Analysis of Music Popularity Prediction Methods Across Different Genres

Leeds Rising, Anthony Lea, John Park

Cornell University

ORIE 4740

Abstract

Predicting the popularity and the success of a song is an age-old problem that has continued to be difficult as modern music preferences and the accessibility of music services continue to change. We wanted to investigate whether or not different genres have different effective popularity predictors and, similarly, if different models tend to be more effective for certain musical genres than others. For the sake of narrowing the scope of our goal and increasing the interpretability of our results, we chose to perform this investigation across five disparate genre groups: hardcore/metal, hiphop/rap, house/electronic, indie/rock, and jazz/blues. Our analysis showed that the most significant predictors varied – although loudness, acousticness, danceability, and speechiness tended to be important predictors for all five genre groups – and their success across different models differed only on one genre group significantly, as Jazz/blues consistently performed better on less flexible models compared to the other genre groups, and worse than the other groups on more flexible models.

I. Introduction

Predicting popular music is no novel task. For decades, different groups have made a wide variety of conjectures about what makes an artist or song popular, and in the last 15 years many computational approaches have sought to help record labels, radio stations, and streaming services predict and/or sign the musicians and songs that receive the majority of the $20-21 billion shared across the music industry.

Despite the many different approaches, few have looked to make different prediction models to serve the different unique subsections of the music industry. Curious to see whether this approach might bring new conclusions to this age-old problem, it occurred to our group that one relatively undocumented approach would be to stratify across different genres. As music lovers, we acknowledged that different genres have inherently different metrics of "quality" defined by their consumer segments, and such metrics are generally considered strong indicators of a song's popularity potential.

The primary difficulty of this approach is that the genre classification of a song is always vague - genre definitions change over time, and songs can be considered to be in multiple genres. Taking this into account, we paid close attention to how we defined different genre groups and what audio features we used as predictors of popularity. Our final dataset covered different aesthetic audio features (loudness, instrumentalness, etc.) as well as features indicating how consumers can interact with a song (danceability, valence, etc.), but intentionally left out metrics such as artist familiarity so that our models would focus on the intrinsic qualities of the songs. We created a number of machine learning algorithms (linear regression, ridge regression, random forest trees, linear discriminant analysis, and K-nearest neighbors) in order to investigate our hypothesis that both the predictive power of these features and the overall success of various models would differ across genres.

II. Data Collection and Preprocessing

We required a dataset with both enough songs across different genres that the performance would not be limited by limited by sample size, as well as a broad range of features that could fully encapsulate the characteristics of a song. We ended up choosing the two following datasets to merge:

- The Spotify audio features dataset (130,000 songs with song characteristics and a popularity rating)
- A comprehensive genres dataset (131,000 songs with genre, from 'Swedish Soul' to 'German hardcore')

After removing extraneous columns to our models (i.e. Spotify track ID, artist name, etc.), we were left with 93,587 observations of 15 columns (14 predictors, and 1 response variable column representing popularity).

With this merged dataset, we first moved into preprocessing our dataset for the different models we planned to use by appropriately converting the data types and standardizing variables that were of different scales. Loudness, for instance, was in the range of 0 to 60 decibels while danceability was in the range 0 to 1, which had the potential to impact the result of some models. A summary of the final dataframe is shown in Figure A1[1].

The next step was creating the genre groups. Choosing different genre groups was a nontrivial task – the genre of a song is inherently ambiguous, and choosing these genre groups in a way that resulted in overlap would make our models even more difficult to distinguish and would likely invalidate our hypothesis of these genres performing well on different predictors. We ended up applying our domain knowledge of what genres have unique characteristics to arrive at five: hardcore/metal, hiphop/rap, house/electronic, indie/rock, and jazz/blues.

The last stage of the preprocessing was to deal with any incomplete data and duplicate rows. We removed any rows with NA values and took the first occurrence of a song's genre even if it was mentioned in multiple genres, leaving us with 80304 values. The only column we filtered for specific values was popularity– if the popularity metric was 0, we conjecture this might have to do much more with the marketing of the artist than the song's characteristics. After removing these songs, we had 78671 total songs – 2799, 8050, 6873, 5165, and 1126 songs for our hardcore/metal, hiphop/rap, house/electronic, indie/rock, and jazz/blues genre groups respectively.

III. Data Exploration

To explore our data more in depth, we initially created a correlation plot (Figure A2) and histograms (Figure A3) to better understand the frequency of values in our different predictors and to ensure that our predictors were all independent. From the histograms, we observed several trends when we plotted the data in their individual genre groups or in aggregate. The acousticness, duration, instrumentalness, liveness, and speechiness of a song are very skewed right, tempo,valence, popularity, danceability, and energy are approximately normal, and loudness is very skewed left.

The skewness of many features was a foreboding sign of how much noise was in our data - while popularity might be normal and any pair of songs usually would have different values, many of the skewed features

[1] Note that figures that are numbered with 'A' are shown in the Appendix

had approximately equal values across many songs, making their effect on popularity harder to differentiate. The top 50 songs across different genres barely had distinct musical characteristics, and looking at our large genre groups with 1000's of songs made these within-genre trends even less clear-cut (see appendix A5). From these radar plots, the top 50 songs per genre appear to be more distinct on average for several predictors (danceability, valence, and popularity) than the overall averages per genre, but given the small sample size no clear trend can be concluded. The plots also suggest that more danceability, lower instrumentalness, and higher valence could accompany a higher popularity score, which is further supported across a larger sample size by the radar plot on the right, which compares the 100 least popular songs, top 100 most popular songs, and average song characteristics.

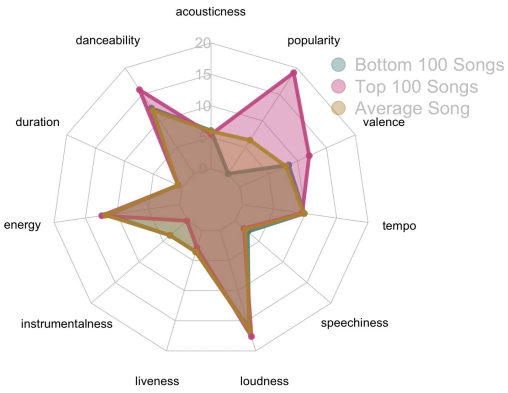


Figure 1: Radar plot of each predictor's score

To pivot into our investigation of what key characteristics might define each genre and to visualize our findings, we performed PCA on the dataset with a subset of 200 observations iterated upon many times to ensure stability of the results. The result is shown on the right, where even though only ~40% of the data's variance is explained by the first two PC's, we are still able to notice some key trends per genre (such as how Jazz/Blues variance is better explained by liveness, tempo, and speechiness than the other genres, which overlap considerably). Two PCs were graphed: PC1, which is dominated by acousticness, energy, instrumentalness, and loudness, and PC2, which is dominated by danceability, duration_ms, speechiness, and valence. While PC1 suggests that ~27% of the variance in our data can be accounted for by audio features, PC2 suggests that ~14% of the variance in our data can be accounted for by more semantic musical qualities that might represent how listeners interact with the song (i.e. how danceable it is). While only explaining ~41% of the variance in our data with these PC's is not ideal to make broad conjectures on the variability of our data in general or within genres, the clear trend with jazz/blues being much better predicted by tempo, duration, and speechiness than the other genres (in particular indie/rock) serves as strong evidence for our hypothesis.

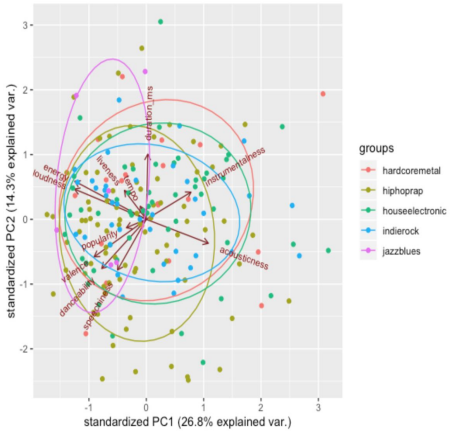


Figure 2: PCA performed on our dataset amongst each genre groups

IV. Methods for Predicting Numerical Popularity

Linear Regression and Ridge Regression

We begin our approaches by predicting the numerical popularity column with a different model for each genre. Given our hypothesis that only a subset of the predictors we have for each genre group will matter for predicting a song's popularity, creating linear models using subset selection to predict popularity is a natural starting point before more flexible models. The first task was to ensure that the linearity assumption was met for each genre group - generally, the data was somewhat linear, showed no correlation in error terms, and showed no significant collinearity. Some genres, however, including jazz/blues, did not obviously pass the check for lack of heteroscedasticity, as shown in the figure on the left. Given that there were no significant failures, we decided to move forward. Before anything else, we moved onto removing outliers and high leverage points, as a key observation during our linear assumption checks was that a number of high leverage data points existed in every group, but hip hop/rap in particular. Upon removal of these points we created our initial linear models for each genre subset.

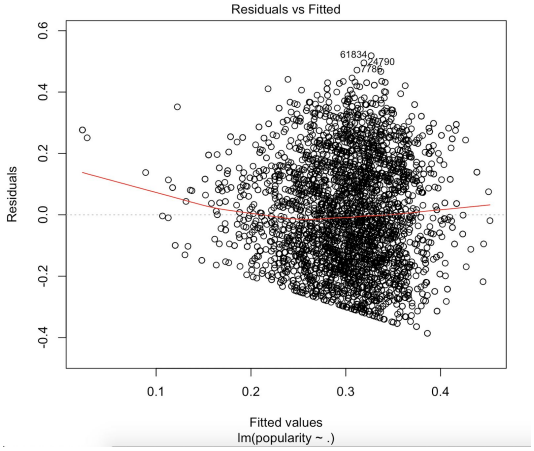


Figure 3: Residual plot for Jazz/Blue Genre

Throughout all the models the following trends emerged: the majority of our predictors were significant at the 0.01 level in each model, but the model suffered from an RSE around .19 and a low R-squared statistic of around 0.05 for each model. The most common significant variables were acousticness, danceability, instrumentalness, and speechiness. Another finding was hardcore metal was the only genre set where acousticness was not significant. We would recommend to not consider this when gauging the popularity of a metal hit. The RSEs for all were tightly bound in the .1857 to .1981 range. The R-squared statistics ranged from .04092 to .07076 indicating a relatively poor explanation in the models of the overall variability.

We saw that there were a lot of insignificant variables so we pursued backwards subset selection. Out of all the subset methods we chose backwards, as we didn't want to overfit or utilize a too computationally expensive method. Also by using subset selection we can capture variables that work together and can be recommended for each genre, as the best group to predict a song's popularity. The amount of variables were chosen by the number returning the lowest Bayesian Information Criteria and Mallows' Cp. The following results were returned:

Genre	Variables	R^2	MSE
All	Acousticness, Duration, Energy, Instrumentalness, Loudness, Speechiness, Tempo	.0381	.1908
JazzBlues	Acousticness, Danceability, Speechiness	.0337	.1943
HiphopRap	Acousticness, Danceability, Energy, Instrumentalness, Loudness, Speechiness, Tempo	.0487	.1987
HCmetal	Energy, Instrumentalness, Loudness, Liveness, Speechiness	.0406	.1817
HouseE	Acoustioness, Energy, Instrumentalness, Loudness, Speechiness, Tempo	.0386	.1906
IndieRock	Acoustioness, Loudness, Speechiness, Tempo	.0312	.1868

Figure 4: Recommended variables for each genre along with R^2 and MSE results

Finally, we decided to use Ridge regression to see if the problems facing our standard linear model could be mitigated by regularizing some of the less important parameters even after subset selection. Ridge regression was chosen over Lasso because we knew that our predictors all had a similar range of values and - as we had already performed subset selection - we had need to take advantage of Lasso's ability to shrink coefficients to zero and thus perform subset selection by itself. The results of performing ridge regression on the different genres is shown on the right, where it is clear that all genres were significantly impacted by Ridge Regression's regularization in respect to their corresponding linear models except for the hardcore and metal group.

Genre	MSE	R^2	Smallest Coefficient
Hardcore Metal	.0337	.0331	.3
Hip Hop Rap	.0340	.0235	.016
House Electronic	.0342	.0180	.017
Indie Rock	.0400	.0310	.05
Jazz Blues	.0400	.0298	.07

Figure 5: Results of ridge regression performed on each genre

Random Forest

With linear regression and ridge regression performing with relatively high error on both the training and test set, we thought random forest would be a logical next step to try and predict the popularity column with a higher flexibility model. We chose to perform random forest as our correlation plot from the data exploration phase showed that some predictors were somewhat correlated, and given the poor performance in linear and ridge regression, higher accuracy at the cost of some interpretability over the standard regression tree was appealing. This intended accuracy did show, as per the table on the left where every genre group had an MSE better than Ridge Regression's MSE and significantly better than linear regression. Jazz/blues stands out as having a somewhat higher MSE than the other models that stay around the same range, but given the difference is not that significant this may be attributed to Jazz/blues smaller sample size than the other genre groups. Jazz/blues also stood out with the importance of its variables to the random forest model - every other genre group except Jazz/blues ranked key as the most important variable, but Jazz/blues ranked key at 5th instead. Other than this, the predictors speechiness,

Genre	MSE
Jazz Blues	.0305
Hardcore Metal	.0221
Hip Hop Rap	.0212
House Electronic	.0170
Indie Rock	.0202

Figure 6: Results of random forest performed on each genre

acousticness, and loudness were the top 3 most important predictors for the almost all genre groups with only small deviations (Jazz/blues had danceability instead of speechiness, as pictured on the left. The other plots of important predictors are shown on

V. Methods for Predicting Categorical Popularity

Linear Discriminant Analysis

Another idea we expanded on was assigning each popularity score to a "bucket" categorical variable and applying classification models on that transformed dataset. Any song popularity metric is inherently "wishy washy" - changing a popularity score by a few points will affect our regression model, but these scores are computed somewhat arbitrarily (number of streams/time period can be impacted by the timing of the song release, for instance). To reduce the amount of noise this vague popularity score brought to our model, we transformed the popularity column to a categorical variable with 10 buckets: a 0-9 is in bucket 0, 10-19 in bucket 1, etc. Because the predictor classes are all well separated and the histograms of the features indicate that the predictors are somewhat normal, linear discriminant analysis is anticipated to be much more stable than logistic regression and thus a better contender for providing lower test error. Given the lack of distinct covariance matrices for our different genres, Quadratic Discriminant Analysis did not make sense to apply. The results are shown to the right: jazz/blues outperformed the other genres in training, and similarly outperformed the other genres in the test set by a fairly wide margin of ~0.05.

Genre Group	Training Classification Error	Test Classification Error
Hardcore/metal	0.1676768	0.2222222
Hiphop/rap	0.1698401	0.1872146
House/electronic	0.1758285	0.2015873
Indie/rock	0.189781	0.193952
Jazz/blues	0.1529988	0.1512195

Figure 7: Results of LDA performed on each genre

KNN

KNN is great for datasets in which the ratio of predictors to observations is quite low - with only 13 predictor columns and a minimum 1126 observations for each genre group, KNN was a rational choice for a high flexibility model. Testing a range of values for K and running the KNN algorithm within each genre group, we found effective models when we set K to 1 (highest amongst all genre groups) as can be seen in the table on the left.

Genre Group	Prediction Accuracy	K value
Hardcore and Metal	0.9821429	K = 1
Hip Hop and Rap	0.9950311	K = 1
House and Electronic	0.9992727	K = 1
Jazz and Blues	0.9601770	K = 1
Indie and Rock	0.9922556	K = 1

Figure 8: KNN results on each genre

VI. Conclusion

In our analysis, we found that more flexible models generally performed better in both the regression and the classification setting at the cost of interpretability. Given our insights from the data exploration stage, this likely can be attributed to how much noise was in our data and the fact that no clear linear (or other) trend seemed to exist. As a result, the MSE values for our less flexible models were significantly high (~0.2 on average for our linear models, 20% of the entire range of popularity scores), but higher flexibility models like random forests performed significantly better (averaging ~0.02 MSE for each genre). Even the classification technique LDA - failed to perform well on average, but KNN with low K values proved to function extremely well across the board.

Going back to our hypothesis, we had two primary goals: to understand the differences between models for the genre groups on an inference level (if certain predictors had more predictive power in some genre groups than others) as well as the prediction level (did certain models work better on some genres than others). On the inference side of our hypothesis, while our inflexible and interpretable models did not perform as well as we would have hoped across the board, the statistically significant variables that they identified (aside from loudness, acousticness, danceability, and speechiness, which usually rounded out the top four of most models) differed a considerable amount (energy generally held far more predictive power for hiphop/rap than hardcore/metal, for instance) supporting our original idea that certain predictors will be less relevant to some consumer segments than others. On the prediction side, many genres remained indistinct, but one outlier - Jazz Blues - tended to perform better on the less flexible models and worse on the more flexible ones than our other genre groups.

Moving forward, more efforts could be taken to mitigate the issue of noise in our data. Our data exploration suggested that this noise was significantly reduced when looking at the top echelon of songs in each genre. While this might be due to smaller sample size, our group conjectures that popularity rankings may be much more meaningful for songs in the upper echelon than others - true break-out songs from unknown artists (i.e. Mo Bamba by Sheck Wes) have the power to become popularized due to their excellent song characteristics (in Mo Bamba's case a very catchy tune) rather than the already existent popularity of the artist, which may be the primary source of noise. As such, it could be helpful to look at limiting our dataset to songs that were considered "breakouts" on Spotify over the years, and thus avoiding the potential for the artist's popularity to add noise.

Bibliography

"An Example Track Description." An Example Track Description | Million Song Dataset, millionsongdataset.com/pages/example-track-description/.

Cabreira, Jonathan. "A Music Taste Analysis Using Spotify API and Python." Medium, Towards Data Science, 11 Feb. 2020, towardsdatascience.com/a-music-taste-analysis-using-spotify-api-and-python-e52d186db5fc.

Chen, Yiyi, et al. "Show Me What You Got Song Popularity Prediction Using FMA Dataset." Berkeley.

Csathy, Peter. "The Future Of Music: Where It Is Today & Where It's Going In The Next Decade." Forbes, Forbes Magazine, 18 Feb. 2020, www.forbes.com/sites/petercsathy/2020/02/02/the-future-of-music-where-it-is-today--where-its-going-in-the-next-decade/#402dab45707e.

Döring, Matthias. "Radar Plots." Datascienceblog.net: R for Data Science, 13 Nov. 2018, www.datascienceblog.net/post/data-visualization/radar-plot/.

Eduardo. "Spotify's Worldwide Daily Song Ranking." *Kaggle*, 12 Jan. 2018, www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking.

"How to Interpret a Regression Model with Low R-Squared and Low P Values." Minitab Blog, blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values.

Pham, James, et al. "Predicting Song Popularity." Department Of Computer Science Stanford University, cs229.stanford.edu/proj2015/140_report.pdf.

Sam, Tobi. "Do Hit Songs Have Anything in Common?" Medium, Towards Data Science, 27 Apr. 2020, towardsdatascience.com/do-hit-songs-have-anything-in-common-37599940590.

Smith, M. "Exploring the Spotify API with R." *Github.io*, msmith7161.github.io/what-is-speechiness/.

Tomigelo. "Spotify Audio Features." Kaggle, 14 Apr. 2019, www.kaggle.com/tomigelo/spotify-audio-features/version/3.

Appendix

Figure A1: Complete Data Frame

Column Name	Data Type	Explanation
track_name	Factor	Name of the track/song
duration	Num	Duration in seconds, normalized to the range 0..1
artist_familiarity	Num	Artist familiarity score in 0..1 where higher value indicates more familiarity
artist_hotttnesss	Num	Artist hotness score in 0..1 where higher value indicates more hot
year	Int	Year song was released
acousticness	Num	Acousticness rating in 0..1 where higher value indicates more acoustic
danceability	Num	Danceability rating (how suitable the song is for dancing) in 0..1 where higher value indicates more danceability
energy	Num	Energy rating (measure of intensity and activity) in 0..1 where higher value indicates more energy
instrumentalness	Num	Instrumentalness rating (measure of the lack of vocals) in 0..1 where higher value indicates more instrumental
key	Factor	Key rating in whole numbers 0..1 where the value corresponds to the major/minor key of the song (major/minor dictated by *mode*)
liveness	Num	Liveness rating (presence of an audience) in 0..1 where higher value indicates more live
loudness	Num	Loudness rating in -60..0 (loudness in decibels) normalized to the range 0..1
mode	Factor	Mode value is either 0 or 1 indicating modality of the track (1 is major, 0 is minor)
speechiness	Num	Speechiness rating in 0..1(presence of spoken words) where higher values indicates more speechiness
tempo	Num	The overall beats per minute (BPM) of the song, normalized to the range 0..1
time_signature	Factor	Indication of how many beats are in each measure, ranging from whole values in 0 to 5
valence	Num	Valence rating in 0..1 (measure of how positive the music is) where higher value means more valence
popularity	Int	Spotify popularity metric from 0..100 based off number of streams / the amount of time the song has been out, normalized to the range 0..1 with 1 being most popular
genre	Factor	Associated genre of the track, where tracks with multiple genres given multiple rows

Figure A2: Overall Correlation Plot

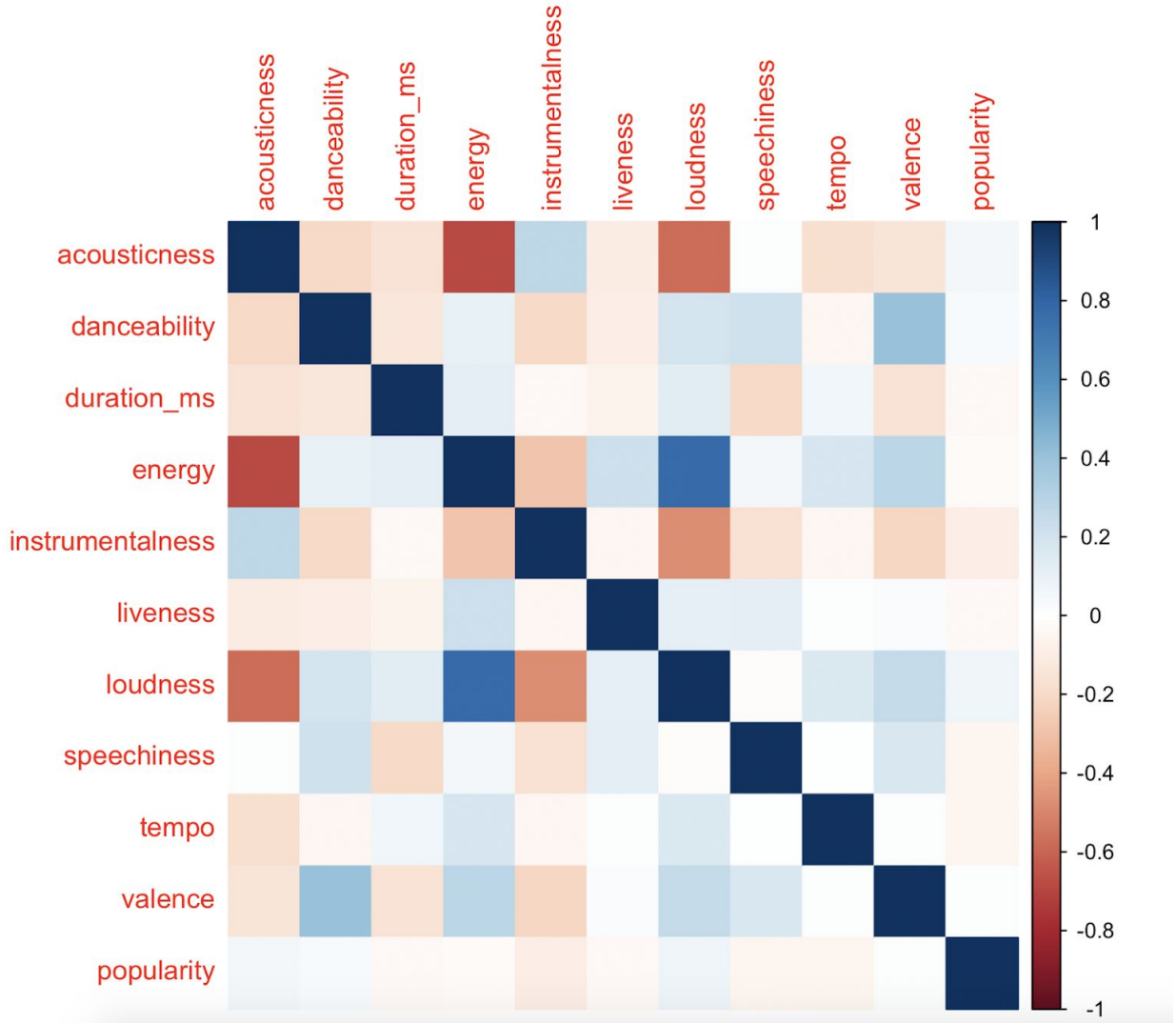
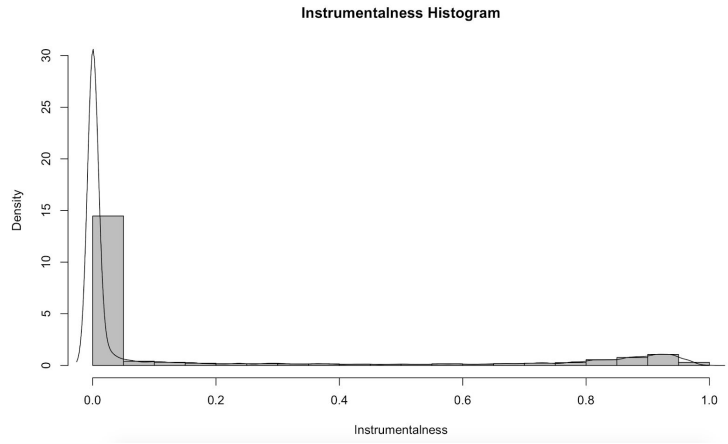
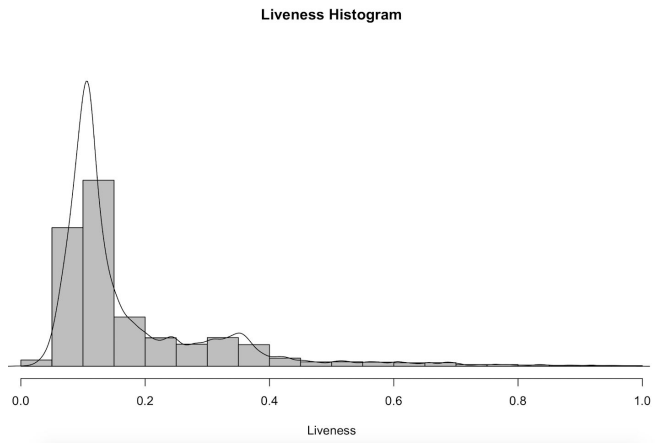
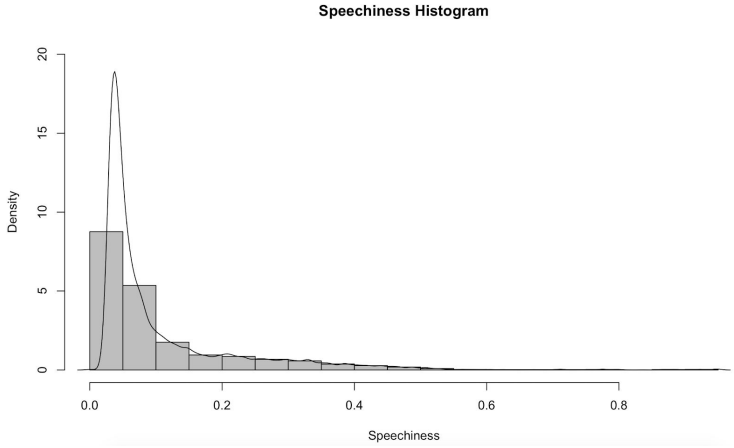
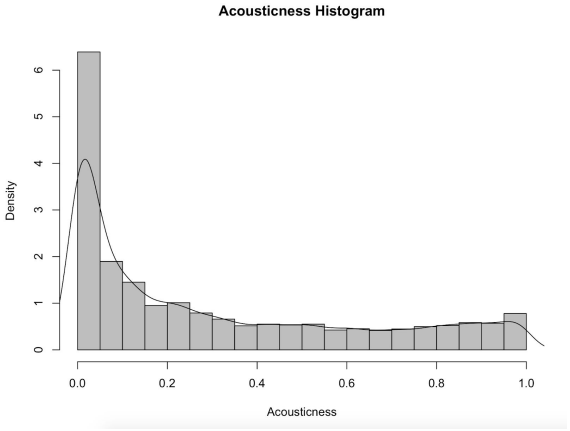


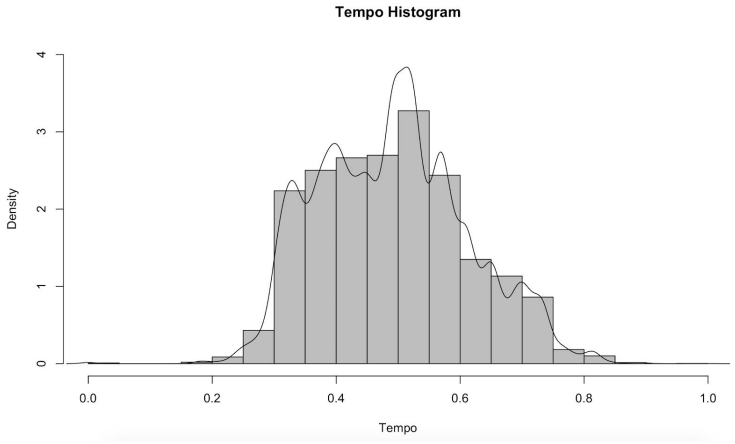
Figure A3: Overall Histograms

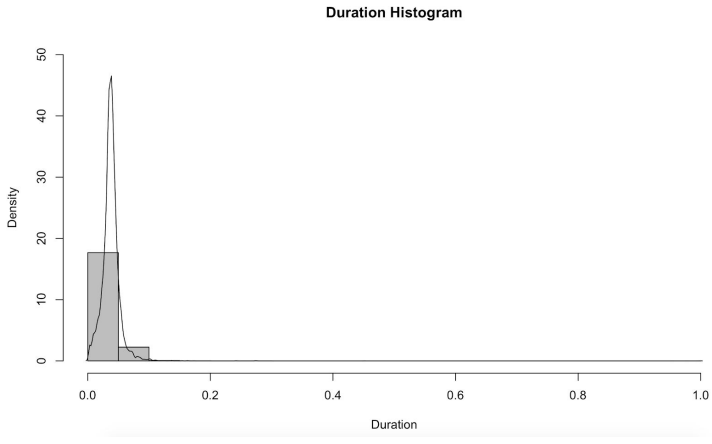


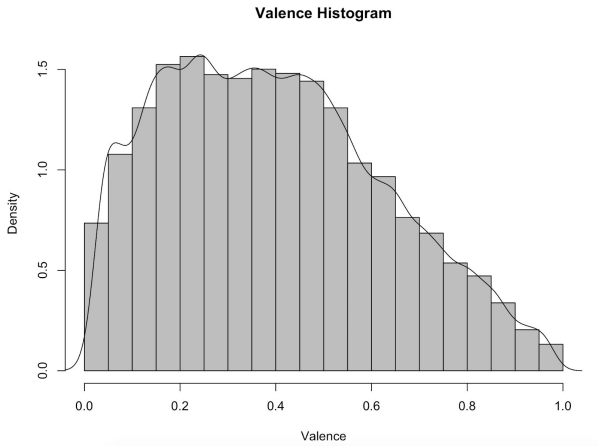


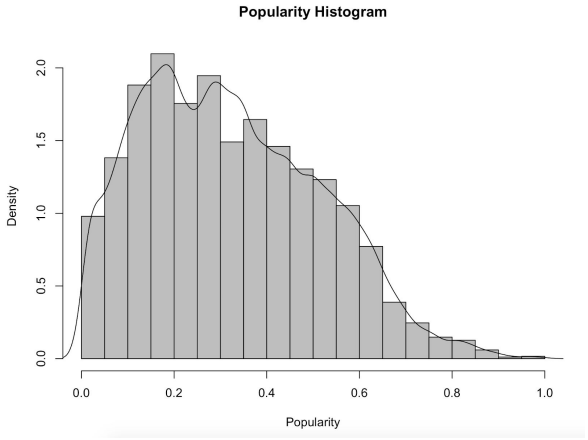


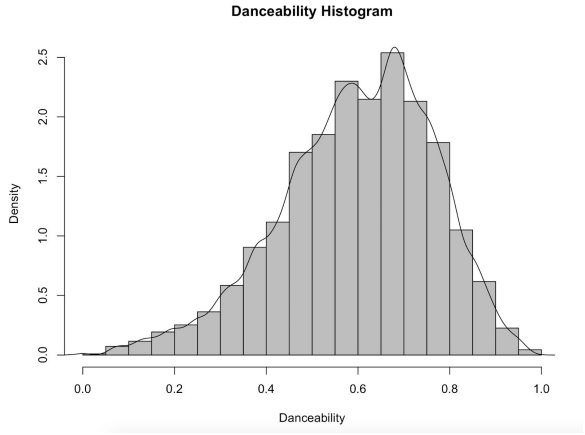


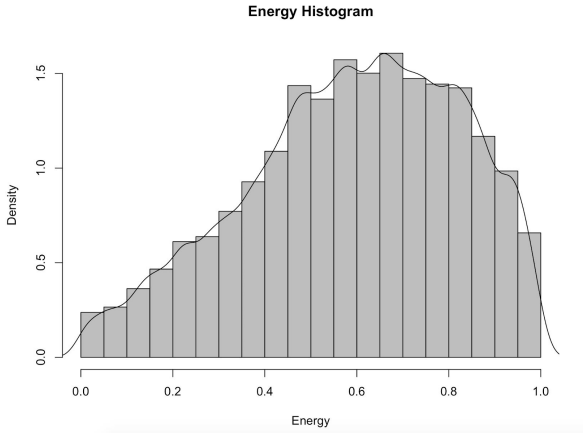





Valence Histogram
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
Valence


Popularity Histogram
Density
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
Popularity


Danceability Histogram
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.2
0.4
0.6
0.8
1.0
Danceability


Energy Histogram
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
Energy

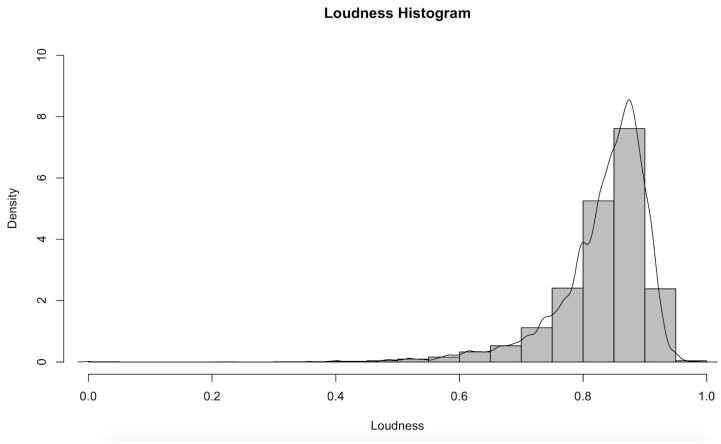

Loudness Histogram
Density
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
Loudness

Figure A4: Linear Model BIC's hardcoremetal

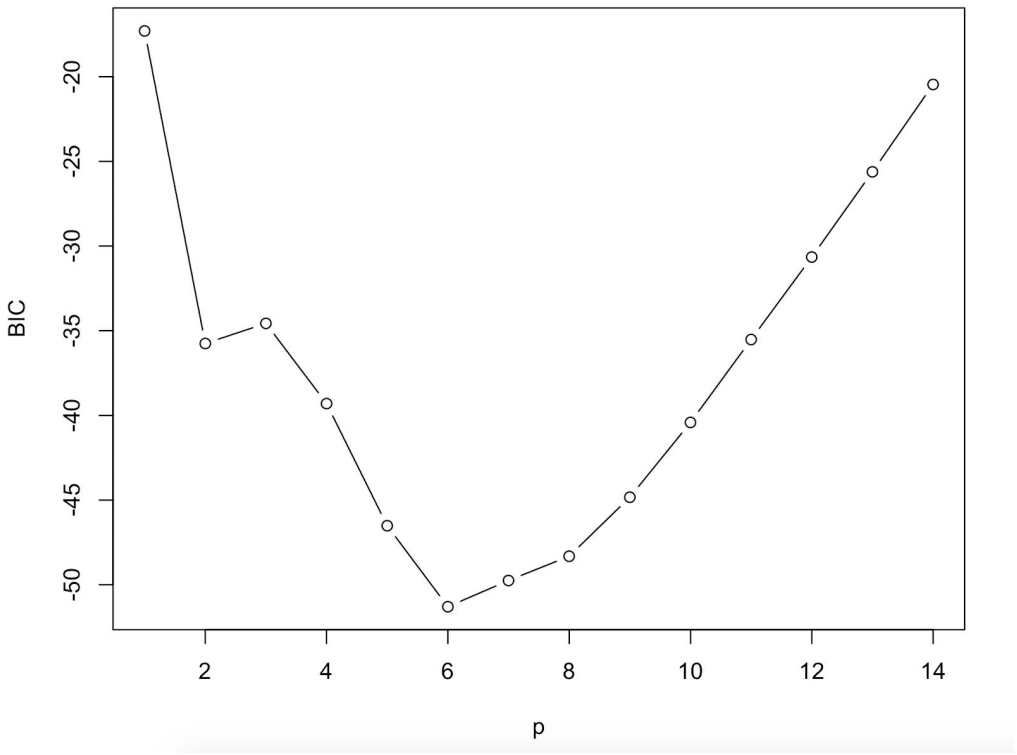


Figure A5: Radar Plots

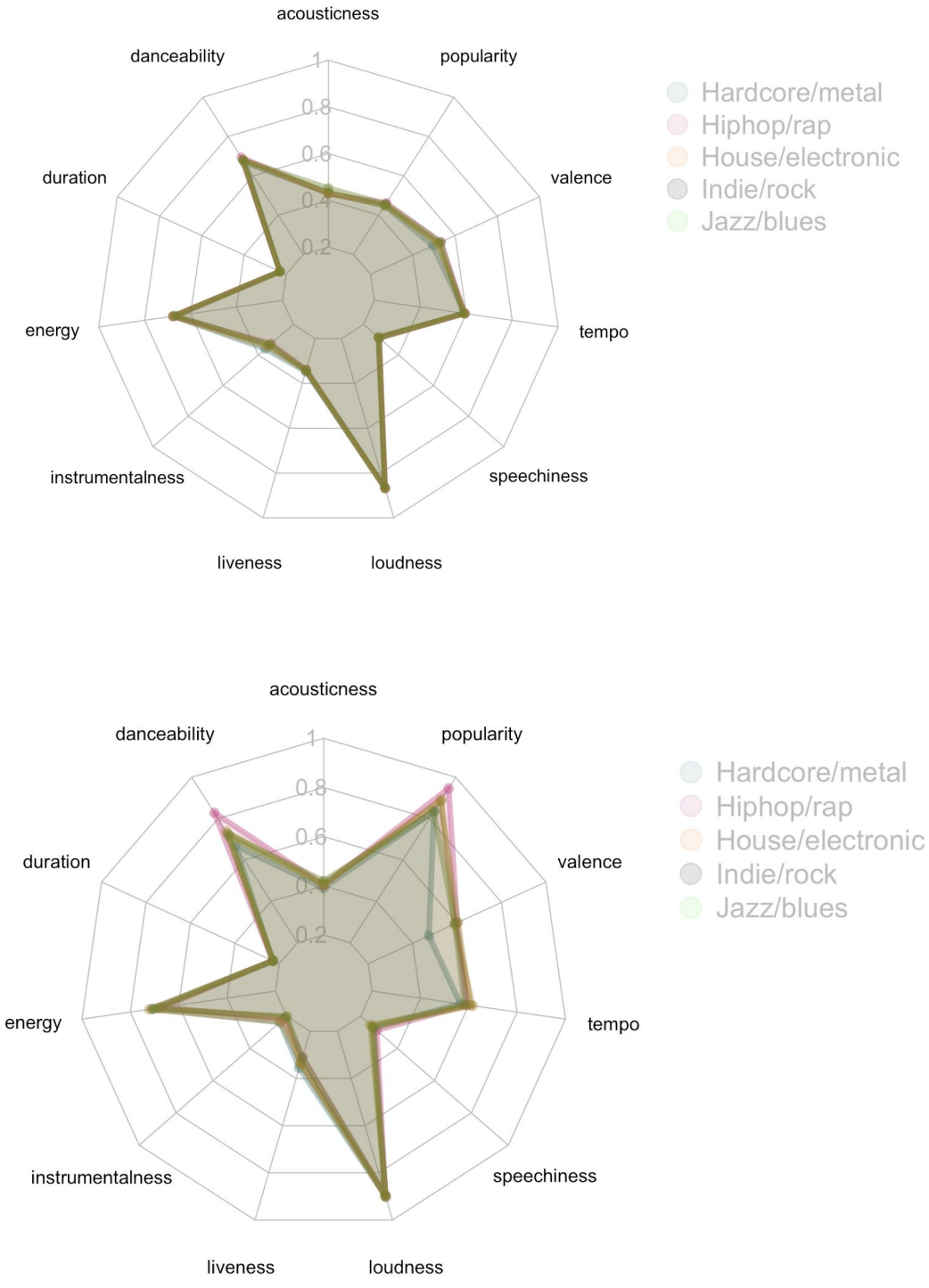
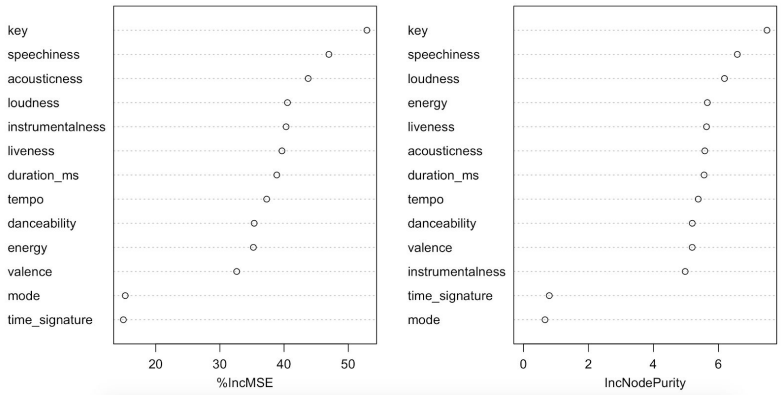
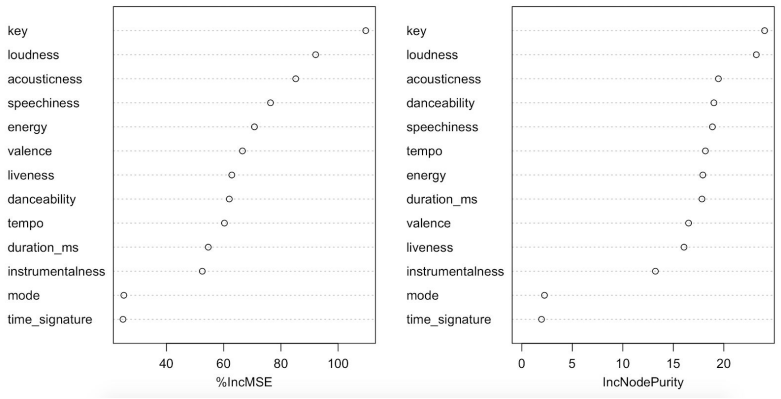


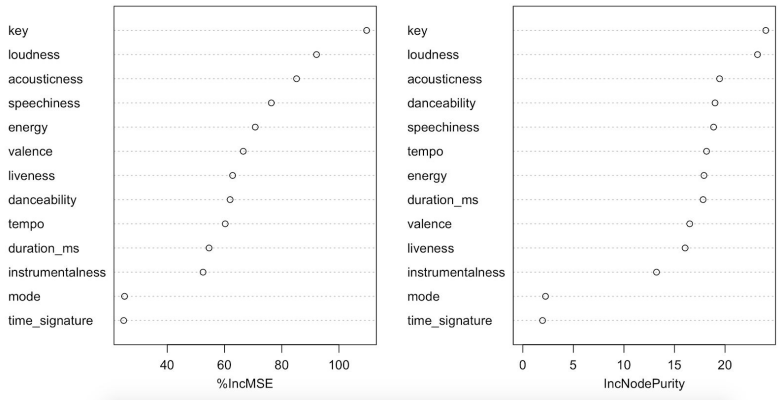
Figure A6:



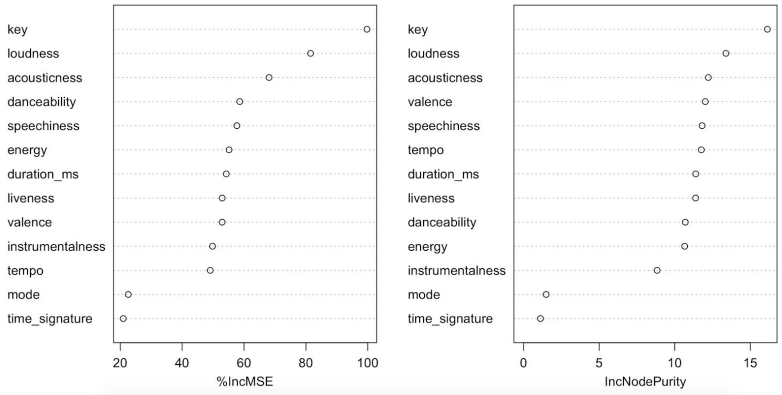
Hardcore/Metal



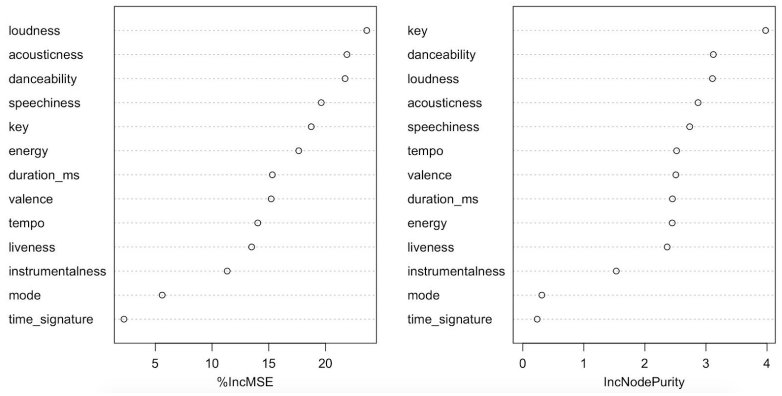
HipHop/Rap



House/Electronic



Indie/Rock



Jazz/Blues

Figure A7:

	%IncMSE	IncNodePurity
acousticness	43.75235	5.5804959
danceability	35.34042	5.1948365
duration_ms	38.85867	5.5590041
energy	35.20926	5.6583875
instrumentalness	40.31142	4.9763060
key	52.90332	7.4937732
liveness	39.66693	5.6327753
loudness	40.52515	6.1863424
mode	15.24682	0.6635764
speechiness	46.98207	6.5810739
tempo	37.28568	5.3788427
time_signature	14.95320	0.7950463
valence	32.61463	5.1921252

Plot showing importance of each variable in Hardcore/Metal

	%IncMSE	IncNodePurity
acousticness	21.886643	2.8707185
danceability	21.723898	3.1207494
duration_ms	15.313539	2.4511671
energy	17.639268	2.4474146
instrumentalness	11.333457	1.5310975
key	18.741798	3.9774805
liveness	13.486167	2.3647499
loudness	23.634580	3.1062929
mode	5.602686	0.3114423
speechiness	19.621386	2.7339316
tempo	14.027173	2.5195011
time_signature	2.233152	0.2353867
valence	15.209645	2.5068376

Plot showing importance of each variable in Jazz/Blues

	%IncMSE	IncNodePurity
acousticness	85.17995	19.464044
danceability	61.95364	19.016444
duration_ms	54.59915	17.834590
energy	70.74311	17.919890
instrumentalness	52.50584	13.226149
key	109.66902	24.038322
liveness	62.83156	16.057157
loudness	92.14955	23.217192
mode	25.06990	2.241856
speechiness	76.36300	18.869477
tempo	60.25225	18.175960
time_signature	24.76917	1.950481
valence	66.54858	16.510564

Plot showing importance of each variable in HipHop/Rap

	%IncMSE	IncNodePurity
acousticness	84.25273	17.401833
danceability	64.79879	15.203378
duration_ms	61.91307	16.007028
energy	70.77637	15.806950
instrumentalness	59.15104	12.829609
key	123.19571	21.221849
liveness	64.20783	15.246321
loudness	89.59229	18.543594
mode	27.78869	1.879464
speechiness	75.79767	16.997003
tempo	67.26429	17.120557
time_signature	24.87108	1.559937
valence	68.42004	14.924725

Plot showing importance of each variable in House/Electronic

	%IncMSE	IncNodePurity
acousticness	68.08353	12.219586
danceability	58.60745	10.697592
duration_ms	54.25348	11.391272
energy	55.15224	10.656712
instrumentalness	49.81931	8.833018
key	99.73830	16.127559
liveness	52.94016	11.379614
loudness	81.50461	13.385920
mode	22.57869	1.485759
speechiness	57.68171	11.807747
tempo	49.06309	11.758809
time_signature	20.95285	1.112562
valence	52.93001	12.018599

Plot showing importance of each variable in Indie/Rock