UMBC

Hit Song Prediction with Spotify Acoustic Data, Phase 2

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

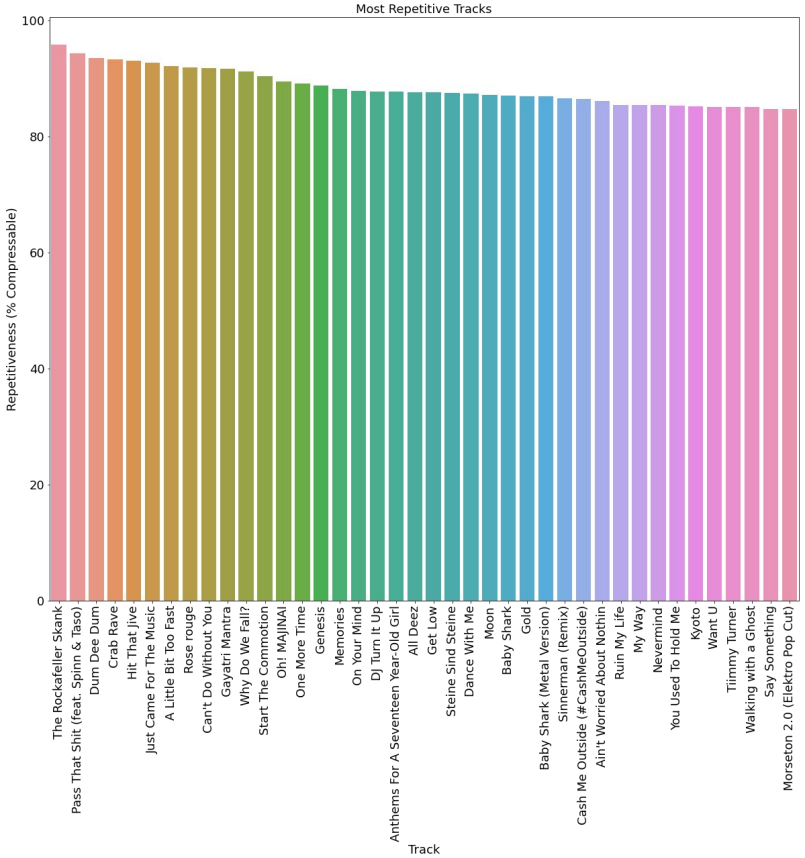
Lyrics

- Lyrics for each song were scraped using the Genius API
- Managed to get lyrics for ≈ 10,000/12,000 songs
- Dropped rows with no lyrics

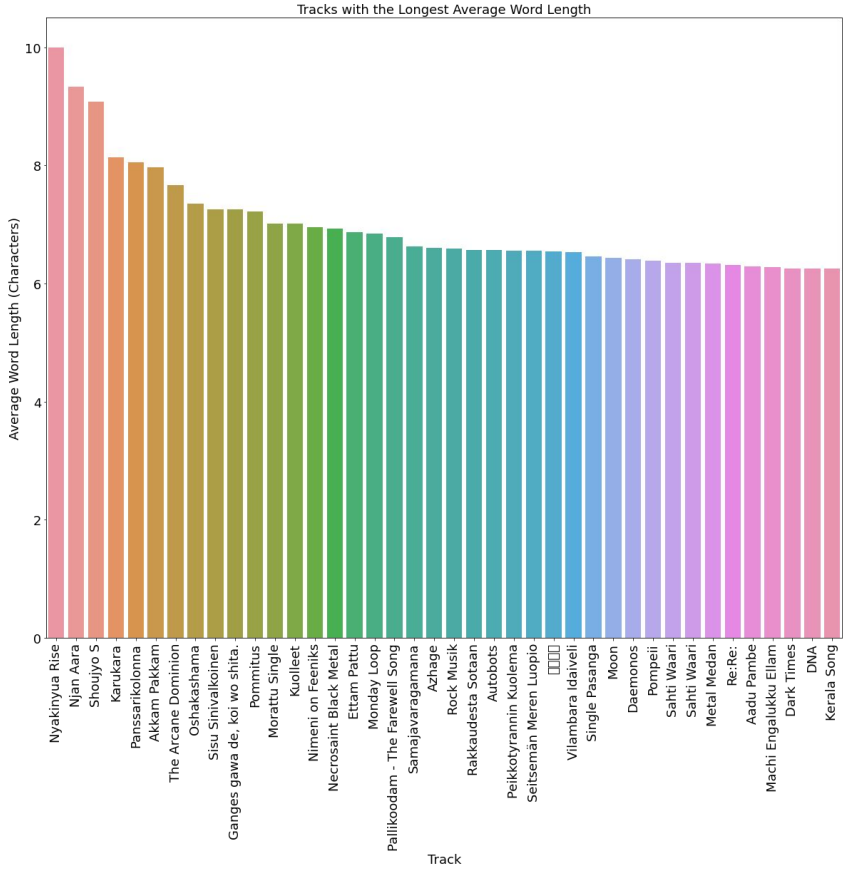
NLP Features

- Repetitiveness score:
  - How compressible are the lyrics?
- Average word length
- Average word uniqueness score
  - Are the lyrics using commonly used words or rarer words?
- Textblob polarity
- Topic Model Features

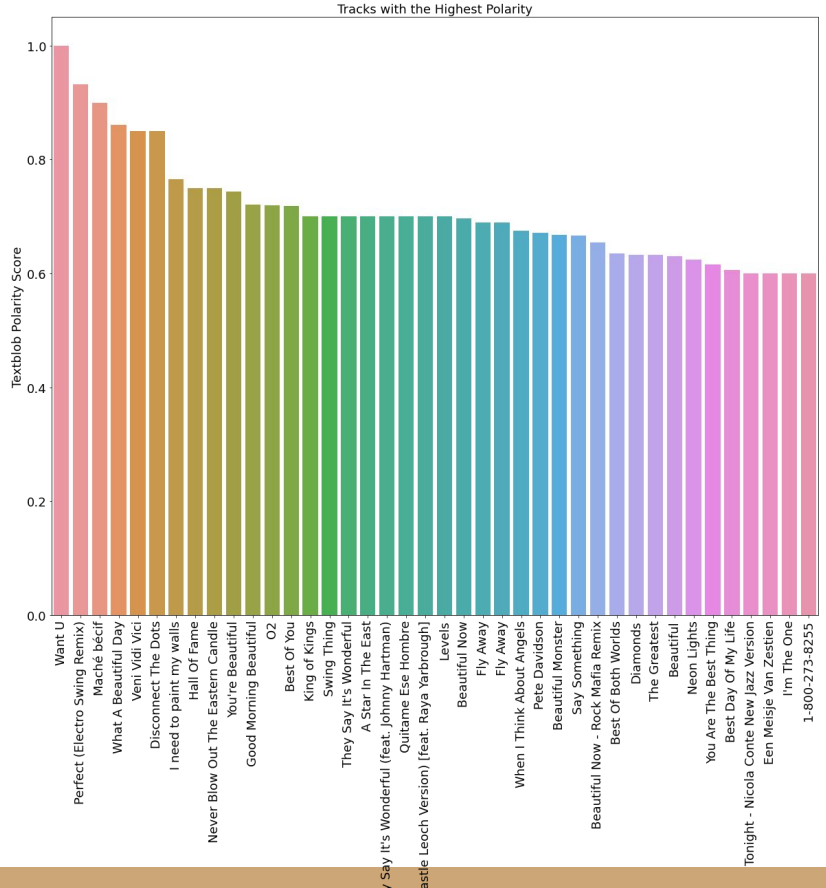
Most Repetitive Tracks



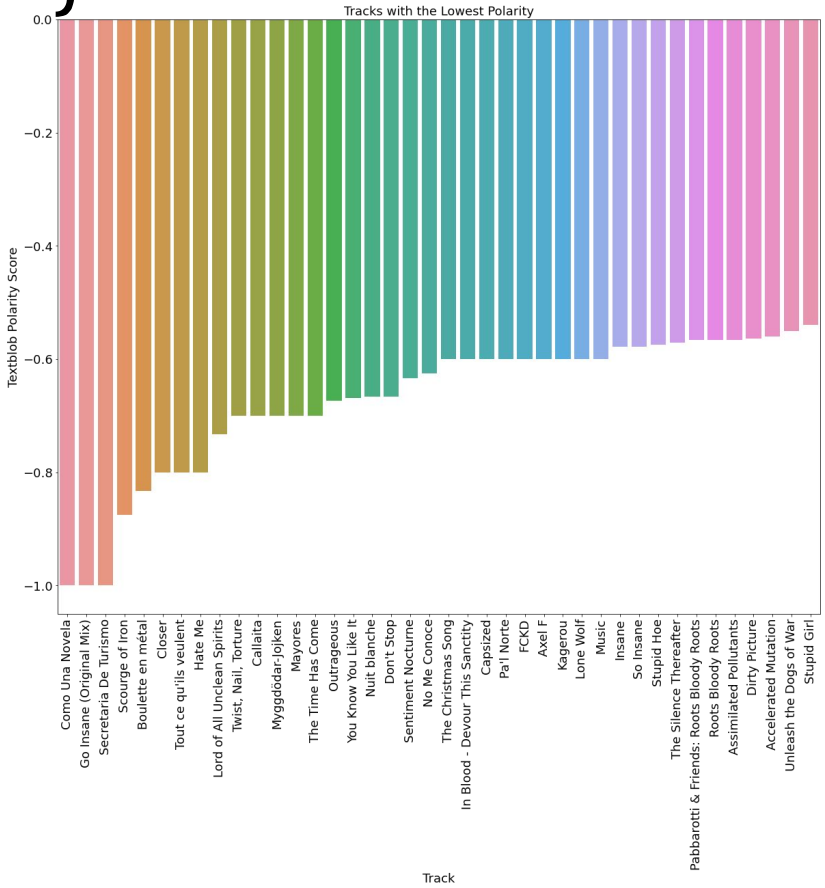
Longest Avg Word Length



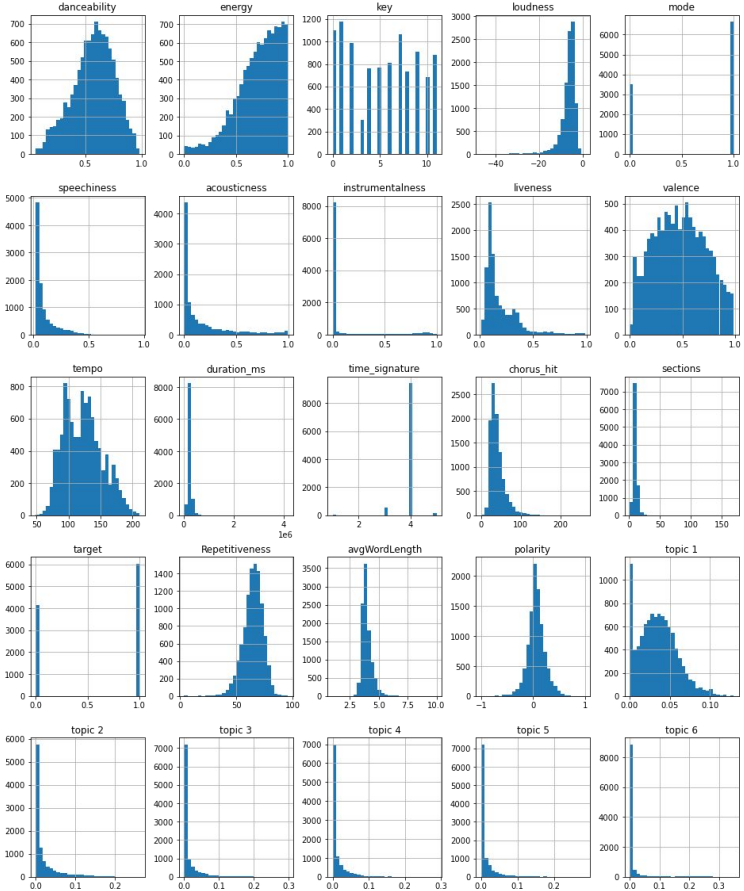
Highest Polarity



Lowest Polarity

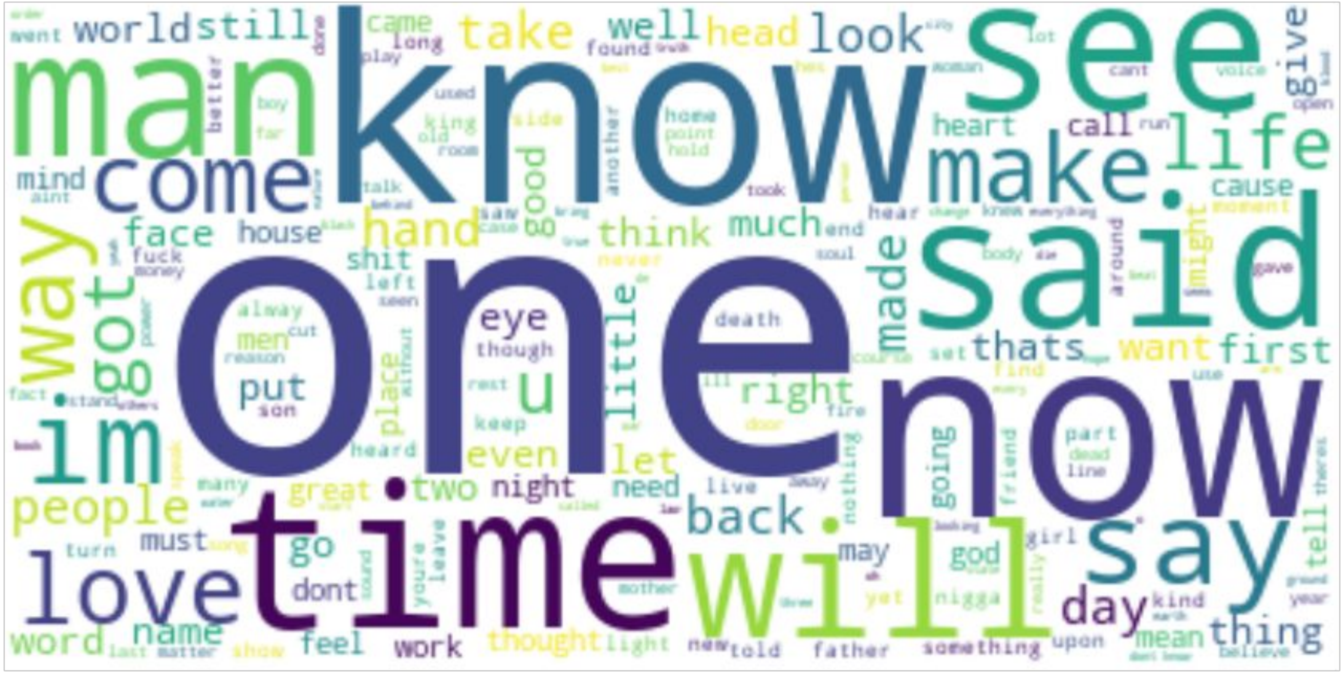


Distributions



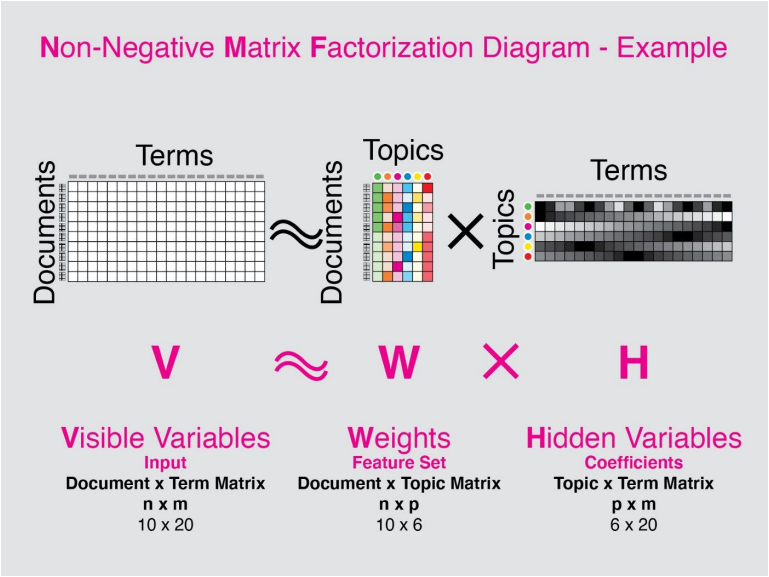
Word Cloud for Hits

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
```

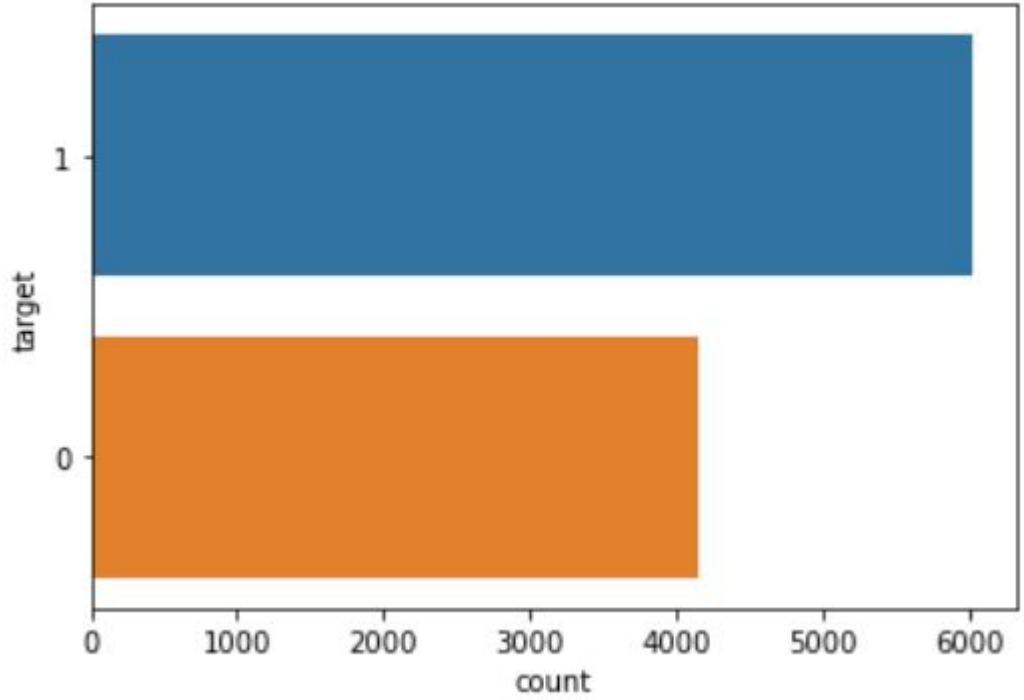
```
For topic 5 the words with the highest value are:
love      3.190413
baby      0.487834
heart     0.233368
way       0.192219
need      0.173950
ooh       0.167016
know      0.164750
girl      0.134040
want      0.130561
ill       0.111428
Name: 4, dtype: float64
```

```
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
```

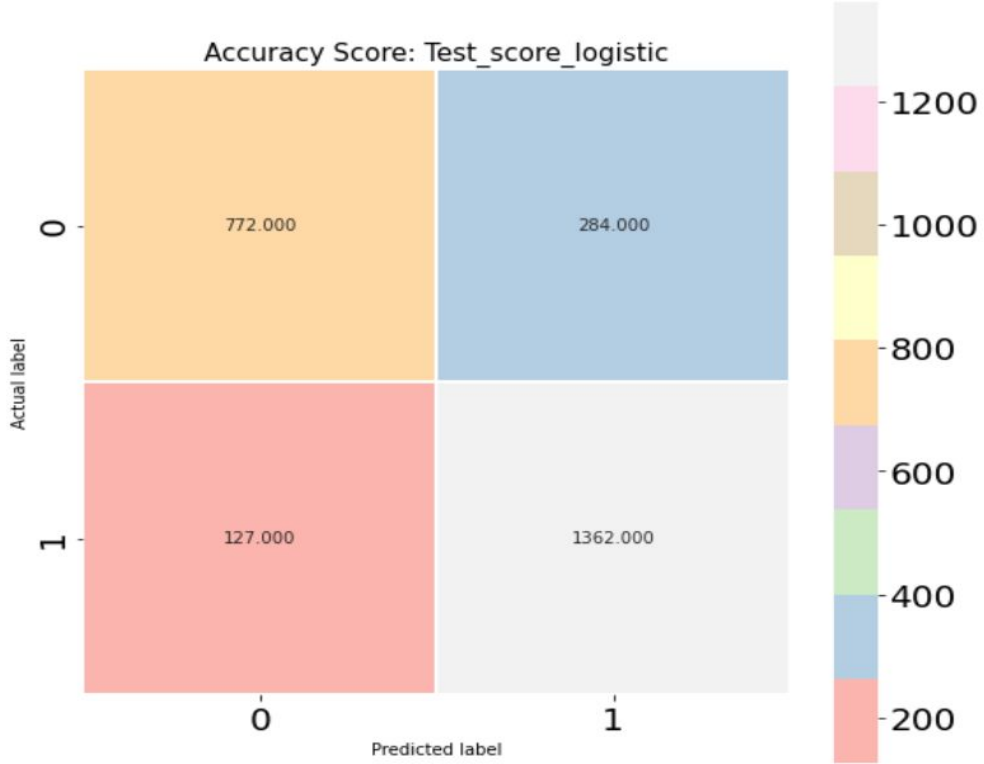
NMF Features

	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

Count of the hits and flops



Logistic Regression - Confusion Matrix & Scores



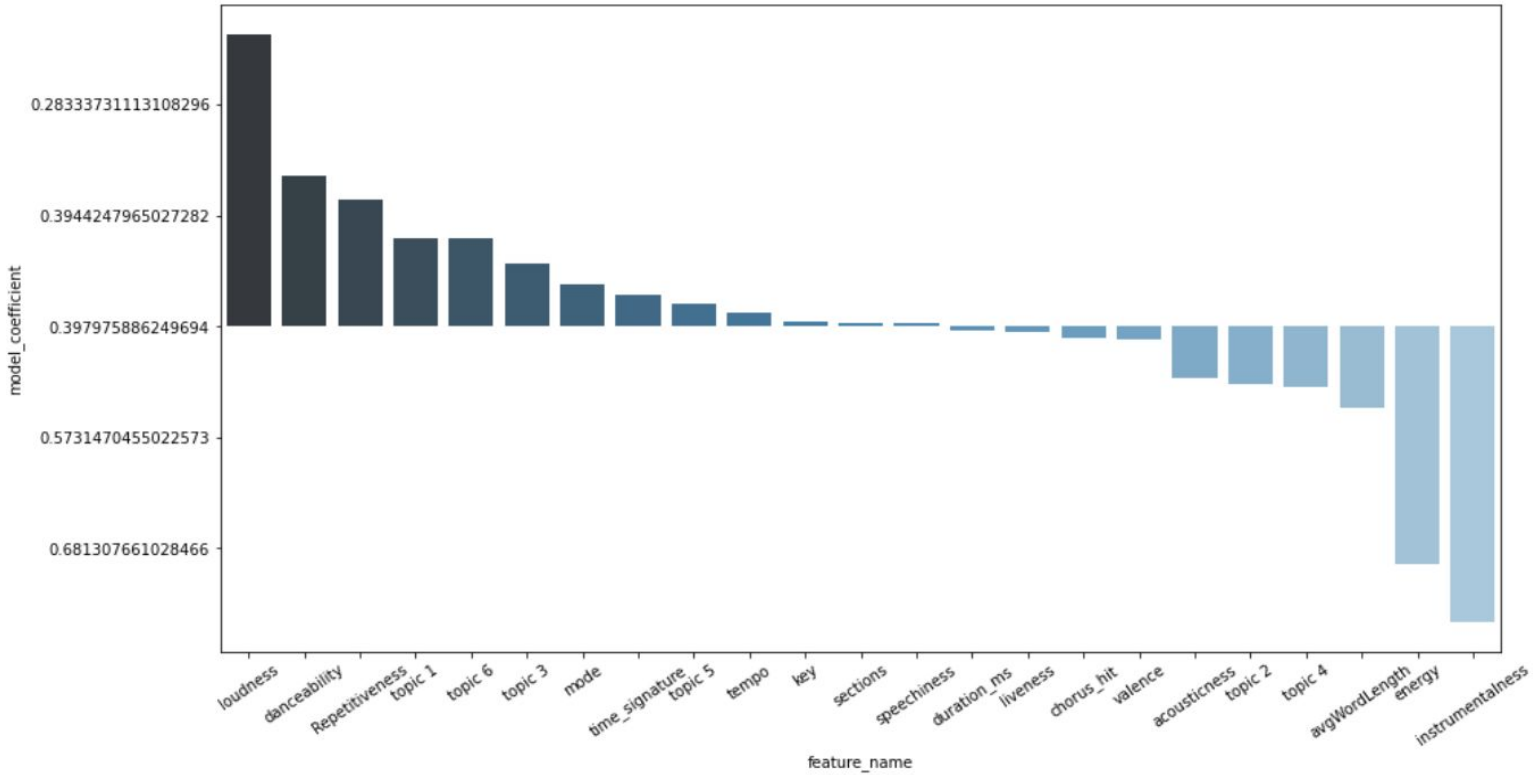
Accuracy: 0.84

Precision: 0.83

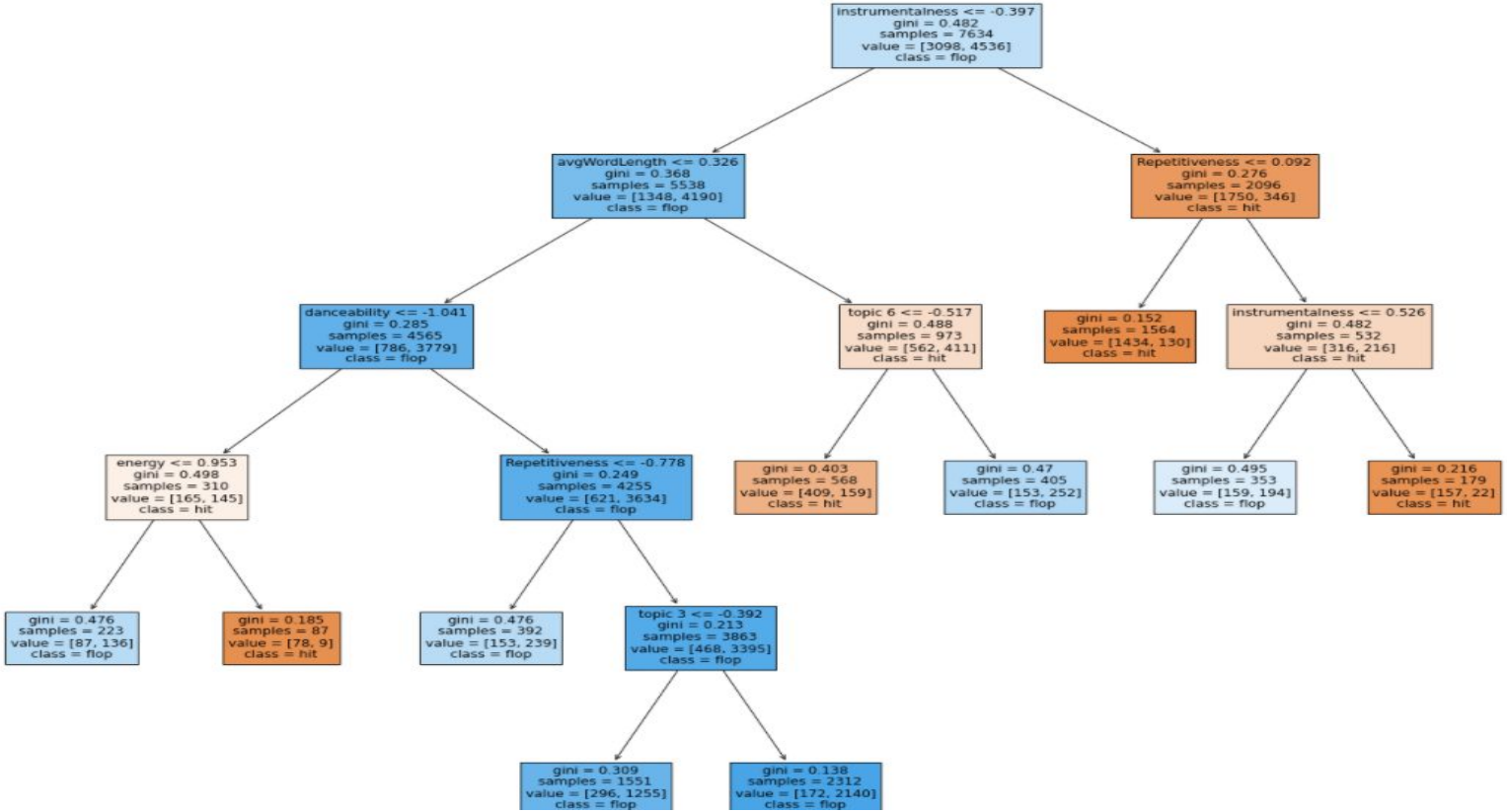
Recall: 0.91

F1: 0.97

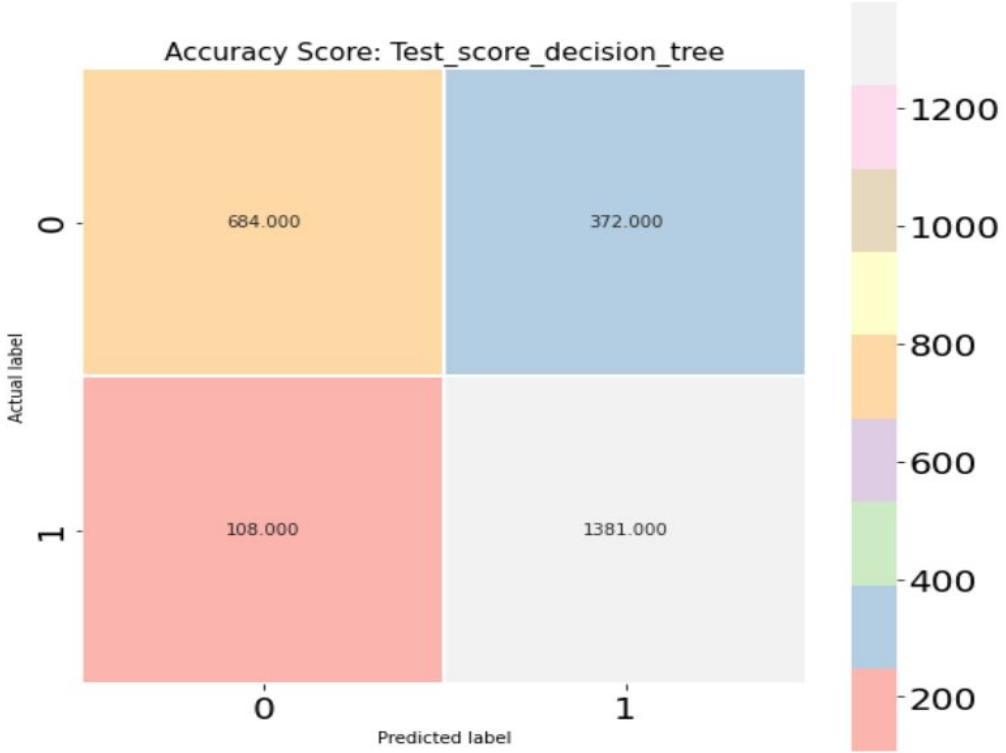
Logistic Regression - model coefficients



Decision Tree - Visualization



Decision Tree - Confusion Matrix & scores



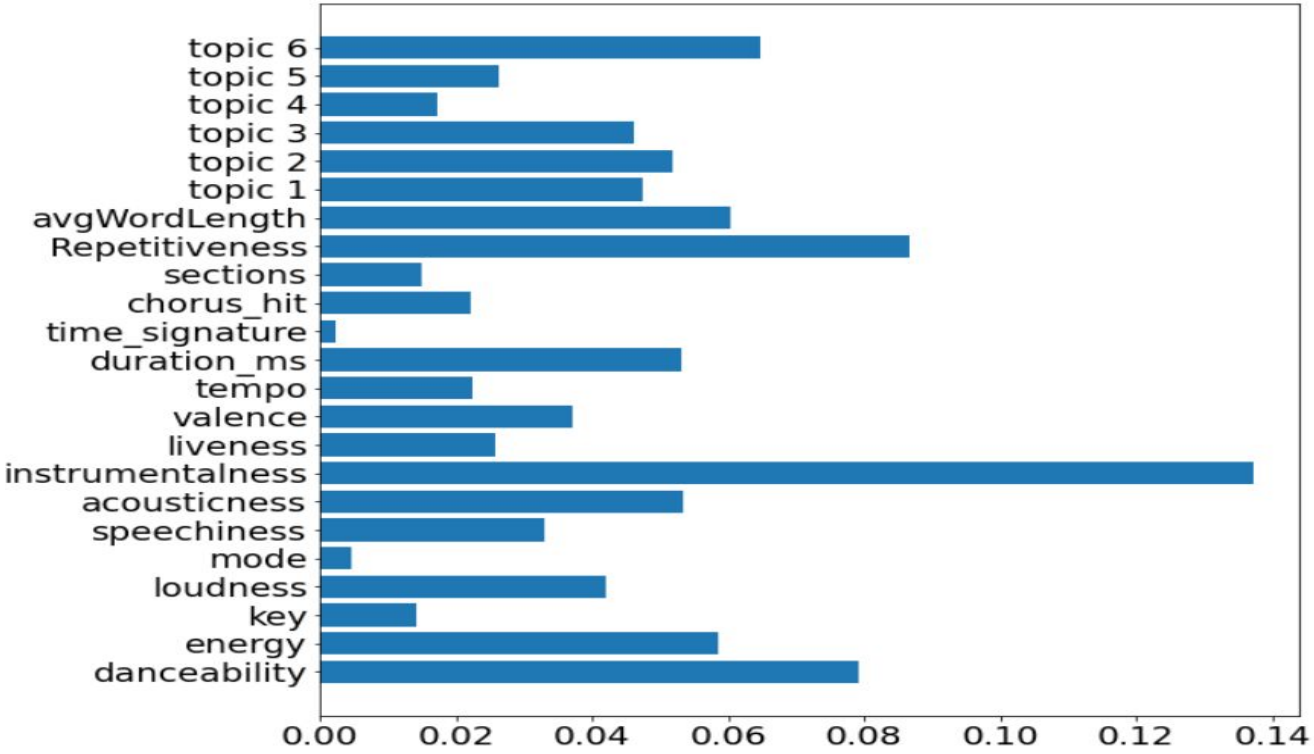
Accuracy: 0.81

Precision: 0.79

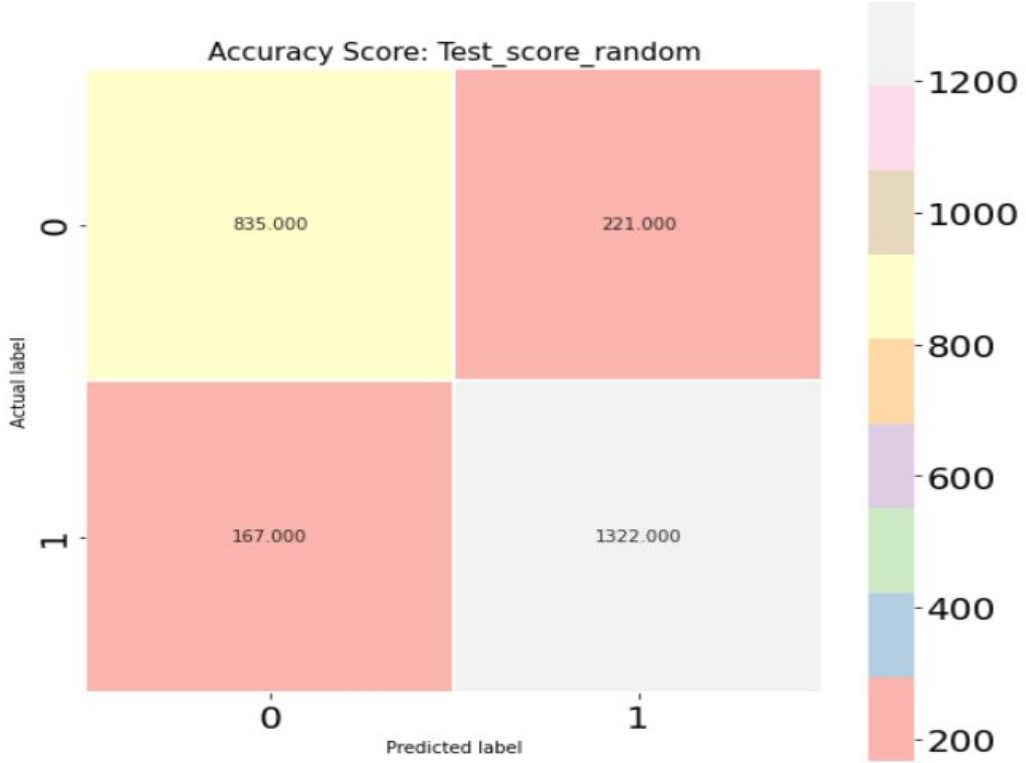
Recall: 0.93

F1: 0.85

Random Forest - Feature Importance



Random Forest - Confusion Matrix & scores



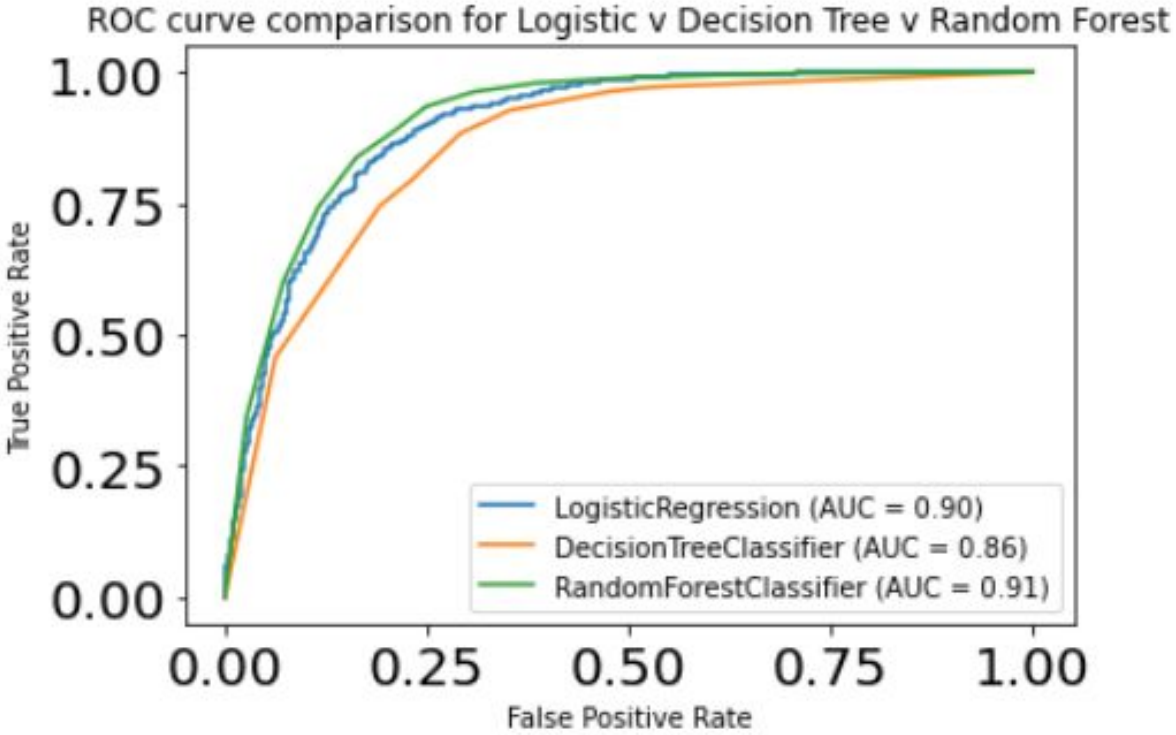
Accuracy: 0.85

Precision: 0.86

Recall: 0.89

F1: 0.87

ROC(Receiver Operating Characteristics) Curve



Next Steps

- Recover more lyrics for missing songs
- Experiment with number of topics for NMF, try other topic model approaches
- Hyperparameter tuning
- Ensemble (stack models)
- Compare model results w/ & w/o NLP features, w/ & w/o acoustic features
- Change the target variable (scrape data for exact placement on hot 100)