Bop or Flop?
A Data-Driven Approach to Predicting Hit Songs

Catalina Munoz, Dr. Robert Buscaglia
Department of Mathematics and Statistics

Abstract

Billboard Music's Year-End Hot 100 showcases the most popular songs of each year based on sales and streaming. Many musicians strive to appear on the Billboard charts as a sign of their success in the industry. Artists encourage fans to stream their music on platforms such as Apple Music, Youtube, and Tidal. However, the most popular streaming service is Spotify which garners streams from over 456 million people across the world. This research will explore the use of Billboard and Spotify to predict the success of songs. Billboard will be used to identify the Top 100 songs each year from 1980 - 2022. The Spotify Web API in R will be used to extract audio features of songs. These features include danceability, acousticness, energy, liveness, loudness, tempo, duration, speechiness, and valence. Using these two platforms, a dataset of thousands of "hit" and "non hit" songs will be collected, where a hit is defined by a song that graces the Year-End Hot 100. The goal of this project is to use the insights provided by Billboard and Spotify to identify the audio features that contribute the most to song success. Machine Learning will be implemented to estimate models that predict when a song will be a hit. Using cross-validated machine learning models, variable importance will be explored to elucidate what song features are most influential to a song making the Year-End Hot 100.

Introduction

Music is everywhere. People listen to it for entertainment, to connect with others, and to relax, myself included. As time has gone by, the rise of technology has also. Music consumption has moved from CD's, vinyl, and cassette tapes to digital streaming services, such as Spotify. Spotify provides a vast collection of songs from every genre and nearly every artist. They also supply many audio features of each song which can provide insights as well.

On the other hand, since the 1930's, Billboard has been tracking music by radio play and sales [1]. For music enthusiasts and professionals, the Billboard charts are a reliable source of information for the most popular songs. In particular, the Year-End Hot 100 chart encapsulates America's favorite songs of each year.

I chose to research the impact that Spotify audio features have on a song's Billboard chart placement because I have an interest in music and am curious in the factors of a song's success. Further, the ability to find features that make up the hit songs can have a positive impact on the music industry by aiding record labels to identify music artists that may make them profit. The objective of this project is to use machine learning algorithms to predict hit songs that grace the Year-End Hot 100 and see which characteristics are more prevalent in popular songs. Researching this topic offers me an opportunity to explore my interests while gaining knowledge.

Data Collection

- ○ Compiling hit songs that appear on the Billboard charts
  - ➢ Accumulated song titles and artist names from Wikipedia using ***rvest*** in R
  - ➢ Found Spotify playlists of the Year-End Hot 100 Billboard songs for each year
  - ➢ Collected the Spotify URI for each playlist
  - ➢ In R, I used the *get_playlist_audio_features()* function from the ***spotifyr*** [2] package to get a dataset of the hit songs
- ○ Compiling non-hit songs that do not appear on the Billboard charts
  - ➢ Identified a large Spotify playlist with a variety of songs from different genres and artists
  - ➢ Input the Spotify URI in R to get the audio features of each song
- ○ Data Polishing
  - ➢ Merged both datasets together
  - ➢ Added a Hit column, where a value of 1 indicates a hit song, 0 otherwise
  - ➢ Removed values that didn't have a track.id
  - ➢ Removed duplicate songs by their track.id
  - ➢ Condensed the dataset by removing unnecessary columns

Spotify Audio Features

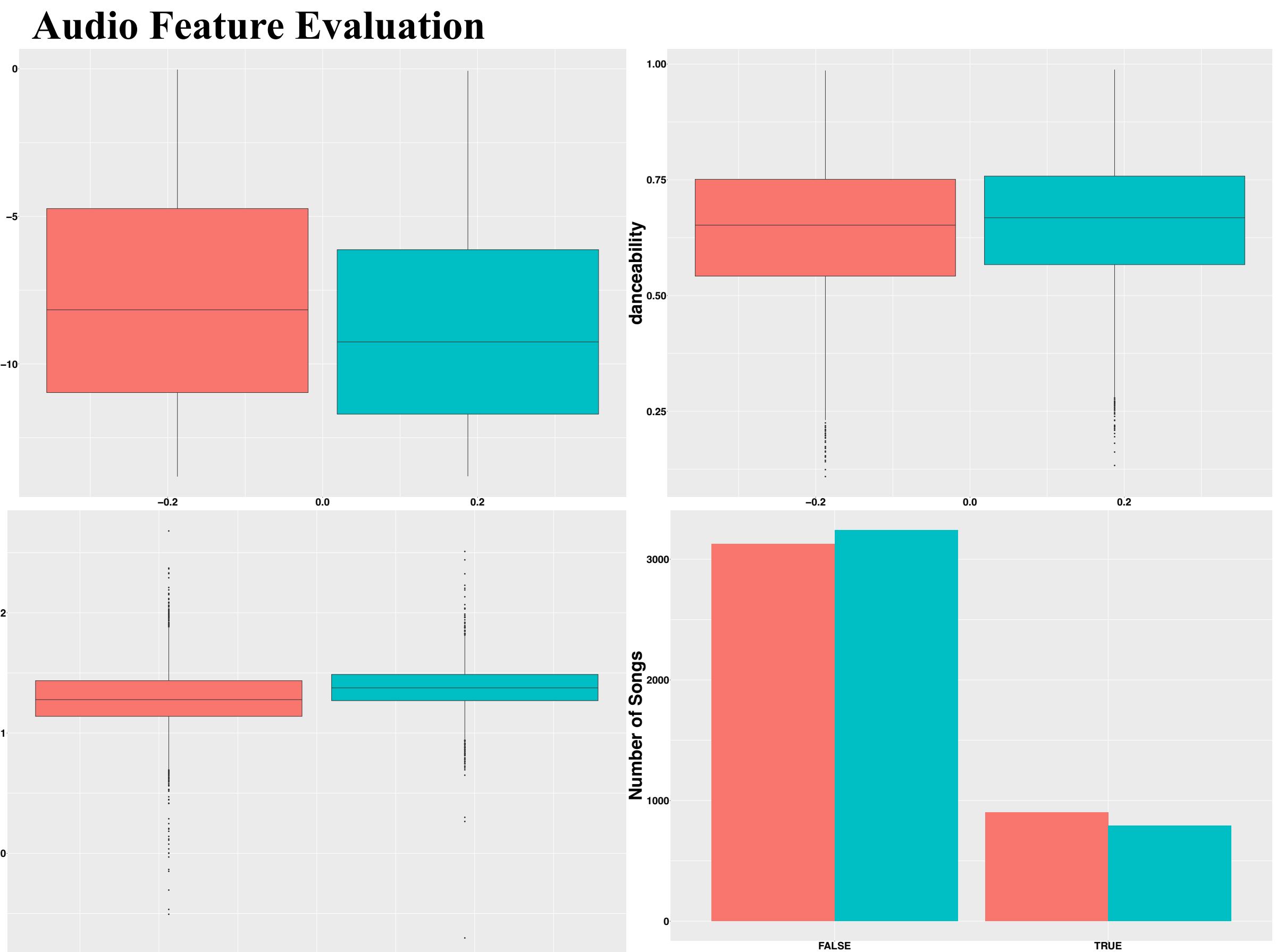
The Spotify Web API [3] provides many audio features that contribute to a song's makeup. The following table summarizes the ones that are being used for analysis.

Variable	Description
Hit	Did the track appear on the Year-End Hot 100 Billboard Charts from 1980-2022? 1 = Hit, 0 = Non-hit
danceability	Value from 0 to 1 on how suitable a track is for dancing based on a combination of tempo, regularity, beat strength, and rhythm stability
energy	Energy value from 0 to 1 of a perceptual measure of intensity and activity based on dynamic range, perceived loudness, timbre, onset rate, and general entropy
loudness	Value from -60 to 0 of loudness of the track in decibels
speechiness	Value from 0 to 1 that represents the presence of spoken words in a track. A higher value makes up a track of more spoken words
acousticness	Value from 0 to 1 on whether a track is acoustic. Higher values represent a higher confidence of an acoustic track
instrumentalness	Value from 0 to 1 that predicts if a track has vocals. Higher values represent less vocal content
liveness	Value from 0 to 1 that represents the presence of a live audience in a track. A higher value indicates an audience
valence	Value from 0 to 1 describing the positiveness of a track. Songs with a higher valence score sound happier and cheerful.
tempo	Estimated tempo of each track in beats per minute.
time_signature	Time signature of a track based on beats per bar
explicit	Does the track contain explicit words? (True/False)
key_name	Key each track is in. If no key was detected, value is -1
mode_name	Modality of the track (major/minor)
key_mode	Key mode of the track
duration_min	Length of the track in minutes
track.popularity	Popularity value from 0 to 100 calculated using an algorithm based on how recent and how many times each song has been played.

TABLE 1 Descriptive table of variables used in analysis

Comparison of Audio Features

Starting off, I explored the relationships of each variable with the variable of interest, Hit. By comparing the Spotify features to the outcome variable, I was able to see which features were more associated with the presence of a song on the Billboard charts. I used statistical tests such as association and t-tests to investigate the relationships between these variables. Numerical variables were used to explore the difference in means between hit and non-hit songs, while categorical features were used to determine if there was a significant association to the Hit variable. This approach gives insights into features that may contribute to a song's popularity and factors that may be important in predictive modeling.



Audio Feature Evaluation

FIGURE 1 Visualization of variables that were most significant in their respective tests. Blue indicates hit songs, red is for non-hits

Analysis

- ○ **Logistic Regression**

10-fold cross-validation was used to evaluate the performance of many different models. Model selection was performed using stepwise regression and lasso. The lasso method is used for variable selection that keeps the most important predictors while shrinking the coefficients of less significant variables to zero.

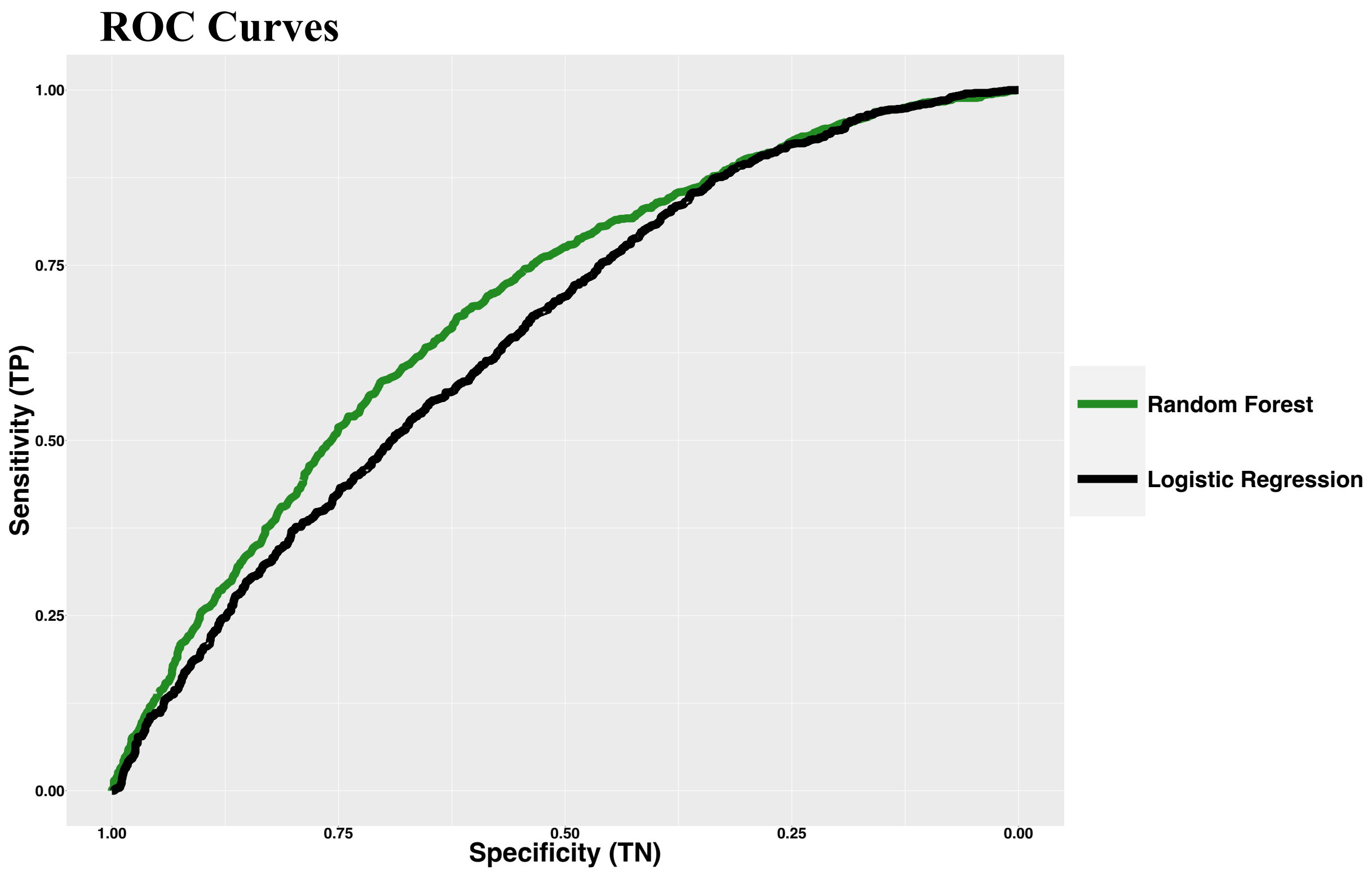
Forward and backwards stepwise regression were also employed for model selection. Forward stepwise starts with the null model and adds predictors until its performance metric is minimized. Backwards does the opposite and removes predictors until the metric is optimal. Two-way interactions between the following variables were evaluated: danceability, energy, speechiness, acousticness, valence, instrumentalness, liveness, time_signature, explicit, and duration_min. Interactions between these predictors can help capture the combined affect on hit songs that individual variables may not be able to explain on their own.

- ○ **Random Forest**

3-fold cross validation was implemented to produce a sensitivity analysis based on a set of 60 different hyperparameter combinations which included ntree, mtry, and max.nodes. Ntree sets the number of trees that will be built in the random forest, mtry determines the number of variables randomly sampled at each split in a tree, and max.nodes is the number of terminal nodes that each tree can have. This system was repeated for each model to estimate its performance over two metrics.

Results

To evaluate the Logistic Regression and Random Forest models, two metrics are being used: classification accuracy and area under the curve (AUC). AUC measures the model's ability to differentiate between positive and negative classes by calculating the area under the receiver operating characteristic curve (ROC). Values closer to 1 suggest that the model can better classify samples. Figure 2 displays the ROC curves of both models depicting their performance in correctly classifying positive and negative cases.



ROC Curves

FIGURE 2 ROC curve of Logistic Regression and Random Forest Models

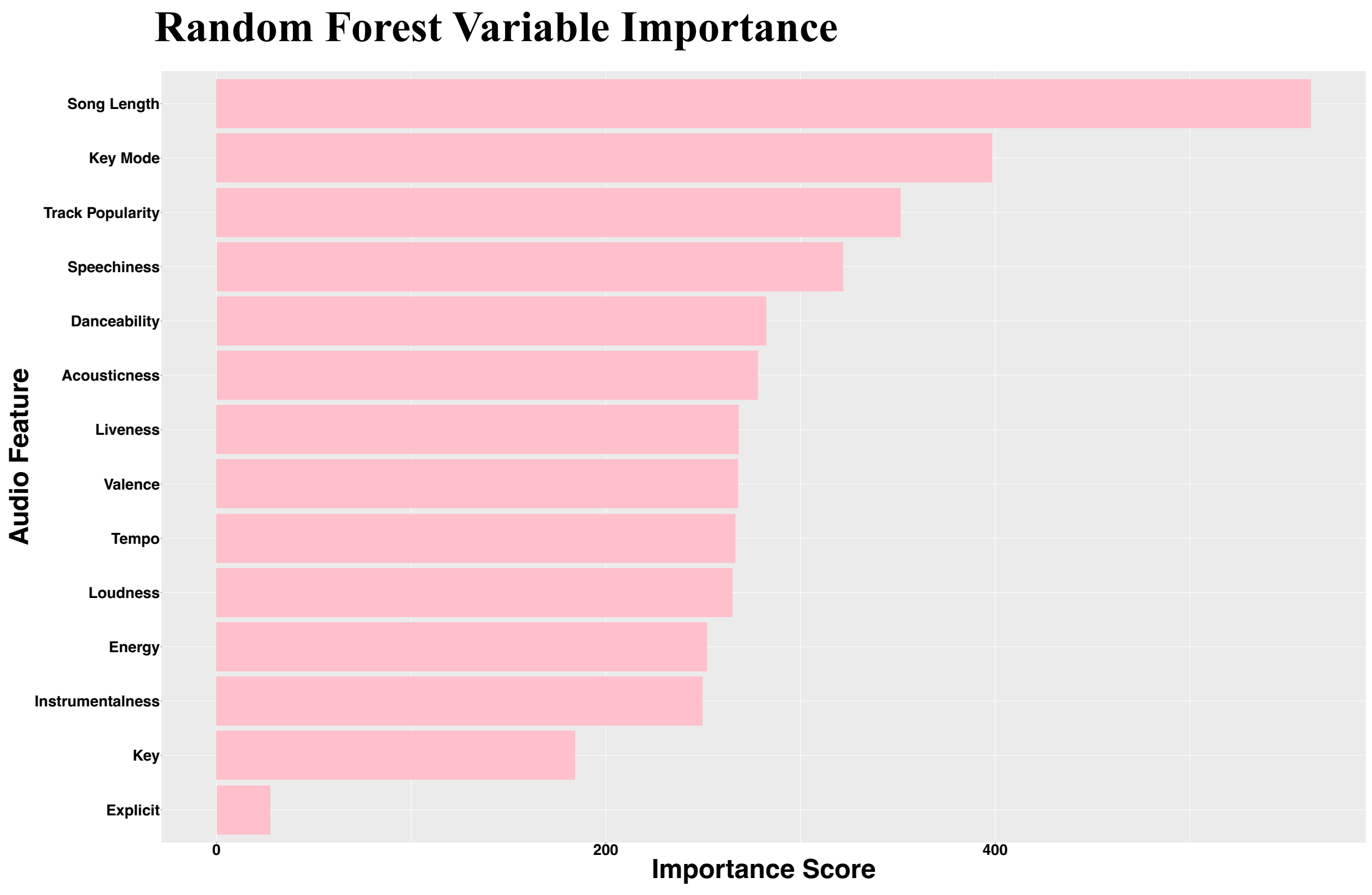
Classification accuracy measures the percentage of correctly classified observations out of the total number of observations. The values below show the top performing logistic regression and random forest models

- ○ Logistic Regression
  - ➢ Classification Accuracy: 62.15%
  - ➢ AUC = .65
  - ➢ Model contained 30 predictors, including interaction terms
- ○ Random Forest
  - ➢ Classification Accuracy: 65.70%
  - ➢ AUC = .70
  - ➢ Hyperparameters: ntree = 1000, mtry = 4, max.nodes = 5

Variable Importance

Variable Importance is a measure that determines the influence of each variable in the outcome of the response we are trying to predict. In this case, the variable importance gives insight into the Spotify audio features that may influence hit songs.

For classification trees, the importance score is the mean decrease in Gini Index. Gini Index measures the likelihood that a data point will be incorrectly classified based on class distribution of a node. Variables that have the biggest reduction in Gini Index are the most "important" variables.



Random Forest Variable Importance

FIGURE 3 Variable Importance plot of optimal random forest model.

The visualization in figure 3 graphs the importance of each variable in a random forest model from highest to lowest. The top audio features that have the largest decrease in Gini index for predicting hit songs are the length, key mode, Spotify popularity, and speechiness of a track.
The audio feature that doesn't contribute as well to this random forest model is whether a track has explicit lyrics.

Conclusion and Future Work

In conclusion, predicting hit music is complex task that is influenced by many factors, both within and outside the scope of this dataset. However, this research has displayed the ability to predict hit music by using a random forest model with an accuracy rate of 65.70%. The Spotify audio features that are among the most important for predicting hit music are the length, key mode, Spotify popularity, and speechiness.

In terms of future work, incorporating more data, such as Spotify monthly listeners or other artist demographics may improve the model. Further, choosing a different Billboard chart that updates weekly or includes additional songs could provide more insights. Lastly, evaluating model performance on a specific genre or time period may be a possible direction for future research.

Acknowledgements

Thank you to the Mathematics and Statistics department for providing me with the academic knowledge and skills necessary to carry out this research and thank you to UGRADS for providing me with the opportunity to present my research.

References

[1] Billboard. (n.d.). Year-End Charts: Hot 100 Songs. Retrieved from: https://www.billboard.com/charts/year-end/

[2] Katz J. spotifyr: R wrapper for accessing Spotify's Web API [software]. Version 2.1.1. Comprehensive R Archive Network. Available from: https://cran.r-project.org/package=spotifyr

[3] Spotify. Get Audio Features for Several Tracks. In: Spotify Web API [Internet]. Stockholm: Spotify AB; c2014–2021. Available from: https://developer.spotify.com/documentation/web-api/reference/get-several-audio-features