Predicting Song Genre

...

Kevin Bengtsson, Lucius Bynum, Drew Summy
25 April 2018

Presentation Outline

1. Motivation/Introduction
2. Exploratory Data Analysis
3. Methods
    a. Initial Results
    b. Model Improvement
4. Final Results and Conclusions

Project objective:
Given lyrical and musical data, predict a song's genre

Motivation

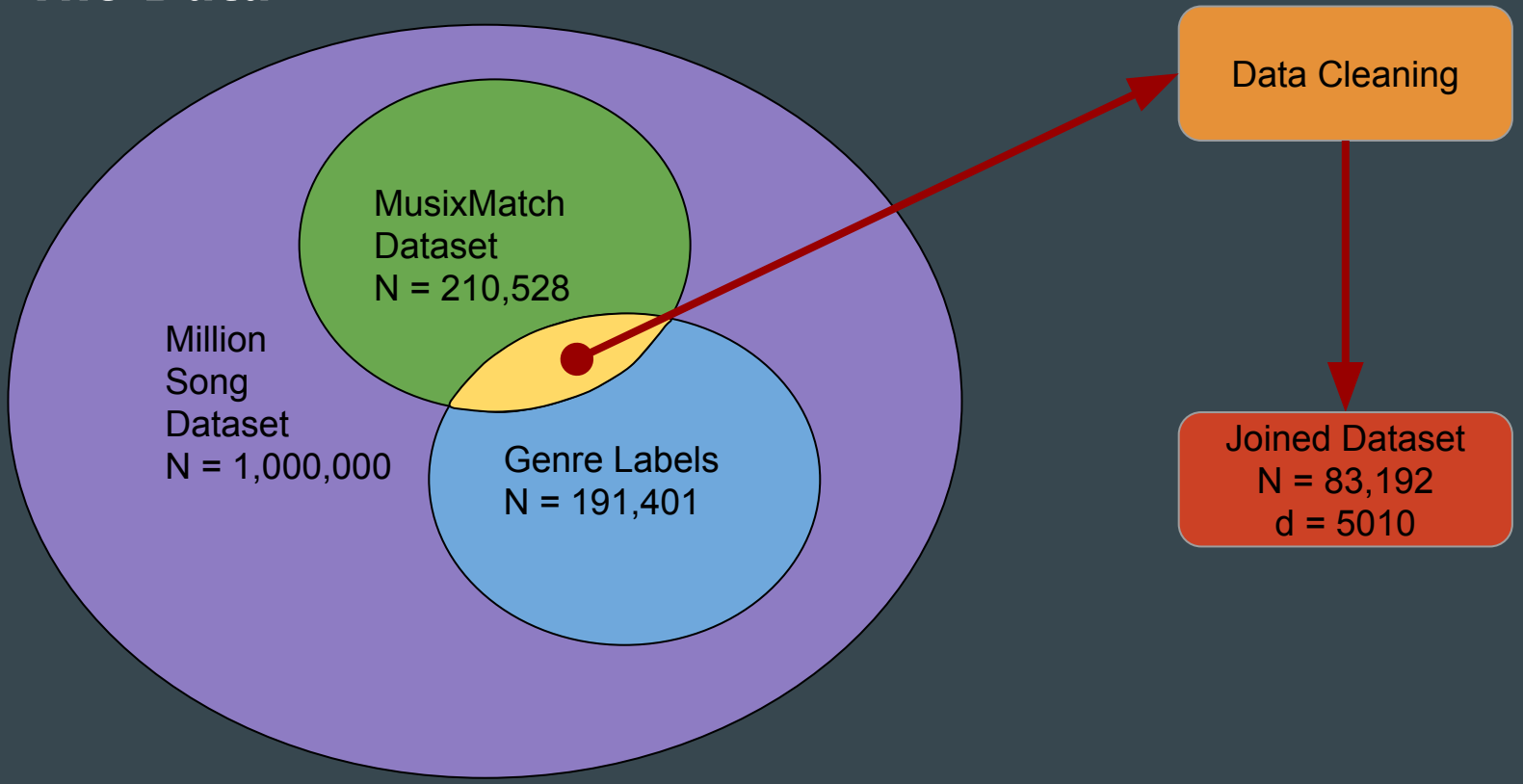
We like music!

Existing businesses

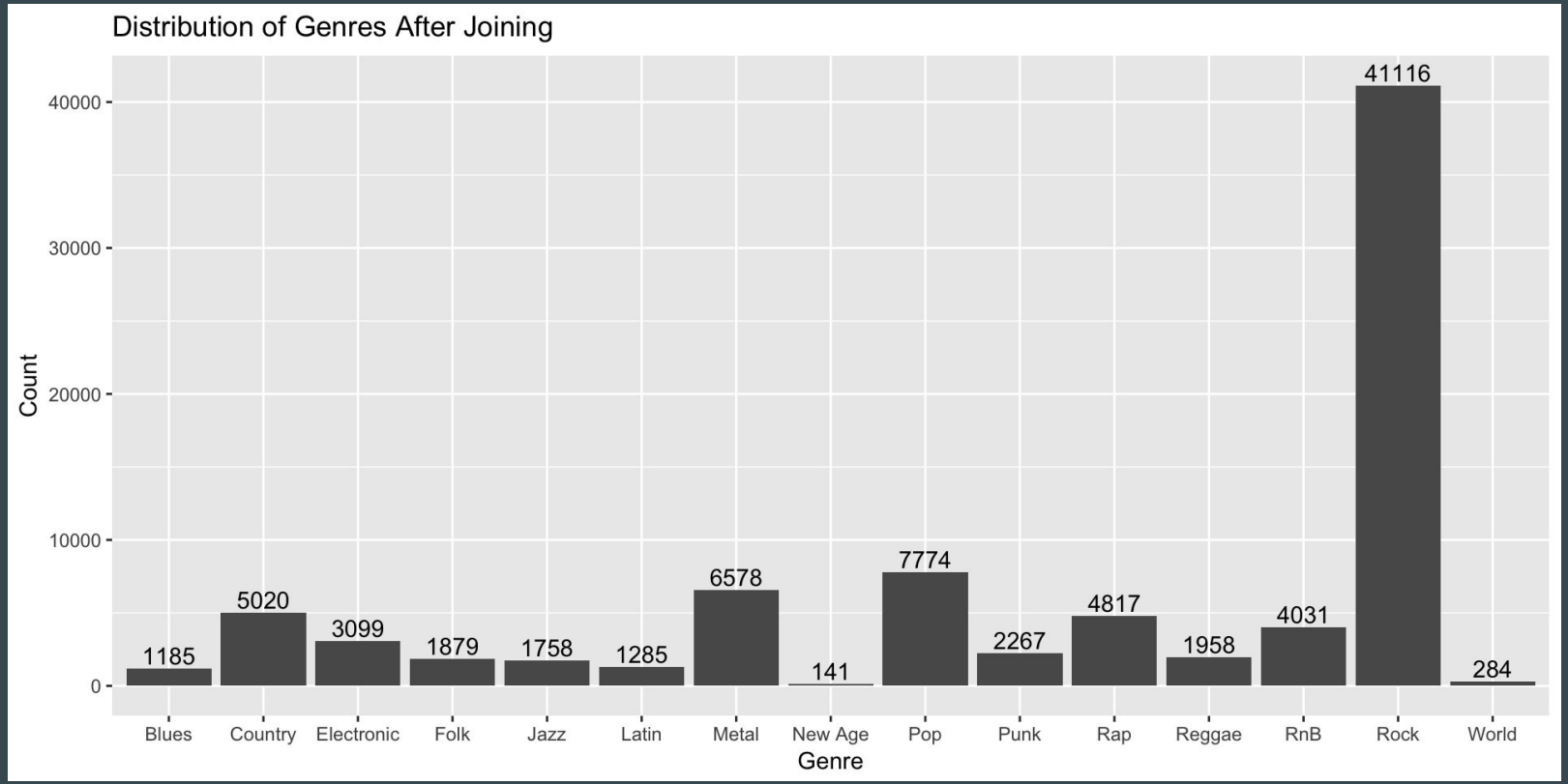
Understanding genres

Exploratory Data Analysis

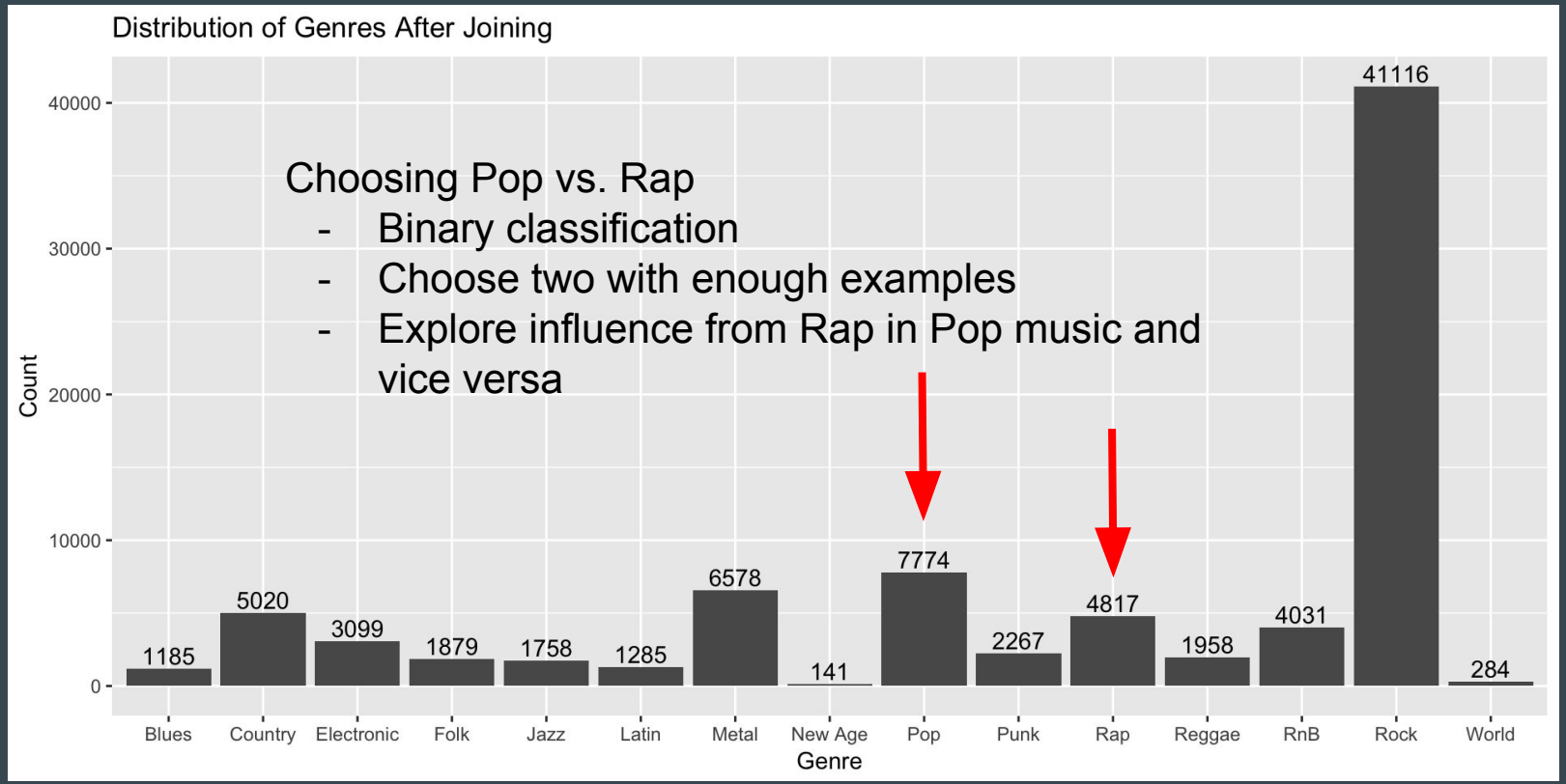
The Data


Million
Song
Dataset
N = 1,000,000
MusixMatch
Dataset
N = 210,528
Genre Labels
N = 191,401
Data Cleaning
Joined Dataset
N = 83,192
d = 5010

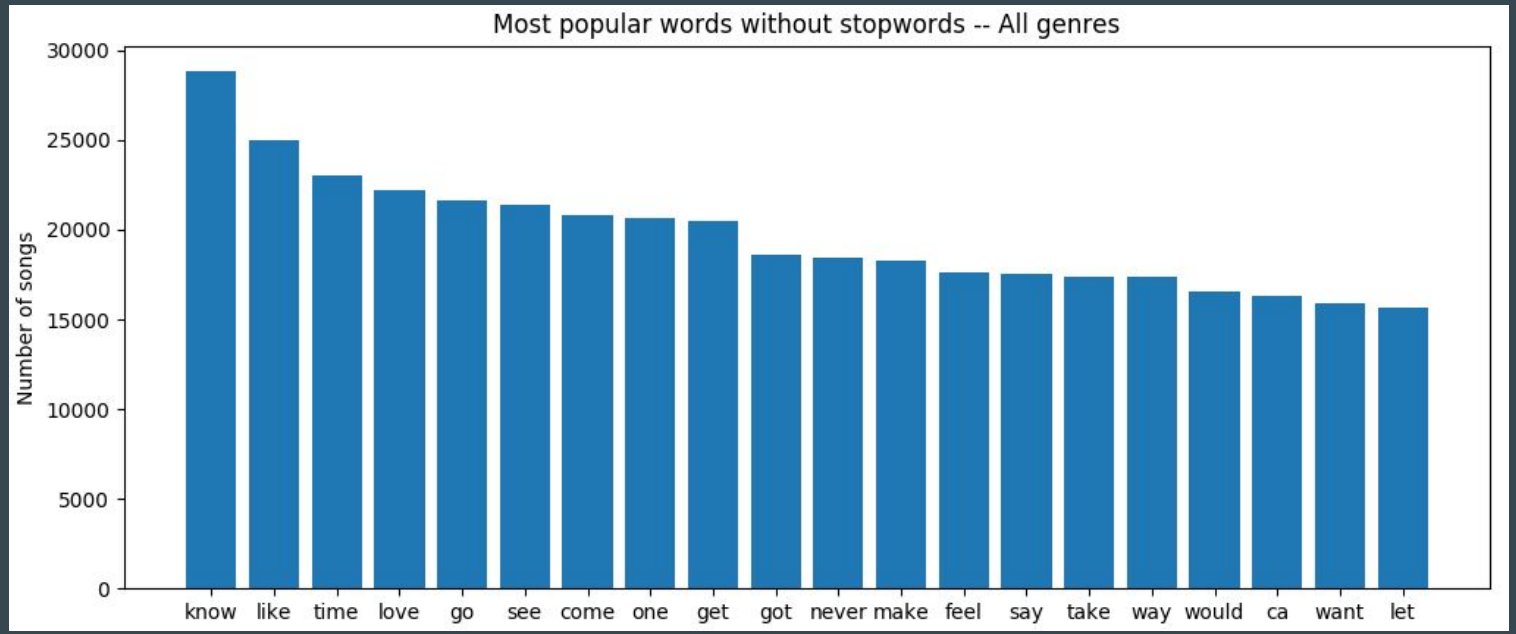
Distribution of Genre Labels After Joining



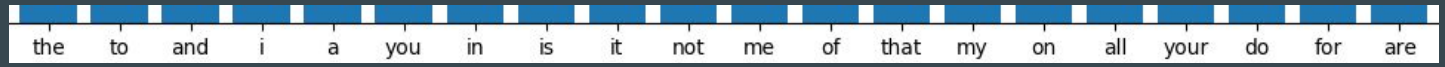
Distribution of Genre Labels After Joining



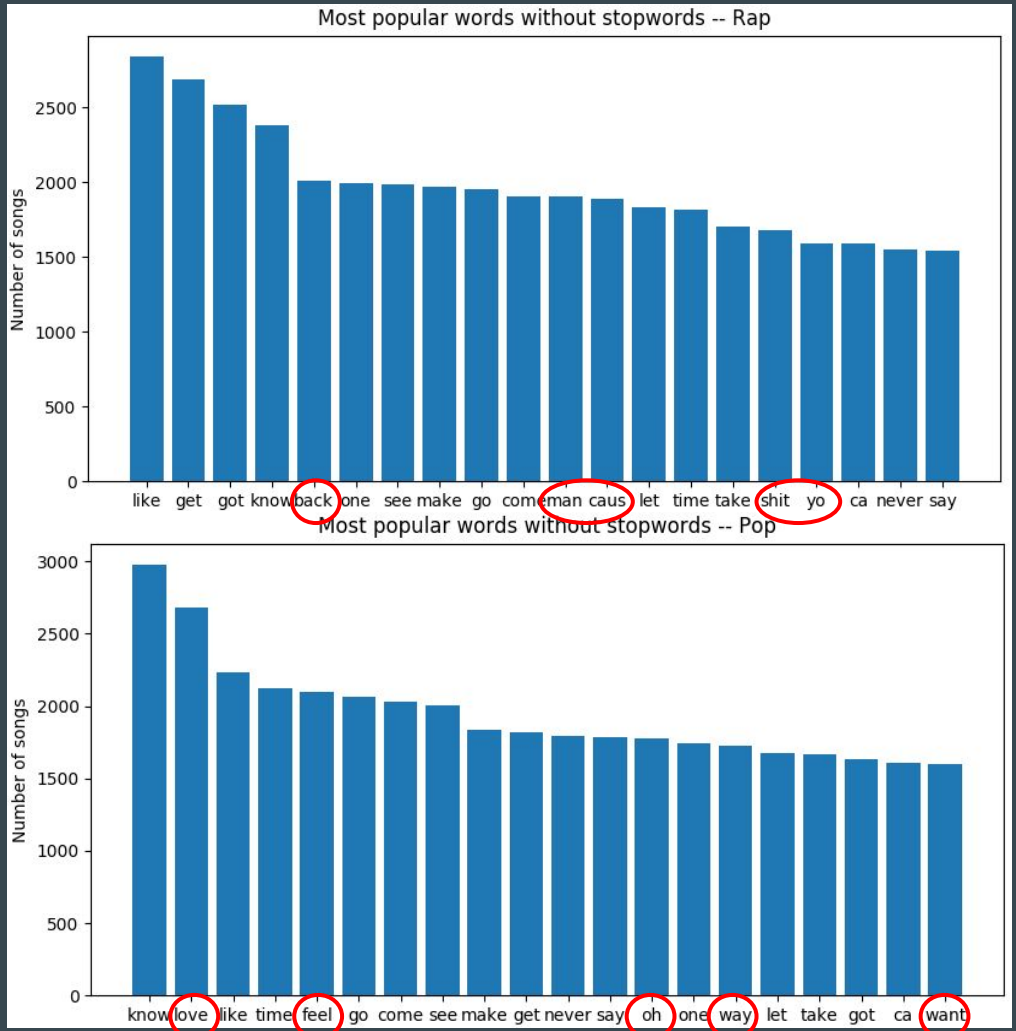
Word Popularity Across the Lyric Data



With stopwords:



Word Popularity - Rap vs. Pop

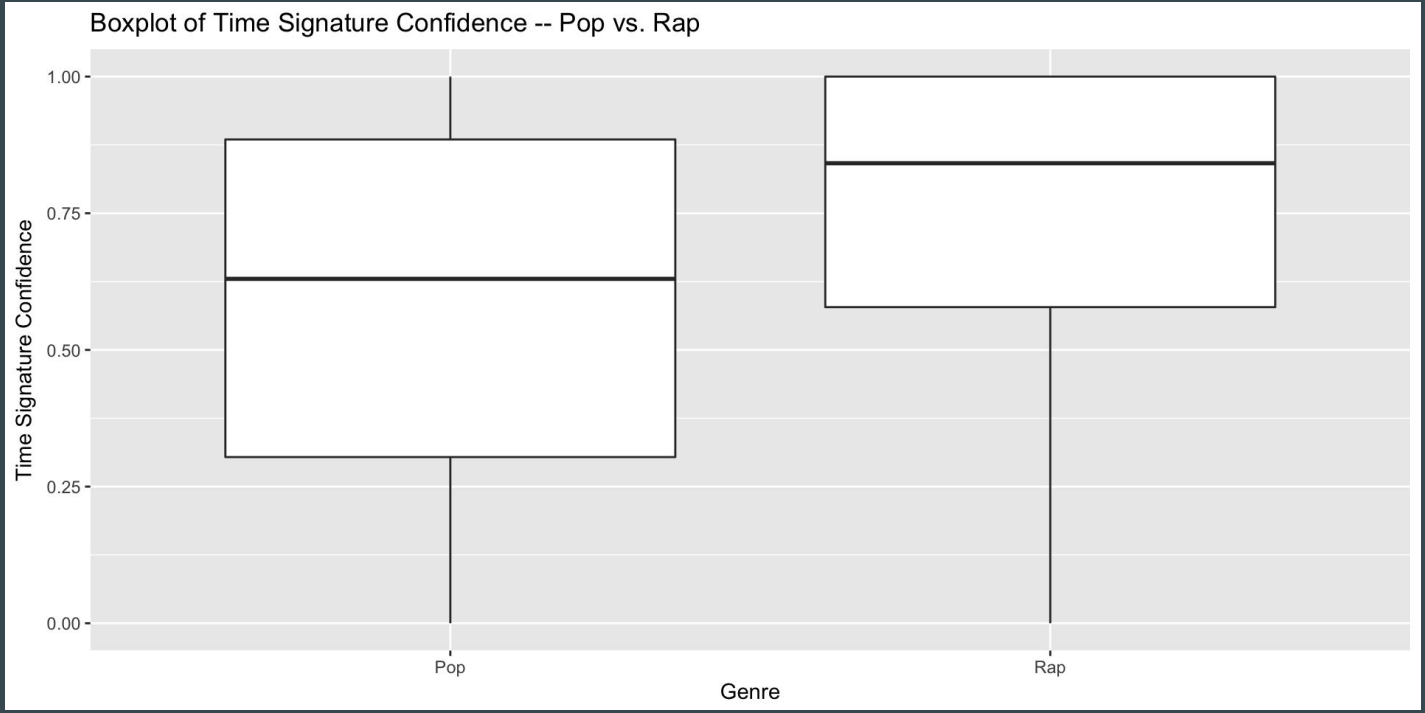


MSD Data Cleaning

- 46 features
  - Usable: loudness, tempo, key, etc.
  - Not-so-usable: bars_start (variable length), danceability/energy (0 for all examples), etc.

- Reduced to:
  - key
  - key_confidence
  - mode
  - mode_confidence
  - time_signature
  - time_signature_confidence
  - duration
  - end_of_fade_in
  - loudness
  - tempo

MSD Data -- Time Signature Confidence



Prediction Methods

Modeling Approach

Fit Initial Models

- Random Forest
- Logistic Regression
- Linear SVM
- RBF SVM

Initial Models and Challenges

1. Problem: very poor performance
   - Not enough data for some classes
   - Large class imbalance

   Solution: binary classification

2. Problem: long training time (many hours)
   - Lots of features

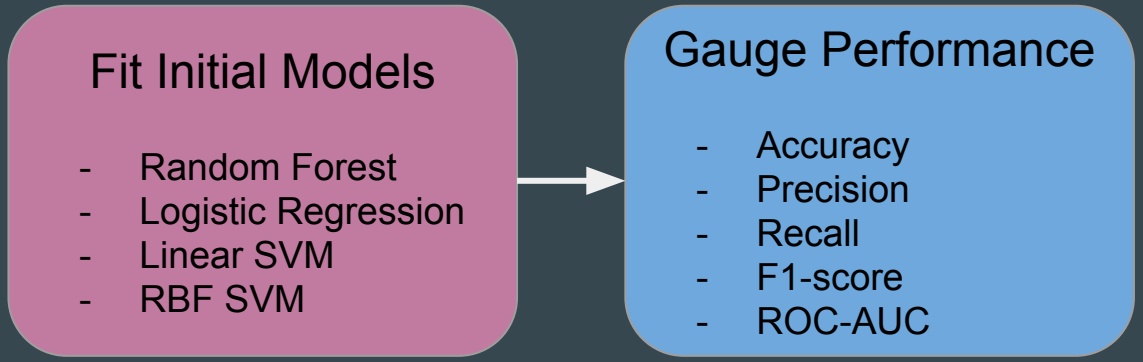
   Solution: train on a subset

Modeling Approach

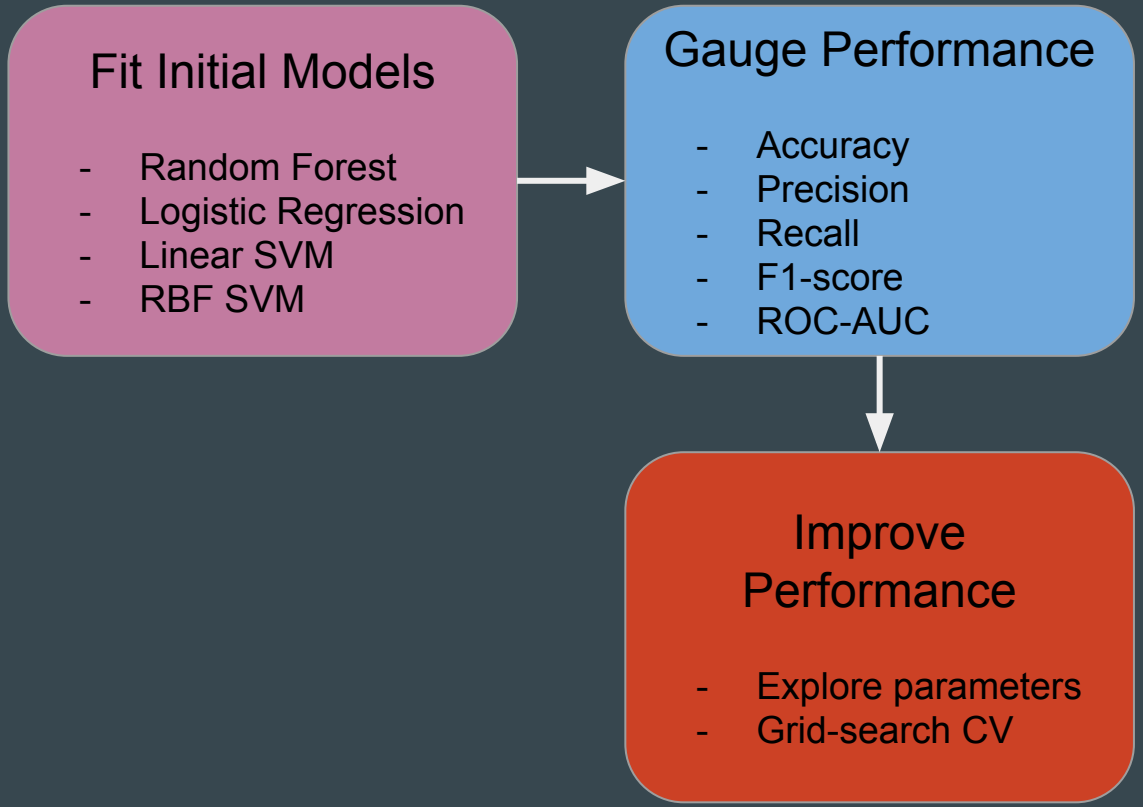
Fit Initial Models

- Random Forest
- Logistic Regression
- Linear SVM
- RBF SVM

Modeling Approach



Modeling Approach

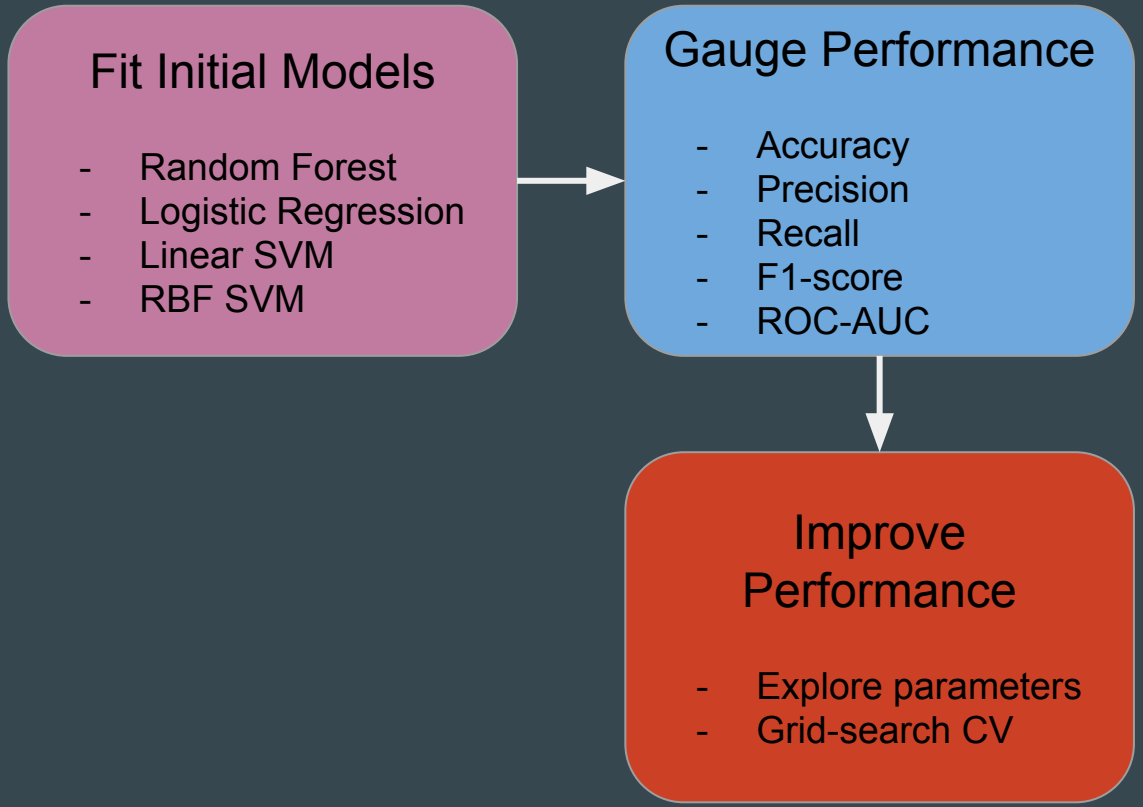


Model Improvements

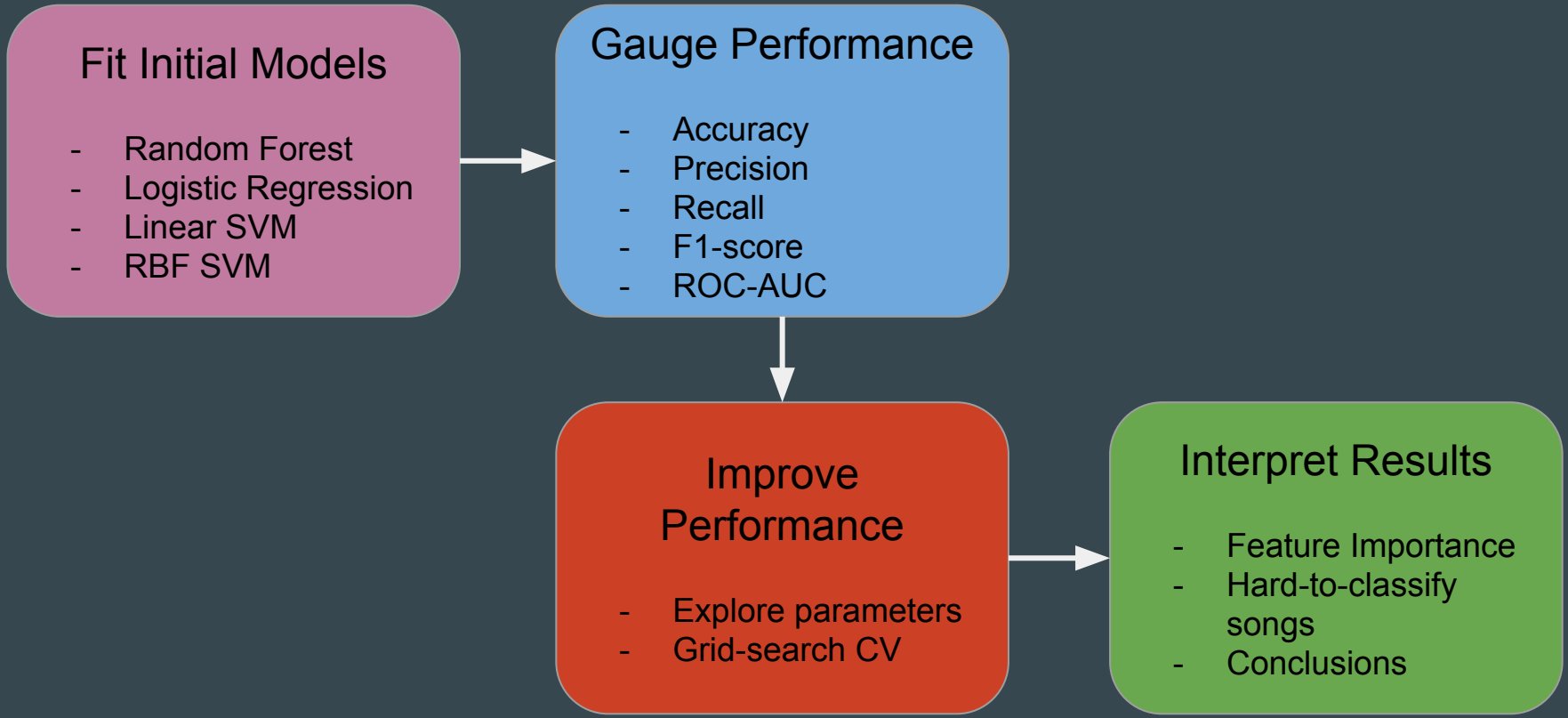
Identify optimal model parameters using grid search cross-validation

Random Forest	Logistic Regression	Linear SVM	RBF SVM
Number of estimators Max depth Max features Class weight Criterion (gini vs. entropy)	C Class weight	C Class weight	C Class weight Gamma

Modeling Approach



Modeling Approach

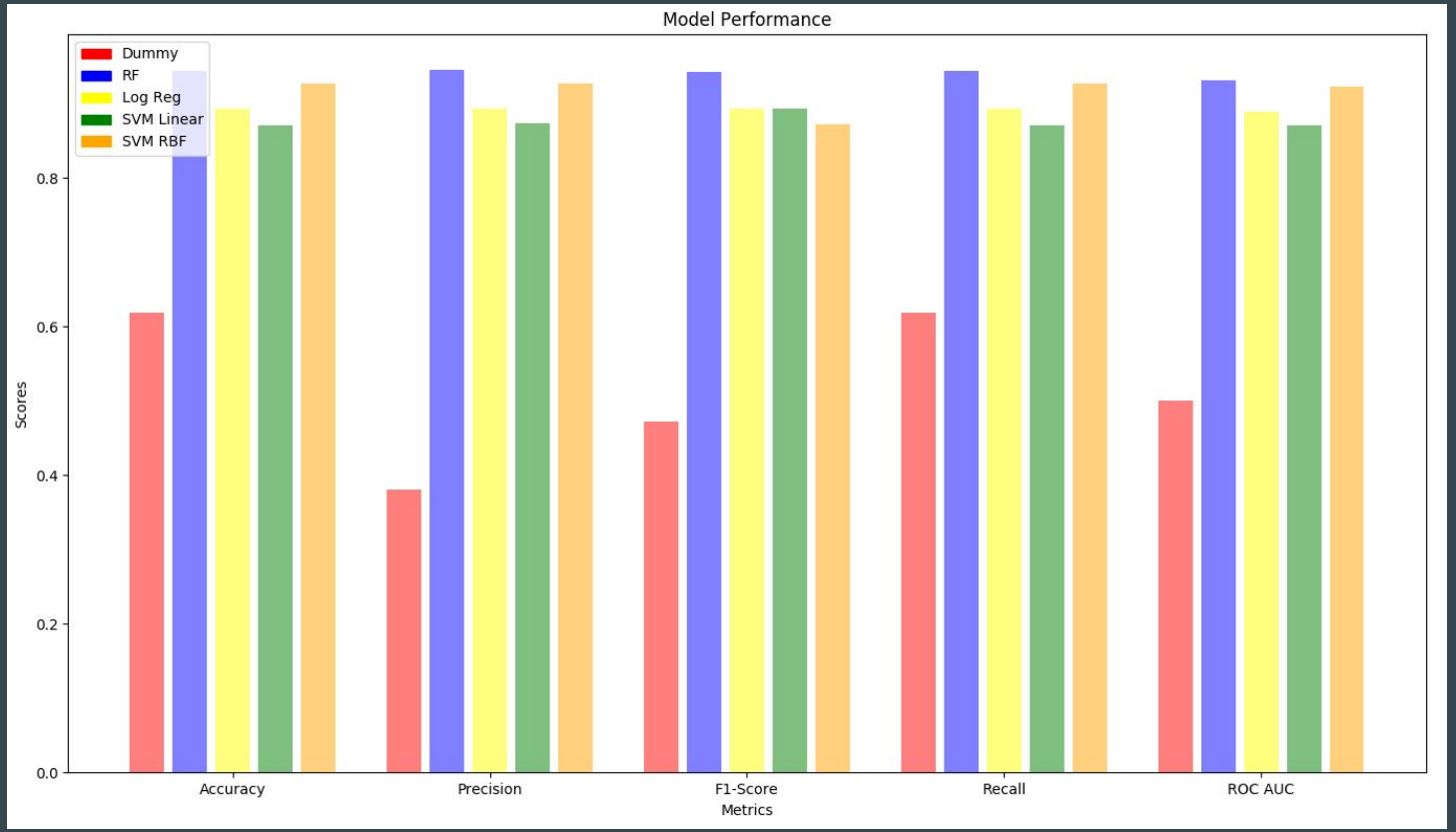


Results

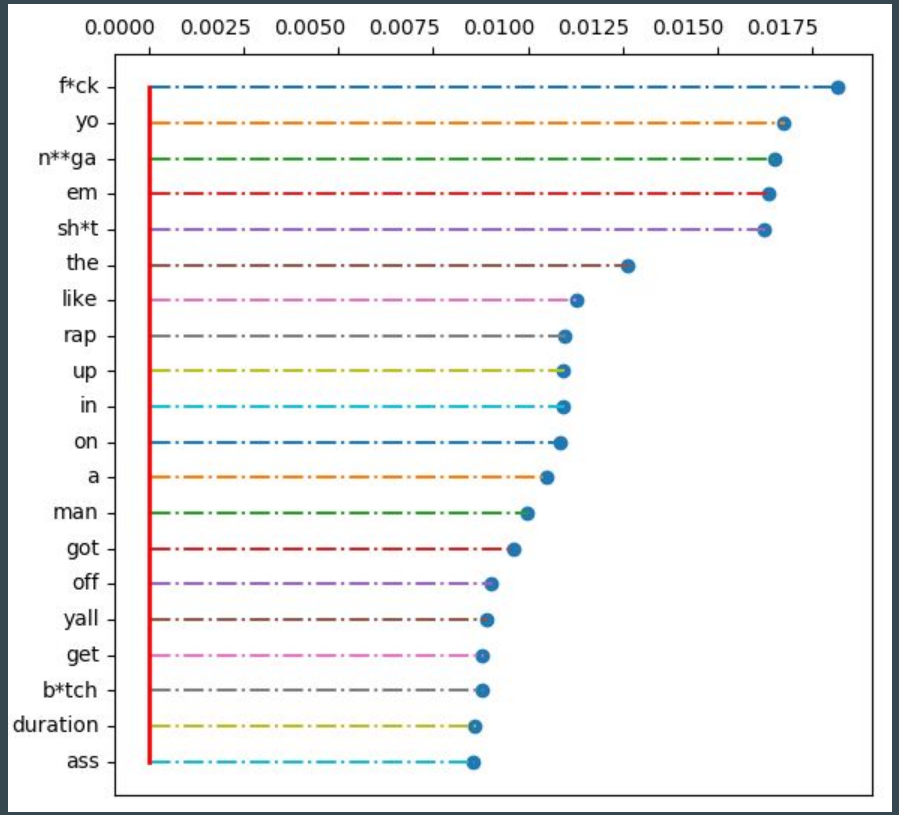
Results

	Accuracy	Precision	F1-score	Recall	ROC AUC
Dummy Classifier	0.618	0.381	0.472	0.618	0.500
Random Forest	**0.944**	**0.946**	**0.943**	**0.944**	**0.931**
Logistic Regression	0.893	0.893	0.893	0.893	0.889
Linear SVM	0.871	0.874	0.872	0.871	0.870
RBF SVM	0.927	0.927	0.927	0.927	0.923

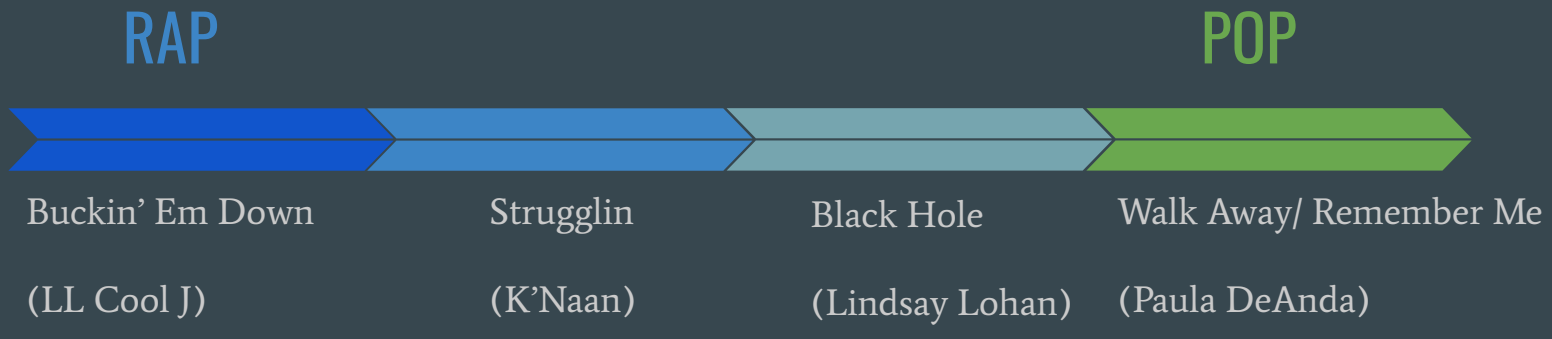
Results


Model Performance
Dummy
RF
Log Reg
SVM Linear
SVM RBF
Scores
0.0
0.2
0.4
0.6
0.8
Accuracy
Precision
F1-Score
Recall
ROC AUC
Metrics

Feature Importance -- Random Forest



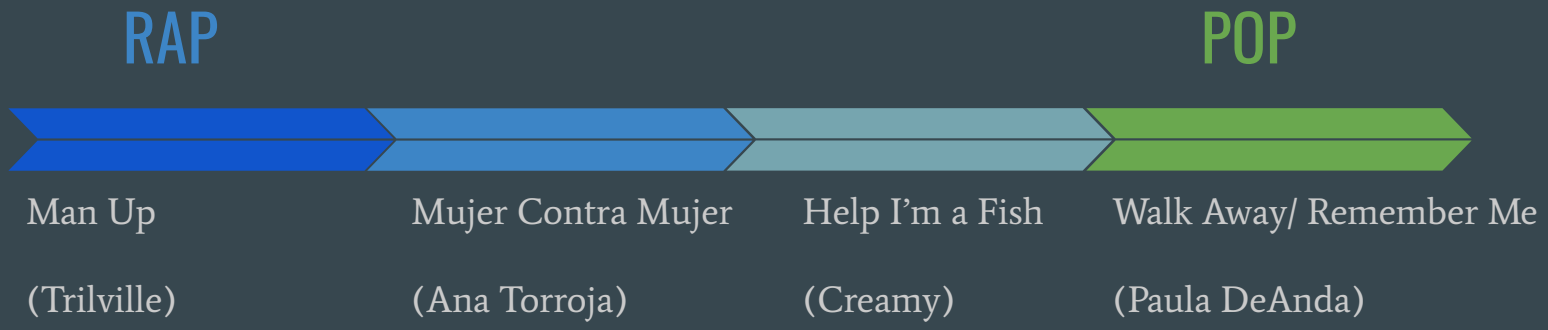
Most/Least Confident Classifications - Linear SVM



Decision Function Value:

10.7479 0.5992 -0.001 -8.2353

Most/Least Confident Classifications - Logistic Regression



Probability According to Model:

Interesting Results

Walk Away / Remember Me (Paula DeAnda)

I'm gonna remember you
You're gonna remember me

I saw you with your new girl just yesterday
And I feel that I must confess
Even though it kills me to have to say
I'll admit that I was impressed

Physically just short of perfection
Gotta commend you on your selection
Though I know I shouldn't be concerned In the back of my mind I can't help but question

Does she rub your feet
When you've had a long day
And scratch your scalp
When you take out your braids?
Does she know you like to play PS2
Till six in the morning like I do

Interesting Results

I'm a little yellow fish

I'm a little yellow fish in the deep blue sea,
Won't somebody help me!
I'm a little yellow fish in the deep blue sea,
Won't somebody save me!

Do you wanna know how living is beneath the waves?
Do you wanna know how everything I knew was changed?

It wasn't such a big commotion,
I just had a drop of magic potion.
And with a wriggle... a twist... a splash... and a splish...
I was a fish!
Help me!

Help I’m a Fish (Creamy)

Conclusion

Things We Learned

- Pop and rap are distinguishable from each other!
- Profanity is key indicator of rap music (in this dataset)
- For RF, genres can be determined primarily by lyrics

Future Work

- Use all of the data for grid search
- Try other genre pairs
- Multiclass modelling
- Use audio instead of audio characteristics
- Dimensionality reduction for faster training

Acknowledgements

- Prof Wu!
- Grutoring team

Summary

Rap-Pop Classifiers

	F1-score
Dummy Classifier	0.472
Random Forest	**0.943**
Logistic Regression	0.893
Linear SVM	0.872
RBF SVM	0.927

Feature Importance