

DATA MINING

GENRE CLASSIFICATION USING SONG LYRICS

BREA BEALS | PETER EUSEBIO | LOLA JOHNSTON
JERRY SHA | ROY XIE

03.12.20





AGENDA

Project Intro
Data Exploration
Data Preparation
Modeling
Evaluation
Deployment
Conclusion

PROJECT SCOPE



PROJECT OBJECTIVE

Explore the feasibility of predicting genres through song lyrics

TACTICS

Use data mining techniques to clean, prepare, model, and evaluate results (NLP).

DATA SOURCE

**Lyrics.com - 50K songs | Kaggle
Genres from Spotify API**

PROJECT SCOPE



PROJECT OBJECTIVE

Explore the feasibility of predicting genres through song lyrics

TACTICS

Use data mining techniques to clean, prepare, model, and evaluate results (NLP).

DATA SOURCE

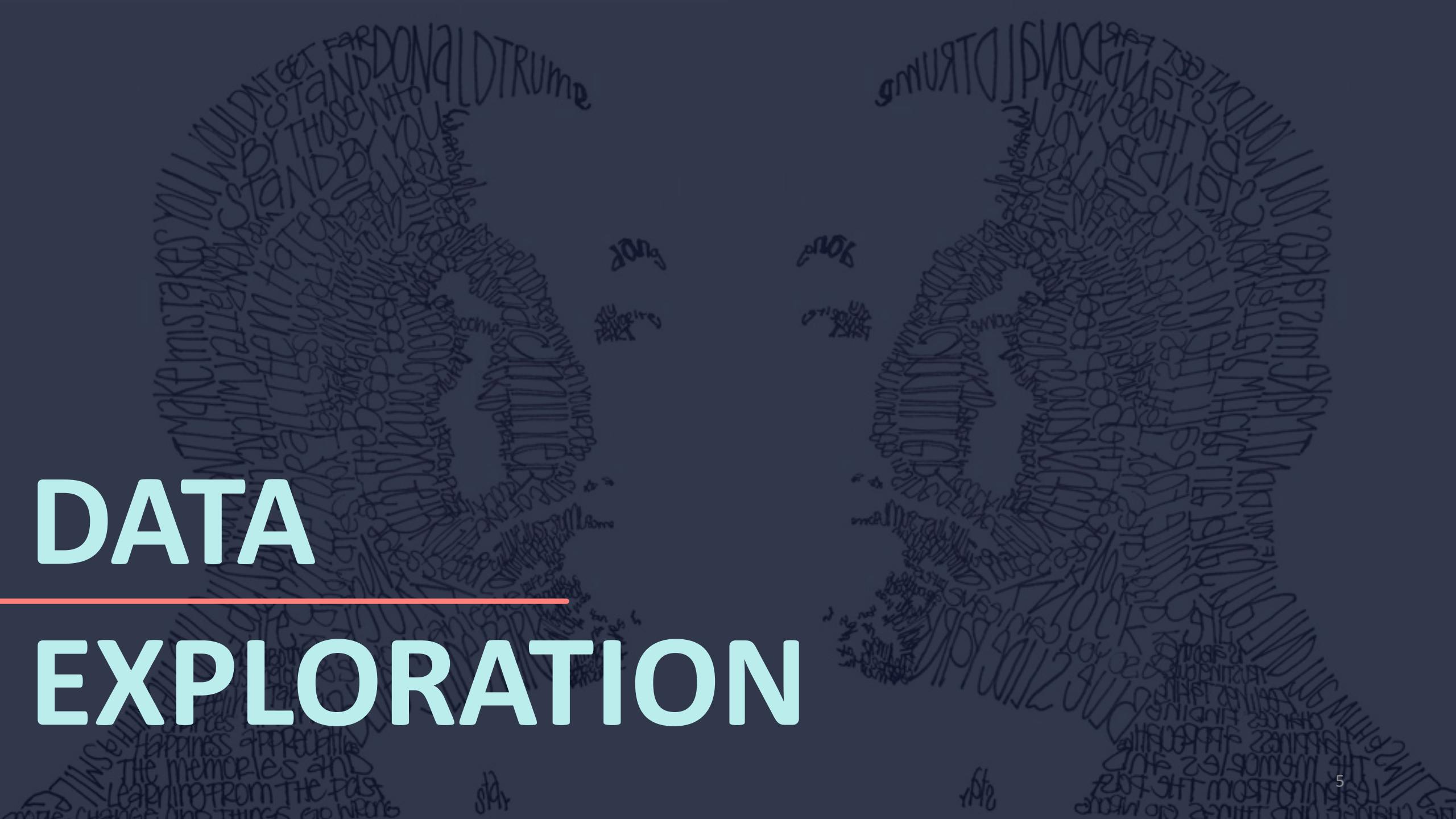
**Lyrics.com - 50K songs | Kaggle
Genres from Spotify API**



WHAT DOES SUCCESS LOOK LIKE?

Find a relationship between lyrics and genre

DATA EXPLORATION



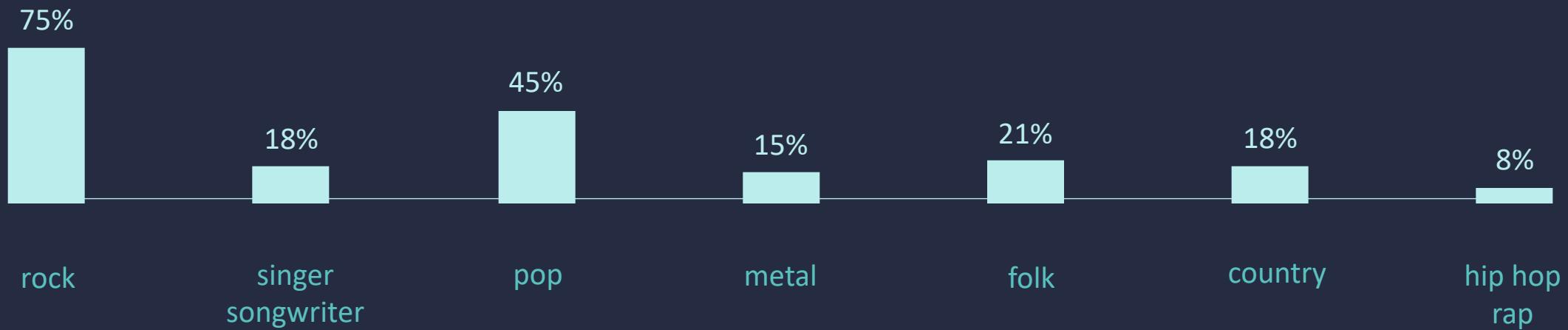
An initial look at our data.

MOST LIKE THE OTHER

- Singer-songwriter and Folk
- Singer-songwriter and Country
- Folk and Rock

MOST UNLIKE THE OTHER

- Hip-Hop/Rap and Rock
- Metal and Pop
- Rock and Pop
- Country and Pop



An initial look at our data.

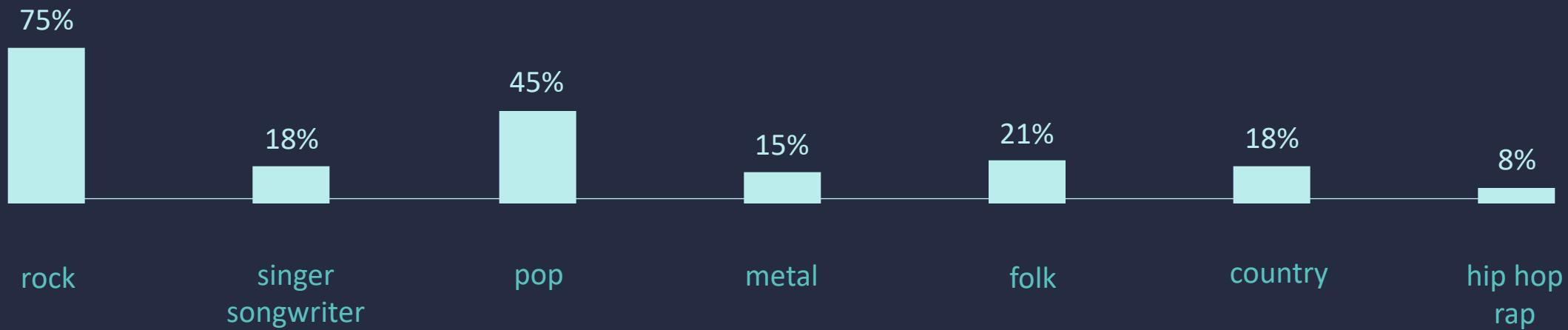
MOST LIKE THE OTHER

- Singer-songwriter and Folk
- Singer-songwriter and Country
- Folk and Rock

MOST UNLIKE THE OTHER

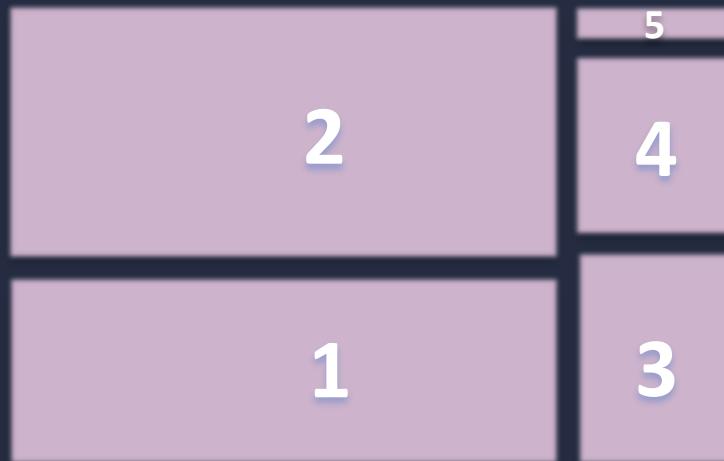
- Hip-Hop/Rap and Rock
- Metal and Pop
- Rock and Pop
- Country and Pop

**SONG GENRES
ARE NOT
MUTUALLY
EXCLUSIVE**



An initial look at our data.

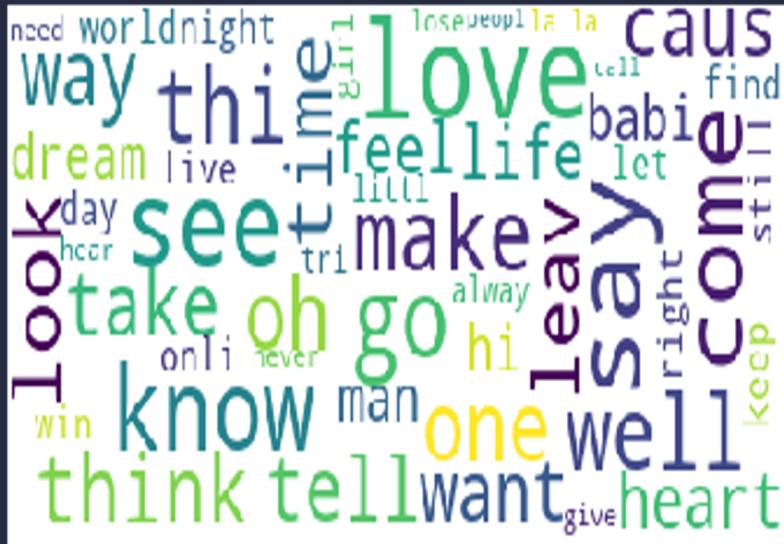
Songs with
multiple genres



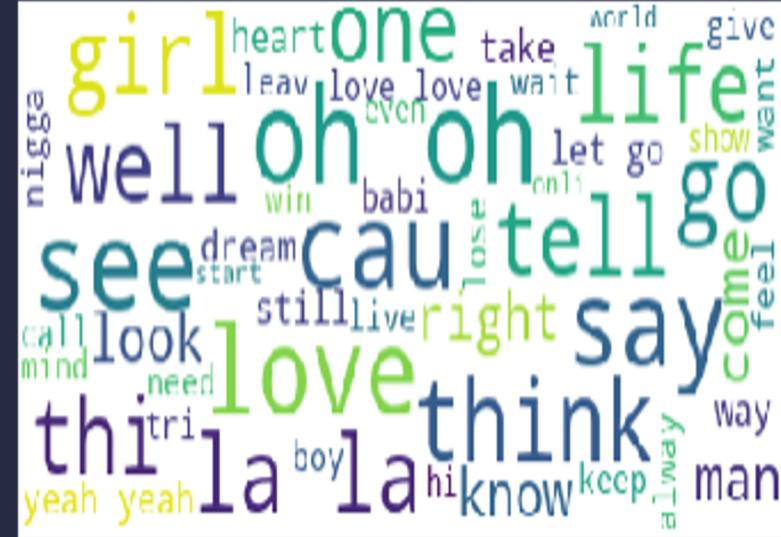
Correlation



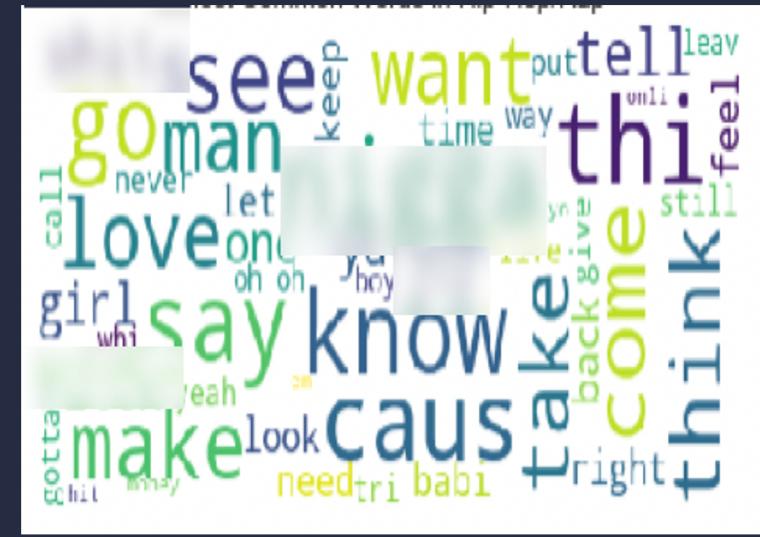
Words By Genre



SINGER-SONGWRITER



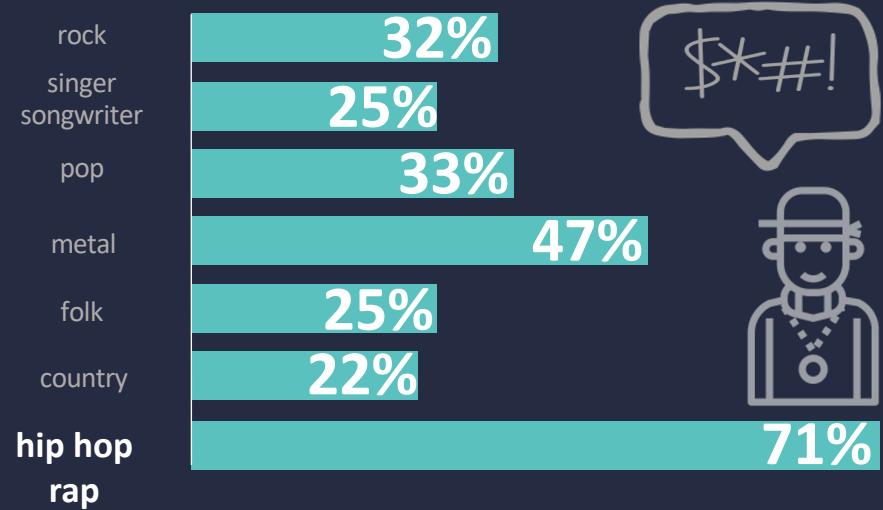
POP



HIP-HOP

Potential Future Exploration

Profanity by genre



`predict` and `predict_prob` from
`profanity_check`

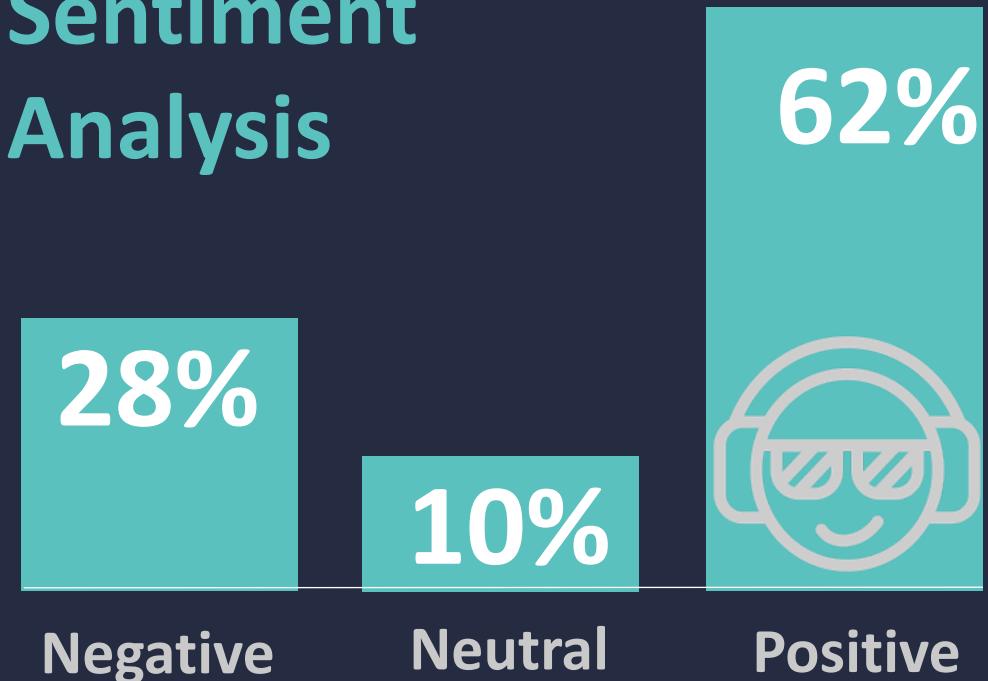
Potential Future Exploration

Profanity by genre



`predict` and `predict_prob` from
`profanity_check`

Sentiment Analysis



`SentimentIntensityAnalyzer()`
`'vader_lexicon'` from `nltk`

DATA CLEANING

Cleaning

Pre-Processing

TEXT FORMATTING

Non-lyric text line breaks

Square brackets vs. parenthesis

LANGUAGE

Non-English lyrics

Probability of English lyrics

langdetect



A hero is
someone who
understands
the
responsibility
that comes
with his
freedom.

-Bob Dylan

Cleaning Pre-Processing

PRE-PROCESSING DATA

Text formatting

Tokenization

`simple_preprocess`

`gensim.utils`

LEMMATIZATION

`WordNetLemmatizer`

`nltk.stem`

REMOVING STOP WORDS

`stopwords nltk.corpus`

Cleaning Pre-Processing

PRE-PROCESSING DATA

Text formatting

Tokenization

simple_preprocess

gensim.utils

LEMMATIZATION

WordNetLemmatizer

nltk.stem

REMOVING STOP WORDS

stopwords nltk.corpus



*"Take a silver dollar and put it in your pocket,
\nNever let it slip away. \nAlways be a man,
not a boy gone astray. \nWhen ya get half cra-
zy from the August heat*

```
""take', 'silver', 'dollar', 'put', 'pocket', 'never',  
'let', 'slip', 'away', 'alway', 'man', 'boy', 'go',  
'astray', 'ya', 'get', 'half', 'craz', 'august', 'heat', '
```

Word2Vec

The word2vec objective function causes the words that occur in similar contexts to have similar embeddings.

What is it?

- A two layer neural network to generate word embeddings (mapping of words) given a text corpus



“Woman” “Man”

0.73
0.89
-1.67
1.32
0.36
-1.49
2.71

0.52
0.76
1.21
0.22
-1.36
0.49
-3.69



Why use it?

- Preserve relationships between words
- Deals with addition of new words in the vocabulary
- Better results in lots of deep learning

Word2Vec

The word2vec objective function causes the words that occur in similar contexts to have similar embeddings.

What is it?

- A two layer neural network to generate word embeddings (mapping of words) given a text corpus

Why use it?

- Preserve relationships between words
- Deals with addition of new words in the vocabulary
- Better results in lots of deep learning



“Woman” “Man”

0.73
0.89
-1.67
1.32
0.36
-1.49
2.71

0.52
0.76
1.21
0.22
-1.36
0.49
-3.69



An Example

The kid said he would grow up to be superman.
The child said he would grow up to be superman.

Thus the words kid and child will have similar word vectors due to a similar context

Doc2Vec

What is Doc2Vec?

- An unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents.

"'take', 'silver', 'dollar', 'put', 'pocket', 'never', 'let',
'slip', 'away', 'alway', 'man', 'boy', 'go', 'astray', 'ya',
'get', 'half', 'craz', 'august', 'heat', 'freez', 'rot', 'road',
'one', 'complain', 'achin', 'feet', 'gonna', 'walk',
'endless', 'highway', 'walk', 'high', 'way', 'till', 'die',
'children', 'goin', 'way', 'better', 'tell', 'home', 'life',
'sweet', 'goodby', 'see', 'detour', 'ahead'..."



Why Doc2Vec?

- We need a vector representation for each lyric, which is a group of words(Paragraph)
- Doc2Vec usually outperforms simple-averaging of Word2Vec vectors

[0.80030984, -1.8968158 , -0.20183174, 2.038949 , 0.6700128 ,
0.80359936, -0.26136535, 0.8038724 , -0.01780331, 1.6143647 ,
-0.66266567, -1.3268974 , 1.4345216 , 1.2054638 , 0.06268109,
-0.29674092, 0.84357643, -0.9477247 , 0.10439444, 0.8167453 ,
-0.02420871, -0.0391538 , 0.11021158, 1.4326952 , 0.01581643,
1.0310534 , -0.2578711 , -0.6832181 , 1.2554826 , 0.5251921 ,
0.09430447, -0.37911037, 1.4898267 , -0.3236018 , 0.9216325 , ...]

MODELING

Understanding the nuances between Multi-label and Multi-class are crucial to identifying the right model.

MULTI-CLASS

More than 2 Classes

Sample is assigned only one class

Classes ARE mutually exclusive



Understanding the nuances between Multi-label and Multi-class are crucial to identifying the right model.

MULTI-CLASS

More than 2 Classes

Sample is assigned only one class

Classes ARE mutually exclusive

movie rating
PG-13



Understanding the nuances between Multi-label and Multi-class are crucial to identifying the right model.

MULTI-CLASS

More than 2 Classes

Sample is assigned only one class

Classes ARE mutually exclusive

movie rating
PG-13



MULTI-LABEL

More than 2 Classes

Samples may be assigned more than one class

Classes ARE NOT mutually exclusive

Understanding the nuances between Multi-label and Multi-class are crucial to identifying the right model.

MULTI-CLASS

More than 2 Classes

Sample is assigned only one class

Classes ARE mutually exclusive

movie rating
PG-13



MULTI-LABEL

More than 2 Classes

Samples may be assigned more than one class

Classes ARE NOT mutually exclusive

movie genres
Action, Adventure, Drama

There are three ways to handle multi-label classification.



- 1) PROBLEM TRANSFORMATION
- 2) ADAPTED ALGORITHMS
- 3) ENSEMBLE METHODS

There are three ways to handle multi-label classification.



- 1) PROBLEM TRANSFORMATION
- 2) ADAPTED ALGORITHMS
- 3) ENSEMBLE METHODS

There are a three ways to handle multi-label classification.

PROBLEM TRANSFORMATION

Transforms multi-label task into several single-label tasks.

BinaryRelevance()

SONGS	GENRES
X ₁	pop folk rap
X ₂	rock
X ₃	pop
X ₄	rock rap
X ₅	rap

There are three ways to handle multi-label classification.

PROBLEM TRANSFORMATION

Transforms multi-label task into several single-label tasks.

BinaryRelevance()

HOW IT WORKS Learns one binary classifier for each label

WORKS Outputs the union of their predictions

	SONGS	GENRES
X ₁	pop	folk rap
X ₂	rock	
X ₃	pop	
X ₄	rock	rap
X ₅	rap	

↓

	pop	rock	folk	rap
X ₁	1			
X ₂	0	1	0	0
X ₃	1	0	0	0
X ₄	0	1	0	1
X ₅	0	0	0	1

There are three ways to handle multi-label classification.

PROBLEM TRANSFORMATION

Transforms multi-label task into several single-label tasks.

BinaryRelevance()

HOW IT WORKS Learns one binary classifier for each label
Outputs the union of their predictions



**SIMPLEST WAY
WIDELY ACCEPTED**



**IGNORES RELATIONSHIPS
BETWEEN LABELS**

	SONGS			GENRES		
X ₁	pop	folk	rap			
X ₂	rock					
X ₃	pop					
X ₄	rock	rap				
X ₅	rap					

	pop		rock		folk		rap	
X ₁	1		X ₁	0	X ₁	1	X ₁	1
X ₂	0		X ₂	1	X ₂	0	X ₂	0
X ₃	1		X ₃	0	X ₃	0	X ₃	0
X ₄	0		X ₄	1	X ₄	0	X ₄	1
X ₅	0		X ₅	0	X ₅	0	X ₅	1

There are a three ways to handle multi-label classification.



ADAPTED ALGORITHMS

Adapts the algorithm to directly perform multi-label classification.

MLkNN()

MULTI-LABEL K- NEAREST NEIGHBORS

There are a three ways to handle multi-label classification.



ADAPTED ALGORITHMS

Adapts the algorithm to directly perform multi-label classification.

ML_kNN()

MULTI-LABEL K- NEAREST NEIGHBORS

HOW IT WORKS

Identifies k nearest neighbor from the training set
Learns info about these neighbors
Determines label

There are a three ways to handle multi-label classification.



ADAPTED ALGORITHMS

Adapts the algorithm to directly perform multi-label classification.

ML_kNN()

MULTI-LABEL K- NEAREST NEIGHBORS

HOW IT WORKS

Identifies k nearest neighbor from the training set
Learns info about these neighbors
Determines label



ACCEPTS MULTI-LABEL CLASSIFICATION AS IS



HATERS WILL SAY THIS METHOD IS LAZY

Not all models are suited for multi-label classification.

RidgeClassifier()

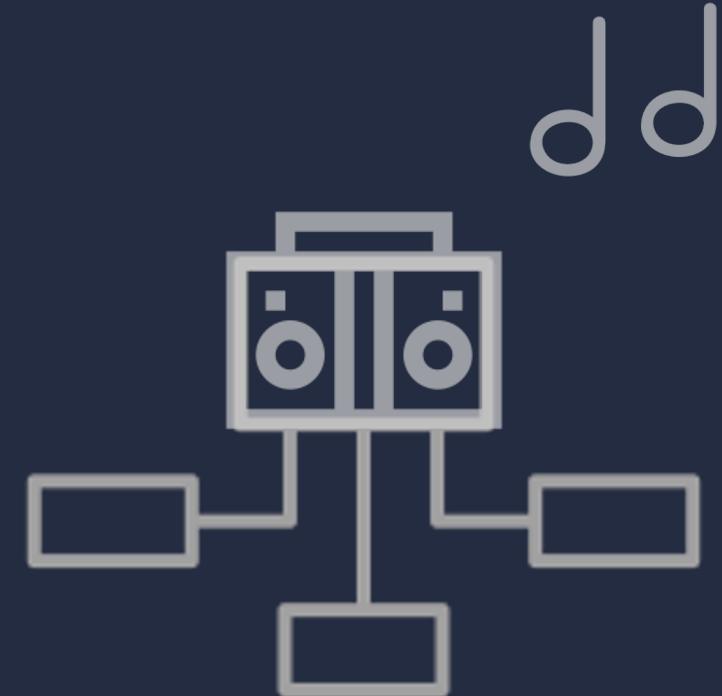
BinaryRelevance()

MLkNN()

GradientBoostingClassifier()

RandomForestClassifier()

LogisticRegression()

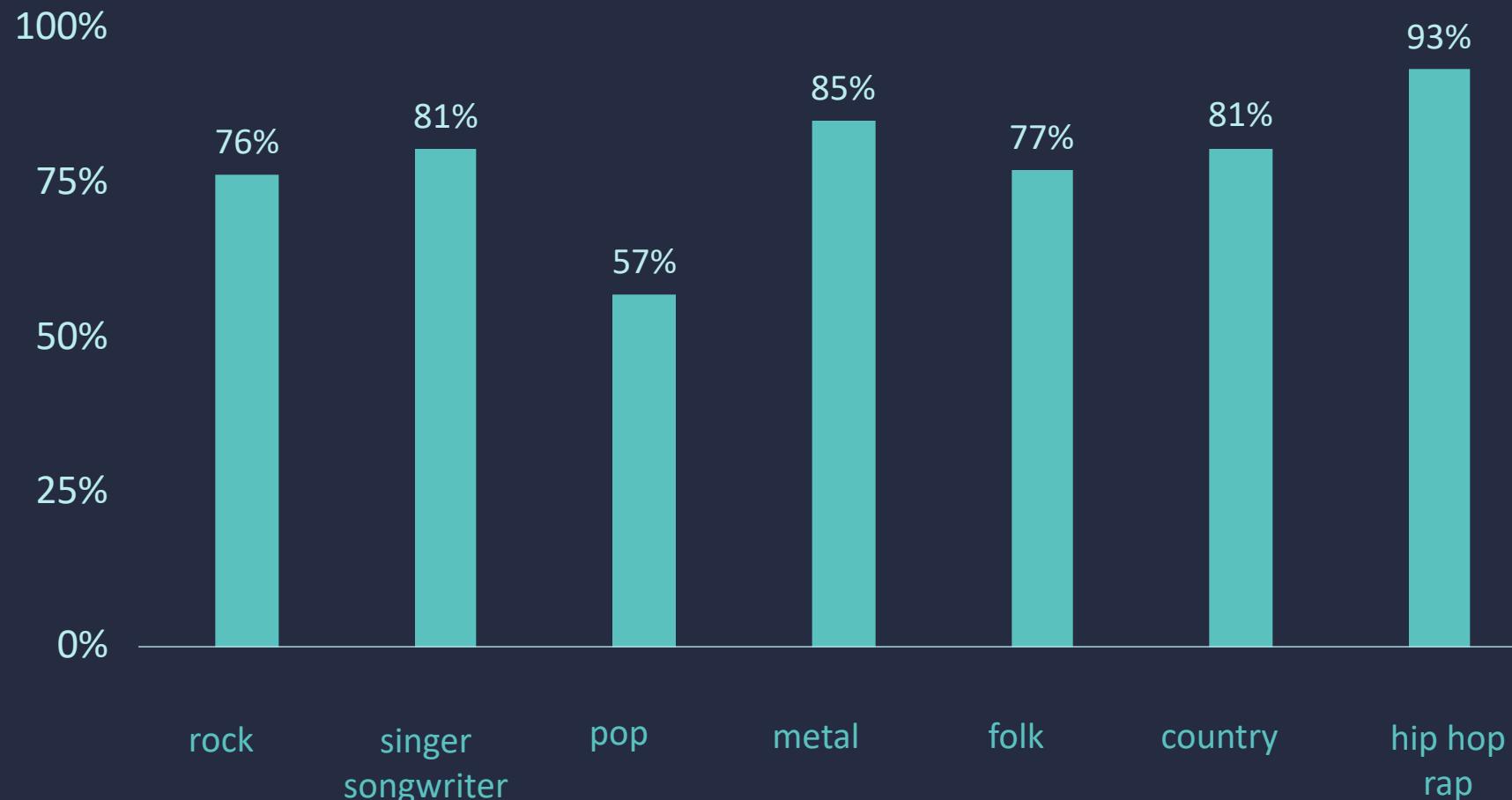


MODELING

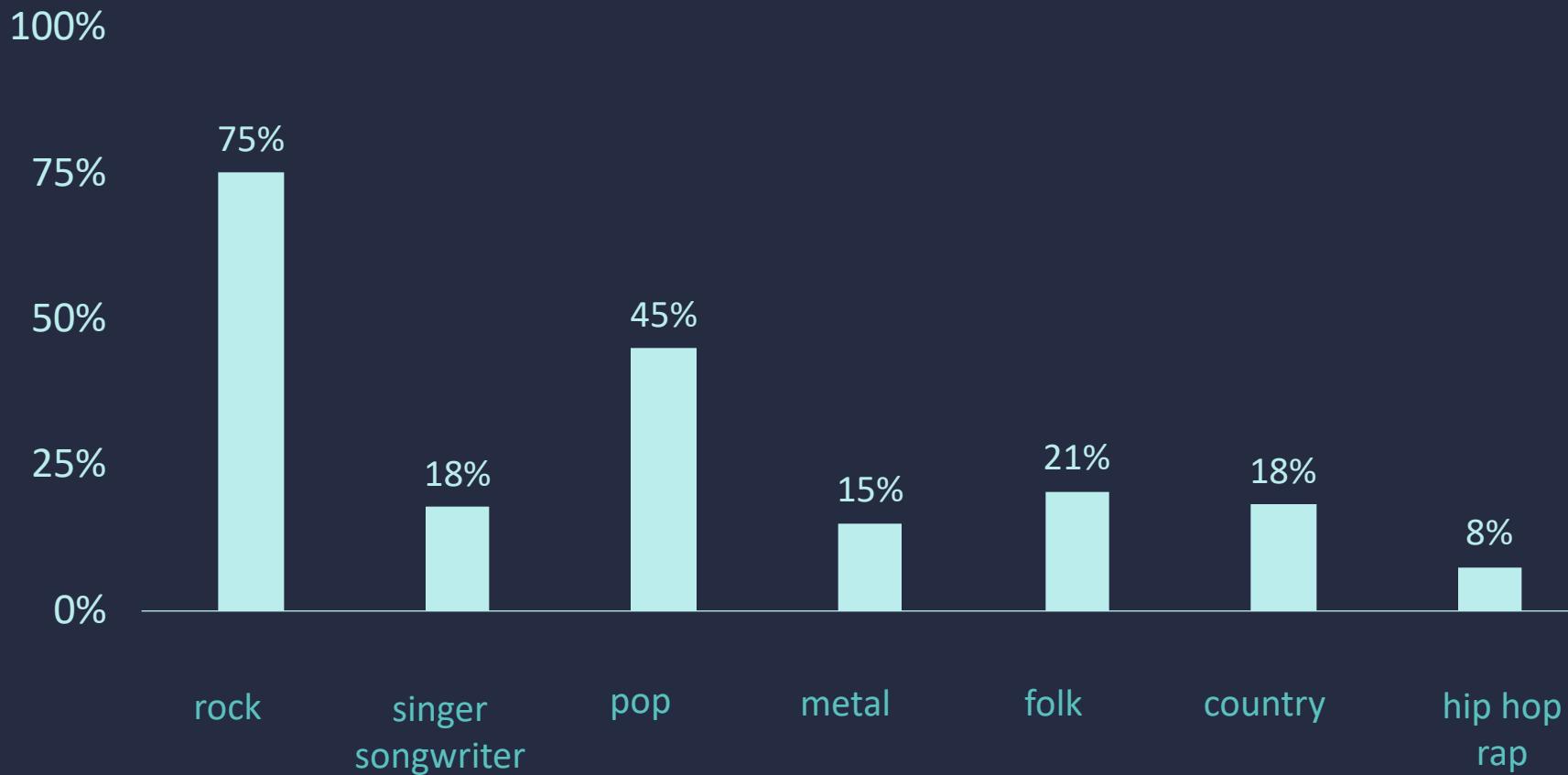
Deep dive

Initial prediction accuracy

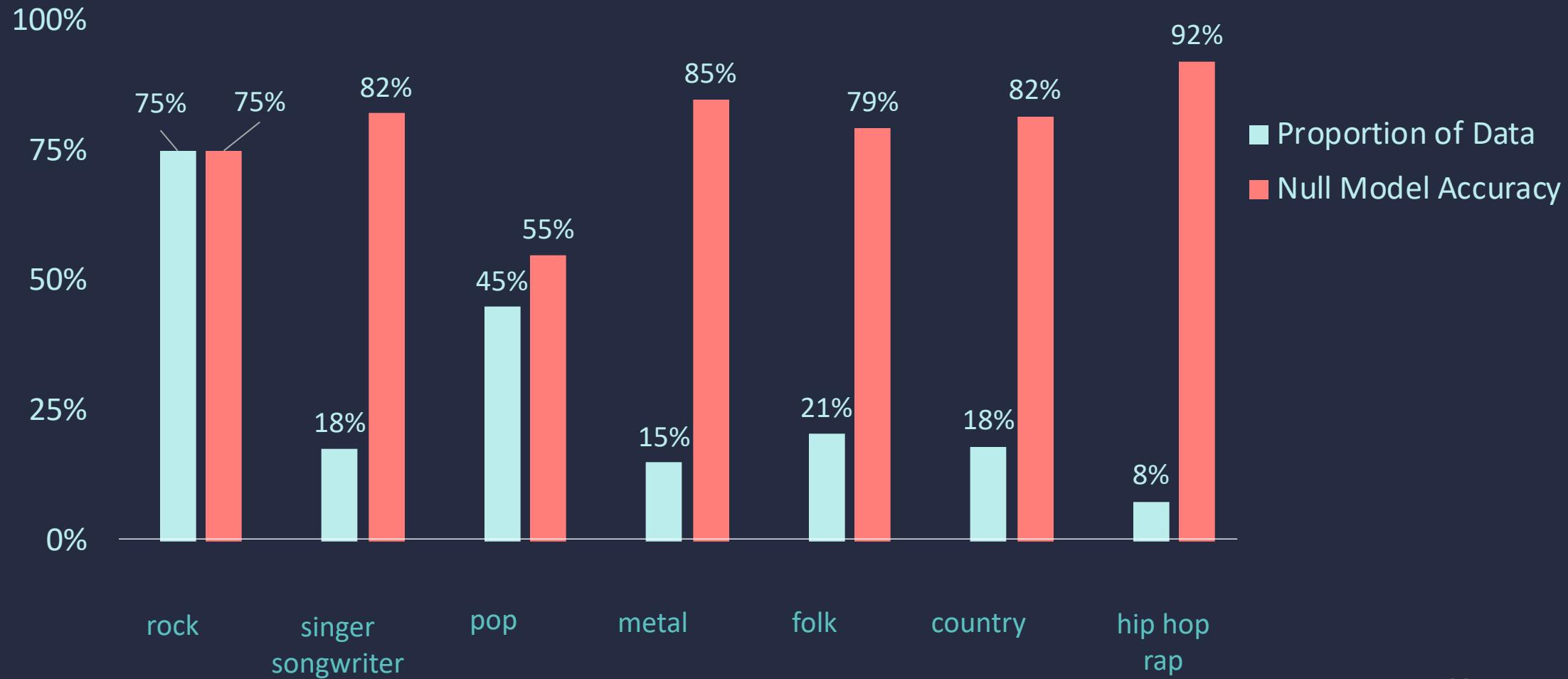
Gradient Boosting



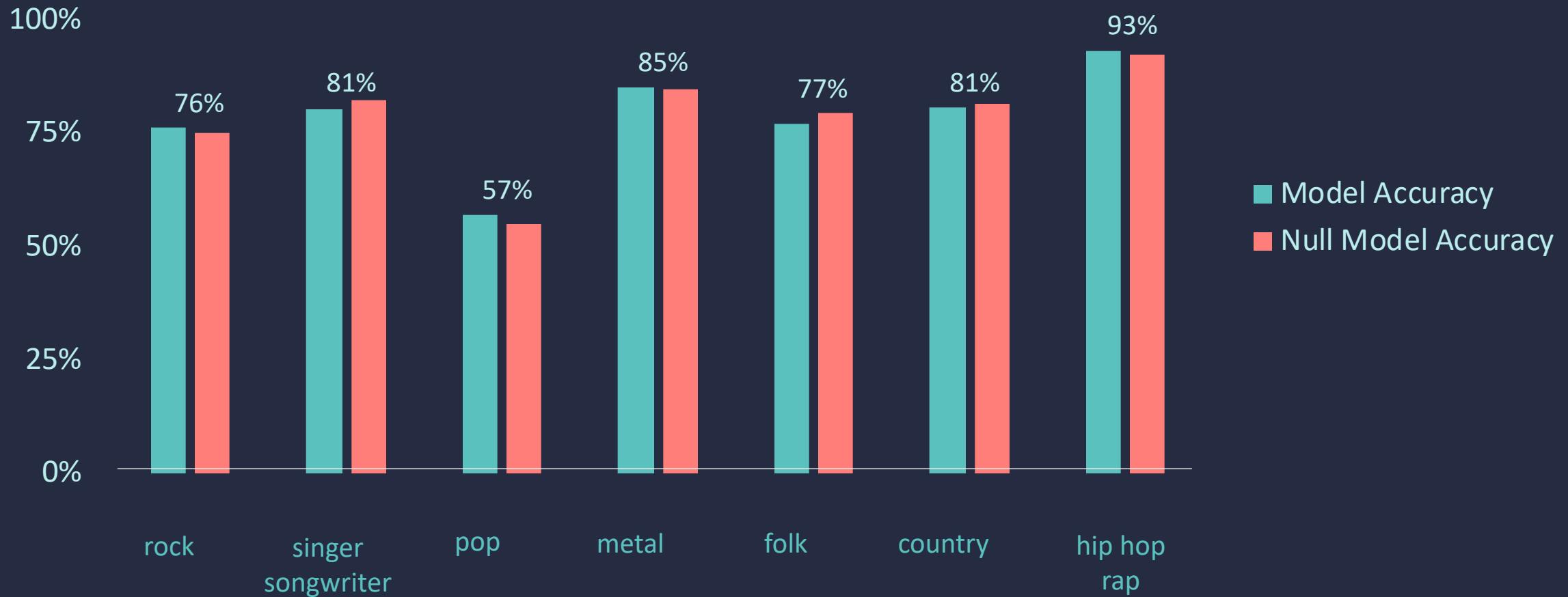
Proportion of songs in each genre



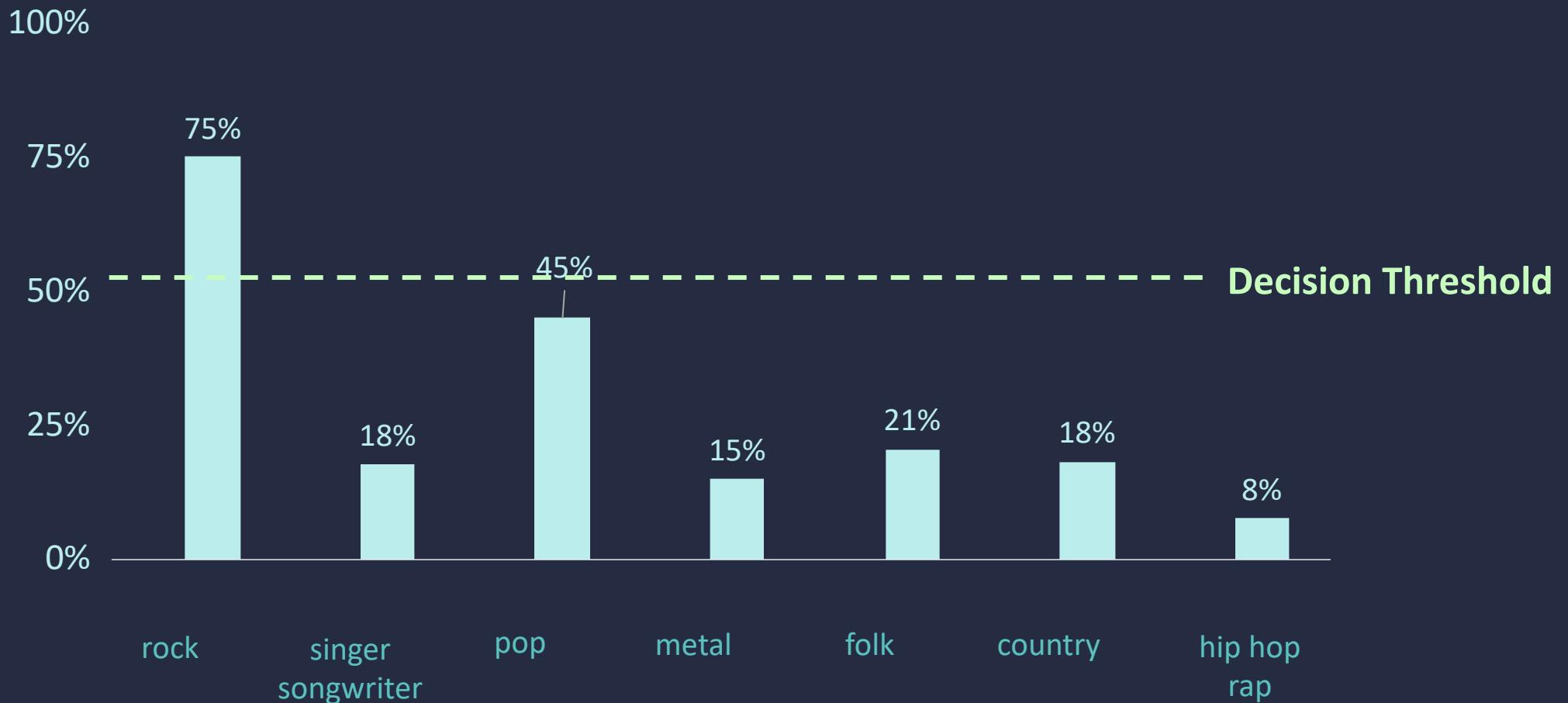
Null Model



Null Model vs Gradient Boosting

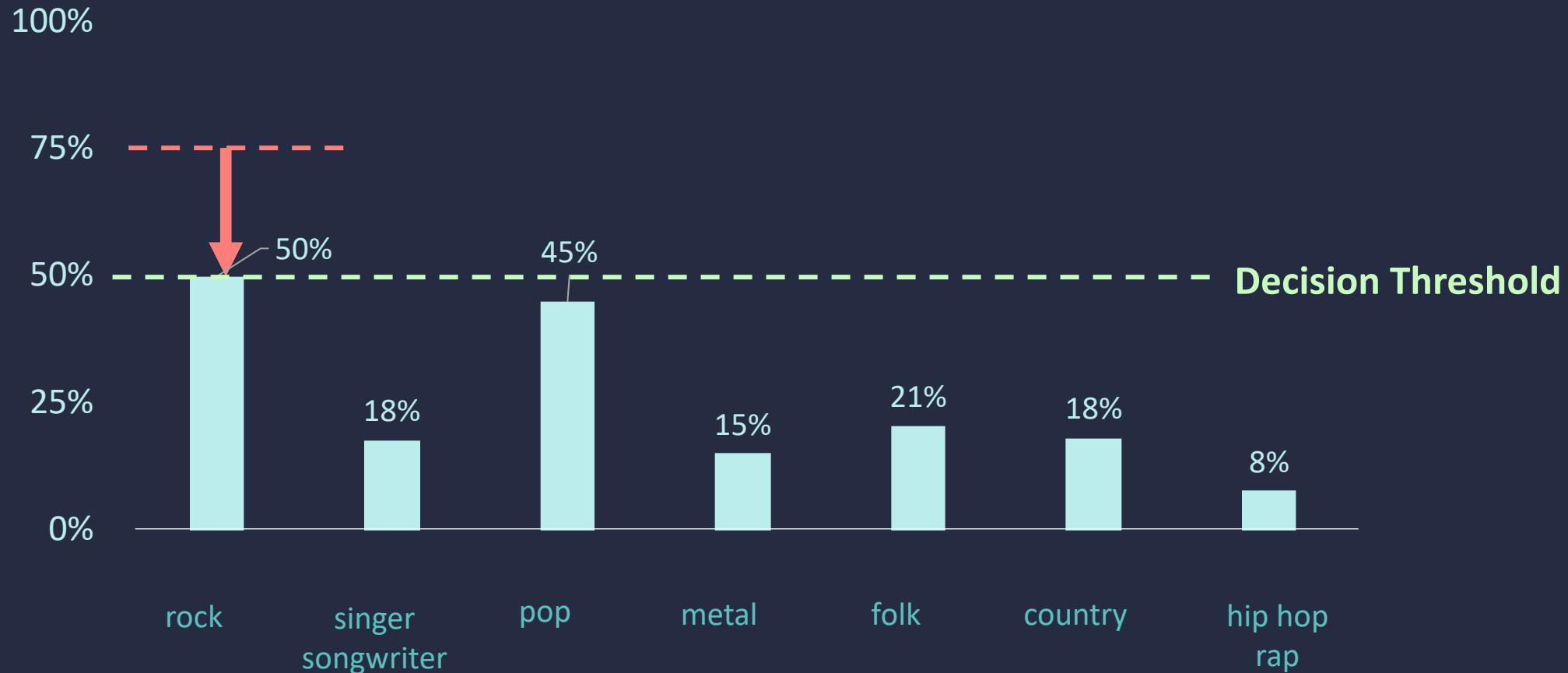


Proportion of songs in each genre



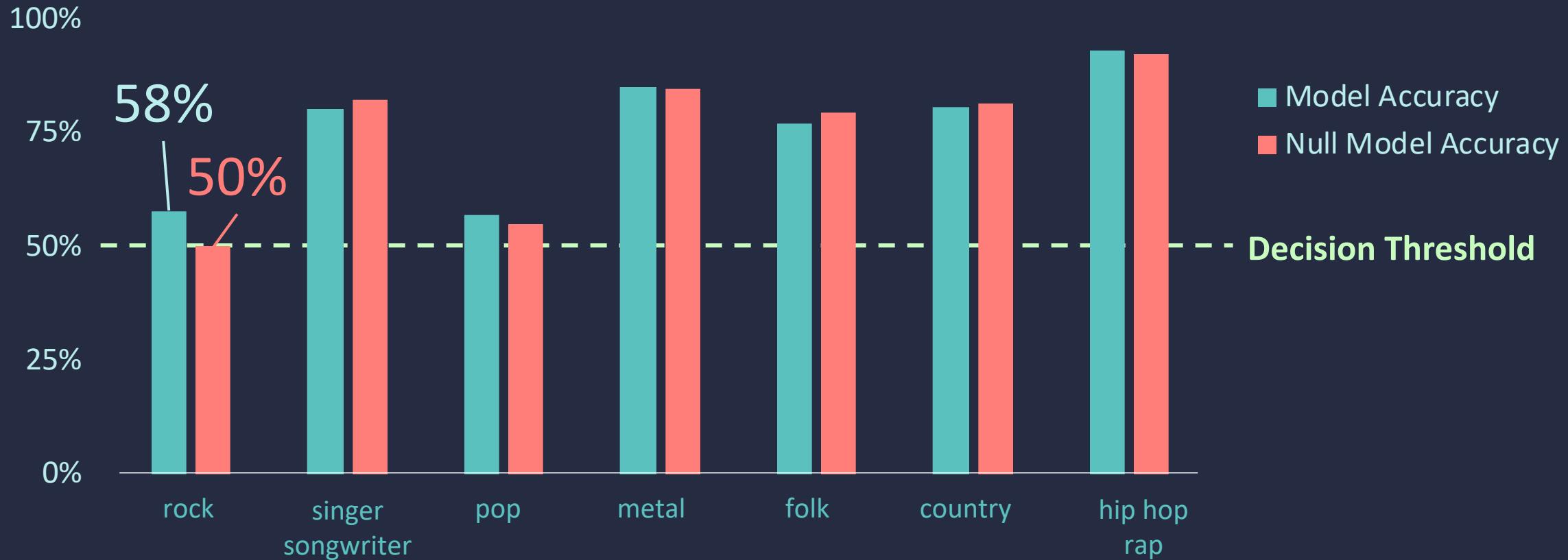
Proportion of songs in each genre

Randomly Undersampling Rock



Null Model vs Gradient Boosting

Randomly Undersampling Rock





EVALUATION

Evaluation Metrics

ACCURACY
PRECISION
RECALL
F1-SCORE
AUC
MATTHEWS
CORRELATION
COEFFICIENT(MCC)

The MCC is in essence a correlation coefficient between the observed and predicted binary classifications

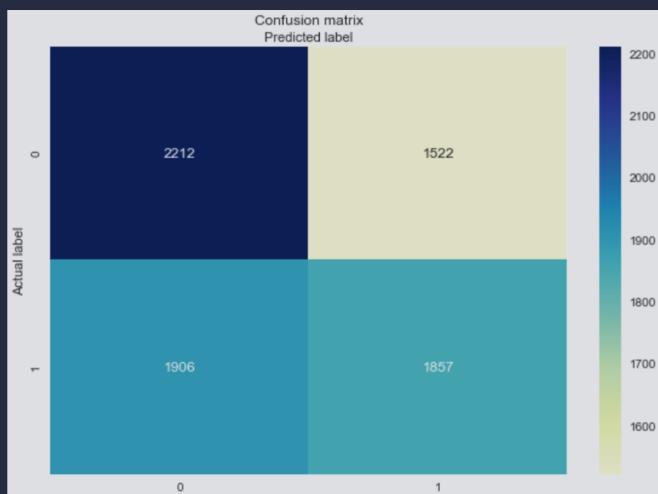
It returns a value between -1 and +1.

A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and -1 indicates total disagreement between prediction and observation.

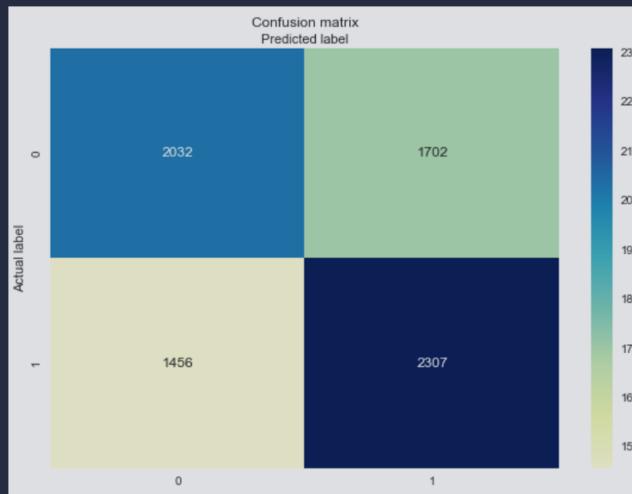
MCC is a useful measure even when the two classes are of very different sizes.

Model Performance Comparison

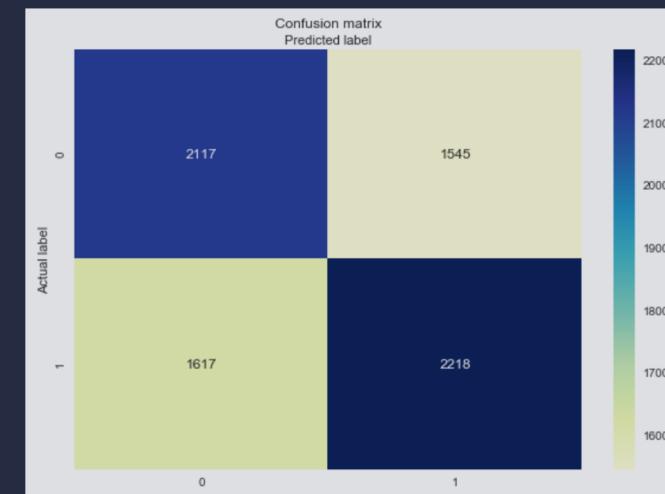
Random Forest



Gradient Boost



Ridge Classification



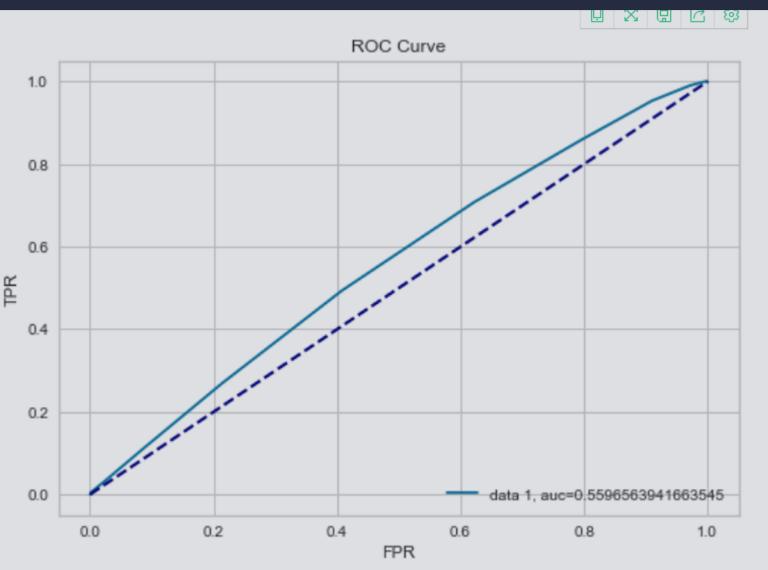
	precision	recall	f1-score
0	0.54	0.59	0.56
1	0.55	0.49	0.52
accuracy			0.54
macro avg	0.54	0.54	0.54
weighted avg	0.54	0.54	0.54

	precision	recall	f1-score
0	0.58	0.54	0.56
1	0.58	0.61	0.59
accuracy			0.58
macro avg	0.58	0.58	0.58
weighted avg	0.58	0.58	0.58

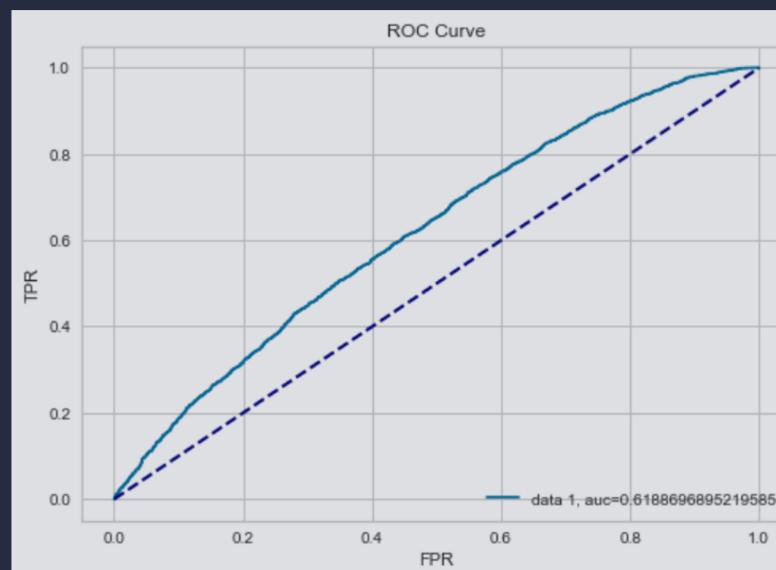
	precision	recall	f1-score
0	0.54	0.59	0.56
1	0.55	0.49	0.52
accuracy			0.54
macro avg	0.54	0.54	0.54
weighted avg	0.54	0.54	0.54

Model Performance Comparison

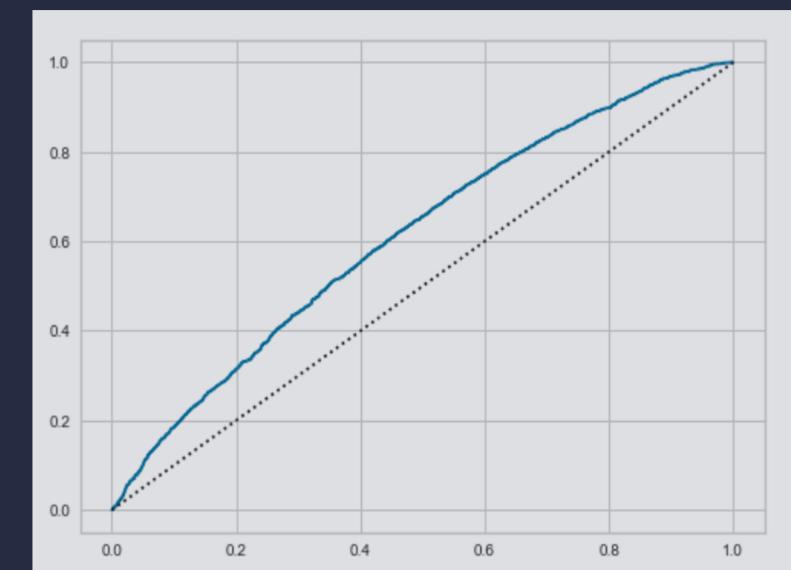
Random Forest



Gradient Boost



Ridge Classification



Accuracy: 0.54
AUC: 0.56
MCC: 0.086

Accuracy: 0.58
AUC: 0.62
MCC: 0.158

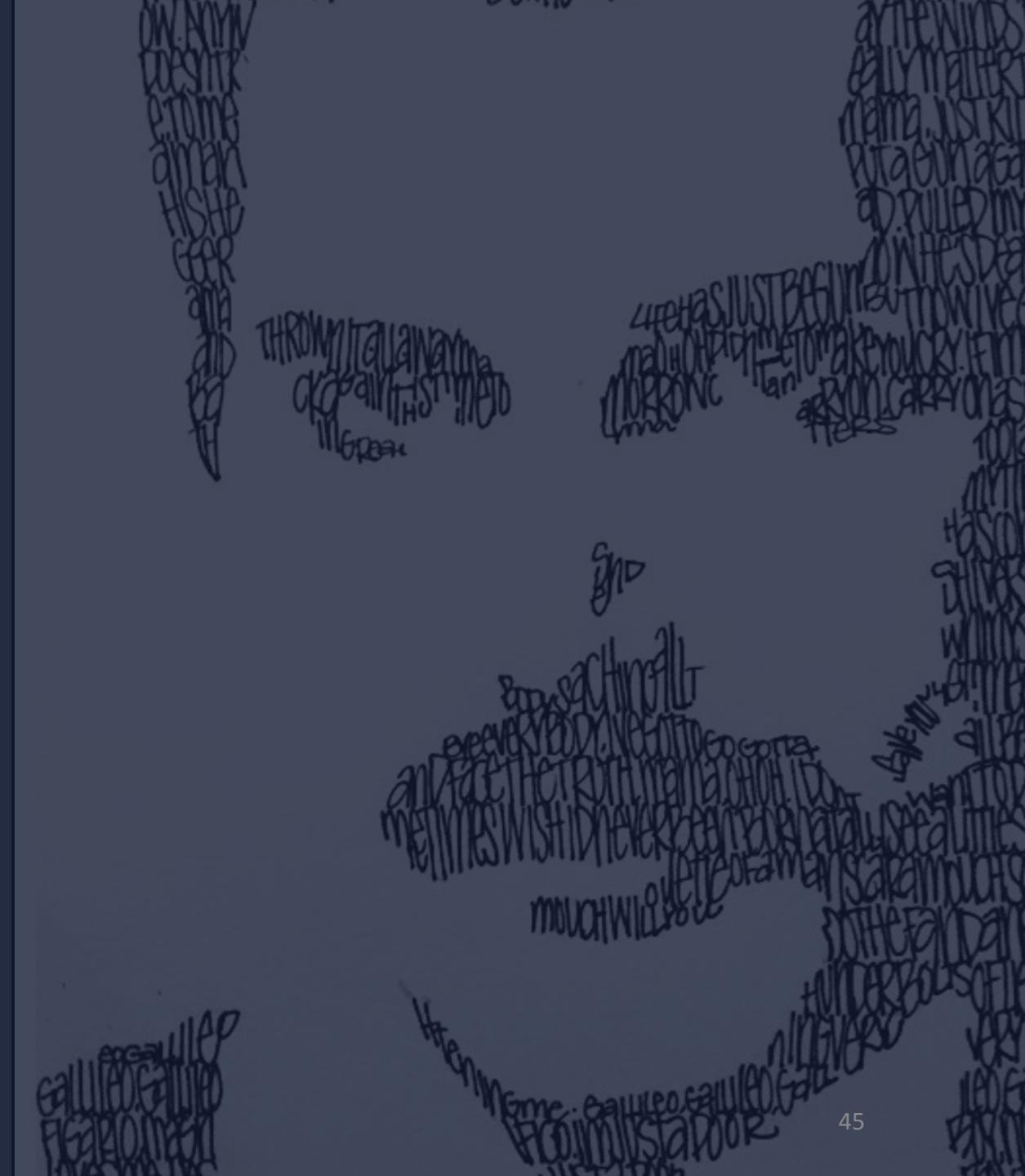
Accuracy: 0.58
AUC: 0.58
MCC: 0.156

Testing our Models on New Data

Data: 3,800 lyrics from MetroLyrics
with song title, artist name, and genre
already

- cleaned, preprocessed and applied doc2vec

Dropped all songs with genres not in
our genre set



Testing our Models on New Data

- Ran on Gradient Boost Model to predict “Rock” Genre
- Accuracy: 0.52
- f1-score: 0.62
- Precision: 0.58
- Recall: 0.67

Confusion Matrix	Actual Rock	Actual Not Rock
	Predicted Rock	Predicted Not Rock
Predicted Rock	700	352
Predicted Not Rock	497	230

CONCLUSION

Balancing Multi-label Data is an unresolved problem in the industry.



CHALLENGE

Common balancing techniques
handle single labels

Labels are not independent

Balance within/between labels

Balancing Multi-label Data is an unresolved problem in the industry.



CHALLENGE

Common balancing techniques handle single labels

Labels are not independent

Balance within/between labels

SOLUTION

REMEDIAL - a decoupling and resampling algorithm

SCUMBLE - a measure of dependence between labels

Cleanliness is a virtue – Data Cleanliness is a necessity.

CHALLENGE

Weighting words by occurrence

Stopword selection



Cleanliness is a virtue – Data Cleanliness is a necessity.

CHALLENGE

Weighting words by occurrence
Stopword selection



SOLUTION

Compare with alternative
feature extraction methods

Refine stopwords

Additional Learnings



Understanding multi-class versus multi-label

Scaling the amount of available data

Understand metrics of model success

**With a few improvements, our model has potential
for many business use cases.**



BUSINESS USE CASES

Music streaming services
additional metrics for recommendation

**With a few improvements, our model has potential
for many business use cases.**



BUSINESS USE CASES

Music streaming services
additional metrics for recommendation

New artists
identify potential new demographics

With a few improvements, our model has potential for many business use cases.



BUSINESS USE CASES

Music streaming services
additional metrics for recommendation

New artists
identify potential new demographics

Media production
algorithmic song generation

THANK YOU
