# Billboard Hot 100 Hit Prediction

1st Somay Jalan *Roll No: 2022505*

2nd Kartikeya Chhikara *Roll No:2022242*

May 13, 2024

## 1 Introduction

Each year, Billboard publishes its Year-End Hot 100 songs list, which denotes the top 100 songs of that year. The objective of this project was to see whether or not a machine learning classifier could predict whether a song would become a hit

## 2 LITERATURE REVIEW

### 2.1 Significance of Predicting Hit Songs

The ability to predict the success of a song before its release holds immense value for both artists and the music industry. It can inform marketing strategies, optimize promotional efforts, and potentially influence creative decision-making processes.

### 2.2 How does Machine Learning fit into this?

Machine learning algorithms have emerged as powerful tools for analyzing vast amounts of music-related data and extracting patterns that contribute to song popularity. By leveraging features extracted from audio, lyrics, and metadata, these algorithms can learn to predict the commercial success of a song.

### 2.3 Methodologies used for this paper.

- **Data Mining:** Kaggle Dataset and Spotify APIs has been used to obtain songs and features like Loudness, Speechiness, Acousticness etc.
- **Machine Learning Algorithms :** Various Algorithms like Random Forest, Quadratic Discriminant Analysis, AdaBoost etc. has been used to predict the song popularity.
- **Evaluation Criteria:** The dataset created was broken into train and test data set and the test dataset was used for testing accuracy of the model.

# 3 Dataset and Statistics of dataset

## 3.1 Dataset Creation.

Hot 100 songs for each year by billboard was scrapped from Wikipedia[1].
A sample of 19000 songs was downloaded from the Kaggle[2].
Both the databases where combined and all the songs which were taken from Wikipedia dataset was assigned a Top100 value of 1 and songs taken from Kaggle dataset was assigned a Top100 value of 0. The songs which were present in both the datasets, only one copy was taken into the combined dataset and was assigned a Top100 value of 1. The combined dataset contained Song Name, Artist Name, Year and Top100 for each.
Using Spotify API, song name and artist name was fed into the API and Audio Features was retrieved for each song.
Audio Features which were retrieved for each song are - Danceability, Valence, Energy, Tempo, Loudness, Speechiness, Instrumentalness, Liveness, Acousticness.

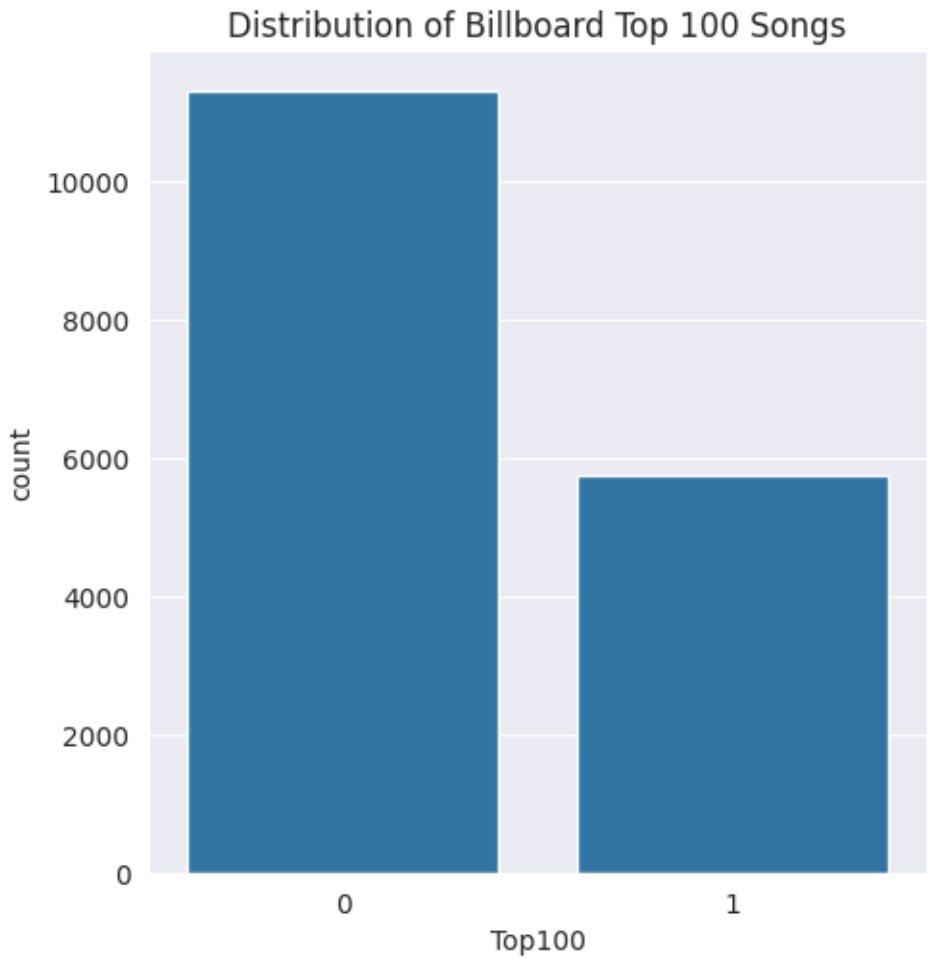


Figure 1: Data Distribution according to top100.

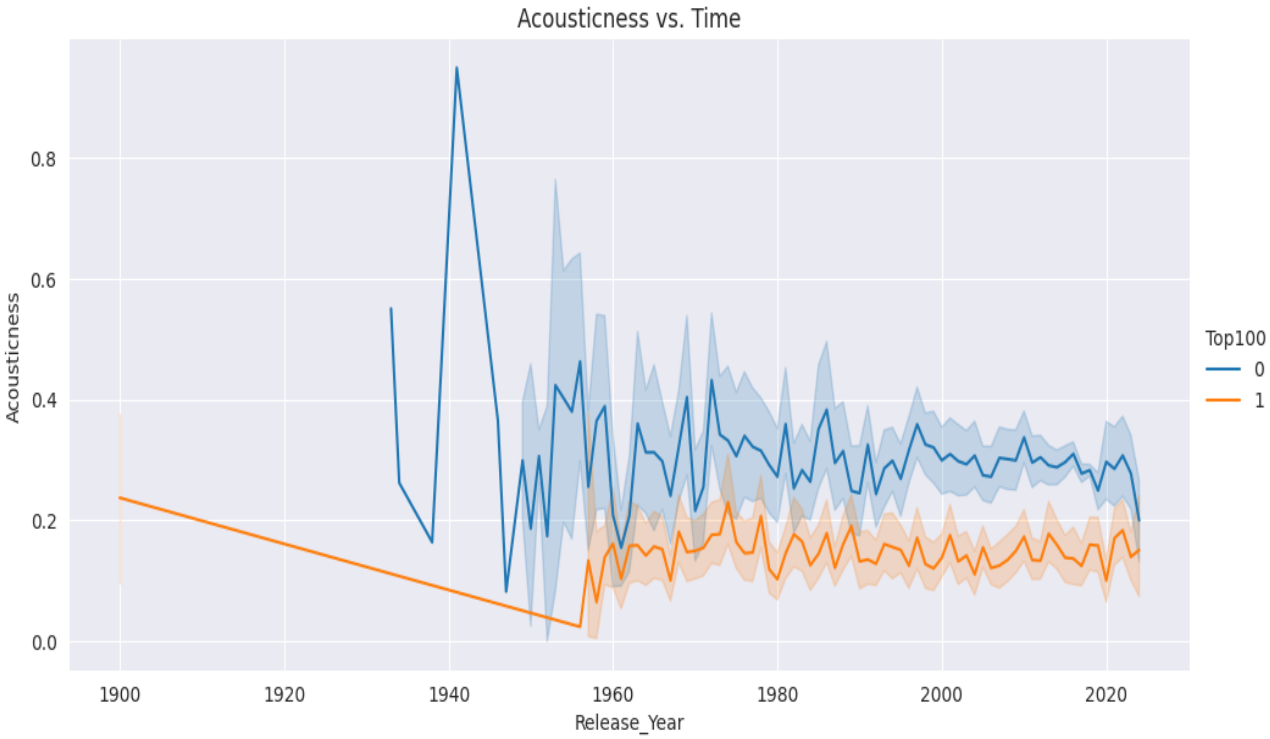


Figure 2: Data Distribution acoustic vs time.

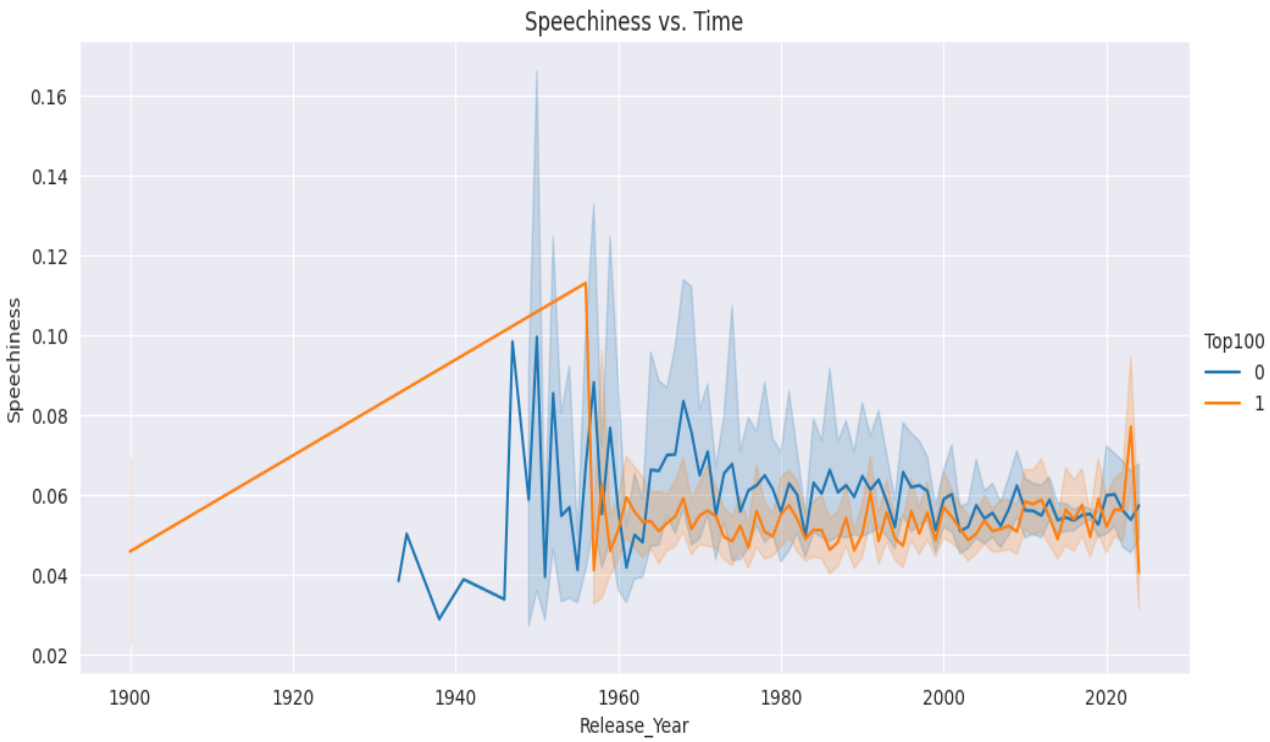


Figure 3: Data Distribution speechiness vs time.

### 3.2 Statistical Analysis for Dataset.

- Distribution of Top100 in Dataset."Fig. 1"
- Distribution of Acousticness vs Time."Fig. 2" It can be seen that the orange line is significantly and consistently below the blue line hence signifying that this can be a useful feature for Machine Learning Algorithm to Decide on.
- Distribution of Speechiness vs Time."Fig. 2" It can be seen that the orange line over the years very close to the blue line and may provide no significance to our decision.

Similar graphs can be generated for all the features but we are not including it as graph of each feature doesn't convey anything and from these graphs which signify the opposite ends of the spectrum - where there is a visible difference in feature for Top100 and not Top100 and where there is no difference in feature for Top100 and not Top100

## 4 Algorithms Used and Results for Each Algorithm

### 4.1 BaseLine

For each Machine Learning we have to set a baseline. It is the accuracy we get if we guess the output.
For base line,
Accuracy: 0.662681669010783
AUC: 0.5

### 4.2 Logistic Regression

Running Logistic Regression for the dataset gives us,
Accuracy: 0.7044069385841538
AUC: 0.7240928276991474

### 4.3 Linear Discriminant Analysis

Running Linear Discriminant Analysis for the dataset gives us,
Accuracy: 0.7025316455696202
AUC: 0.7199107287933564

### 4.4 Quadratic Discriminant Analysis

Running Quadratic Discriminant Analysis for the dataset gives us,
Accuracy: 0.6315049226441631
AUC: 0.7488374659818838

### 4.5 AdaBoost

Running AdaBoost for the dataset with number of trees set to 10000 gives us,
Accuracy: 0.823722456633849
AUC: 0.914884835571218

### 4.6 Random Forest

Running Random Forest for the dataset with number of trees set to 1000 and with maximum depth of each tree set to 10 gives us,
Accuracy: 0.8511486169714018
AUC: 0.9335220067191847

### 4.7 Bagging

Running Bagging for the dataset with number of trees set to 1000 gives us,
Accuracy: 0.8511486169714018
AUC: 0.9345138817021311

### 4.8 Gradient Boosting

Running Gradient Boosting for the dataset with number of trees set to 5000 and minimizing criteria as squared error gives us,
Accuracy: 0.8506797937177684
AUC: 0.93378454017192

### 4.9 K-Nearest Neighbors

Running K-Nearest Neighbors for the dataset with k set to 10 gives us,
Accuracy: 0.8335677449601501
AUC: 0.9173867695430714

## 5 Comparison of results.

Taking accuracy as the measure of how good each model performed, random forest, bagging and gradient boosting algorithm performed similarly good. They had similarly good AUC values also. The confusion matrix for each algorithm was also generated and compared. We are not attaching each individual confusion matrix as it will be redundant but it is provided with the paper in case anyone would like to study more. ROC curves for each algorithm for each is also generated and provided.

## References

[1] Wikipedia Link:
https://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles_of_1999

[2] Kaggle Link:
https://www.kaggle.com/edalrami/19000-spotify-songs