 20%) of the data was held in reserve for validation.

Beyond issues with the data itself, it turned out that the decision tree was a terrible model for the data.

Table 8: Tree size for the decesision trees calculated on different partitions of data

Metadata Included?	Grammy Award	Billboard Hot 100	14 days of 1M Streams
without metadata	2	8	3
with metadata	16	13	1

Some of these trees were concerningly small, in particular the tree fit to the `Grammy without Metadata` data and the tree fit to the `Streams with Metadata` data, which, after CV pruning, was reduced to a single node.

4.3.3 Findings

4.3.3.1 Model fit Plotting the ROC (receiver operating characteristic) curves reveals that some models did significantly better than others.

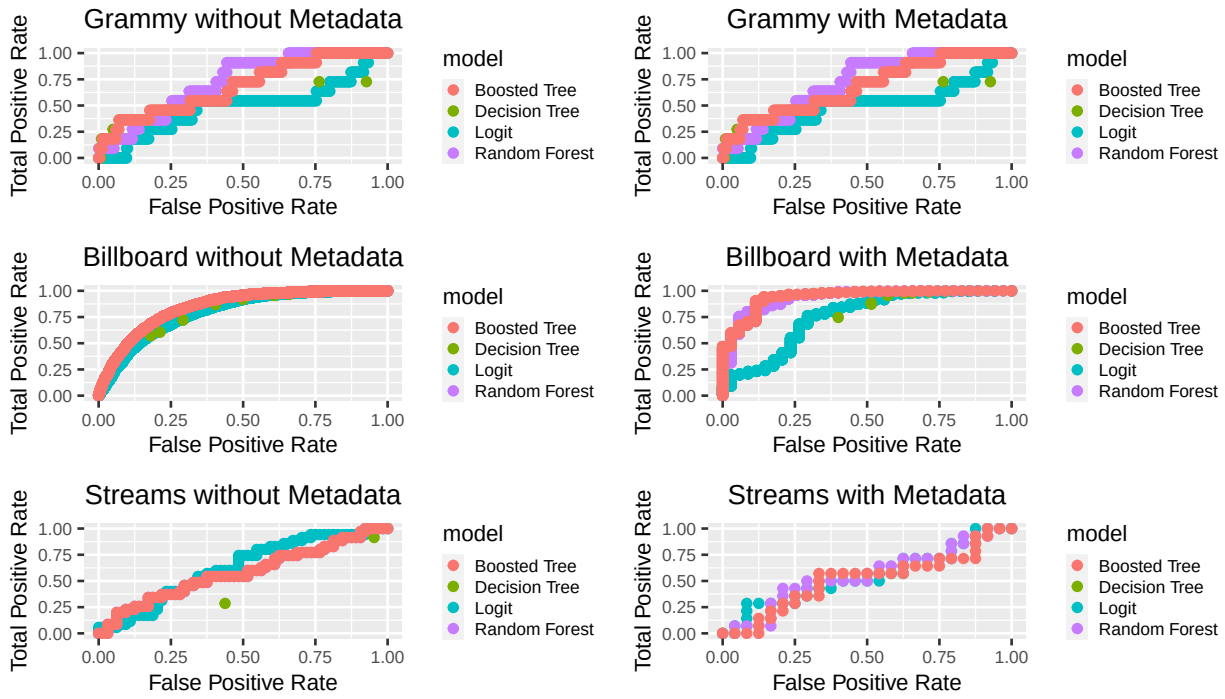


Figure 2: ROC curves of various models

First, the ROC curves show the limitations of the decision tree model. Because of the simplicity of the models, only a scarce handful of values can be plotted for them.

Moreover, though the Grammy prediction models seem to fare slightly better than the models with Streams as a criterion, the neither of the criteria have been able to generate a model with good predictive capacity. Especially with the Streams as a criterion, all of the ROC curves appear very close to the 45 degree line, suggesting that their predictive power is not much better than just randomly guessing. In fact, we can even see that in the logistic regression fit to the `Grammy with Metadata` data, the model does slightly *worse* than random (which, I guess, in theory is a better classifier than one that classifies randomly).

Table 9: Area under the curve of ROC curves

Model	Grammy	Billboard	Streams	Grammy w Meta	Billboard w Meta	Streams w Meta
Logit	0.743	0.798	0.625	0.494	0.754	0.548
Random Forest	0.781	0.845	0.580	0.723	0.934	0.554
Boosted Tree	0.759	0.845	0.580	0.683	0.942	0.515

The models fit to the `Grammy without Metadata` set did decently (perhaps because this set had many more observations than the data set with metadata.) Moreover, the Random Forest classifier on the `Grammy with Metadata` set performed reasonably well, but considering how small the sample size was and how poorly the other models performed, we can conclude that perhaps this is just random chance.

What is immediately visible is that the models fit to the data with `Billboard Hot 100` as a criterion fared much better than the other models. This could potentially be indicative of the fact that there is more coherence in the factors that get a song to this list, or it could be indicative of the fact that, with more and less lopsided data, the models can be better fit. Additionally, for whatever reason, in the `Billboard without Metadata` data, the random forest ROC curve is completely identical to the boosted ROC curve. Regardless, this increased classification accuracy means that the findings from the Billboard data should be taken more seriously than the findings from the other models.

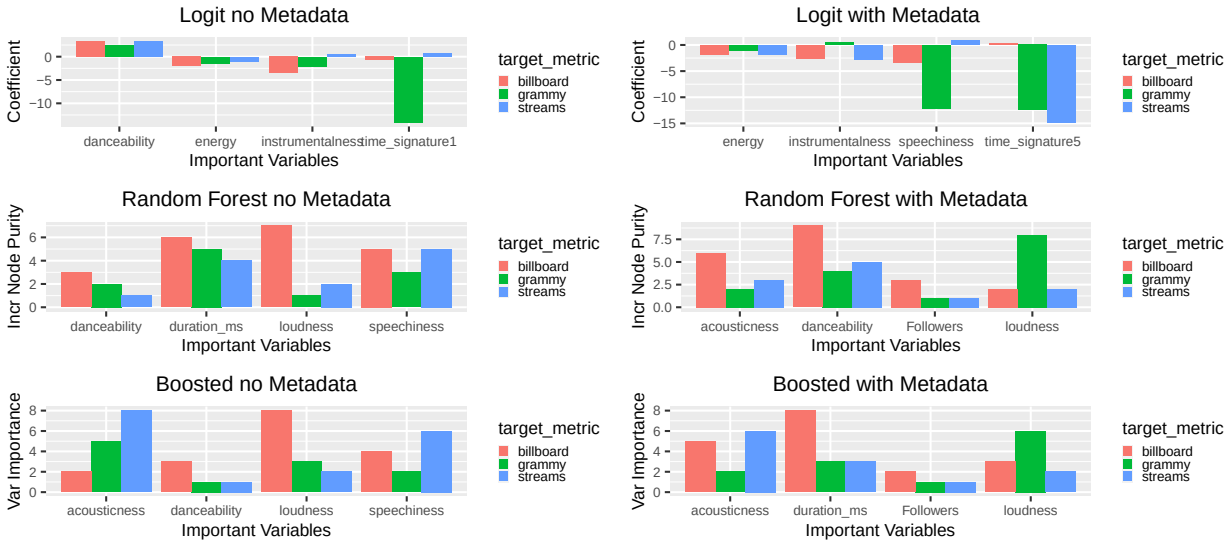


Figure 3: Most important variables of each model

4.3.3.2 Model interpretation Figure 3 takes the average ranking of each feature across the three different response variables (Grammy, Billboard and Streams) and then displays the top four ranked variables. The figure confirms what we had seen in the previous section on model evaluation: the ensemble models are much better than logistic regression. There same variables appear several times: `loudness`, `acousticness`, `danceability`, `speechiness`, `duration_ms` and `Followers`. It seems quite intuitive that these variables would have a strong correlation with song success, since successful the most popular genres today tend to be rap (speechiness), singer/songwriter or folk (acousticness), or pop (danceability). Moreover, of course there should be a huge positive correlation between the number of followers of an artist and the popularity of his/her/their music; that correlation was one of the main reasons why we chose to incorporate artist metadata in the first place.

Moreover, it's very possible to see what could have led the logit models astray is the huge negative coefficient that they placed on the different time signatures. As 4/4 time is by far the most common time signature, it makes sense that anything that deviates from the dummy `time_signature4` would have a negative coefficient. However, the fact that the coefficients on `time_signature1` and `time_signature5` are so deeply negative is probably just indicative of a small sample size, with an overwhelming majority of those few observations with this time signature pertaining to the negative class.

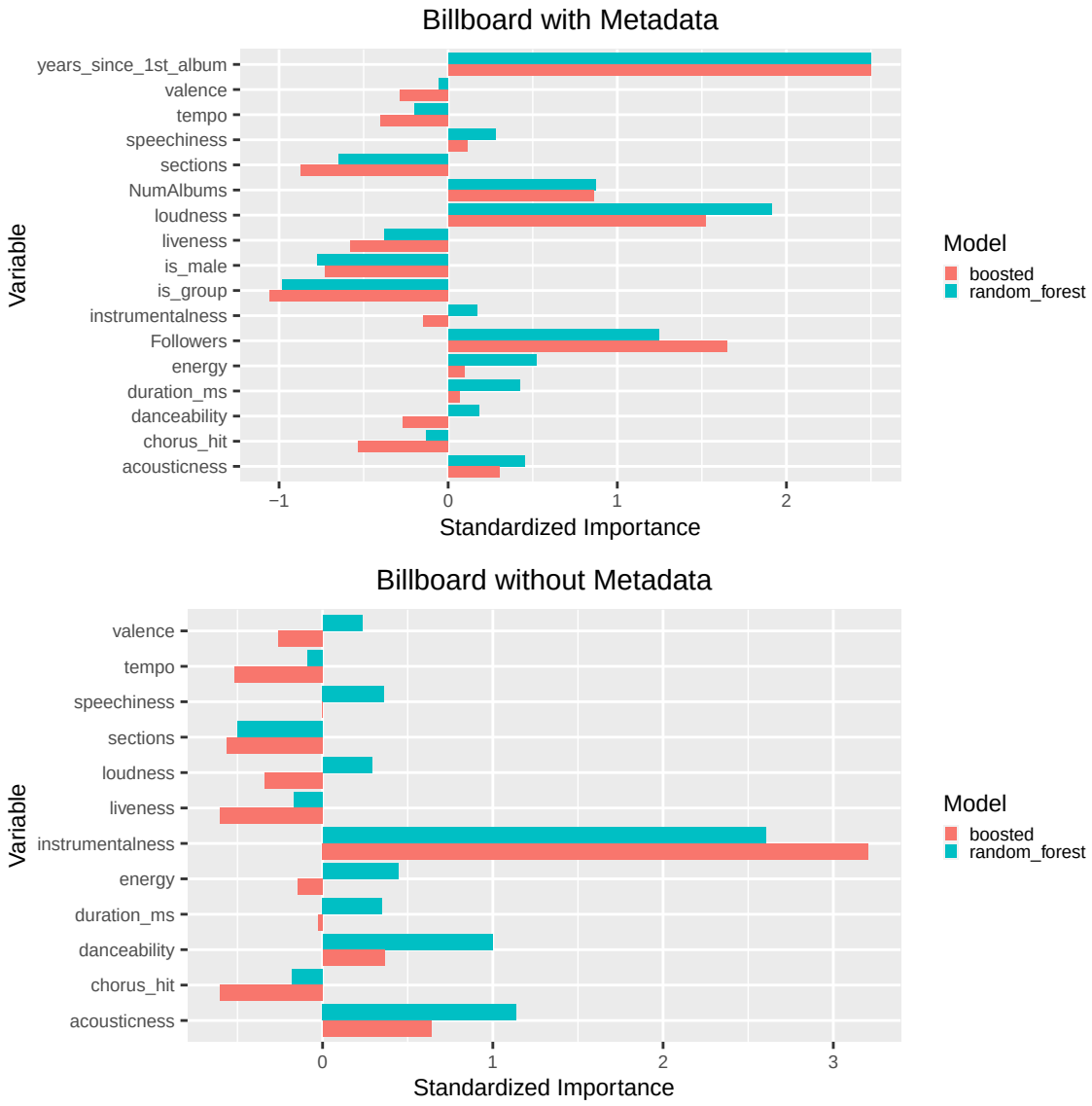


Figure 4: Standardized importance of all vars with Billboard Hot 100 as response

Figure 4 shows right away the general agreement of the random forest and boosted models on the Billboard Hot 100 data, both with and without metadata. It shows that once the logistic regression coefficients are removed from the data, suddenly brand new coefficients emerge as important. On the plot without metadata, `instrumentalness` becomes enormously the most important variable in the models without metadata, as does `years_since_1st_album`, `is_group` and `sections` for the models with metadata.

Table 10: Most positive and negative five variables

Name Most Positive Vars	Standardized Importance	Name Most Negative Vars	Standardized Importance
years_since_1st_album	2.4978659	is_group	-1.0172987
loudness	1.7174758	sections	-0.7595886
Followers	1.4465755	is_male	-0.7514766
NumAlbums	0.8652002	liveness	-0.4789207
acousticness	0.3761792	chorus_hit	-0.3291532

Table 10 averages the standardized variable importance of both the boosted and random forest models on the Billboard data with metadata and takes the top five most positive and negative averages. On the positive side, we can conclude that young musicians who are struggling to succeed in the industry shouldn't give up. In fact, the standardized importance of `years_since_1st_album` is over 2.5 standard deviations from the mean value. In addition, loudness is important, is it may make your music more engaging, and increased Followers also translate into increased probability of an artist's song becoming a hit.

On the negative side, the most negative coefficient is that on `is_group`. Potentially this is because it's harder for fans to obssess over a group than it is for them to obssess over a singular artist. Behind `is_group` is the variable `sections`, whose negative coefficient indicates that listeners want simpler songs with fewer sections. The sign and magnitude of the coefficient on `is_male` is also surprising to us, as we assumed that male musicians were much more popular than female musicians. Apparently, the modern music industry has managed to at least start reversing the previous sexist trend. The fourth most negative coefficient, `liveness`, indicates that listeners want their music to sound seamless, as if it were straight from the production studio. Knowing that in Brazil, for example, where the most common recordings are the ones that come from concerts, this is an insightful discovery about global music culture. Lastly, the coefficient on `chorus_hit` suggests that listeners would rather have a chorus occur sooner in a song rather than later. Considering that the chorus is, by far, the most remembered part of a given song, this is not surprising.

This portion of the analysis, which integrates artist metadata into the models, supports the findings in (Askin and Mauskapf 2017) that a song's position on the charts is strongly influenced by artist familiarity. This conclusion is echoed in figure 4, which shows that the number of followers an artist has on Spotify is the second or third most important variable in determining a song's presence in the Billboard Hot 100.

5 Conclusion

Our findings indicate that, although song-level attributes do not explain all of the variance in song popularity, they are heavily indicative of how popular a song may be.

In the unsupervised learning portion of this project, we found that not only can song-level features very distinctly classify songs into very different genres/vibes, however these classifications can be extremely predictive of song popularity. In the extreme case, Cluster 1 (Slow, Soul, Breakup) as over 13 times more popular than Cluster 10 (Background music), as defined by average streams per song. Looking at the subjective descriptions of these groups, it seems very intuitive that Cluster 1 would be more popular than Cluster 10. Moreover, looking at the number of songs per cluster, we can see that it appears that the industry is relatively aware of popularity differences among clusters, as the second most popular cluster, Cluster 9 (Popular, Radio, Head-Nod (Not Dance)), is by far the most represented cluster in the data.

When it comes to model evaluation in the supervised learning portion of this project, we can see that tree methods are a much more adequate tool for prediction than linear methods. Continuous Response shows that regularization did not improve the fit of the model on the data largely, with only a minor improvement. Moreover, Discrete Response showed that, when the data was adequate for modeling, tree methods led to much better ROC curves. In both sections, we see that, in fact, the tree methods and the linear methods even substantially disagree with the *signs* of the coefficients.

In the data without artist metadata, we found in Discrete Response that `instrumentalness`, `danceability` and `acousticness` have the most positive influence in song popularity, and `sections`, `liveness` and `chorus_hit` are among the variables with the greatest negative influence in popularity. This is slightly different from the variables with the greatest positive effect in Continuous Response, as the variables with the greatest positive influence are `acousticness`, `duration_ms`, `sections` and the variables with the greatest negative influence are `instrumentalness`, `liveness` and `tempo`. This is fascinating, because, between the two different kinds of response variables, the coefficients on `instrumentalness` and `sections` changed in a big way. Looking at the findings in Clustering, `duration_ms` and `sections` are defining features of Cluster 6 (orchestral music), which also, as a less popular genre, helps potentially explain the negative coefficients on those variables in the models above. Moreover, `instrumentalness` is very positively associated with identity in Cluster 10, the least popular group, and, to a slightly lesser extent, positively associated with Cluster 7, the second most popular group. This potentially explains some of the heterogeneity in the coefficient on `instrumentalness`. It suggests that this study is, in fact, very complicated, because the treatment effect is *not* homogeneous, or, in other words, more of some features may be beneficial in certain genres, but less so in others.

In conclusion, though there is some agreement in the models that variables like `acousticness` increase the chances of making a song a hit and variables like `liveness` are not so well received by global listeners, the huge variability in the models' findings shows that there is heterogeneity in the treatment effect and that there is no one-size-fits-all method of designing a hit song. Furthermore, with the findings in Discrete Response that artist metadata ia also hugely influential in song popularity, our study suggests that music producers should understand the limitations of the studied features as 'leavers' to increase song popularity.

In future iterations of studies on the impact of song-level features on song popularity, it may make sense to cluster the songs through some unsupervised learning algorithm, as we did, and then run separate models for each individual cluster. This may help achieve homogeneity of the treatment effect, as perhaps song length's impact on song popularity is uniform in one genre and may have a different uniform impact on songs in other genres.

6 Individual Contributions

Ethan Kallett: In charge of the `Discrete Response` methods section and `Data Review`

Sabrina Peltier: In charge of the `Clustering` methods section

Sabhya Raju: In charge of the `Continuous Response` methods section

Rohin Shivdasani: In charge of `Introduction` and `Literature Review`, as well as playing a supporting role on the `Continuous Response` methods section

The `Conclusion` was a collective effort among all teammates. And though certain teammates may have been in charge of a certain portion of the assignment, we all collaborated and helped each other whenever someone was stuck.

Reference

Askin, Noah, and Michael Mauskapf. 2017. "What Makes Popular Culture Popular? Product Features and Optimal Differentiation in Music." *American Sociological Review* 82 (5): 910–44.

Ceulemans, Cedric, and Lionel Detry. 2014. "Does Music Matter in Pop Music? The Impact of Musical Characteristics on Commercial Success and Critics' Ratings," 24–27.

Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. 2006. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* 311 (5762): 854–56.

Nijkamp, Rutger. 2018. "Prediction of Product Success: Explaining Song Popularity by Audio Features from Spotify Data." {B.S.} thesis, University of Twente.

Suh, Brendan Joseph. 2019. "International Music Preferences: An Analysis of the Determinants of Song Popularity on Spotify for the u.s., Norway, Taiwan, Ecuador, and Costa Rica." *CMC Senior Thesis*.