UMBC

Hit Song Prediction with Spotify Acoustic Data, Phase 3

- *Team Members*
  - *Debanjan Chowdhury*
  - *Jagan Sirigiri*
  - *Max Grody*
- *Instructor: Dr. Ozgur Ozturk*

Why Hit Song Prediction?

- Hit songs engage a wide audience, make more money
- Knowing what songs can be hits is valuable:
  - Promote songs that are predicted to be popular to attract more users
  - Record labels are interested in what makes songs a hit
  - Help artists understand ways to write hit songs

Hypothesis

- Spotify EchoNest features and song lyrics can explain a song's ability to be a hit

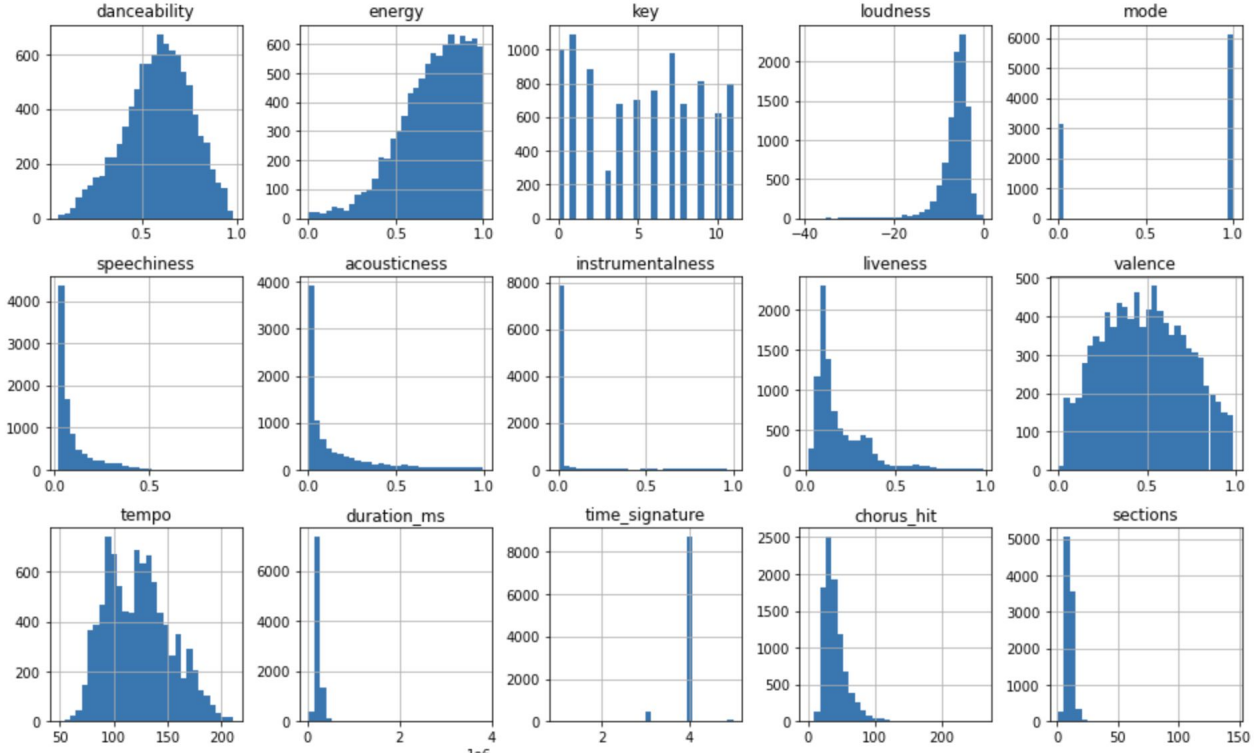
Dataset

- Acoustic and meta features of songs (sourced by author from Spotify and Billboard API)
  - 10,000 rows of songs that have and have not made the Billboard Hot 100 after the year 2000
  - Data includes Spotify Echonest acoustic features
    - Danceability
    - Energy
    - Key
    - Loudness
    - Mode
    - Speechiness
    - Acousticness
    - Instrumentalness
    - Liveness
    - Valence
    - Tempo
  - Target Variable (1 for hit, 0 for not)

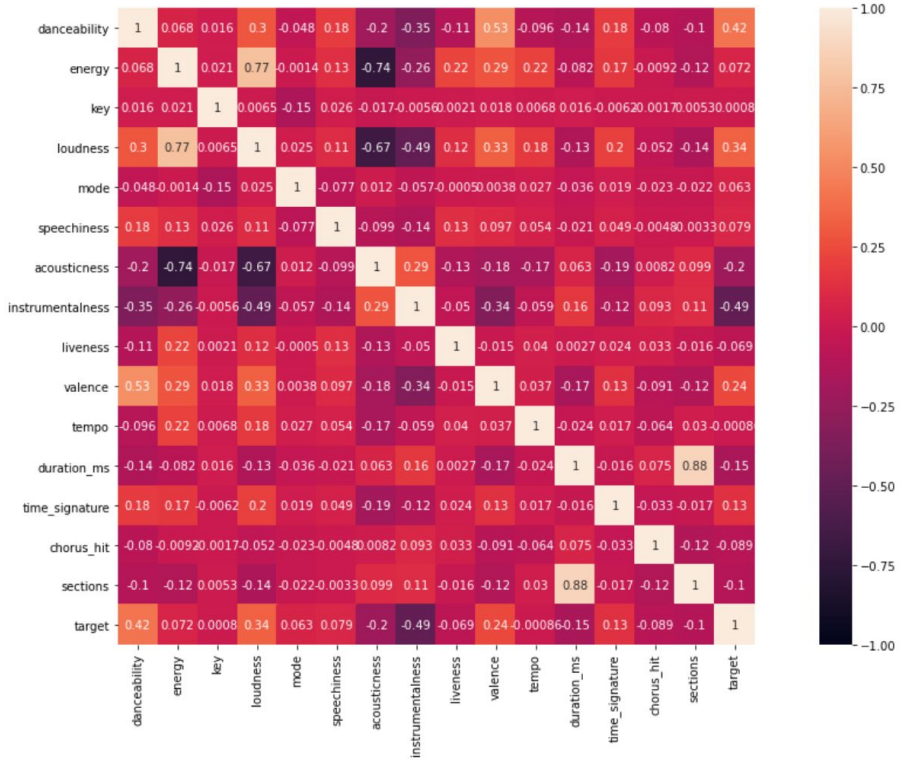
Dataset cont.

- Genius Lyrics
  - Scraped website using Genius API
- AZ Lyrics
  - Songs not found in Genius API were scraped from AZ
- Lyric features:
  - Repetitiveness
  - Average word length
  - Word count
  - Words/duration
  - Repetitiveness*duration
  - Repetitiveness*word count
  - Polarity (how positive or negative a song is)
  - Average Commonness Score

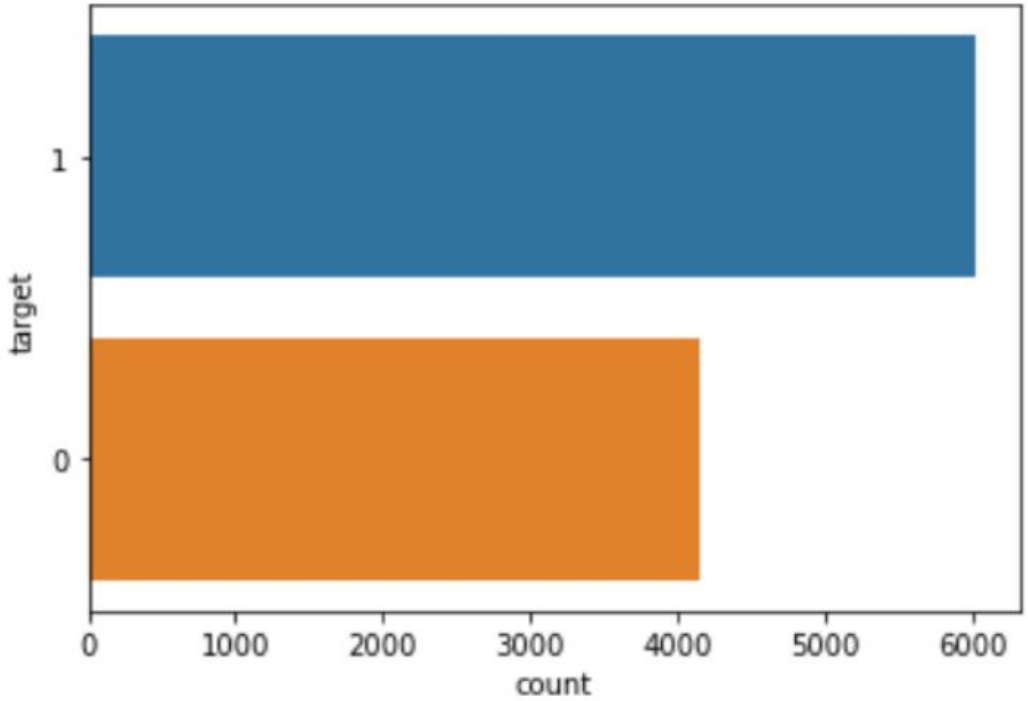
Brief EDA for Acoustic Features



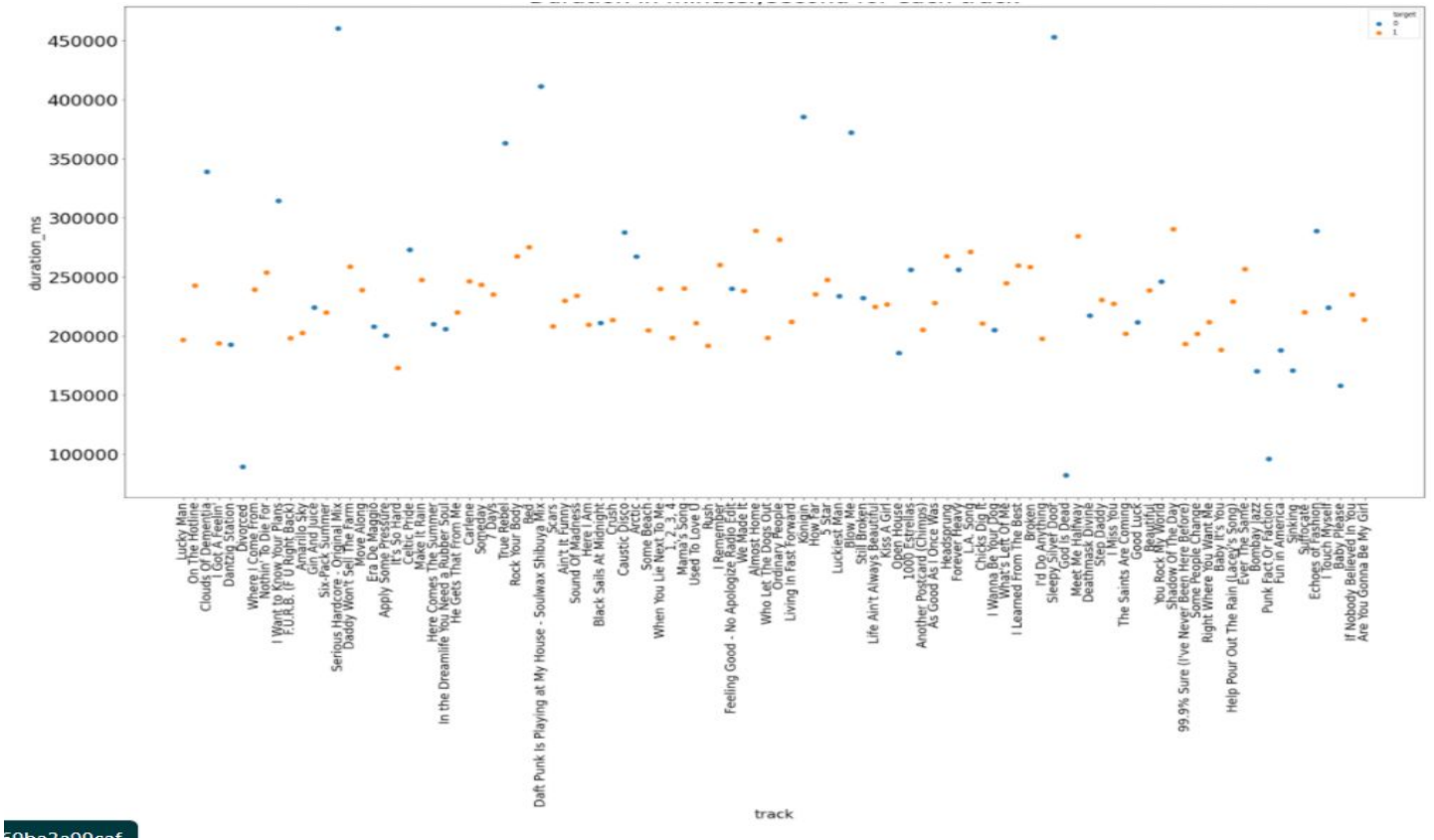
Heatmap showing correlation between columns



Target count

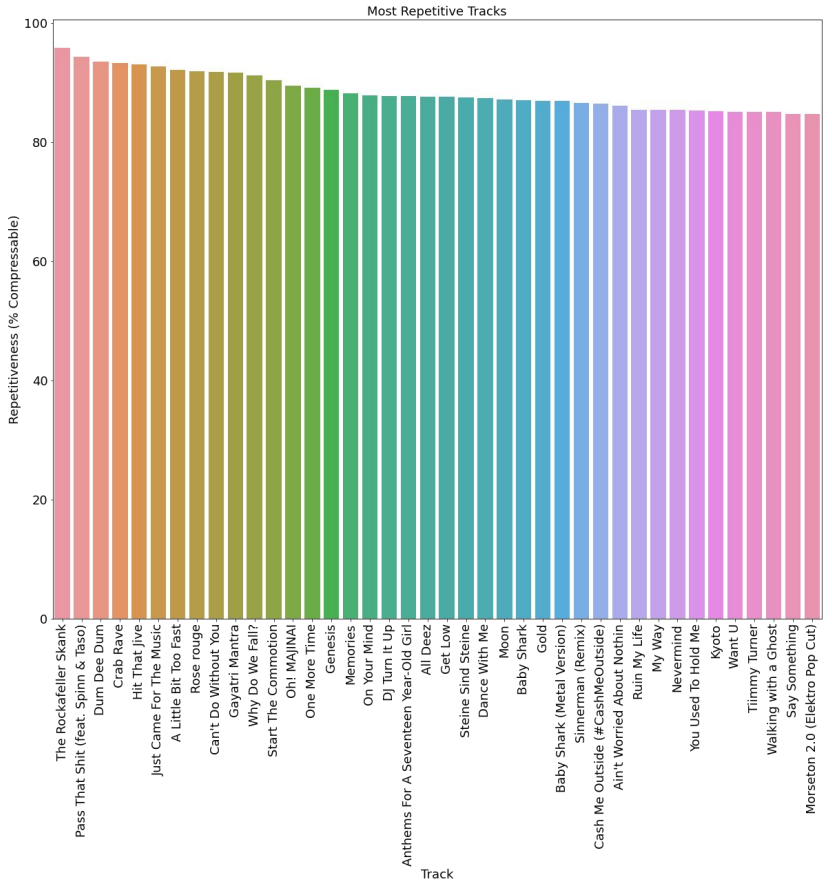


Duration of each track and results



Most Repetitive Tracks

- Zlib (DEFLATE) compression on lyrics
- Repeated character strings replaced with a marker
- Example lyrics:
  - Uncompressed: *“Get back, get back, Get back to where you once belonged.”*
  - Compressed: “*, *, * to where you once belonged.”
- Repetitive songs will be more compressible



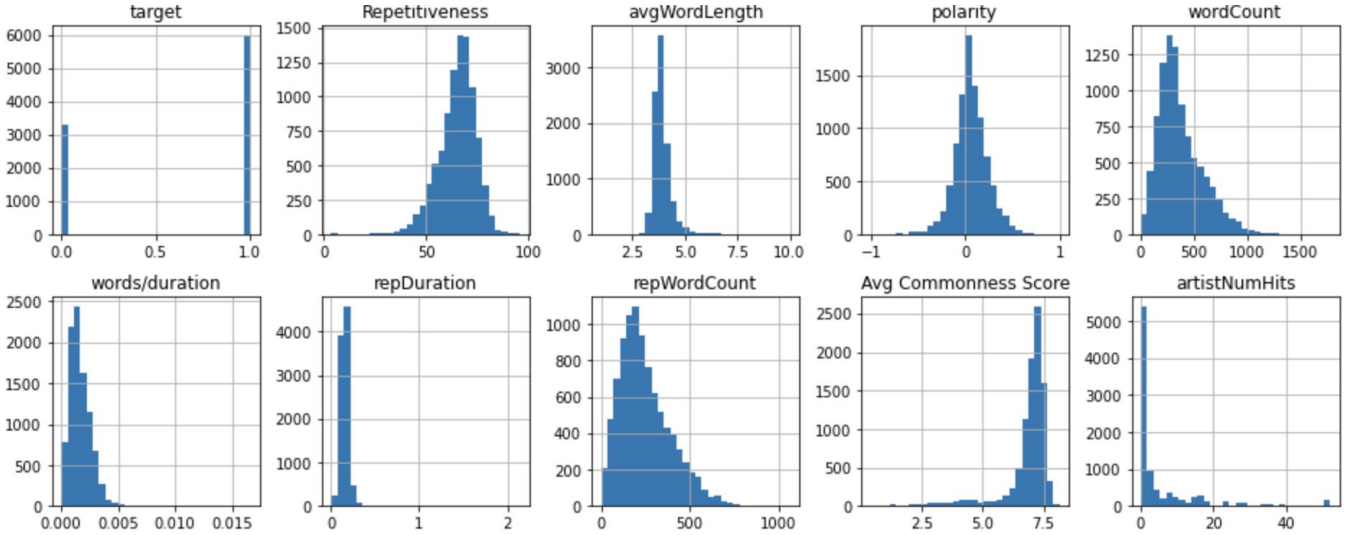
First 3 Verses of Rockafeller Skank

Right about now The funk soul brother Check it out now The funk soul brother
Right about now The funk soul brother Check it out now The funk soul brother
Right about now The funk soul brother Check it out now The funk soul brother

Average Word Commonness

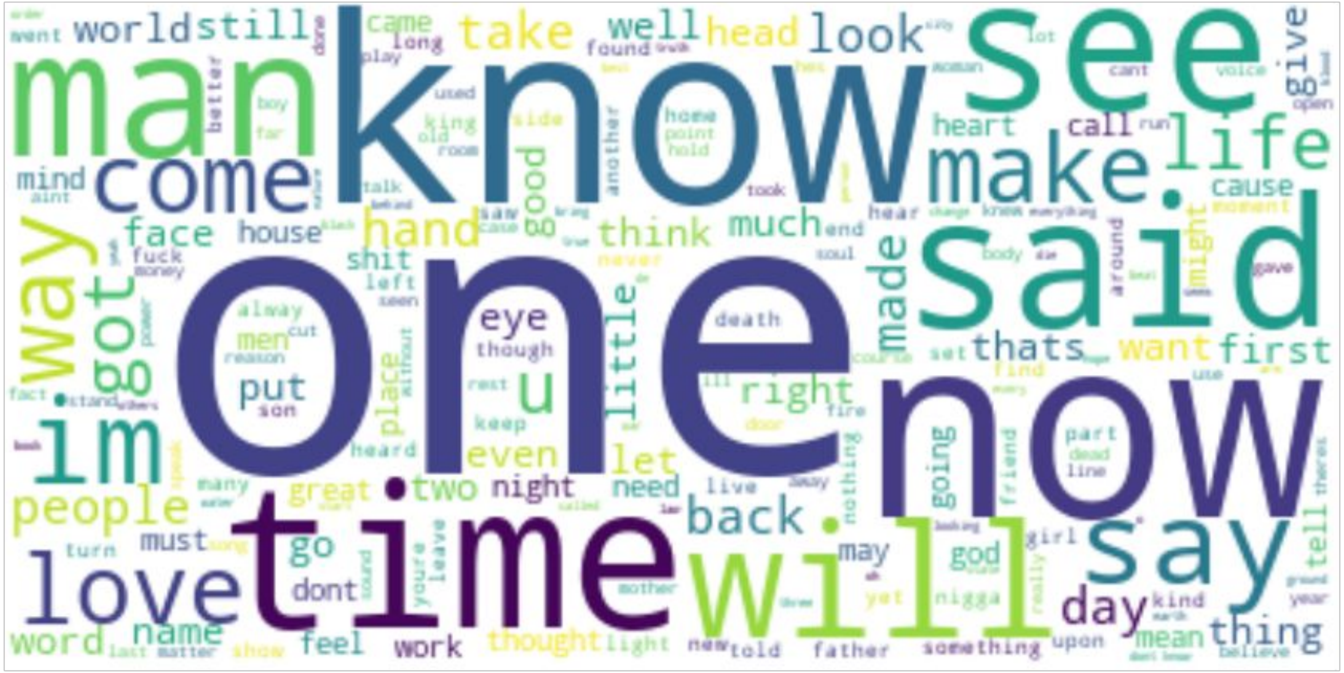
	Word	Frequency	Commonness Metric	Normalised Commonness Metric
0	the	522930	5.718444	10.000000
1	and	263228	5.420332	9.478684
2	to	230490	5.362652	9.377818
3	of	230019	5.361764	9.376264
4	a	223619	5.349509	9.354833

NLP Feature Distributions



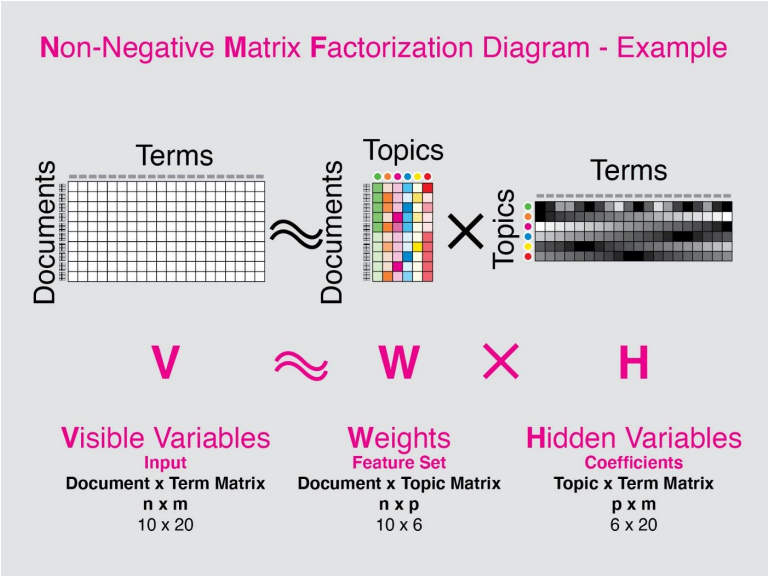
Word Cloud for Hits

Word Cloud for Flops



Non-Negative Matrix Factorization Topic Modelling

- Decompose a document-term matrix
- Identify latent themes



Matrix Decomposition in NMF Diagram Example by Anupama Garla

Topic Model Results

```
For topic 1 the words with the highest value are:
im        1.371411
know      1.033697
dont      1.000849
youre     0.870400
like      0.753556
time      0.682333
baby      0.642504
girl      0.619906
got       0.602030
ill       0.594356
Name: 0, dtype: float64
```

```
For topic 2 the words with the highest value are:
nigga     1.306355
bitch     0.858441
yeah      0.817048
got       0.725090
im        0.699781
like      0.654939
shit      0.578375
fuck      0.545393
aint      0.521717
ayy       0.398391
Name: 1, dtype: float64
```

```
For topic 3 the words with the highest value are:
na          2.407800
wan         1.366314
gon         0.849181
dont        0.323133
tonight     0.172958
ya          0.161721
let         0.149199
im          0.140535
hey         0.132376
baby        0.131704
Name: 2, dtype: float64
```

```
For topic 4 the words with the highest value are:
oh       2.684627
yeah     1.027156
ooh      0.518456
hey      0.397928
baby     0.345286
whoa     0.191745
ohoh     0.166910
let      0.163492
uh       0.158587
girl     0.158295
Name: 3, dtype: float64
```

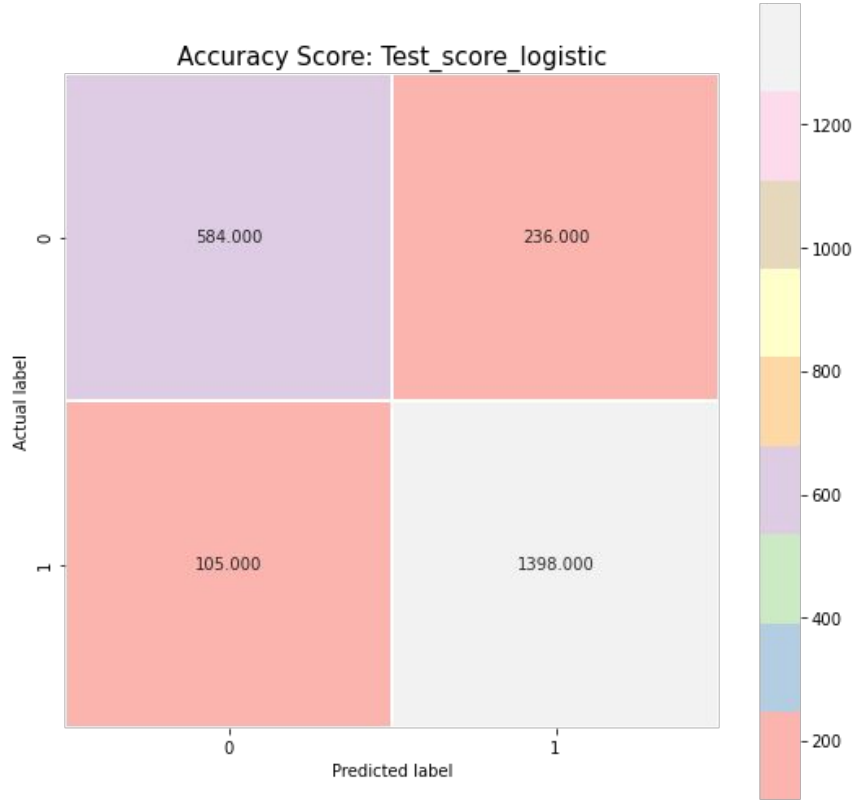
```
For topic 5 the words with the highest value are:
love      3.190413
baby      0.487834
heart     0.233368
way       0.192219
need      0.173950
ooh       0.167016
know      0.164750
girl      0.134040
want      0.130561
ill       0.111428
Name: 4, dtype: float64
```

```
For topic 6 the words with the highest value are:
la      1.717625
que     1.283050
tu      0.608101
te      0.572887
el      0.476627
se      0.397005
mi      0.389155
en      0.384505
lo      0.375088
yo      0.289667
Name: 5, dtype: float64
```

NMF Features

	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6
0	0.054884	0.033882	0.000000	0.000000	0.000000	0.000000
1	0.042156	0.048403	0.019146	0.000000	0.037693	0.004506
2	0.050763	0.011998	0.024909	0.016706	0.000000	0.000000
3	0.036645	0.002149	0.000546	0.011385	0.000000	0.000732
4	0.061251	0.015321	0.000000	0.000000	0.000000	0.000000

Logistic Regression - Confusion Matrix & Scores



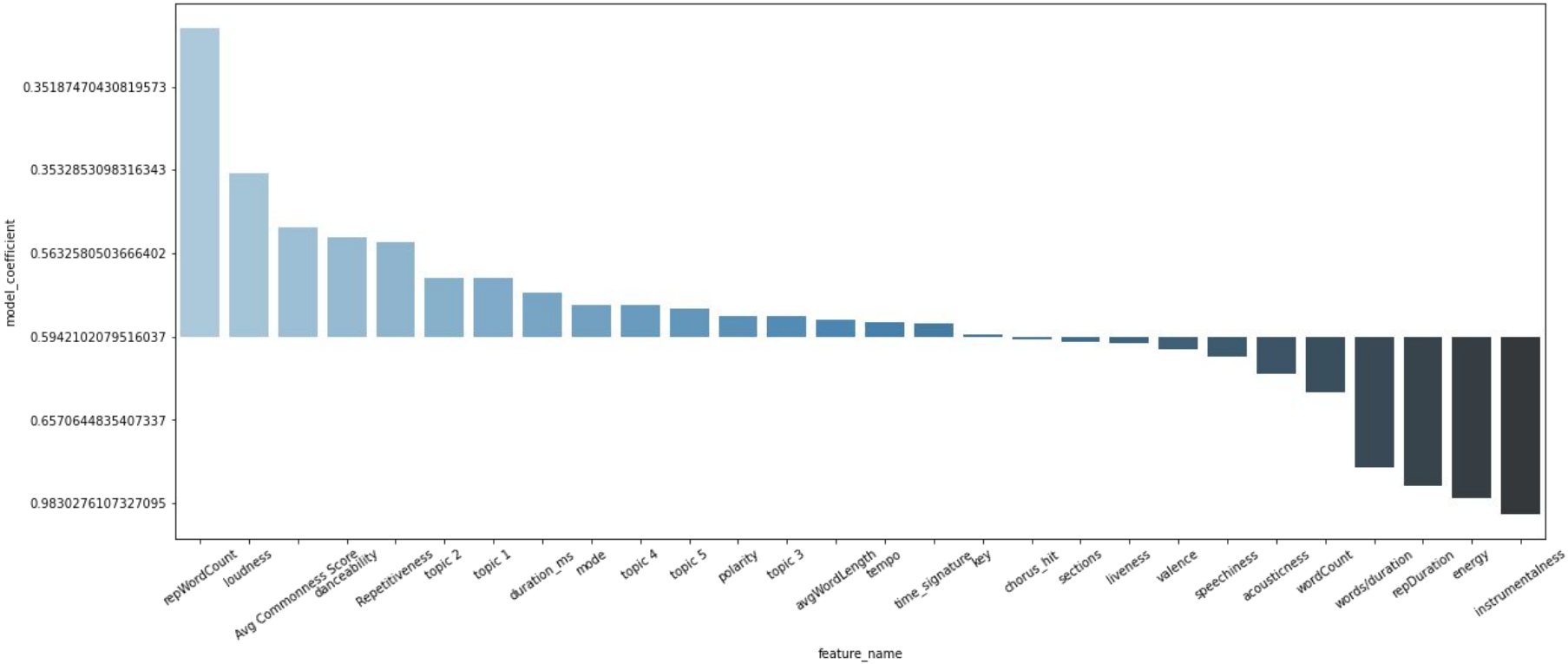
Accuracy: 0.853

Precision: 0.856

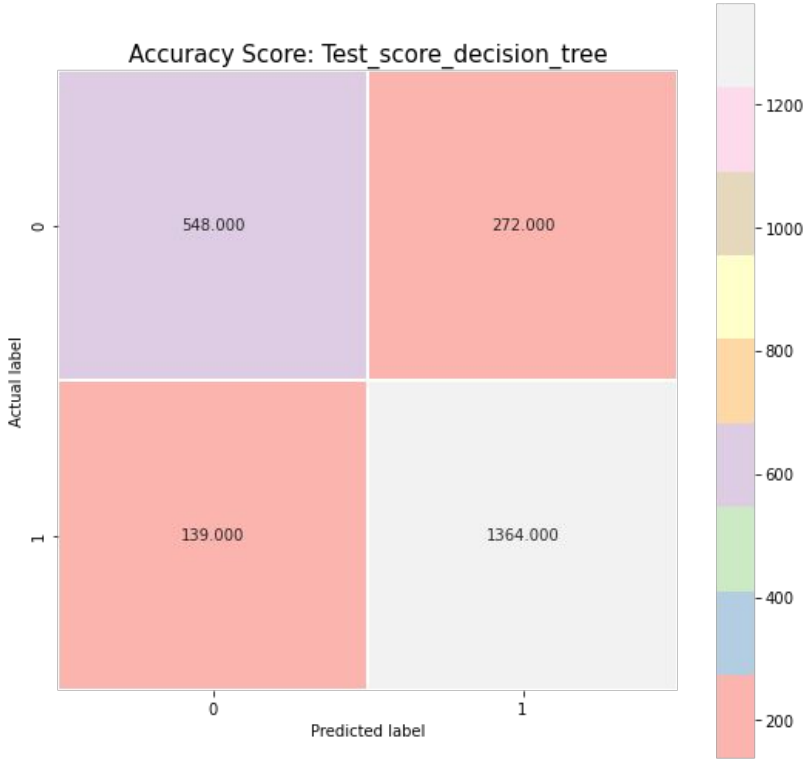
Recall: 0.930

F1: 0.891

Logistic Regression - model coefficients



Decision Tree - Confusion Matrix & scores



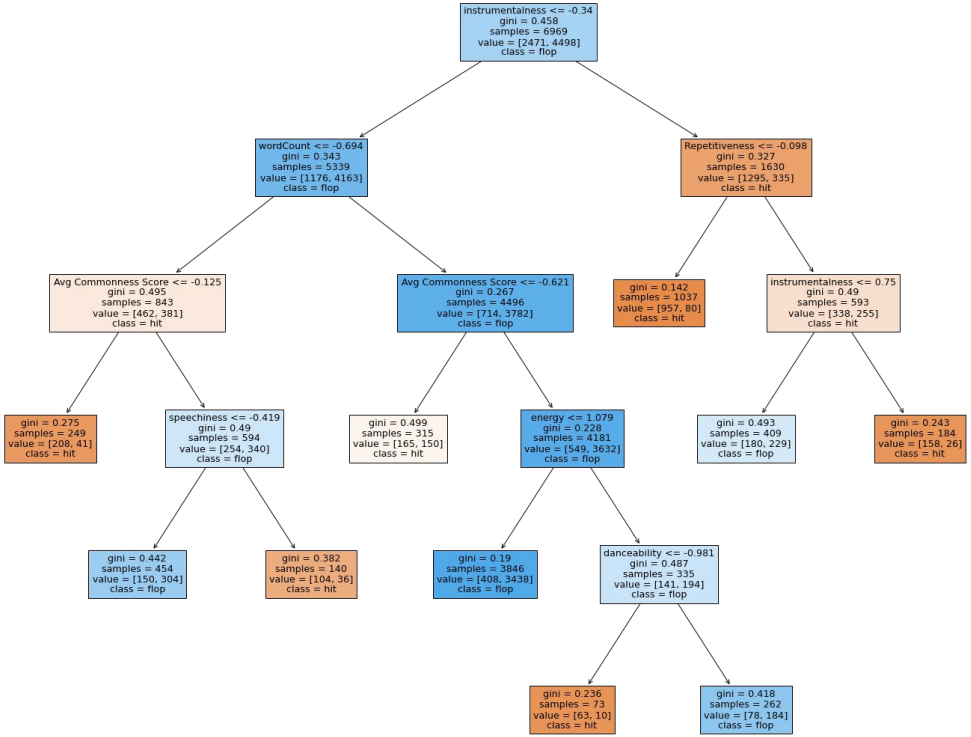
Accuracy: 0.823

Precision: 0.834

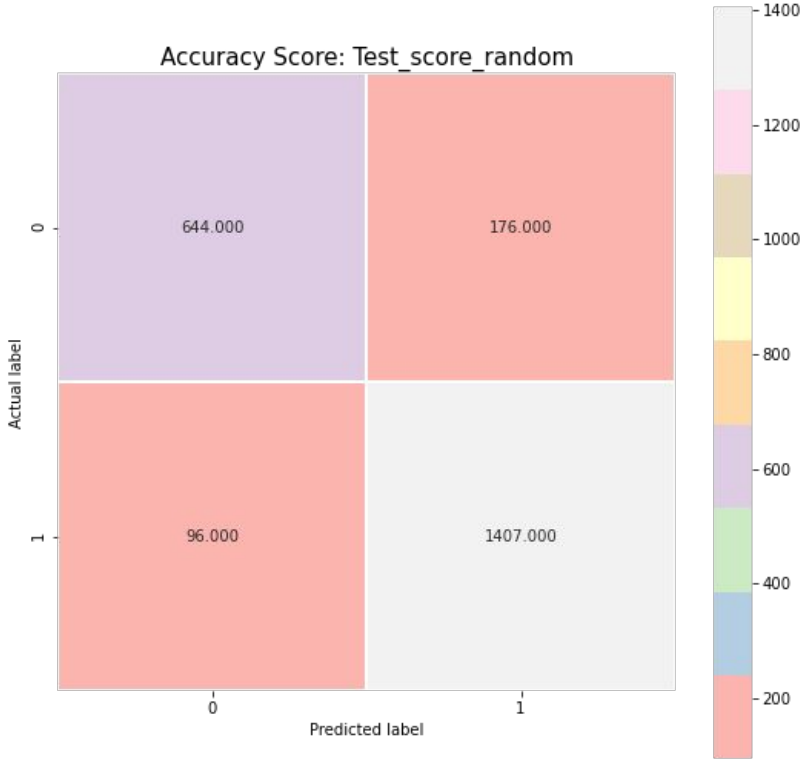
Recall: 0.908

F1: 0.869

Decision Tree - Visualization



Random Forest - Confusion Matrix & scores



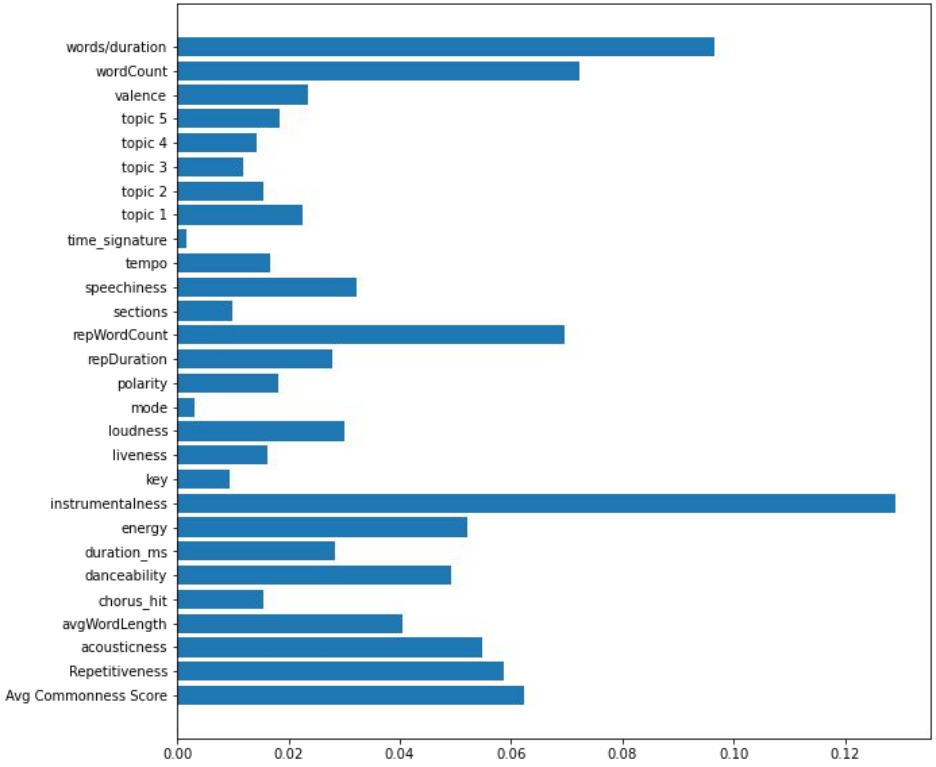
Accuracy: 0.882

Precision: 0.888

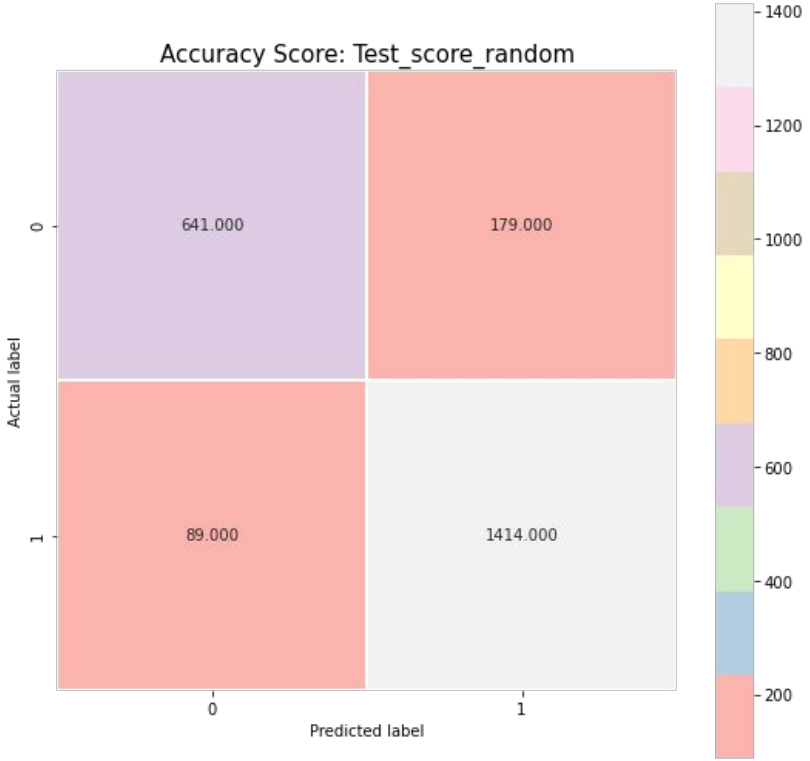
Recall: 0.936

F1: 0.912

Random Forest - Feature Importance



XGBoost - Confusion Matrix & scores



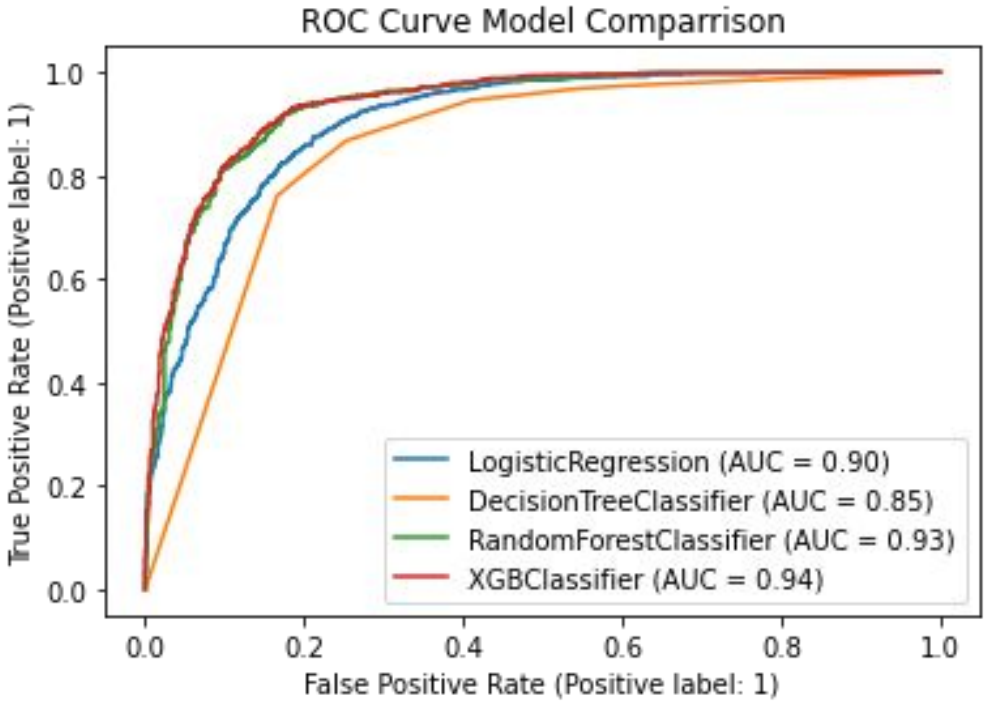
Accuracy: 0.882

Precision: 0.888

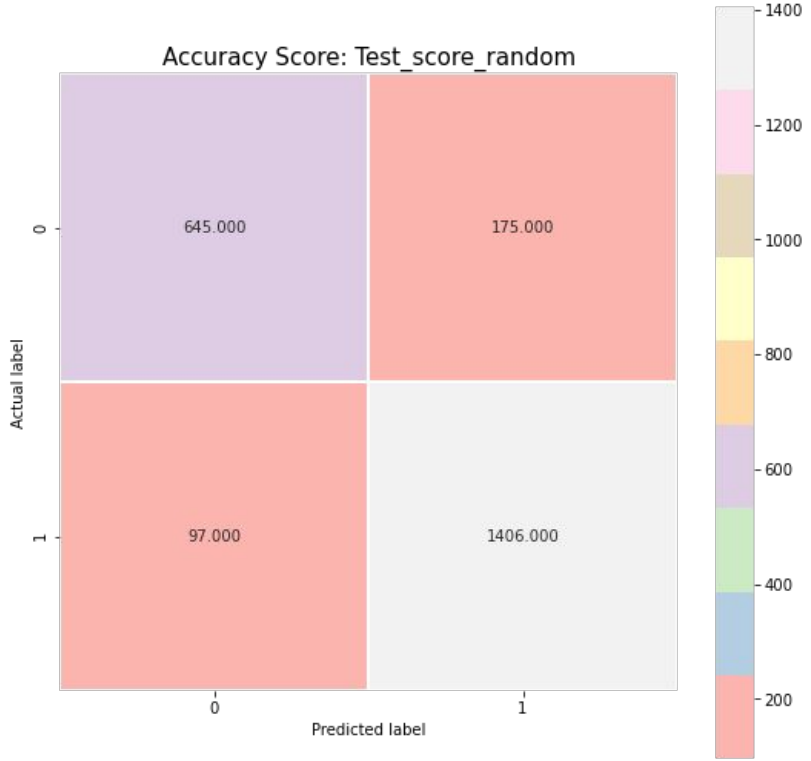
Recall: 0.941

F1: 0.913

ROC(Receiver Operating Characteristics) Curve



Voting Classifier - Confusion Matrix & scores



Accuracy: 0.883

Precision: 0.889

Recall: 0.935

F1: 0.912

Conclusions

- EchoNest Acoustic features and lyrical features do explain a song's ability to be a hit
- The elements of a song that most improve its ability to be a hit:
  - Repetitiveness*word count, loudness, average commonness score, danceability
- The elements of a song that most decrease its ability to be a hit:
  - Instrumentalness, energy, words/duration, word count

Next Steps

- Expand the dataset with more songs (both hits and non-hits)
- Compare model results w/ & w/o NLP features, w/ & w/o acoustic features
- Change the target variable, measure song popularity by different metric than Billboard placement (e.g. spotify's popularity metric)
- Segment study by genre
  - What makes a song in one genre popular might not in another genre
- Create a tool or web app