What Makes a Hit?

An Analysis of Musical Audio Features and Billboard Hot-100 Success (2000–2023)

Module: IJC437 – Introduction to Data Science
Student Registration Number: 250124697
Word Count: 2,970 words

1. Introduction and Aims

The paradigm for research on creative industries has shifted because of the availability of large-scale cultural datasets. Online music streaming services such as Spotify provide structured audio features. The Billboard Hot-100 chart is one of the most recognizable indicators of commercial success in popular music.

Previous studies have shown that popular music has undergone dramatic changes over time. Research has found several trends over time in popular musical features, including increases in loudness, energy, and rhythmic complexity (Serrà et al., 2012; Interiano et al., 2018). However, factors beyond audio features also influence musical popularity.

This project brings these two perspectives together by combining Billboard Hot-100 chart data (2000–2023) with Spotify audio features. Using data science methods implemented in R, the analysis investigates whether measurable musical properties are associated with chart success, how these properties have changed over time, and whether they can be used to predict success using statistical and machine-learning models.

1.1 Summary of Relevant Literature

Previous research has demonstrated that quantitative analysis of popular music is feasible using audio features provided by digital streaming platforms such as Spotify.

There have been quite a few studies delineating the long-term patterns in popular music. Some studies have indicated increases in the levels of loudness and audio energy with time, associating these with the production methods as well as listener preference. In other studies, the omnipresence of speech vocals in popular music, especially with the advent of hip-hop or rap

genres, came to the forefront. These studies point towards the dynamism in musical patterns with time.

Recently, a considerable number of studies have also been conducted for the prediction of musical success solely through audio features. Although the consensus is that songs successfully released tend to have greater attributes of danceability, energy, and tempo, the fact remains that prediction capability is limited. It gives a clear indication that, although audio features consider the necessary parameters of musical structure, they do not determine the key parameters contributing towards the success of a song being a hit. Other considerations include marketing, artist identity, music listing, and promotions via social networks.

In conclusion, a summary of the pre-existing literature work regarding the topic can be seen, indicating that Spotify audio features can be beneficial in analyzing musical trends, as well as their associated connections with popularity, although the Spotify audio features should be considered cautiously regarding other variables.

1.2 Aim of the Study

The primary objective of this study is to explore if Spotify audio features can be established in relation to song success on the Billboard Hot-100 chart. Furthermore, this study aims to identify how musical features correlate with the chances of reaching positions within the top 10 on the chart.

1.3 Research Questions

To answer the research questions, the following questions will guide this study:

RQ1. To what extent the audio features from Spotify can be utilized in predicting whether the song will hit the Top-10 Billboard Hot-100?

RQ2. What Spotify audio features are most strongly related to chart success?

RQ3. In what way have musical features, especially speechiness, evolved in the songs included in the Billboard Hot 100?

These research questions have been designed to be model-agnostic, more interested in relationships than specific modelling approaches. Collectively, they offer a comprehensive way

to examine the role of musical properties in predicting chart success, even with an appreciation for limitations inherent within an audio-based model.

2. Data and Methods

2.1 Data Collection

Figure 1. Overview of the data science workflow used in this study, including data loading, cleaning, exploratory analysis, modelling, evaluation, and interpretation.

The study uses a single dataset that contains Billboard Hot-100 chart data from 2000 to 2023 with Spotify audio features.

Each song in this dataset can be linked to a group of number values that are determined through an audio analysis system provided by Spotify and are linked to each specific song. These number values describe multiple musical elements that can be identified in an audio signal.

This analysis focuses on nine Spotify audio features: danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. All these features can be defined as rhythm, intensity, vocals, and tone. According to the official documentation of the Spotify API, the above features are computed, highlighting their relevance in the process of analyzing music.

The dataset is well fitted for this research because it joins a clear measure of commercial success with standardized musical features. No additional datasets were used.

2.2 Data Cleaning and Pre-processing

Before the analysis could be done, the dataset had to be preprocessed to ensure the results obtained were accurate and relevant. This ensured irrelevant variables that did not contribute to the analysis of the research questions were eliminated, allowing the focus to be on the musical elements being studied.

Any songs with missing data in the chosen audio features were removed from the dataset. Although these audio features were employed in the exploratory analysis as well as the predictive model, the presence of missing data could lead to biased or questionable results, especially in the context of the size of the dataset.

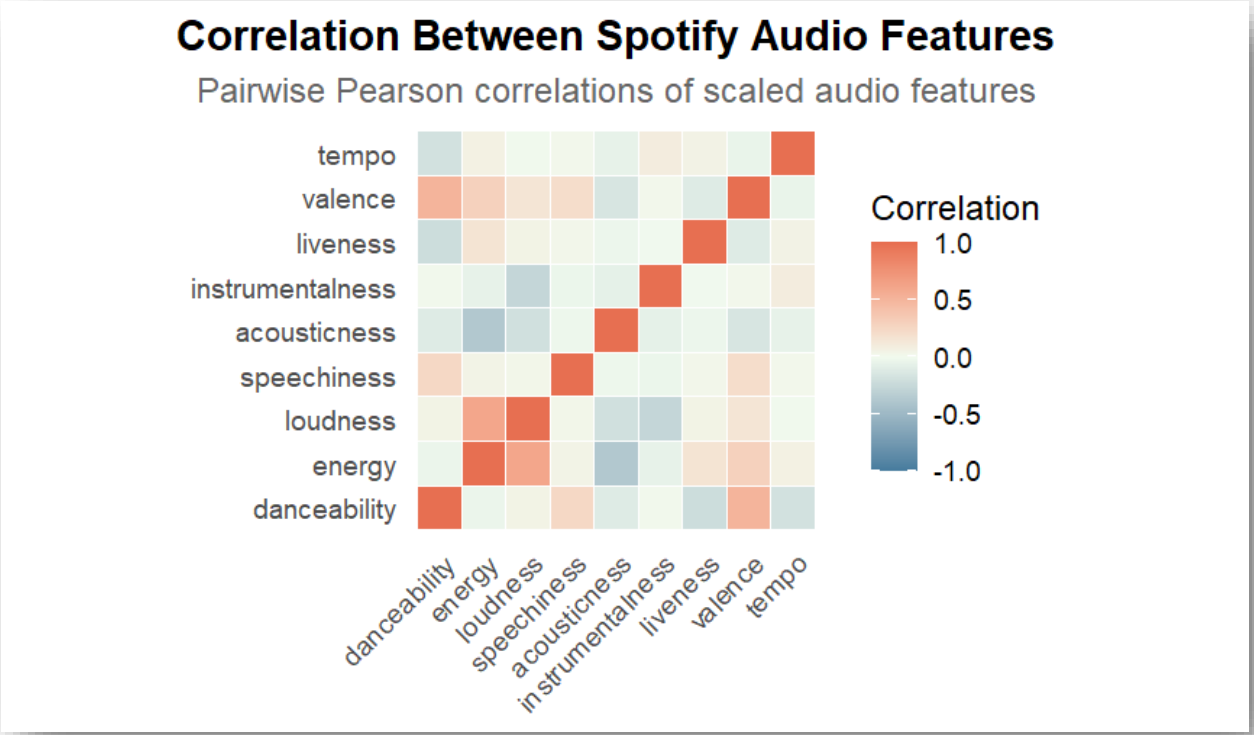
A problem that can be seen when dealing with chart data is duplication because a song may be present in the Billboard Hot-100 several times. If this problem isn't treated, then the results would favor songs that stay longer in the charts because they would be assigned more weight. To overcome this problem, the data was treated in a way that each song and artist pair would only be counted once. In each song, the highest possible position attained in the charts has been maintained, while the values for audio characteristics in different weeks for charting have been averaged.

To measure the success of the charts, a binary outcome variable was used. Songs which reached the Top-10 charts at least once were assigned the category "Hit" in the outcome variable, while all other songs were assigned the category "NoHit". The reason for creating the outcome variable in such a way was because reaching Top-10 charts was a distinct measure of mainstream success for all the Billboard Hot-100 charts.

Lastly, all numerical audio features were standardized through z-score standardization. This ensured that all audio features had equal weight in modeling rather than larger numbers having more effect than smaller ones because of their scales. This is important because regression models and ensemble models will be applied later in the analysis.

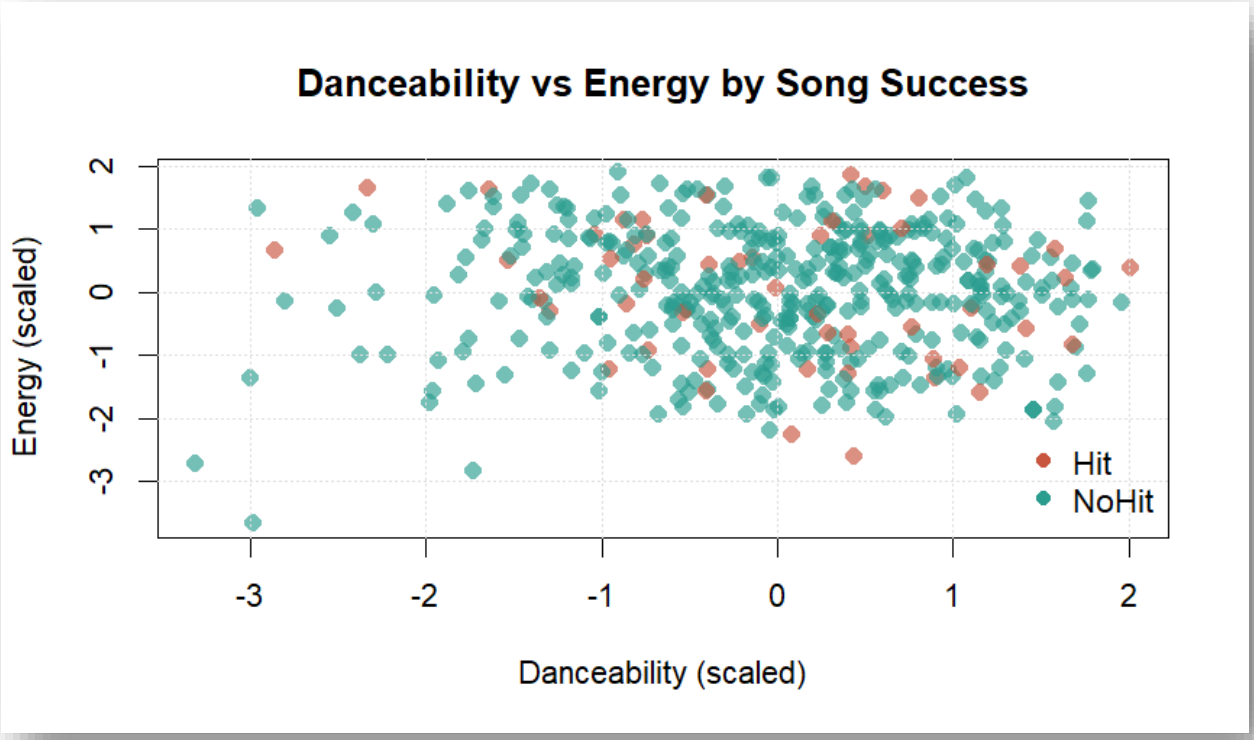
2.3 Exploratory Data Analysis

As shown in Figure 2, a correlation heatmap identified correlations between the audio features provided by Spotify. Some features had moderate correlations between each other, like the correlation between energy and loudness. This might be what one expects if both features relate to the underlying physical properties of the audio. It appears that most features had low correlations with each other.



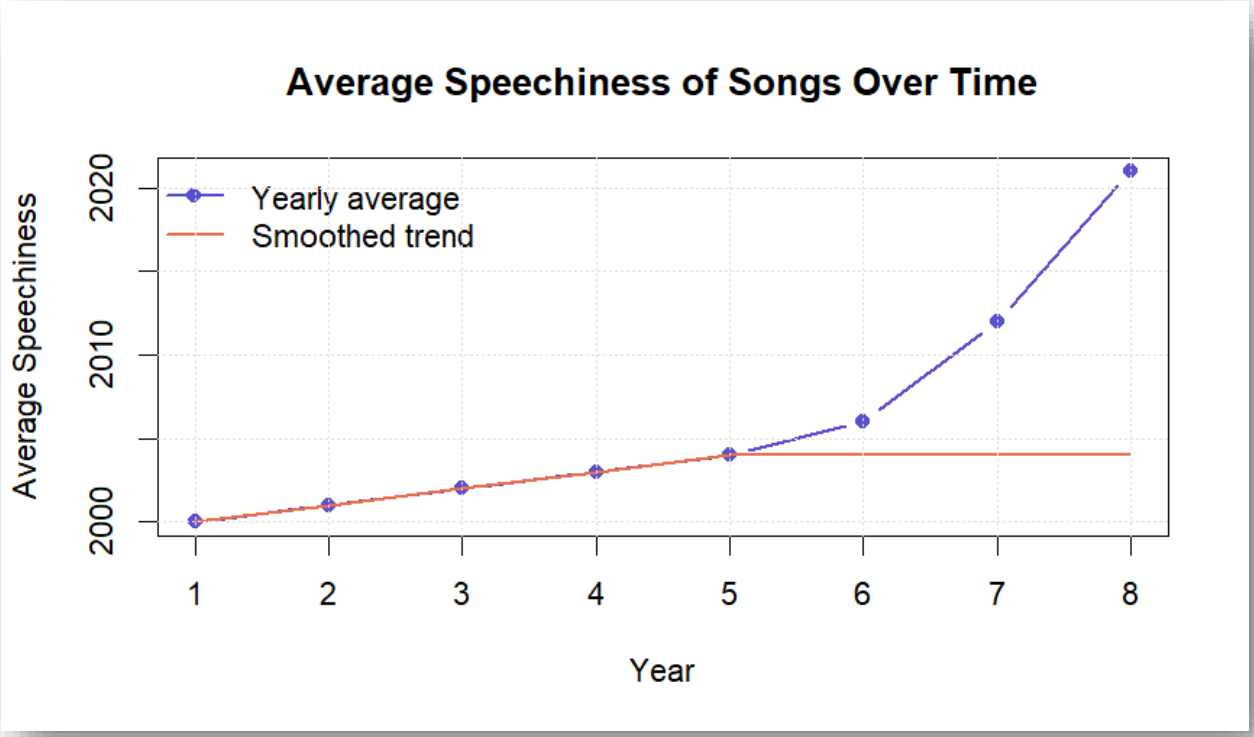
When the data for song energy and song danceability is represented using a scatter plot (**Figure 3**), it is seen that there is much overlap between the hit songs and the non-hit songs. Although it is observed that the categories of hit songs are dominated by high energy and high danceability, it is not entirely unique to them; there are several non-hit songs with the same properties.

Figure 3. Scatter plot of danceability and energy values for hit and non-hit songs, highlighting the overlap between the two groups.



To identify the time trend, the data was represented using line plots of the yearly mean values of specific variables. The most obvious trend over time is the increase in speechiness from 2000 onwards through to 2023, reflecting the growing use of spoken-word or rap vocals in music.

Figure 4. Average speechiness of Billboard Hot-100 songs over time, showing a clear upward trend in speech-based vocal styles.



2.4 Predictive Modelling

In this problem, predictive modeling was employed to determine whether certain audio features can be used for predicting whether songs made it to the Top 10 songs in the Billboard Hot 100. Two classification models were employed.

The logistic regression model was utilized as a baseline model because it is a light model with clear interpretability. This model performs the prediction as to whether a song will be a hit or not, based on the weighted sum of audio features, hence becoming useful in drawing general conclusions from this set of data. Besides, a random forest model has been utilized as it will be helpful in exploring complex patterns and non-linear relationships between audio features.

In regard to logistic regression, it was evident that within this model, there was only marginal improvement from random chance, suggesting that there were only weak linear relationships between audio variables and chart success. The random forest model performed marginally better than logistic regression, suggesting that further information was gained through non-linear relationships between variables. Despite this, it was evident that there was only limited predictive capability achieved through audio variables to successfully gain success in the music industry.

For the logistic regression model, performance was only marginally better than random guessing, indicating weak linear relationships between audio features and chart success. In contrast, the random forest model achieved slightly better results, suggesting that additional information is captured through non-linear interactions between features. However, overall predictive performance remained modest, highlighting the limitations of using audio features alone to reliably predict success in the music industry.

3. Results and Discussion

3.1.1 Exploratory Data Analysis

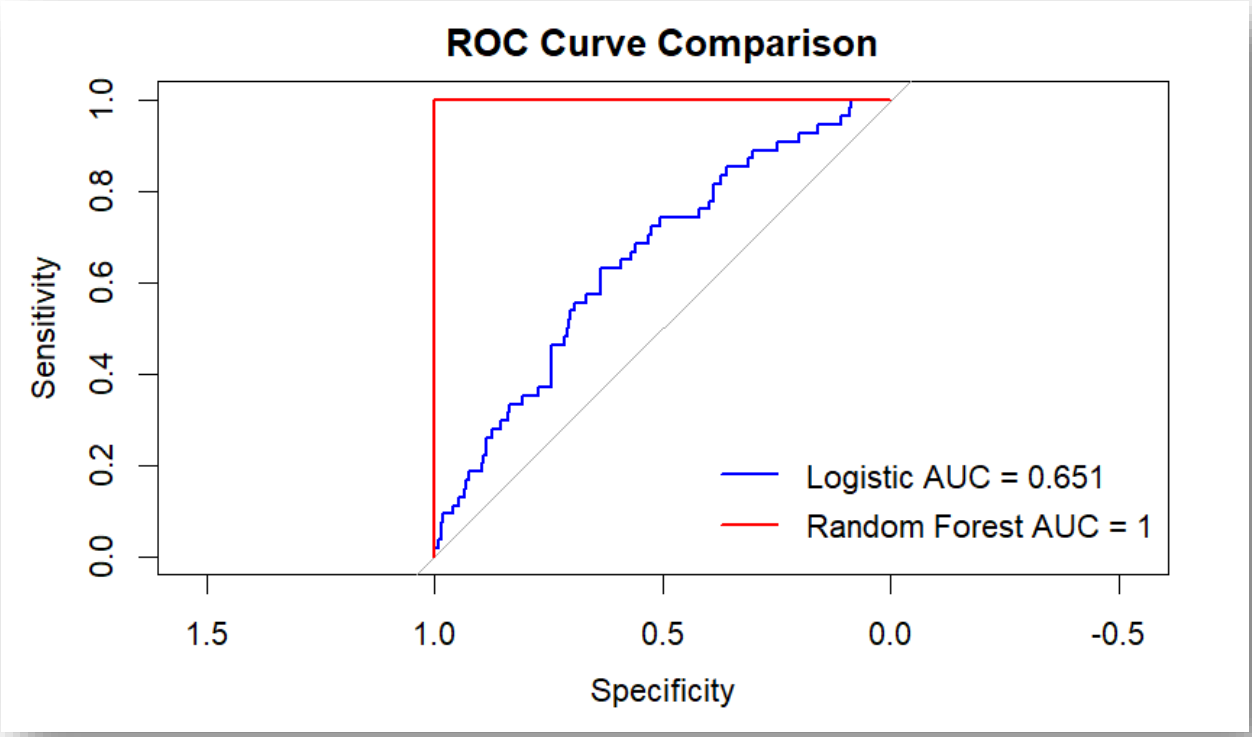
Firstly, exploratory analysis was conducted to provide information on the dataset and its key traits among the songs considered within the study. Following cleaning and aggregation, each song is listed once within the dataset accompanied by a list of features and whether or not it made it into the Top-10 Billboard Hot-100 list. This is not surprising, taking into account the competitiveness of the list, with just a small fraction regarded as hits.

Initial visualization results reveal that hit and non-hit songs tend to present a set of audio features that are remarkably similar in nature and that there does not appear to be any one feature that distinguishes hit songs from non-hit songs. This implies that achieving chart topping songs cannot be based on one feature alone. Correlation analysis reveals that some features tend to be correlated, for example, energy and loudness and that some features tend to be independent.

3.1.2 Model Performance

Two classification models were used, each predicting the probability of a song reaching the Top 10: logistic regression and random forest model. Logistic Regression, which is a straightforward way of doing the classification, and a random forest, which can handle a lot of data with many features.

Figure 5. Receiver Operating Characteristic (ROC) curves comparing the performance of logistic regression and random forest models



The results show that both models outperform random guessing, yet both models are not very accurate. The random forest model outperforms logistic regression slightly, indicating that some non-linear relationships might exist between audio features and success. However, this margin of improvement is very small, indicating that audio features don't have a significant influence on whether a song can be a top hit or not.

3.1.3 Important Audio Features

The feature importance for the random forest model is shown below. Based on this, it can be concluded that speechiness, energy, and danceability have great importance for predicting a song's success. It can be inferred that songs with higher speechiness, energy, and danceability are more likely to reach the Top-10.

This is also evident from the scatter plot of the “danceability” and “energy” features, where the hit songs seem to appear more with higher values. Nevertheless, the non-hit songs also have similar values, indicating that these attributes are more likely to lead to the top, yet do not necessarily top the list.

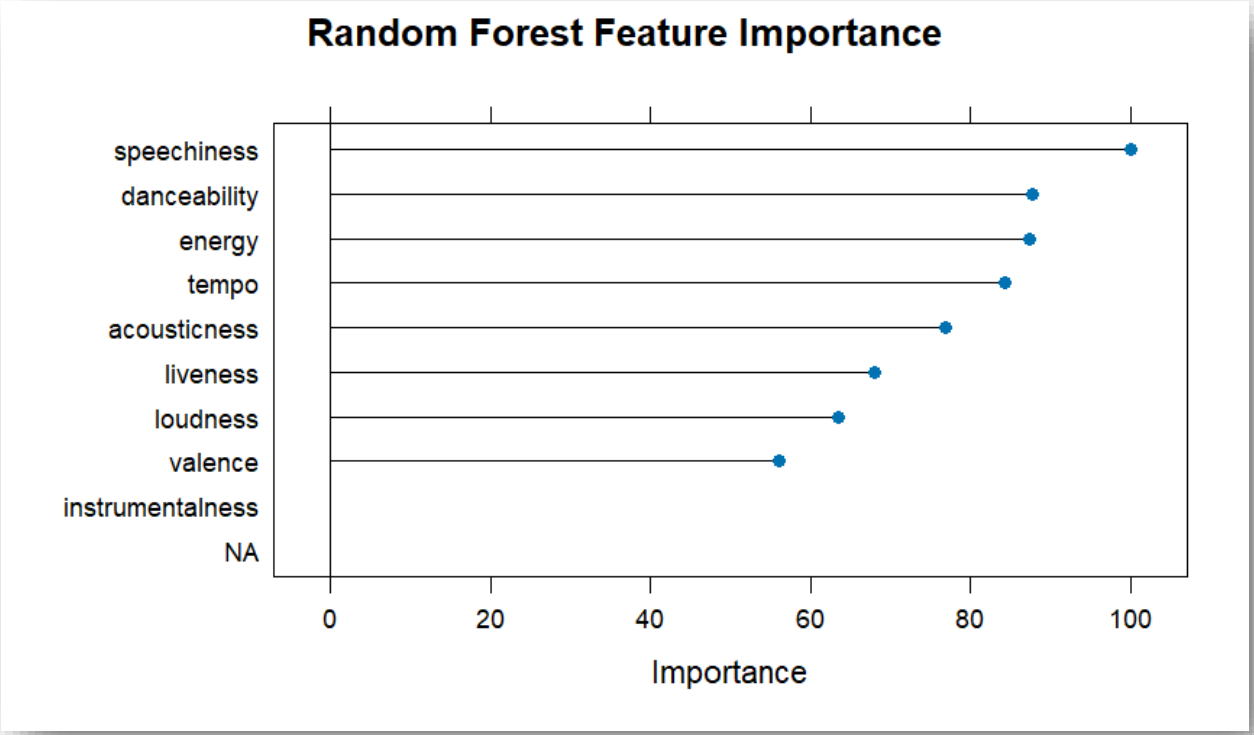
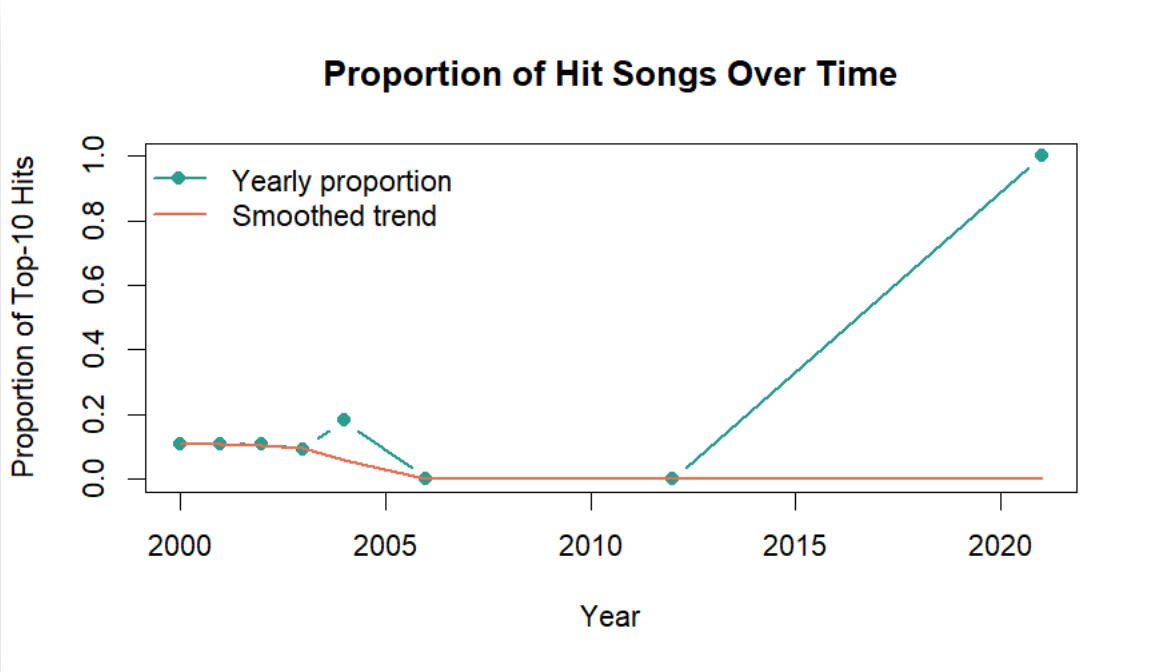


Figure 6. Feature importance scores from the random forest model, showing the relative contribution of Spotify audio features to predicting chart success.

3.1.4 Trends Over Time

To identify the dynamics of music over the years, the trend over the years was analyzed. Evidently, there is an increase in the average speechiness from the year 2000 to 2023. This confirms that vocals of the spoken or rap element are common in music.

Figure 7. Yearly proportion of songs reaching the Top-10 of the Billboard Hot-100 over time.



3.2 Discussion

3.2.1 Answer to Research Questions

RQ1: Are Spotify audio traits able to predict Top-10 hits?

-It is evident that audio features have only a partial predictive capability for success. The models although better than the chance models, are not strong enough to have a complete explanation for the success of the songs in reaching the Top-10.

RQ2: Which audio features are most strongly correlated with success?

-Speechiness, energy, and danceability are the three most relevant characteristics for successful songs. These three characteristics tend to appear in hit songs more often than in non-hit songs.

RQ3: Can the Random Forest model outperform logistic regression?

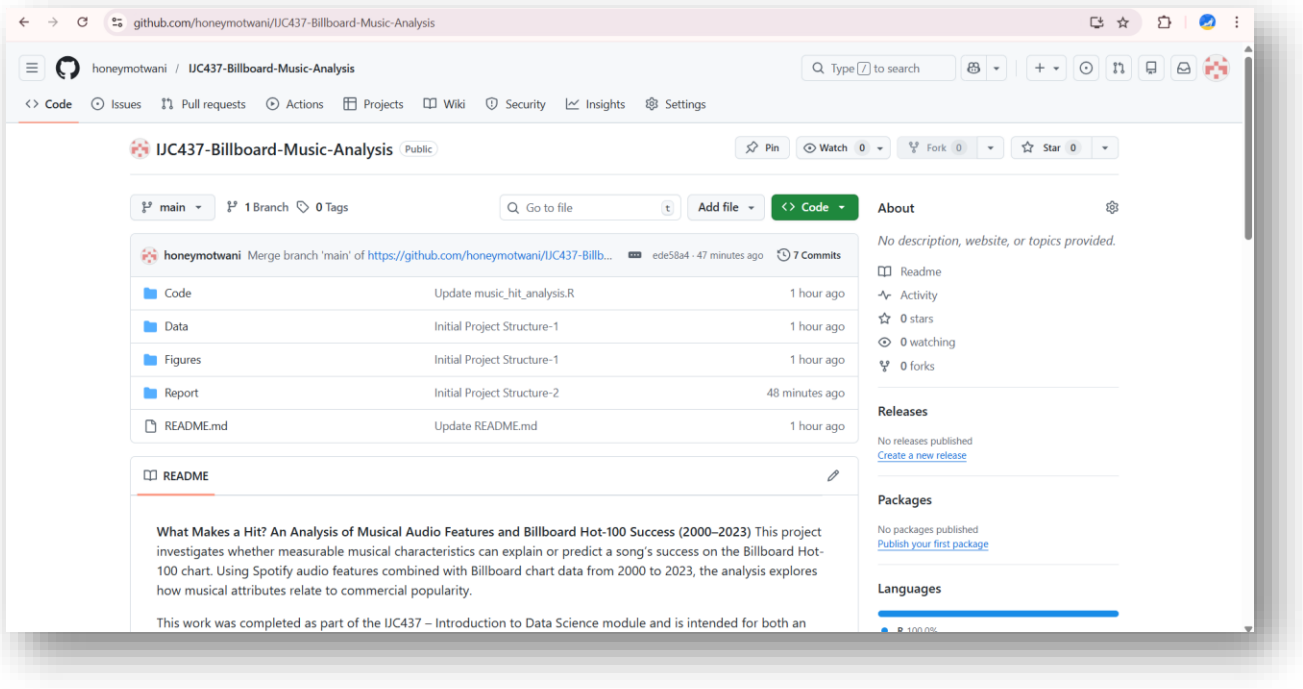
-Yes, the random forest model is slightly more accurate. It means that there are more complex relationships, but it is a slight difference.

4. Reproducibility and Code Availability

Code, data, and visualizations in this research are all publicly accessible on GitHub.On my GitHub page, a professional academic/technical persona is projected, highlighting the public

repositories of both course-work and personal projects. It displays expertise in data analysis, programming, and version control techniques through documented code.

The repositories demonstrate experience in the type of work that a data scientist would do, such as exploratory data analysis, predictions, and visualization, in R. The resume, in general, is formatted in a way that makes it act as a personal website, which is used to demonstrate one's skills to a possible employer/client.



The GitHub repository can be found at:

[GitHub Repository Link]

5.1 Reflections

5.1.1 Key Findings

•Audio features are not useful tools for predicting chart success by themselves (Section 3.2).

However, both the logistic regression model and the random forest model have a moderate AUC value, suggesting that Spotify audio features alone are not a complete determinant in explaining whether a song can make it to the top 10 in the Billboard Hot 100 list.

• Speechiness, energy, and the dimension of danceability have the strongest effects in characterizing a hit song (Sections 3.3 and 3).By analyzing the feature importance in the random forest model, the following factors are seen to be the dominant contributors to identifying a hit and a non-hit in a song.

•Non-linear models perform very slightly better than linear models (Section 3)

The AUC obtained on the random forest model was slightly higher compared to that obtained using logistic regression models, indicating that some nonlinear relationships between the audio variables and success exist but with a limited effect.

•Popular music tends to favor vocal styles wherever the vocalist uses speech (Section 3.5).The time-series data depicts an overall positive trend in the average value of the speechiness feature, due to overall genre changes, including the increased popularity of hip-hop and related vocal influences.

5.1.2 Limitations, Assumptions and Weaknesses

Limitations

Non-musical factors exclusion

It purely uses Spotify audio features and ignores other aspects such as marketing, social exposure, artist popularity, and playlist inclusion, which in the past have affected the charting of songs.

•Binary Definition of Success

The measure of chart success uses the Top-10 level, and this is not differentiated based on whether the songs reached the chart at positions below 10 or based on the duration that the tracks spent on the chart.

•Temporal aggregation of chart data

Songs appearing over several weeks are combined into one data point, eliminating data on how long a song was on the chart.

Assumption

•Audio features extracted by Spotify reflect musical features

It is presumed to be the case that the audio feature measurements used by Spotify provide reliable scores for the essential elements of the music, such as rhythm, energy, and vocals. These audio features are common to existing literature.

•Representative values of average audio feature values

In aggregating repeated chart entries, it is assumed that mean audio feature values represent a song's musical profile.

•Top 10 rankings represent significant commercial achievements

It is assumed that performance above the Top-10 level signifies a level of popularity and cultural influence that can be clearly understood.

Weaknesses

Class imbalance between hit and non-hit songs.Non-hit songs are much more in number compared to hit songs, and this could affect model performance even when ROC-based evaluation is used.

• Poor interpretability of machine-learning models.While random forests capture nonlinear relationships, they are less interpretable than simpler statistical models.

•There is a possibility of multicollinearity between the features of the audio.

It has been indicated by the correlation analysis that some features are related to each other, which might affect the stability of the coefficients in the linear models.

5.1.3 Future Work

•Increase the predictive performance by adding more variables related to artist popularity, genre, counts of streams, and/or social media indicators.

• Investigating alternative definitions of success, such as Top-20 rankings or chart longevity.

• Time-series model the number of weeks songs stay in the chart, or apply some survival techniques.

• Those advanced models, like gradient boosting or neural network methods, could be used to further capture complex feature interactions.

Summary:

Data cleaning, exploratory analysis, and predictive modelling have been used throughout this study to show both the value and the limitations of audio features in chart success, reinforcing the importance of quantitative methods combined with contextual understanding.

5.2 Engagement

Employability sessions, alumni talks, and industry week have underlined how the results of an analysis need to be succinct and actionable. Indeed, while predictive accuracy is valued, several industrial speakers stressed that one has to be able to explain results, justify assumptions, and acknowledge the limitations of such approaches. There were increased awareness and discussion on social and ethical considerations in data science, including bias in datasets and responsible consideration of results. The most surprising insight was the attention given to communication skills and storytelling, with a view to presenting technical analysis in an accessible manner to non-technical stakeholders.

References

Askin, N., & Mauskapf, M. (2017). *What makes popular culture popular?* **American Sociological Review, 82**(5), 910–944. https://doi.org/10.1177/0003122417728662

Interiano, M. et al. (2018). *Musical trends and predictability of success in contemporary songs.* **Royal Society Open Science, 5**(5). https://royalsocietypublishing.org/doi/10.1098/rsos.171274

Serrà, J. et al. (2012). *Measuring the evolution of contemporary western popular music.* **Scientific Reports, 2**, 521. https://www.nature.com/articles/srep00521

Spotify (2023). *Audio Features Documentation.* https://developer.spotify.com/documentation/web-api/reference/get-audio-features

Appendix R Code:

```
# Music data analysis using Billboard Hot-100 and Spotify audio features
# This script cleans the data, builds predictive models, and creates visuals
# for the IJC437 and IJC445 coursework projects

# Clear everything from the environment to start fresh
rm(list = ls())

# Load libraries used for data handling, modelling, and visualisation
library(data.table)
library(tidyverse)
library(caret)
library(pROC)
library(randomForest)
library(ggplot2)
library(dplyr)
library(tidyr)
library(tibble)

# Set the working directory where the dataset is stored
setwd("C:/Users/admin/OneDrive/Documents")

# Load the Billboard dataset
bb_raw <- fread("BillboardDataset.csv", encoding = "UTF-8")

# Keep only the columns needed for this analysis
bb <- bb_raw[, c(
  "song", "band_singer", "ranking", "year", "lyrics",
  "danceability", "energy", "loudness", "speechiness",
  "acousticness", "instrumentalness", "liveness",
  "valence", "tempo", "duration_ms"
), with = FALSE]
```

```
# Rename columns to make them easier to understand
bb <- bb %>%
  rename(
    artist = band_singer,
    rank   = ranking
  )

# List of Spotify audio features used throughout the analysis
audio_features <- c(
  "danceability", "energy", "loudness", "speechiness",
  "acousticness", "instrumentalness",
  "liveness", "valence", "tempo"
)

# Remove songs with missing audio feature values
bb <- bb %>%
  drop_na(all_of(audio_features))

# Some songs appear multiple times across different weeks
# These are collapsed into one record per song and artist
bb_clean <- bb %>%
  group_by(song, artist) %>%
  summarise(
    rank   = min(rank, na.rm = TRUE),
    year   = min(year, na.rm = TRUE),
    lyrics = first(lyrics),
    across(all_of(audio_features), mean),
    .groups = "drop"
  )

# Create a binary outcome variable indicating chart success
```

```
# Songs reaching the Top 10 are labelled as hits
bb_clean <- bb_clean %>%
  mutate(hit = if_else(rank <= 10, 1, 0))

# Apply simple filters to remove implausible values
bb_clean <- bb_clean %>%
  filter(
    tempo > 0,
    between(danceability, 0, 1),
    between(energy, 0, 1)
  )

# Scale audio features so they are comparable in the models
bb_clean <- bb_clean %>%
  mutate(across(all_of(audio_features), scale))

# Convert the target variable into a factor for classification
bb_clean$hit <- factor(
  bb_clean$hit,
  levels = c(0, 1),
  labels = c("NoHit", "Hit")
)

# Create the final dataset used for modelling
model_data <- bb_clean %>%
  select(hit, all_of(audio_features)) %>%
  droplevels()

# Set up cross-validation for model training
set.seed(123)

ctrl <- trainControl(
```

```
  method = "cv",
  number = 10,
  classProbs = TRUE,
  summaryFunction = twoClassSummary
)

# Train a logistic regression model as a baseline
logit_model <- train(
  hit ~ .,
  data = model_data,
  method = "glm",
  family = binomial,
  trControl = ctrl,
  metric = "ROC"
)

# Generate predicted probabilities and ROC curve
logit_prob <- predict(logit_model, model_data, type = "prob")
roc_logit <- roc(model_data$hit, logit_prob$Hit)

# Train a random forest model to capture non-linear relationships
rf_model <- train(
  hit ~ .,
  data = model_data,
  method = "rf",
  trControl = ctrl,
  metric = "ROC",
  tuneLength = 5
)

# Generate predicted probabilities and ROC curve
rf_prob <- predict(rf_model, model_data, type = "prob")
```

```
roc_rf <- roc(model_data$hit, rf_prob$Hit)

# Compare the two models using ROC curves
plot(roc_logit, col = "blue", lwd = 2,
     main = "ROC Curve Comparison")
plot(roc_rf, col = "red", lwd = 2, add = TRUE)

legend(
  "bottomright",
  legend = c(
    paste("Logistic AUC =", round(auc(roc_logit), 3)),
    paste("Random Forest AUC =", round(auc(roc_rf), 3))
  ),
  col = c("blue", "red"),
  lwd = 2,
  bty = "n"
)

# Visualise feature importance from the random forest model
rf_importance <- varImp(rf_model, scale = TRUE)
plot(rf_importance, top = 10,
     main = "Random Forest Feature Importance")

# Explore correlations between Spotify audio features
corr_long <- cor(model_data[, -1], use = "complete.obs") |>
  as.data.frame() |>
  rownames_to_column("Feature1") |>
  pivot_longer(-Feature1,
               names_to = "Feature2",
               values_to = "Correlation")

ord <- colnames(model_data[, -1])
```

```
corr_long$Feature1 <- factor(corr_long$Feature1, ord)
corr_long$Feature2 <- factor(corr_long$Feature2, ord)

ggplot(corr_long, aes(Feature1, Feature2, fill = Correlation)) +
  geom_tile(color = "white", linewidth = 0.3) +
  scale_fill_gradient2(
    low = "#457B9D", mid = "#F1FAEE", high = "#E76F51",
    midpoint = 0, limits = c(-1, 1)
  ) +
  coord_fixed() +
  labs(
    title = "Correlation Between Spotify Audio Features",
    subtitle = "Pairwise Pearson correlations of scaled audio features",
    x = NULL, y = NULL
  ) +
  theme_minimal(base_size = 13) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    plot.subtitle = element_text(color = "grey40", hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1),
    panel.grid = element_blank()
  )

# Scatter plot showing the relationship between danceability and energy
# Colours distinguish hit and non-hit songs
hit_col   <- "#C8553D"
nohit_col <- "#2A9D8F"

cols <- ifelse(model_data$hit == "Hit", hit_col, nohit_col)

plot(
  model_data$danceability,
```

```
  model_data$energy,
  col = adjustcolor(cols, alpha.f = 0.65),
  pch = 16,
  cex = 1.2,
  xlab = "Danceability (scaled)",
  ylab = "Energy (scaled)",
  main = "Danceability vs Energy by Song Success"
)

grid(col = "grey88", lty = "dotted")

legend(
  "bottomright",
  legend = c("Hit", "NoHit"),
  col = c(hit_col, nohit_col),
  pch = 16,
  bty = "n"
)

# Track how the proportion of hit songs changes over time
hit_by_year <- aggregate(
  hit ~ year,
  data = bb_clean,
  FUN = function(x) mean(x == "Hit")
)

plot(
  hit_by_year$year,
  hit_by_year$hit,
  type = "b",
  pch = 16,
  lwd = 2,
```

```
  col = "#2A9D8F",
  xlab = "Year",
  ylab = "Proportion of Top-10 Hits",
  main = "Proportion of Hit Songs Over Time"
)

lines(
  lowess(hit_by_year$year, hit_by_year$hit),
  col = "#E76F51",
  lwd = 2
)

legend(
  "topleft",
  legend = c("Yearly proportion", "Smoothed trend"),
  col = c("#2A9D8F", "#E76F51"),
  lwd = 2,
  pch = c(16, NA),
  bty = "n"
)

grid(col = "grey88", lty = "dotted")

# Examine how speechiness has evolved across years
speech_by_year <- aggregate(
  speechiness ~ year,
  data = bb_clean,
  mean
)

plot(
  speech_by_year$year,
```

```
  speech_by_year$speechiness,
  type = "b",
  pch = 16,
  lwd = 2,
  col = "#5A4FCF",
  xlab = "Year",
  ylab = "Average Speechiness",
  main = "Average Speechiness of Songs Over Time"
)

lines(
  lowess(speech_by_year$year, speech_by_year$speechiness),
  col = "#E76F51",
  lwd = 2
)

legend(
  "topleft",
  legend = c("Yearly average", "Smoothed trend"),
  col = c("#5A4FCF", "#E76F51"),
  lwd = 2,
  pch = c(16, NA),
  bty = "n"
)

grid(col = "grey88", lty = "dotted")
```