SonicWave Productions Song Popularity Prediction Technical Report

Methodology, Results, and Interpretations

Rohan Krishnan

STAT 1361

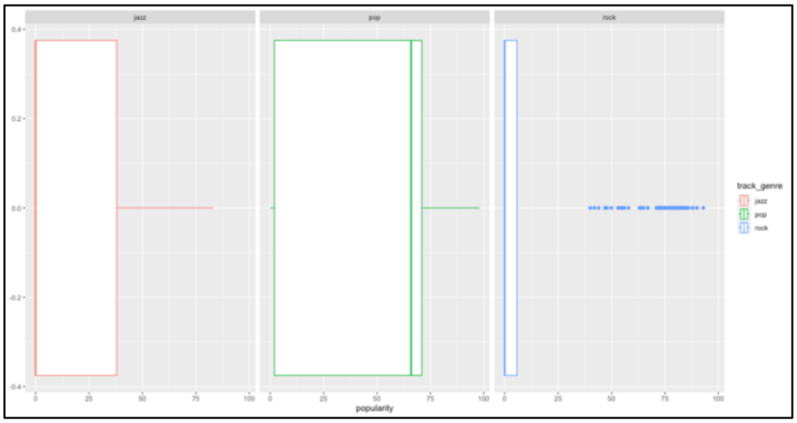
April 16, 2024

Introduction

SonicWave Productions is a growing company seeking to gain headway in the music industry. The music industry is valued at 14.34 billion as of 2024 and is projected to consistently grow in the following years (Statista, 2024). To gain a competitive advantage and grow SonicWave's market share, it is imperative to understand what factors are most important in creating a popular song.

As a Data Science Consultant , I was hired to predict the popularity of songs from rock, jazz, and pop genres. I was provided a data set with 1200 observations across 19 variables that encompassed various song metrics. Such a model would empower their team of music professionals to swiftly identify songs that are either undervalued or overvalued in the market, facilitating strategic decisions in song selection, promotion, and distribution. In this report, I will highlight the exploratory analysis I conducted to understand the relationships within the data, how I cleaned the data, the models I developed and how they performed, and my final takeaways regarding how the models should or should not be utilized.

Exploratory Analysis



I first looked at how popularity differed when compared with the categorical variables in the data set, specifically track genre and time signature. As is shown in the boxplot figure to the right, pop (green) has most of the popular songs while jazz (red) has mostly lower and middle rated songs and rock (blue) has almost entirely lower rated songs, with its higher rated songs falling outside of 1.5*IQR of its bounds. I then looked at how popularity differed by time signature. It appears that there is no real difference between time signatures of 3, 4, and 5; though time signatures of 1 seemed to have a larger density of lower popularity longs. Another interesting finding was that songs that were not explicit tended to have a higher density of lower ranked songs and very few extremely high ranked songs while explicit songs tended to have either lower ranked or extremely high ranked songs. Overall, very few of the variables appeared to follow a normal distribution. There also appeared to be some level of collinearity between energy and loudness. However, since there are only a small number of variables to work with, I elected to keep both in the model and address the issue via model selection techniques.

Data Oddities

The data set did not have any missing values. There were also no outliers in the popularity values as they ranged from 0-100. However, upon examining the time signature variable, I found that there was only one value for time signatures of 1 and no values for time signatures of 2. Because these values were so small, I elected to remove them entirely from the dataset so that the other variables could be properly considered.

Summary of Models Considered:

Multiple models were considered for predicting the popularity of songs across the rock, jazz, and pop genres. Each model tried, a small description of the model, and their final test MSE is reported below:

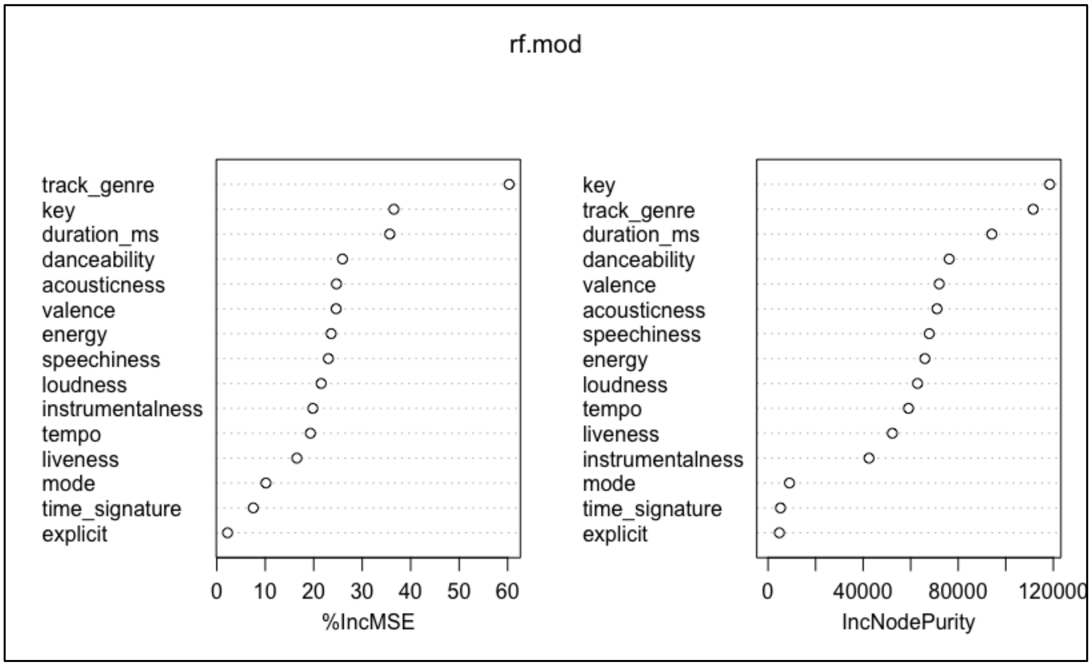
Model	**Description**	**Mean Squared Error**
Multiple Linear Regression	Simple MLR with all 15 variables	877.32

Best Subset Selection	BSS algorithm, chose output with only genre variable	862.09
LASSO Regression	Chose 11 variables including genre	857.25
After-LASSO OLS	OLS ran with variables chosen by LASSO	1058.71
Ridge Regression	Largest effect was genre	877.89
Bagged Random Forest	Full size trees: genre was most important	707.13
Random Forest	**Mtry value selected as p/3 ~ 5, genre was most important**	**694.76**
BART Model	Bayesian Additive Regression Tree model	847.19
Boosted Tree	Normal boosted tree	919.81

Overall, the random forest with an mtry value of 5 and ntree value of 501 (odd number to resolve any ties) performed the best by the metric of mean squared error. The bagged random forest performed similarly well but all the other models were significantly worse. I chose to use MSE as a measure as it is a standard measure of predictive error for regression tasks. For explainability to senior leadership, I would convert it to root mean square error, which would be in the same units as the popularity variable.

Most Important Variables and Variable Importance:

Upon examining relevant variables across all models, the biggest takeaway is that the genre of the track is extremely important in predicting song popularity. In the MLR, only duration, key, valence, track (specifically the pop genre dummy) were highly significant predictors of popularity. In the best subset selection, the final model only contained the track genre variable (specifically the pop dummy), indicating that whether a song is a pop song alone is effective in predicting a songs popularity. Both the LASSO and ridge regressions included track genre as a variable in their final models after cross validation to find the optimal lambda value. However, the most useful measure of variable importance came from the random forest model. As shown in the model to the right, track genre causes the largest increase in MSE when removed from the model, indicating it is a highly important variable. Overall, this can inform how SonicWave Production allocates its investments towards artists.



Challenges Faced and Model Trustworthiness

Although we were able to highlight the importance of genre in predicting popularity, it should be noted that none of the models performed spectacularly well. I would not recommend using any of these models in isolation as a surefire predictor of a song's popularity. The best performing model had an MSE of 694.76, which corresponds to an average discrepancy of 26 popularity points in predictions.

To improve the final model further, I believe we should focus on collecting more high-quality data with more information about each song as well as the producing artists and audience reception. For example, the artists' monthly listeners, number of pre-saves, and monthly view of the artists' profiles before release may have been very useful in predicting a song's popularity. With the inclusion of these variable and more observations, we could likely greatly increase the performance of the random forest model.

Conclusion

Overall, I would not recommend using any of these models in isolation to predict potential songs' popularities. However, these models have allowed us to learn that pop songs are by far the best performing songs in terms of popularity, which can inform SonicWave Productions' focus in the coming quarter.

Final Note

There were several exploratory plots that I could not include due to space restrictions. To view these plots, please reference the Basic EDA section of the code. Additionally, I have provided a visual representation of each model's MSE and how they compare to each other in the MSE Graph section of the code.