Predicting Hit Songs on Spotify

A combination of Lyrics and Audio Features

Warner Music Group: Ryan Faus
NYU Center for Data Science: Preet Gandhi, Kenia Saenko, Stella Sun, Wenjie Sun, Ben Zhang

WARNER MUSIC GROUP

NYU Center for Data Science

Abstract

In this research, we are using lyrics, audio features, and producer / songwriter information to predict music's popularity before a song is released.

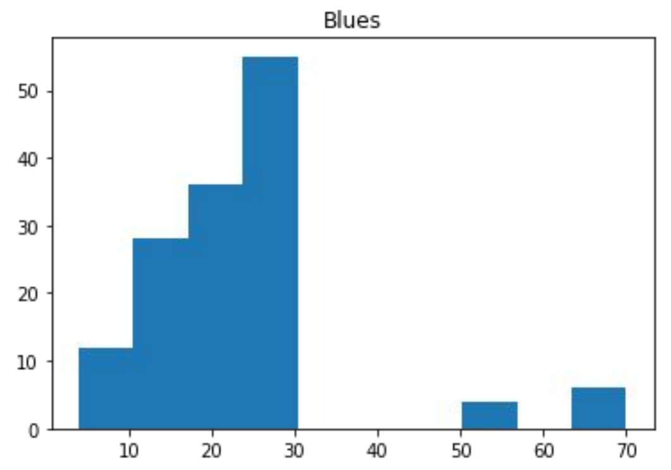
Introduction

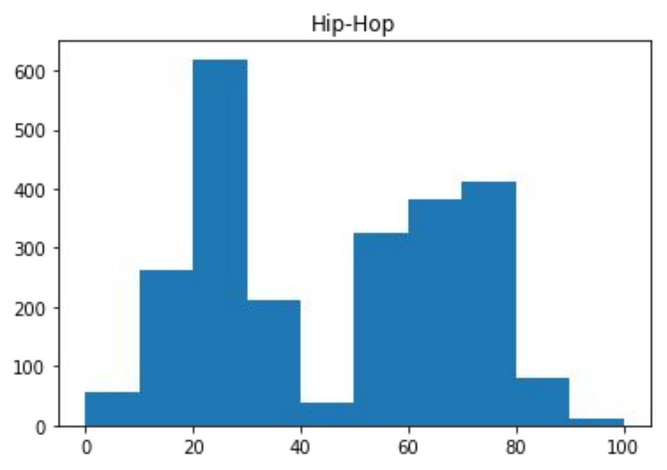
Problem statement

Given a track before it is released to public, how do we know if it would be a "hit" song on Spotify, which is the biggest music streaming platform that has 170 million worldwide users. Without any marketing strategy or promotion, only from the track attributes: title, artist, producers, lyrics, melodies, is it sufficient to predict its success?

Challenges

- Industry AUC score doesn't go above <u>60</u>
- Strict copyright from obtaining a large dataset
- Spotify doesn't share with WMG information except their own productions
- Highly Imbalanced Data (6% Hit Song) and different distributions by genre





Data Source

WARNER MUSIC GROUP

GENIUS

Spotify

Audio

Lyrics

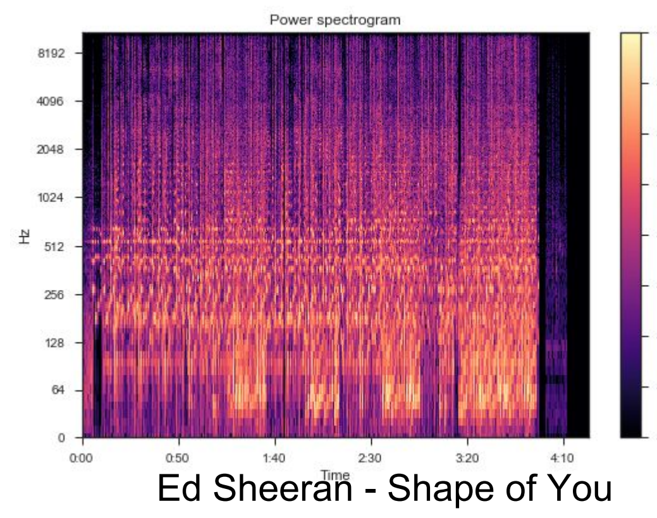
Producer Songwriter

Genre

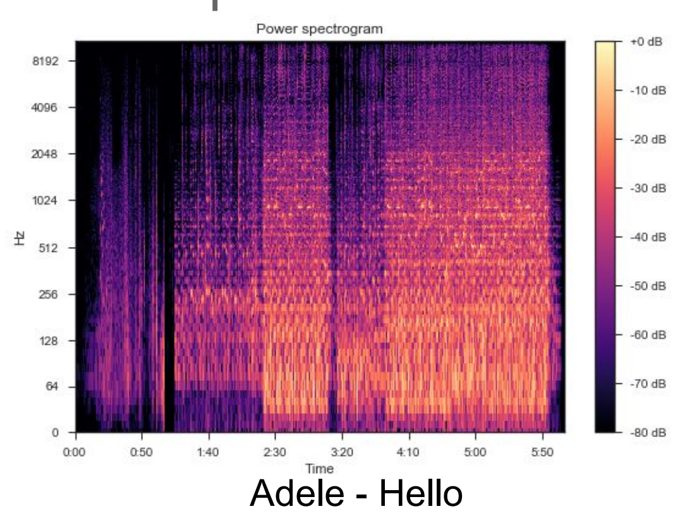
Data Processing

Audio

Melspectrogram Loud & Repetitiveness



Ed Sheeran - Shape of You



Adele - Hello

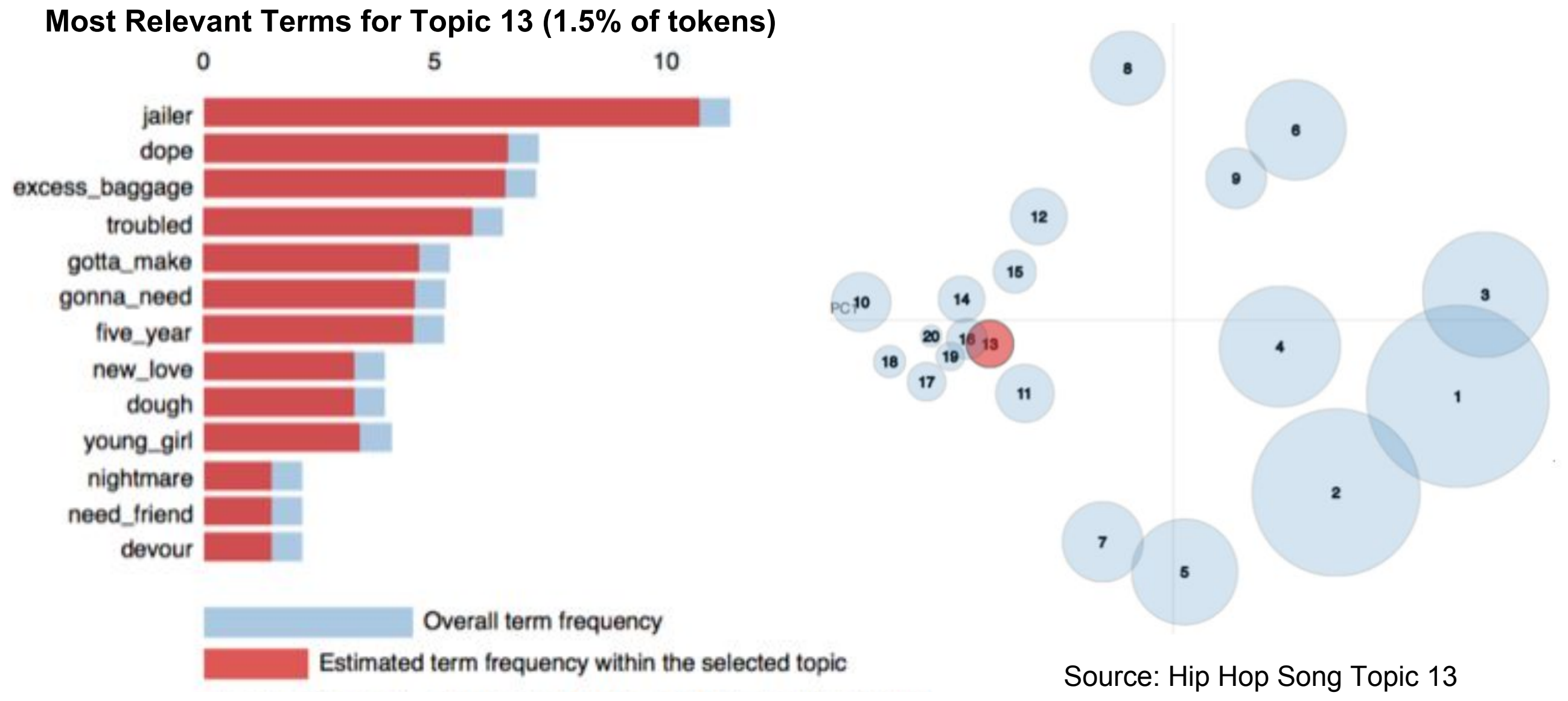
Song Tags Vocal & Instrument feature

🎸 🎹 👩‍🎤 & 47 other features

Lyrics

Topic Modeling (LDA) from Gensim

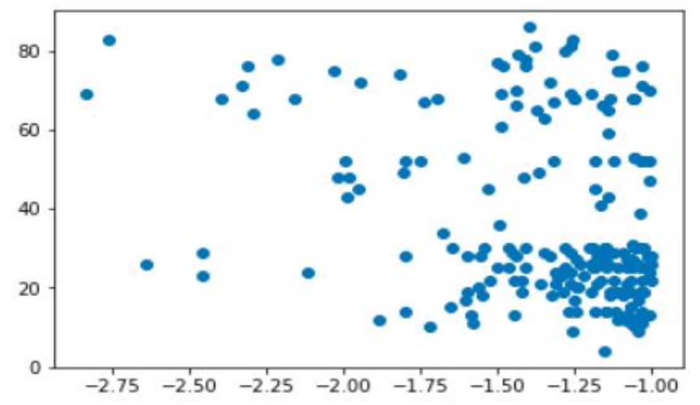
This hip hop topic 13 can inferred to be drug/crime related from the words associated



Source: Hip Hop Song Topic 13

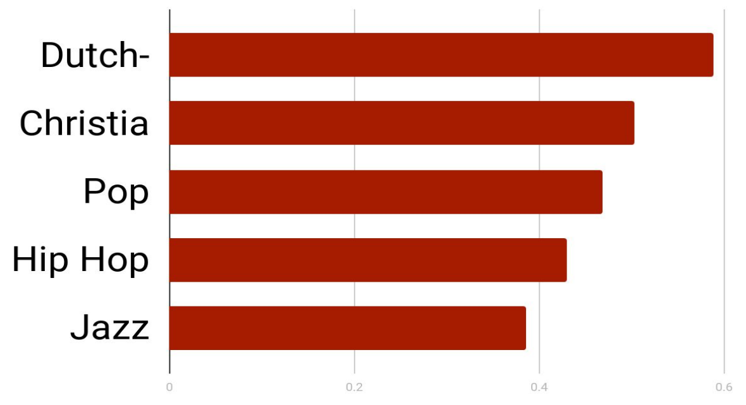
Sentiment Analysis from SentimentR

Negative sentiment is significantly correlated with the popularity score



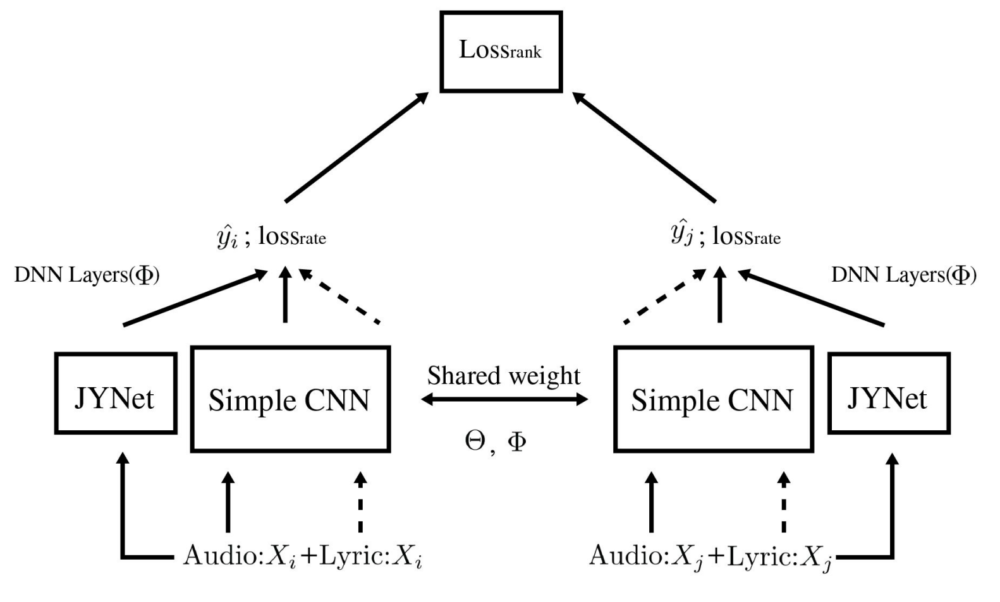
Repetition Analysis with Lempel-Ziv Algorithm

Songs with more repetitive lyrics are more likely to become a hit song



Methodology

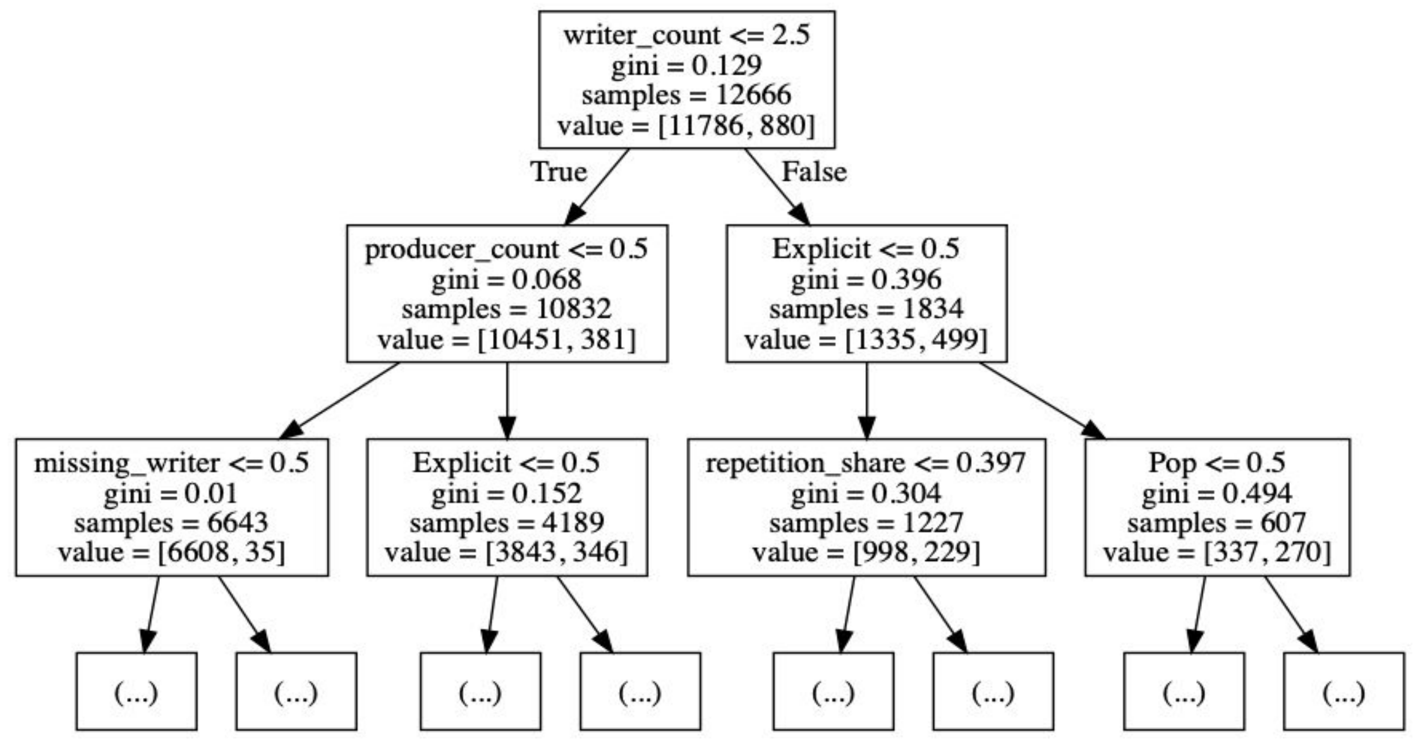
Siamese CNN



Siamese CNN with audio, lyric &tag features

Decision Tree

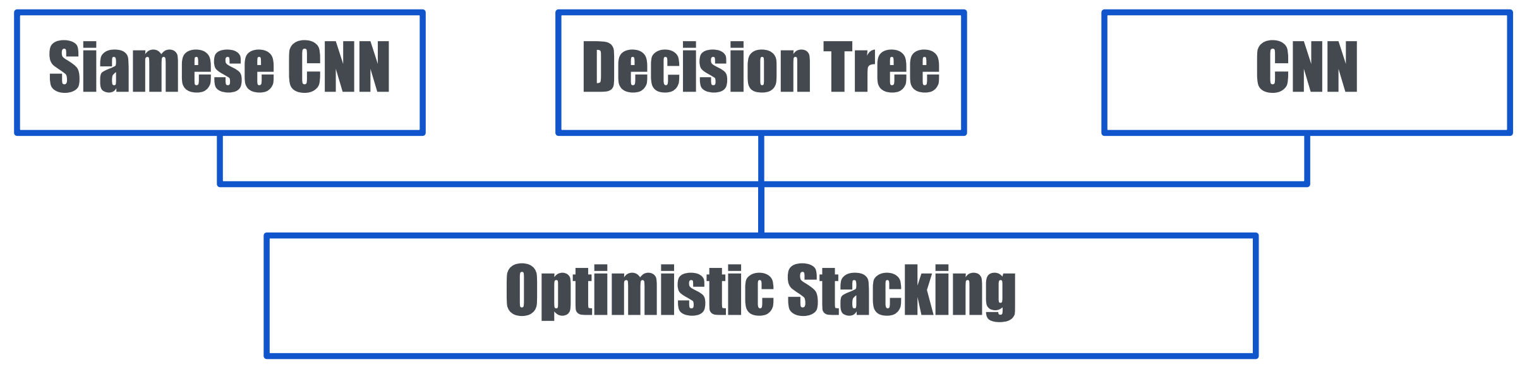
- Spotify Audio Features and Scraped Data
- 5 minimum leaf split and 20 maximum depth achieved maximum recall / F1 score



Convolutional Neural Network

- A pre-trained word embedding from FastText, the CNN takes the embedded lyrics and predict a binary class
- After running 20 epochs, training loss dropped from 0.26 to 0.008, and the model achieved a recall of 33%, which is much better than the baseline model
- Predictions on the test set is later being used in the final model.

Optimistic Stacking



Siamese CNN

Decision Tree

CNN

Optimistic Stacking

Result

Conclusion

- Audio features from Spotify are not predictive
- Number of songwriters and producers is correlated with popularity. The more songwriters and producers lead to higher score.
- More popular songs are more repetitive and they are often explicit
- Pop and hip-hop songs are more popular than blues and country
- Combining predictions from multiple models in "optimistic stacking" leads to the best results
- We achieved **<u>40% recall</u>** and **<u>25.7% F-1 score</u>** in the final model, improved from a baseline model Decision Tree with 5% recall and 8.9% F-1 score.

Forthcoming Research

In the future work, we would like to implement the embedded lyrics from CNN and audio features into a deep learning model, and integrated all other features and train the model with an attention mechanism.

Acknowledgments

The authors thank Ryan Faus for providing data sources and advice, and thank Warner Music staff for technical assistance. We also thank Professor Brian McFee and Professor Sam Bowman for their help on the model.

Reference

1.Topic Modeling in Python with Gensim. (2018, December 04). Retrieved from https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

2. Hit song prediction for pop music by siamese cnn with ranking loss https://arxiv.org/pdf/1710.10814.pdf?fbclid=IwAR12GZ44S2KpciaTHNZhT1U0zAja1XhLtiR03bqptO1lwIA6E20P60f6V0c