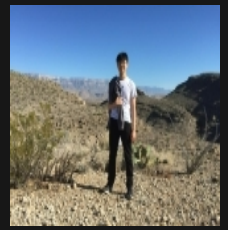
Predicting the Next Big Hit using Spotify

The Team

Anvesh Karangula

Artist

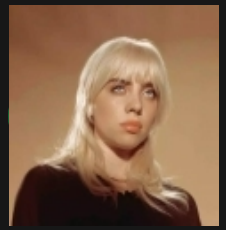


David Gong

Artist

Ian McIntosh

Artist



Raffaele Mannarelli

Artist

Raghav Vaidya

Artist

Upgrade

D. Chakrabarti

The Agenda

#	Title	Tasks	Slide Number	
1	Introduction and Problem Statement	Explain what we are trying to solve	4	02:00
2	The Dataset	Detail the dataset we are using	5	02:00
3	Exploratory Data Analysis (EDA)	Describe what do we observe	7	02:00
4	Solution and Insights	Share what we have learned	9	02:00
5	Conclusion and Further Studies	Discuss what can be improved	13	02:00

Upgrade

D. Chakrabarti

Introduction And Problem Statement

For years labels have been wondering what exactly it is that makes a song a hit. Many of the companies we work for have produced countless hits but also countless misses, and never know which category the song will fall into until it is released. They want a way to be able to tell for certain whether a song will be a hit before they release it to the public, thereby taking the guesswork out of selecting which songs to release.

Play

Follow

The Dataset

01 **Artist**: Artist Name

02 **Track**: Name of the Song

03 **Sections:** The number of sections the track has

04 **Time Signature:** Estimated overall time signature of the track

05 **Danceability**: How suitable a track is for dancing based on musical elements including tempo, rhythm stability, beat strength, and regularity

06 **Energy**: A perceptual measure of intensity and activity

07 **Key**: Estimated overall key of the track (0 = C, 1 = C?/D?, 2 = D, and so on)

08 **Loudness**: The overall loudness of a track in decibels (dB)

09 **URI**: The unique resource identifier for each song

10 **Mode**: Major is represented by 1, minor is 0

11 **Speechiness**: Speechiness detects the presence of spoken words in a track

12 **Acousticness**: A confidence measure of whether the track is acoustic

13 **Duration**: The duration of the track in milliseconds

14 **Liveness**: Detects the presence of an audience in the recording

15 **Valence**: A measure describing the musical positiveness conveyed by a track

16 **Tempo**: The overall estimated tempo of a track in beats per minute (BPM)

17 **Chorus Hit**: Estimate of when the chorus starts for the track

18 **Instrumentalness**: Predicts whether the track contains vocals

The Dataset

Target Variable

Popularity

Whether a song is a hit or a flop. If the song is a hit, then its value will be 1, and if the song is a flop, its value will be 0

The condition of a track being a hit is that it has been featured at least once in the weekly Billboard Hot 100 list.

Anything else is considered a flop.

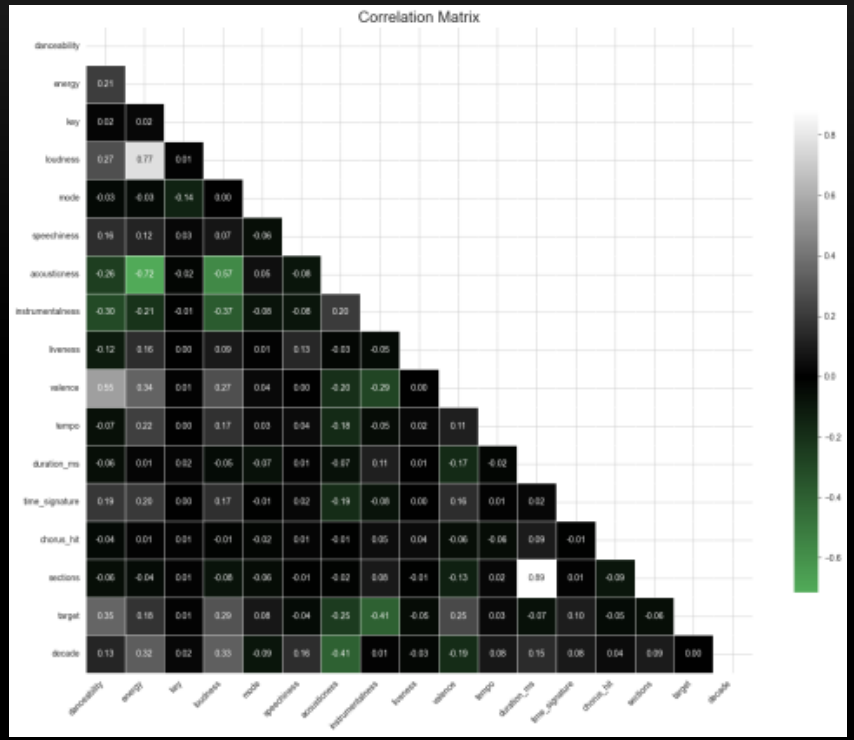
Exploratory Data Analysis

Top 5 features positively correlated with target:

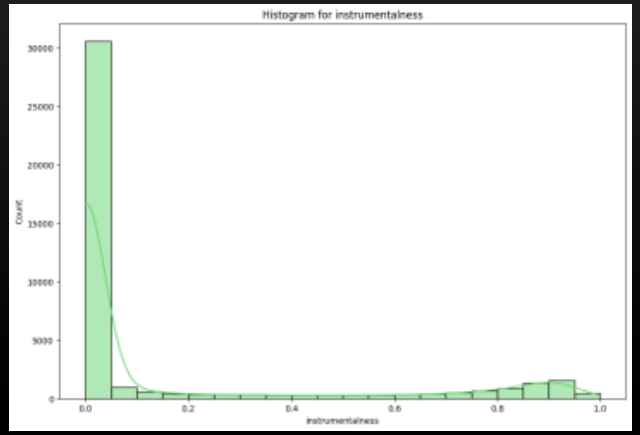
danceability	0.346097
loudness	0.286034
valence	0.251147
energy	0.177142
time_signature	0.104884

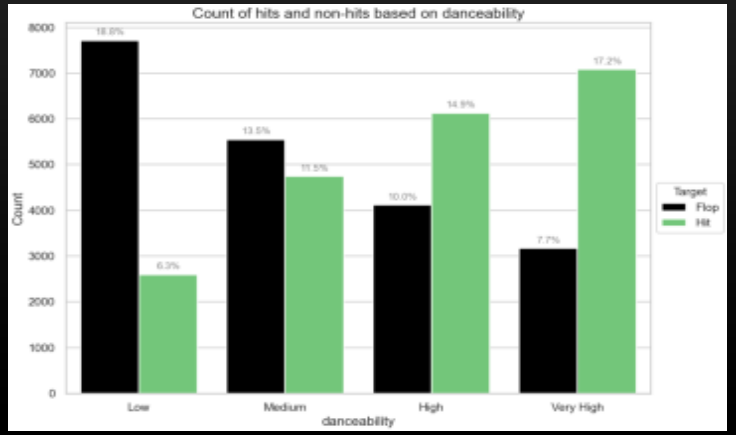
Top 5 features negatively correlated with target:

instrumentalness	-0.407638
acousticness	-0.246036
duration_ms	-0.073820
sections	-0.059997
liveness	-0.051445



Exploratory Data Analysis

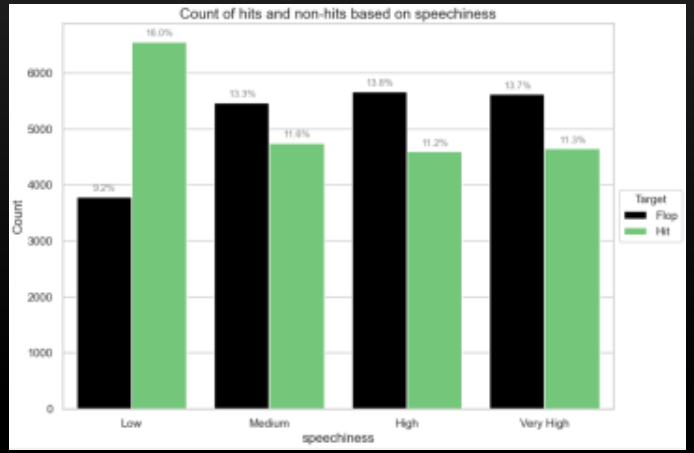


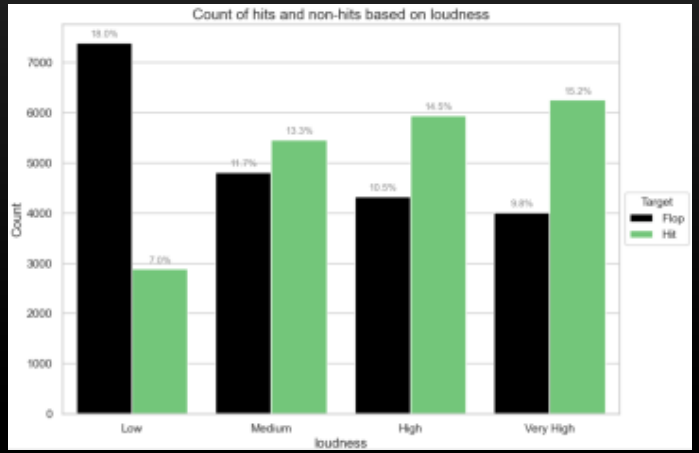


- High majority of songs scored very low on instrumentalness (~70%)
- Models considered feature very important
- Not unique feature to most hit songs because the majority will have words
- So we removed instrumentalness before running model

- Songs with high danceability scores were more often hits
- Songs with low danceability scores were more often flops
- Danceability appears to be very important variable in predicting whether a song is a hit

Exploratory Data Analysis





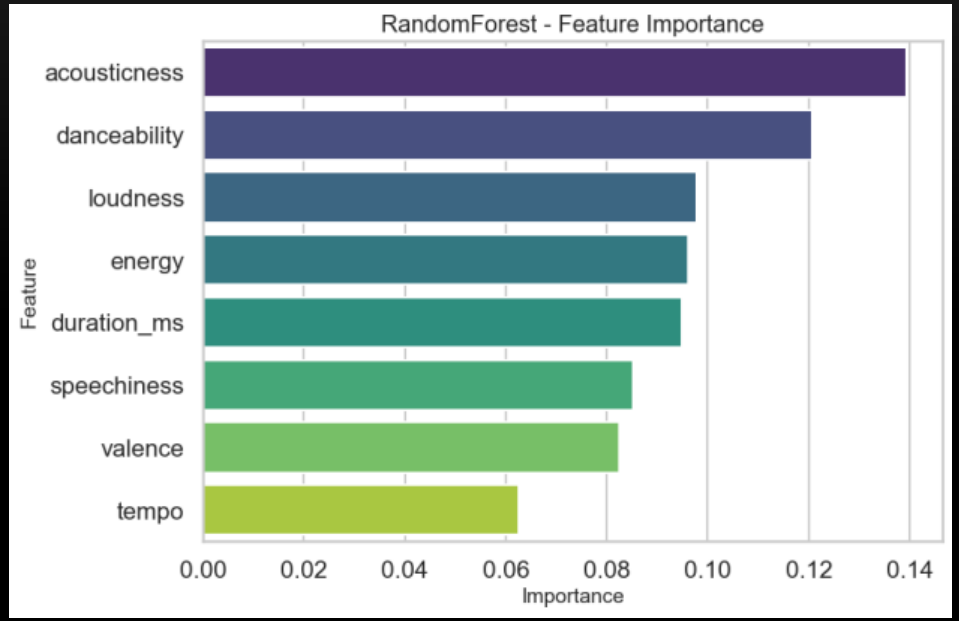
- Songs with low speechiness scores were more often hits
- Songs with high speechiness scores were more often flops
- This indicates that more users tend to prioritize the music over the lyrics

- As the songs' loudness increased, so did the chance of them becoming a hit
- Softer songs seem to flop more often
- One reason for the popularity of loud music could be that loud music is more often played in social events such as parties, marriages, gatherings etc.

Solution and Insights

Random Forest

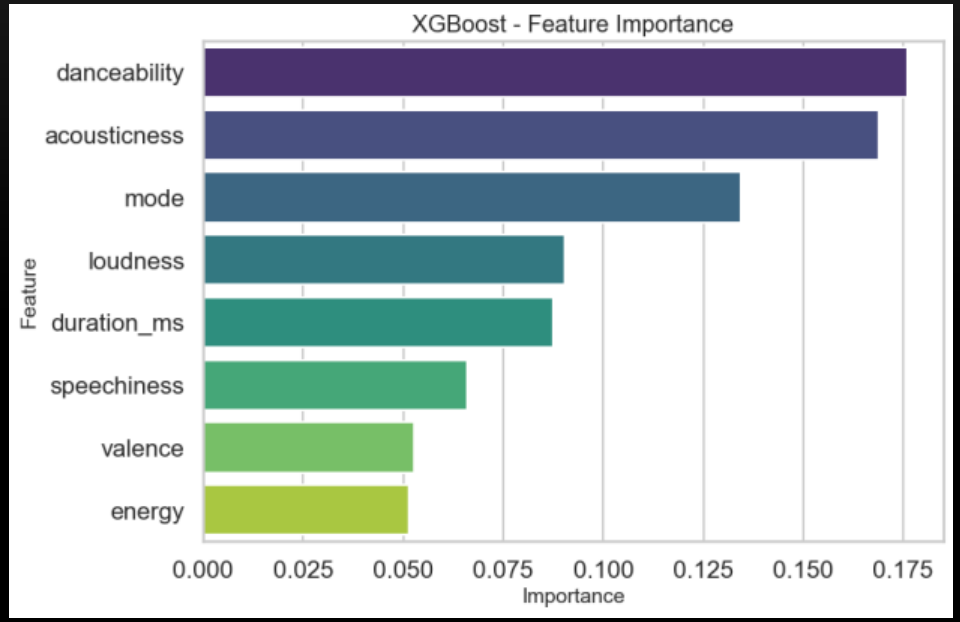
- Accuracy 76%
- Acousticness most important
- Danceability and Loudness conclude top 3



Solution and Insights

XGBoost

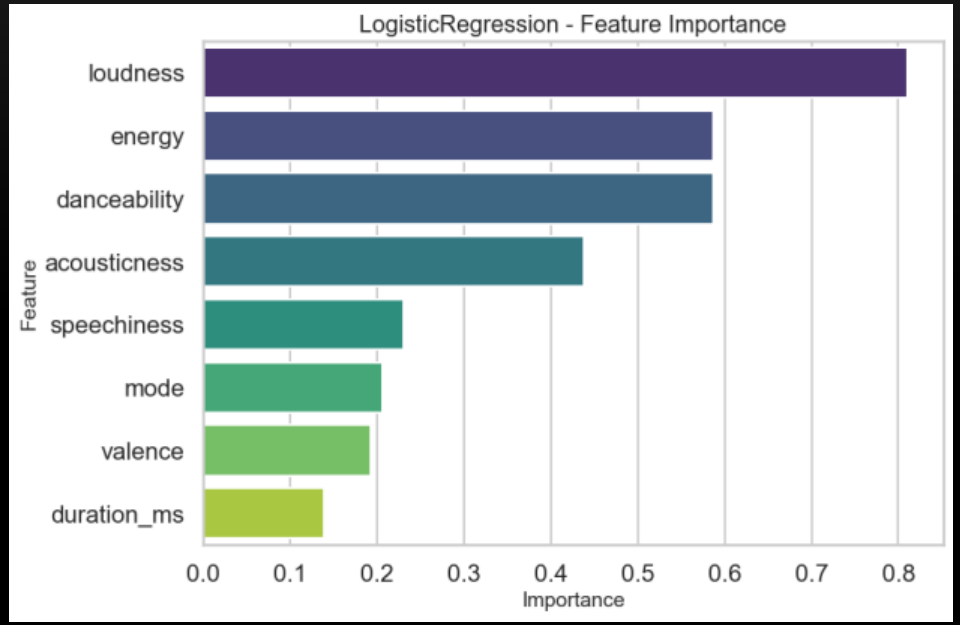
- Accuracy 75%
- Danceability most important
- Acousticness also within the top 3



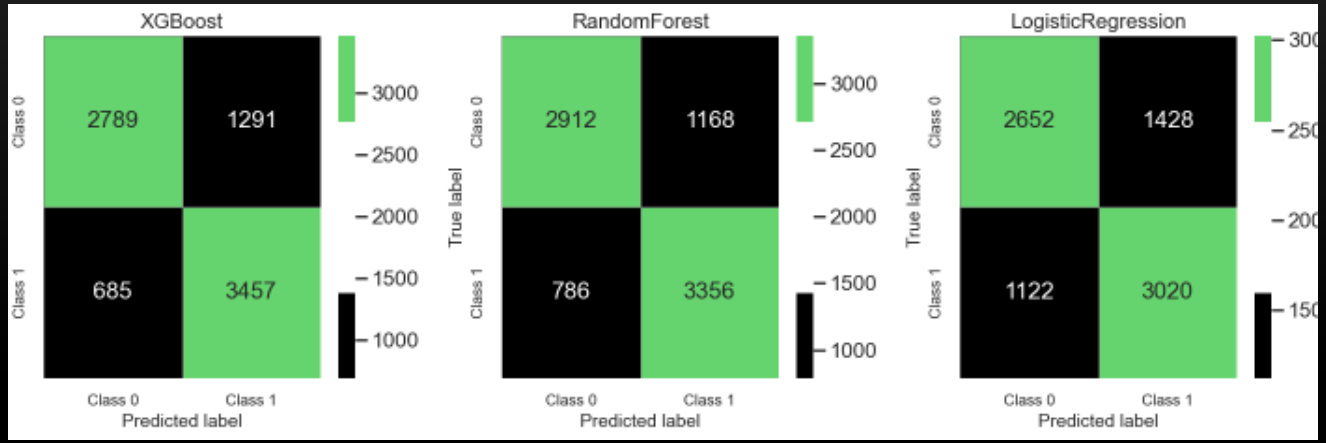
Solution and Insights

Logistic Regression

- Accuracy 68%
- Loudness by far the most important
- Danceability is within the top 3



Solution and Insights



Random Forest (most accurate model): 76.23%

XGBoost : 75.97%

Logistic Regression: 68.[illegible]

Conclusion and Further Studies

- Random forest model is chosen with an accuracy of around 76%
- Songs with more Danceability and Acousticness tend to be more popular
- App/Web app will be rolled out to predict the popularity based on the data input

Limitations:

- Oversampling - since the number of flop is equal to number of hits
- We can't predict how popular a song is going to be, just whether it is popular or not.

Conclusion and Further Studies

Further Studies:

- Add more dimensional data
- Group by artist, genre, release period to better predict the values
- Predict the number of streams and the amount of revenue that will be generated

Thank You

And so ends our awesome presentation :(

This is so sad, Alexa play Despacito.

Despacito

Luis Fonsi, Daddy Yankee

0:23 -3:25