The Million Song Dataset

Making predictions based on artist and audio information

Sam Lundberg

Problem Statement

- Using audio & artist features of a song, can a song be predicted to become popular or not?
- Using the audio features, can a songs genre be predicted?
  - Adding in additional artist information, does that improve the results?

Background

- Million Song Dataset developed for machine learning
- Each song contains audio measurements
  - Tempo, duration, bpm, acoustical measurements, key signature, mode, etc.
- Each song also contains artist information
  - Artist name, location, year, genre tagging, title, algorithmic based fields such as artist familiarity

EDA

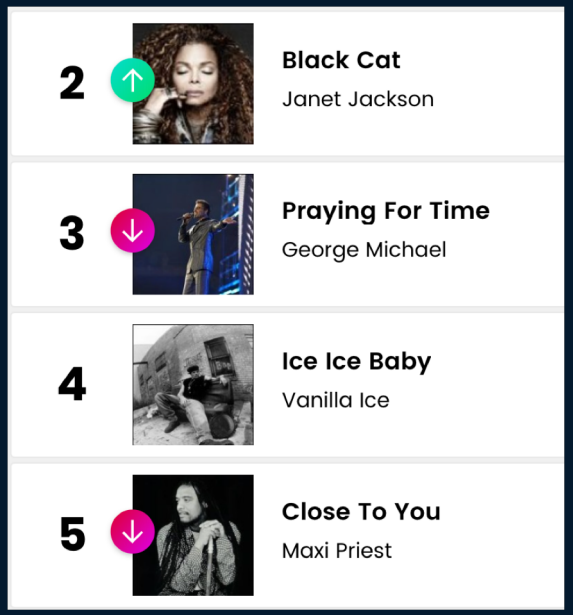
- 10k sample used
- Missing data
  - Year
  - Artist hotness
  - Location
- Audio features in arrays
  - Corresponding column of confidence interval
  - Created a median value of fields x C.I.

EDA

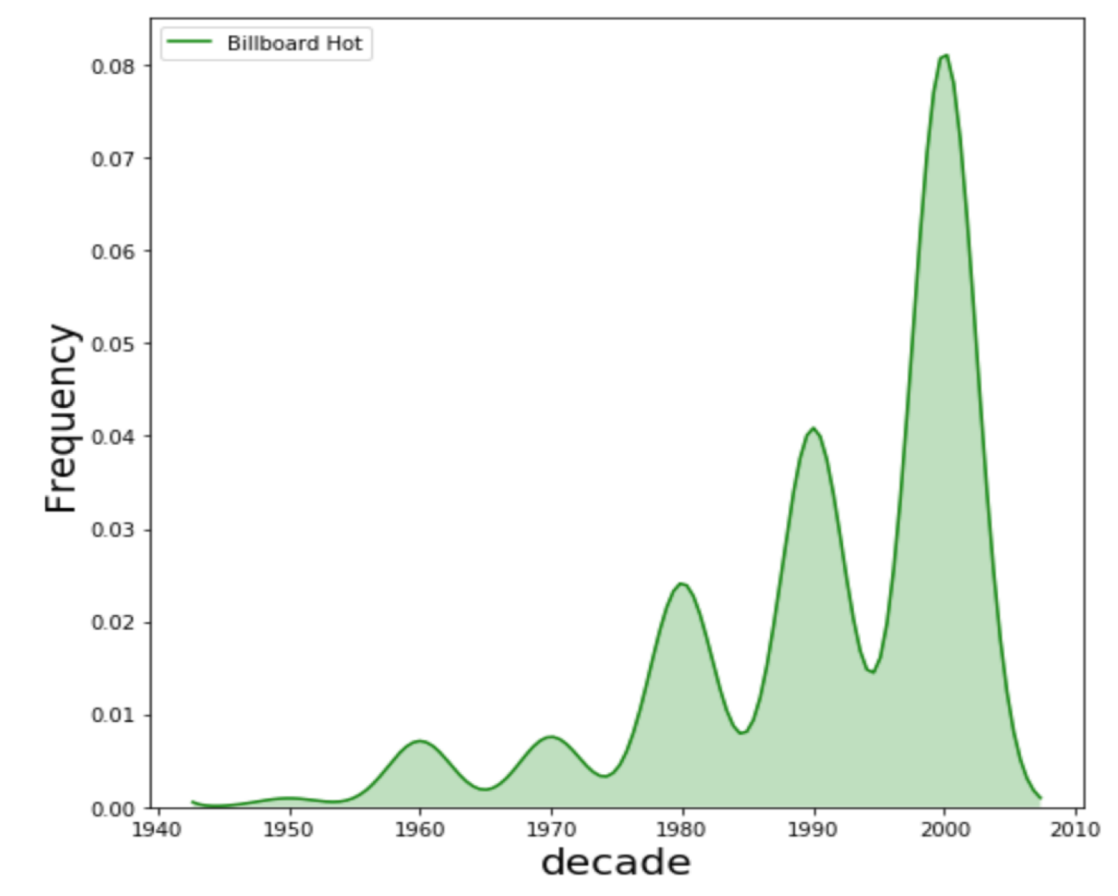
- Genre tagging
  - Multiple tags per artist in array
  - C.I. and Frequency of genre tagging
  - 3500+ unique genres
  - Found the highest valued genre and that became the genre for each song
  - Narrowed it down to 457 genres
  - Problematic in genre identification

EDA

- Older songs missing a lot of info
- Scraped Billboard Hot 100 charts from 1958-2011
- Matched Hot songs from BB to MSD
- MSD 1922-2011
- Sample size cut over 50%



EDA


Billboard Hot
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
1940
1950
1960
1970
1980
1990
2000
2010
decade

Predicting a Hot Song

- Created target variable of BB_Hot
- Used audio and artist information as predictors
- Label-encoded fields such as artist name and genre

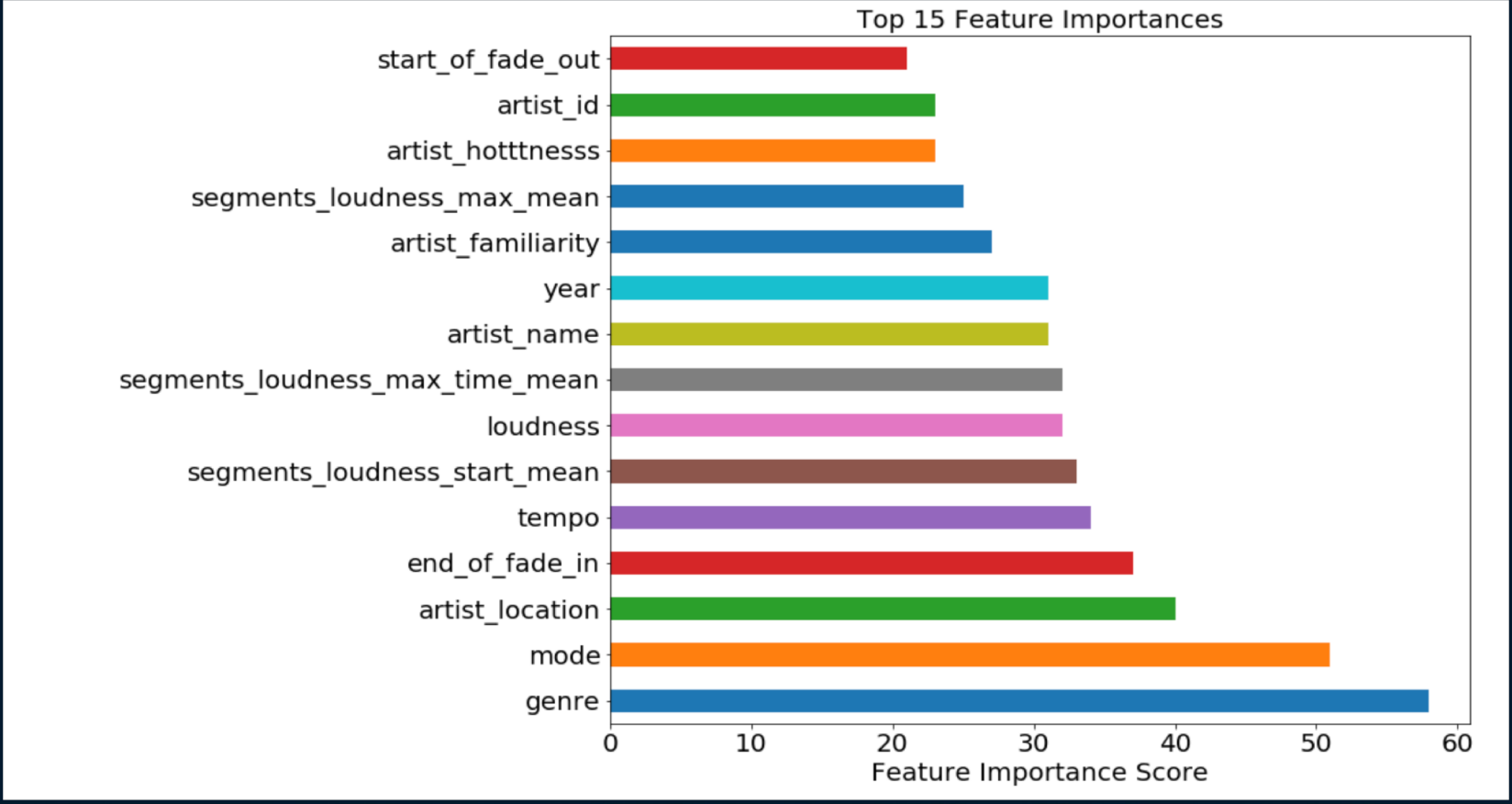
Predicting a Hot Song

- Unfamiliar audio fields:

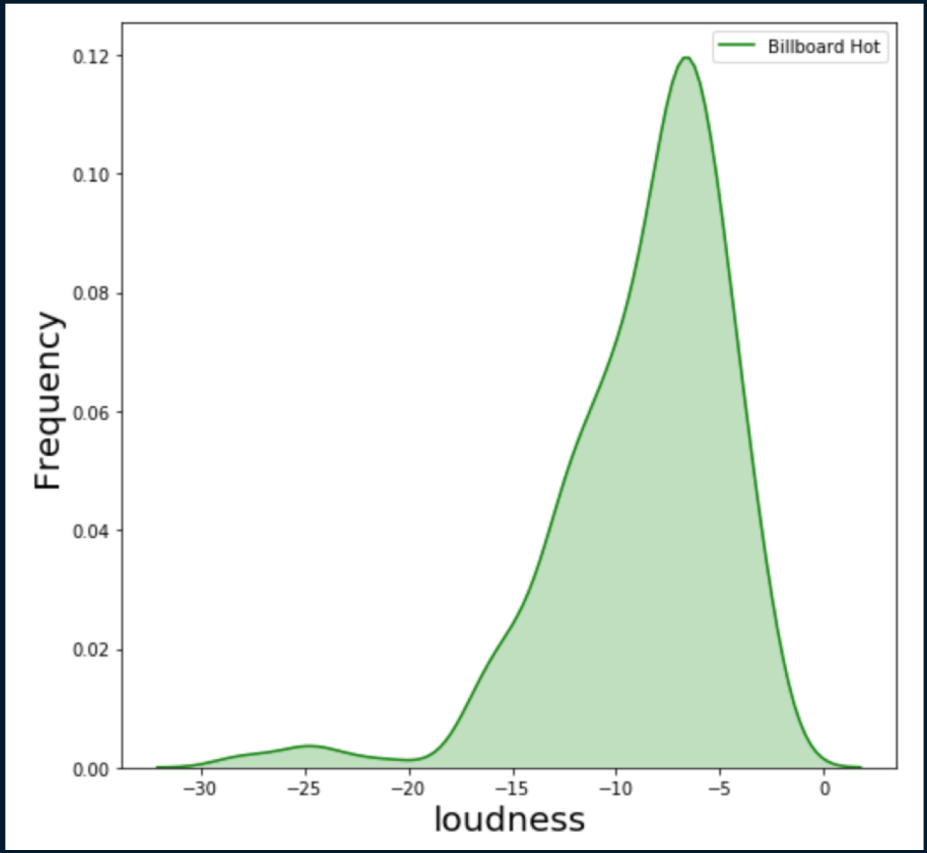
sections start	largest grouping in a song, e.g. verse
segments loudness max	max dB value
segments loudness max time	time of max dB value, i.e. end of attack
segments loudness max start	dB value at onset
segments pitches	chroma feature, one value per note
segments start	musical events, ~ note onsets
segments timbre	texture features (MFCC+PCA-like)
start of fade out	time in sec
tatums start	smallest rhythmic element
mode	major or minor

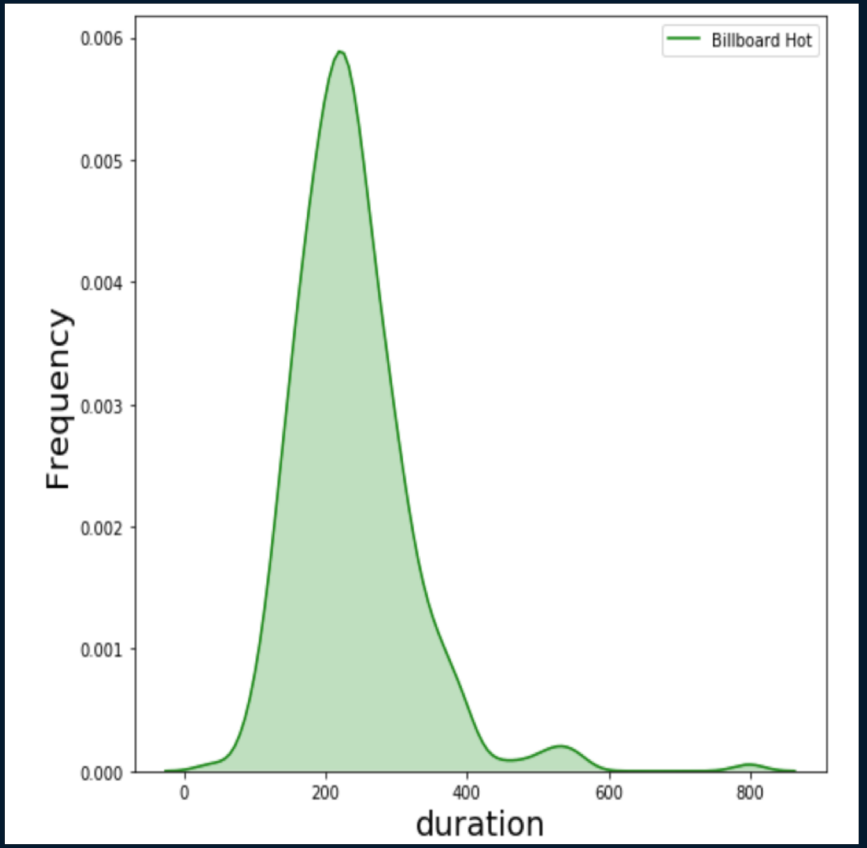
Predicting a Hot Song

- What features are important?

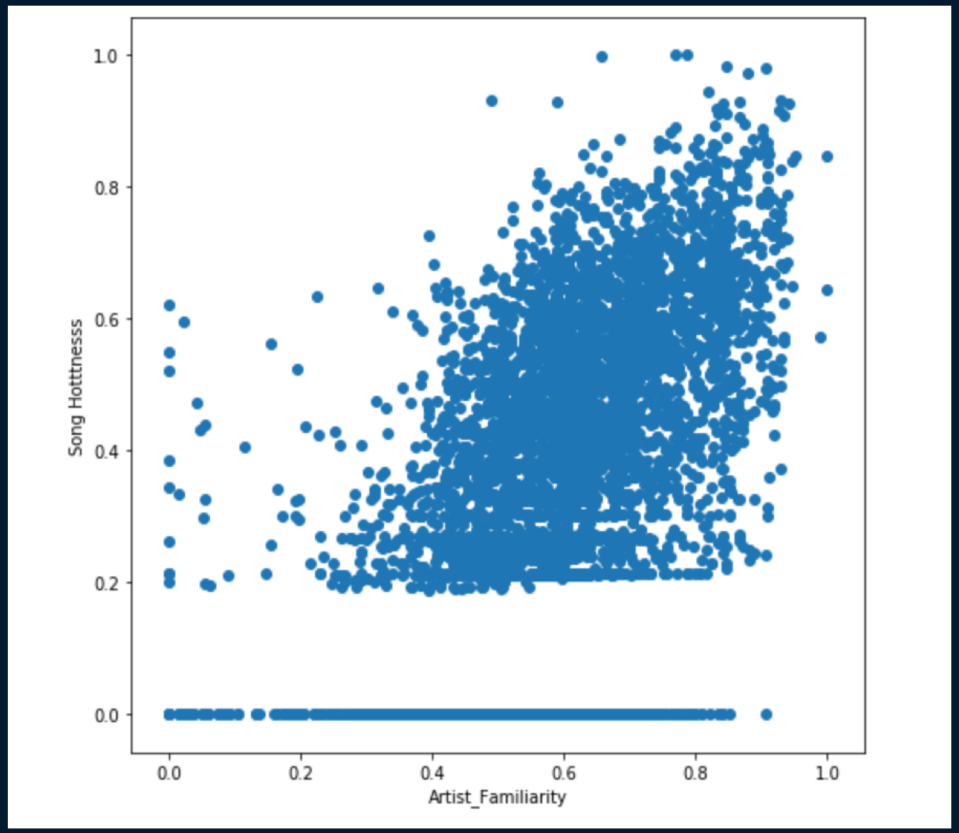


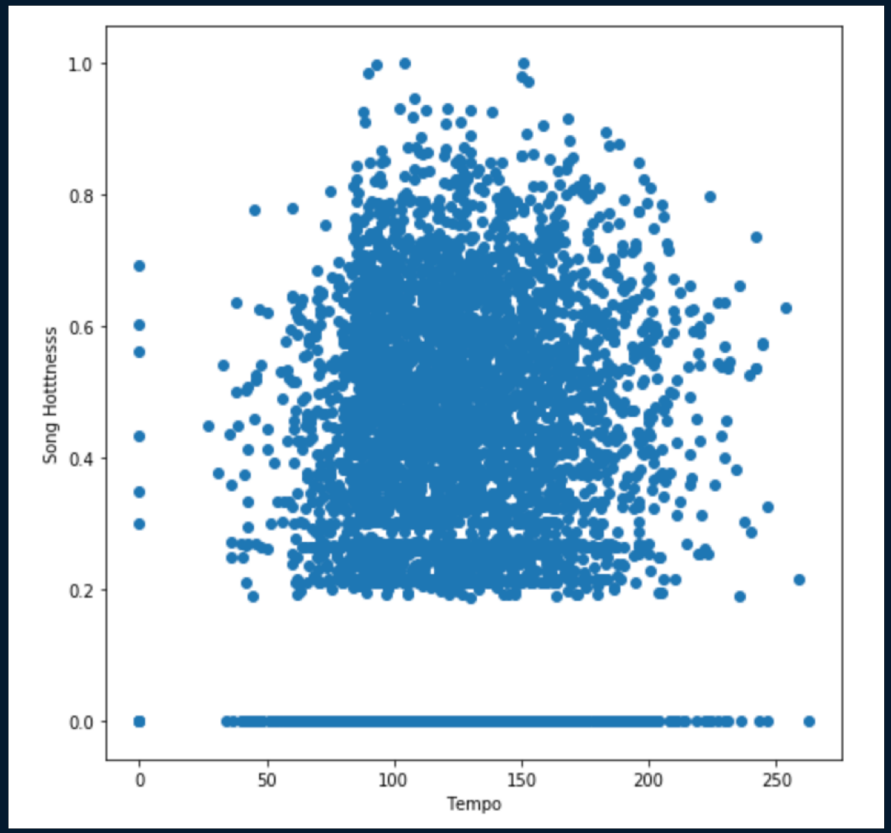
Predicting a Hot Song



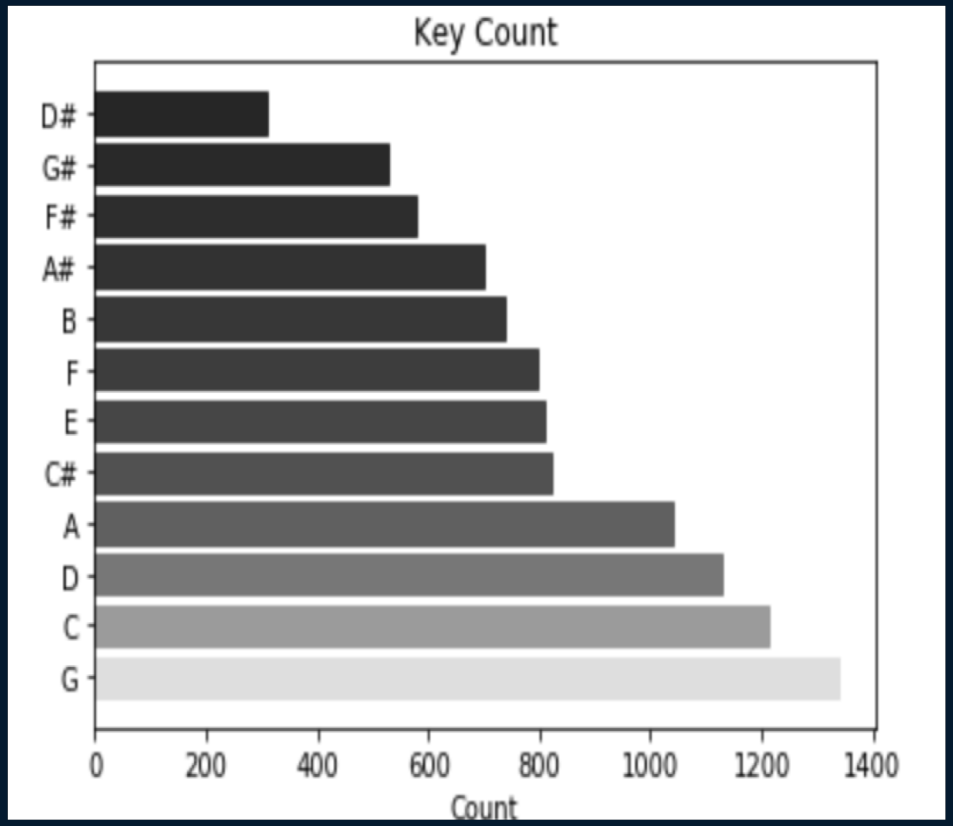


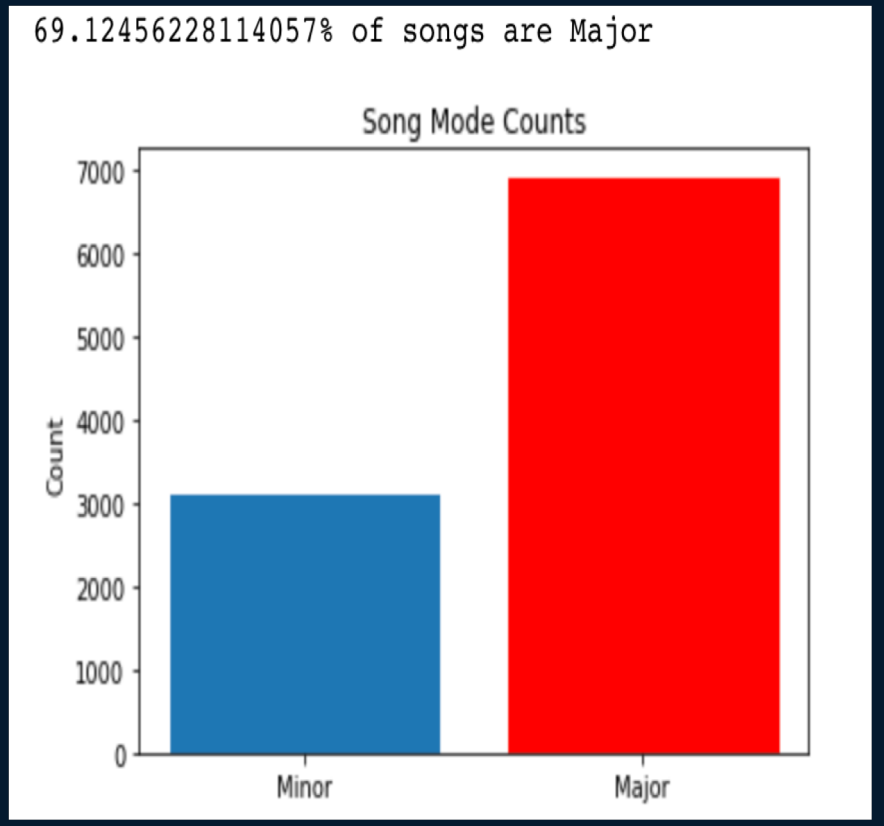
Predicting a Hot Song





Predicting a Hot Song





Predicting a Hot Song

Artists that show up as a Billboard Top 100 artist have averages of:

- Tempo 126 BPM
- Duration 240.6 seconds
- Key averages key of F, but this is a deceptive stat, with Std. Dev puts them more in line with D or A
- Mode .69
- Loudness -9.1 decibels
- Artist Familiarity .65 AND
- Artist Hotttnesss .45 shows there are a lot of one hit wonders, or artists that don't have a lot of hits

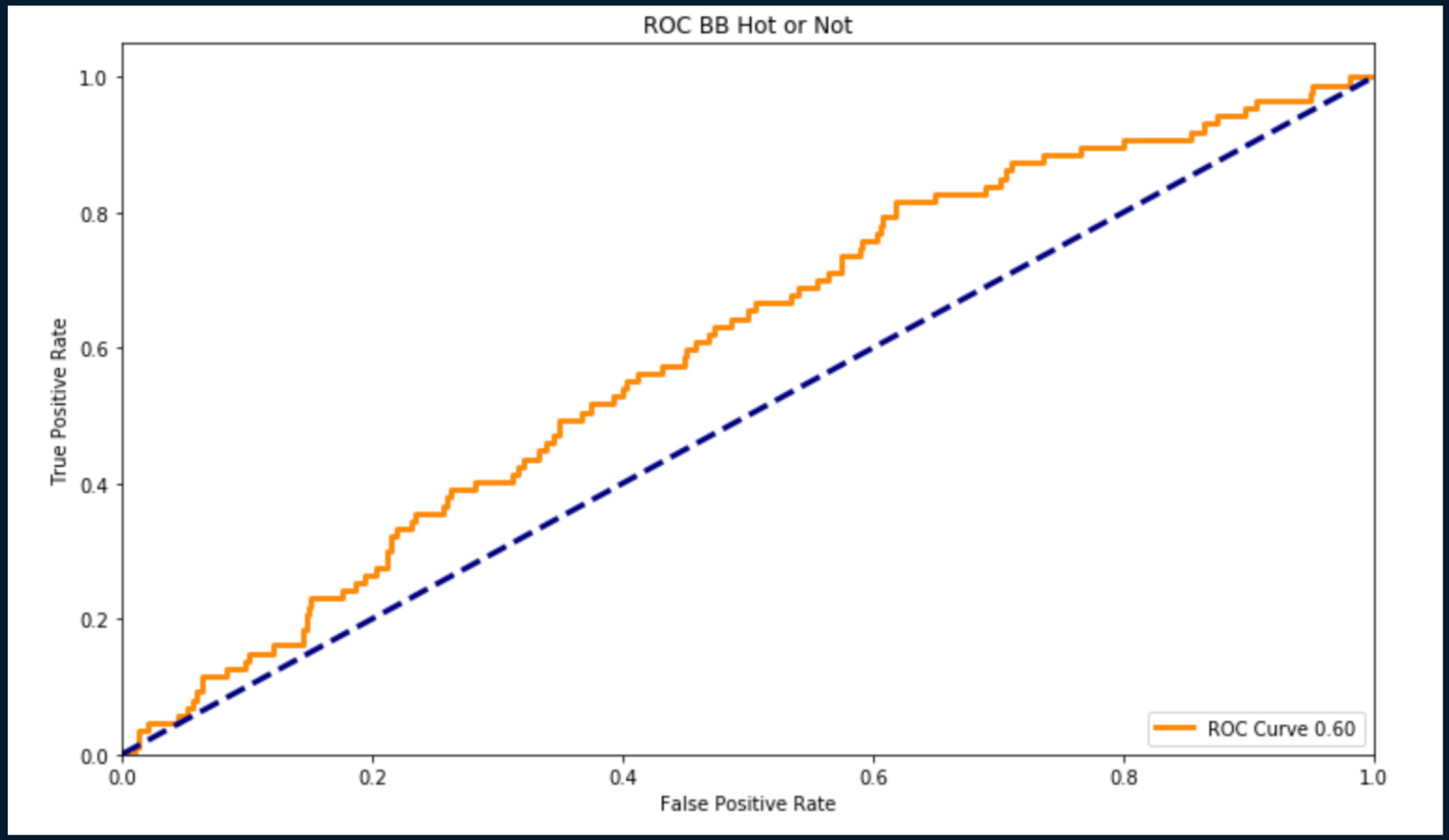
Predicting a Hot Song

Predicting a Hot Song-Modeling

- Modeling revealed consistent, but poor results:

Model	Score
Random Forest	0.552278
XGB	0.547926
Support Vector Machines	0.546122
Logistic Regression	0.543926
KNN	0.542076

Predicting a Hot Song-Modeling



Predicting a Hot Song-Modeling

- Optimize for Sensitivity. Want to predict a hot song more first and foremost. Boost True Positives
- After tuning:
- [317 272]
  [34 53]
  - Sensitivity = .61
  - Accuracy = .54

Predicting a Hot Song-Conclusions

- Very difficult to predict if a song will be hot or not
- Certain characteristics are apparent
- Music is structured
- Better to predict if a song is NOT hot?

Predicting a Songs Genre

- Next question
  - Can a songs genre be predicted?
- Audio features only
- Add in artist based features

Predicting a Song's Genre-Audio

- Compared 2 artists to that of their genre:

Feature	Aerosmith	Blues Rock Genre
Fade Out	302.12	251.35
Tempo	121.13	121.35
Bars	40.03	24.64
Beats	57.19	52.44
Key	5.58	5.30
Mode	0.75	0.23
Tatums	31.93	36.49
Loudness	-7.24	-9.71
Duration	312.31	260.70

Feature	Sugar Minott	Roots Reggae Genre
Fade Out	214.10	237.71
Tempo	155.97	136.38
Bars	18.33	20.13
Beats	60.37	68.66
Key	4.91	5.43
Mode	0.75	0.58
Tatums	60.23	66.42
Loudness	-9.75	-10.16
Duration	222.28	247.42

Predicting a Song's Genre-Audio

- 457 unique genre tags
- Not very effective in predicting a genre just on audio features on full genre list

Model	Score
Logistic Regression	0.410868
XGB	0.402253
Random Forest	0.351889
ADA	0.341286

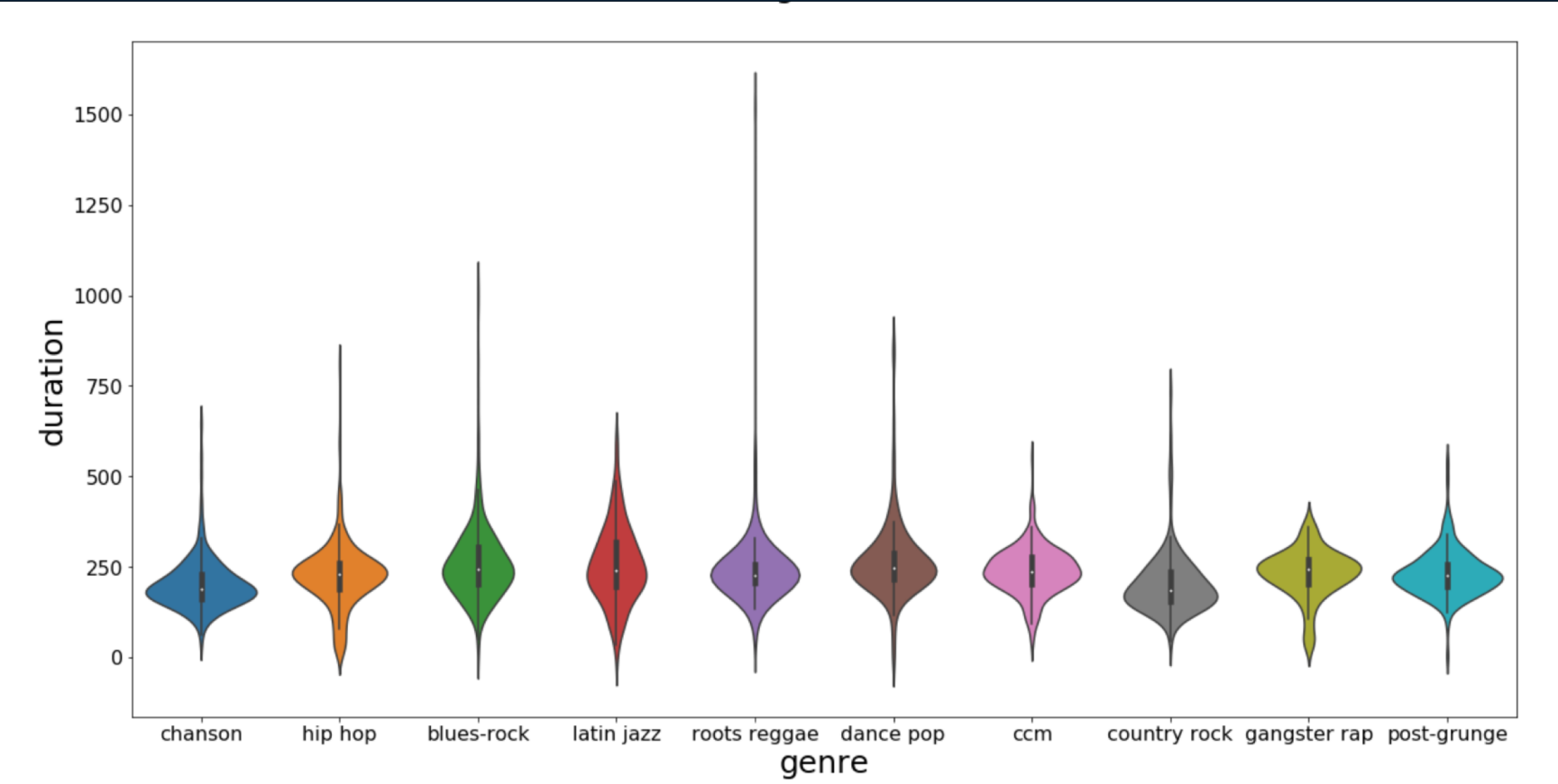
Predicting a Song's Genre-Audio

- 457 unique genre tags too many
- Isolated the top 10 genres:

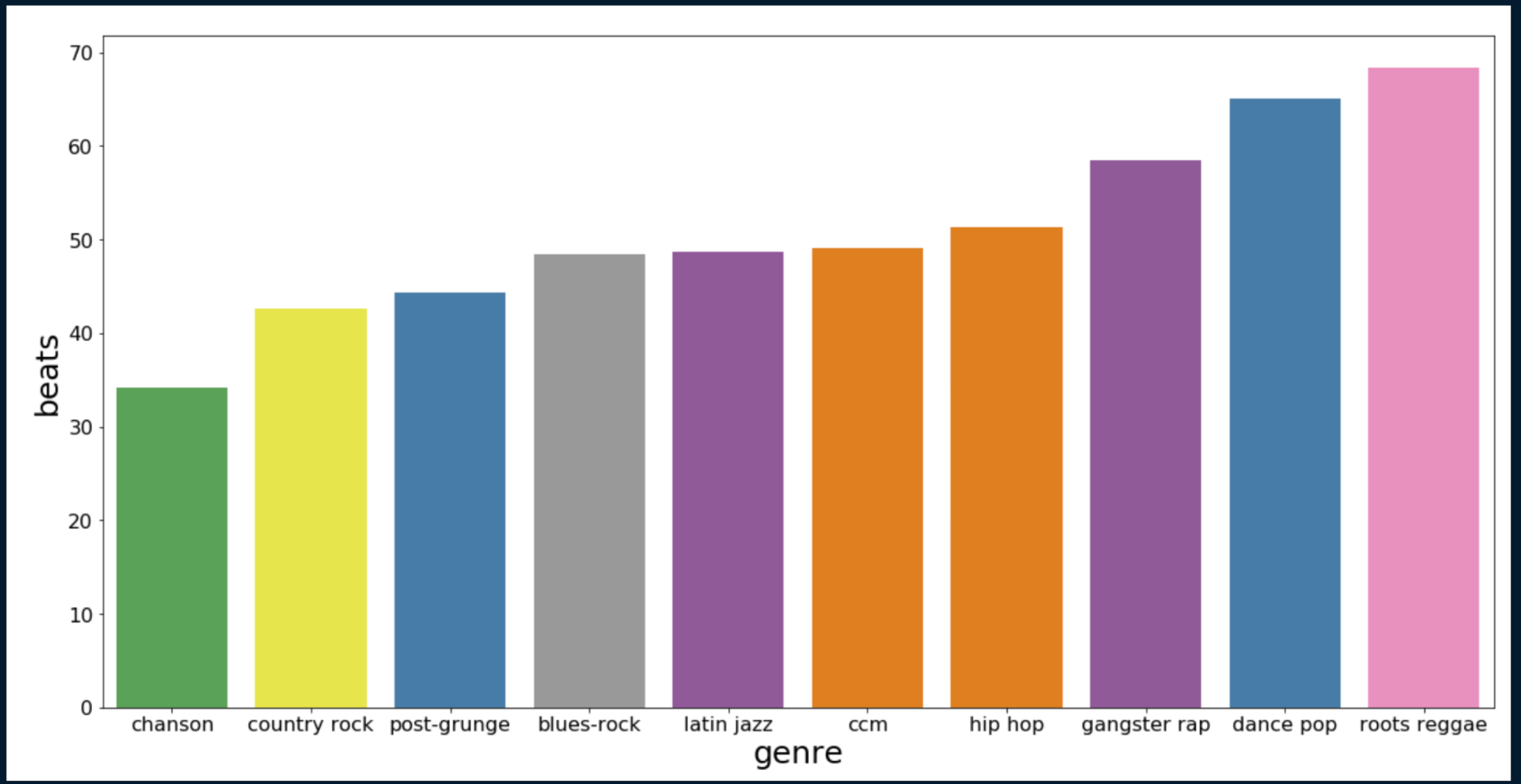
blues-rock	346
hip hop	346
ccm	255
chanson	208
country rock	156
latin jazz	150
post-grunge	146
dance pop	141
gangster rap	134
roots reggae	131

- Unfortunately shrunk the dataset to around 2000 records

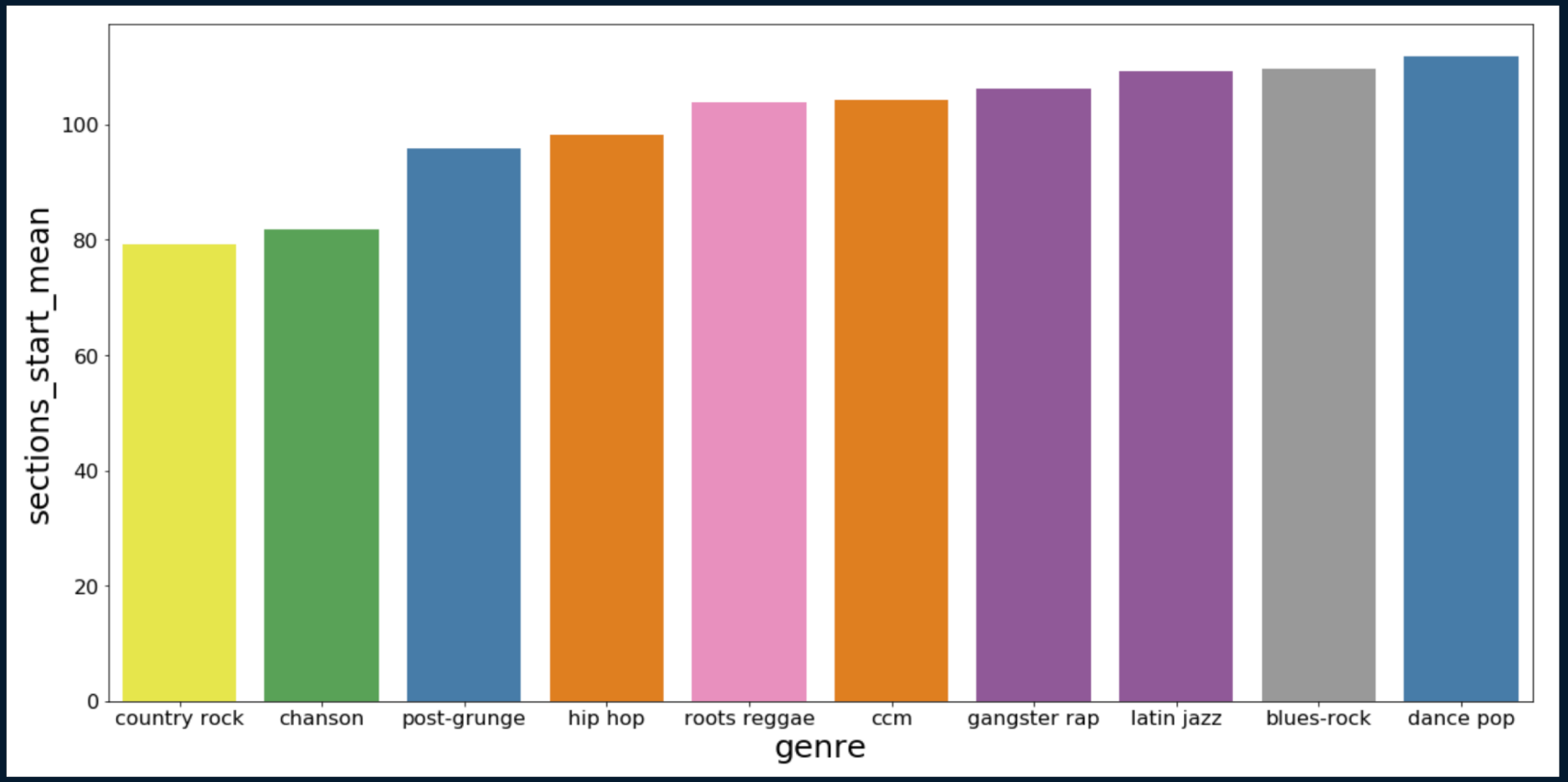
Predicting a Song's Genre-Audio



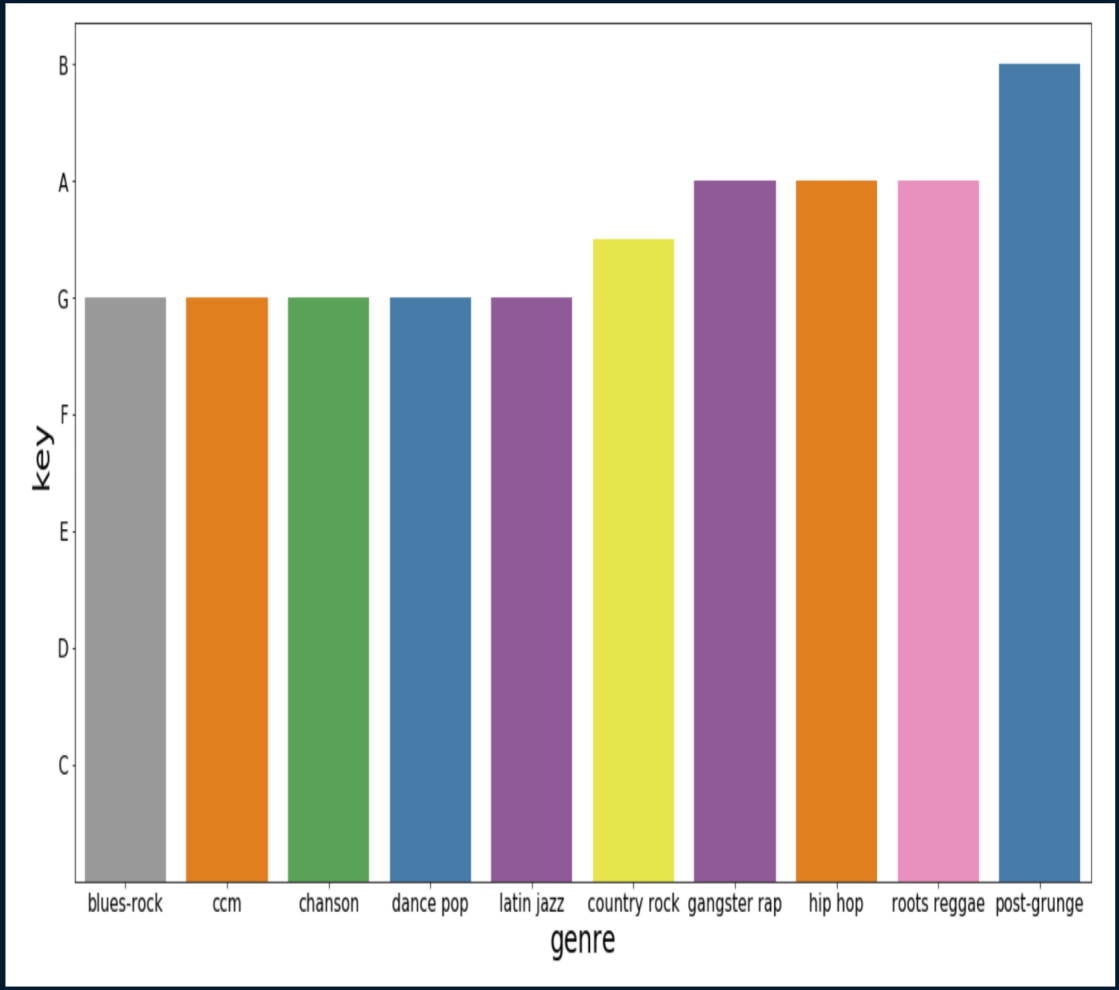
Predicting a Song's Genre-Audio

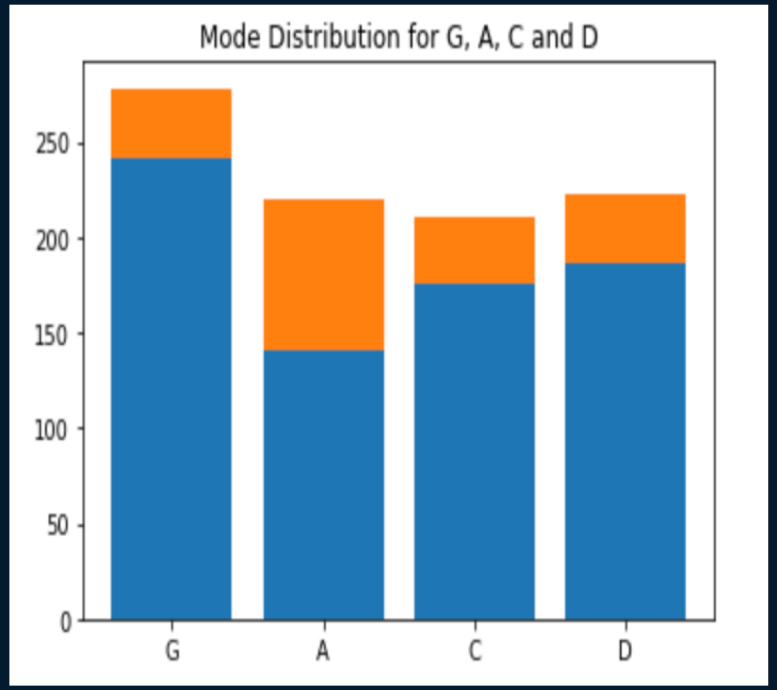


Predicting a Song's Genre-Audio



Predicting a Song's Genre-Audio





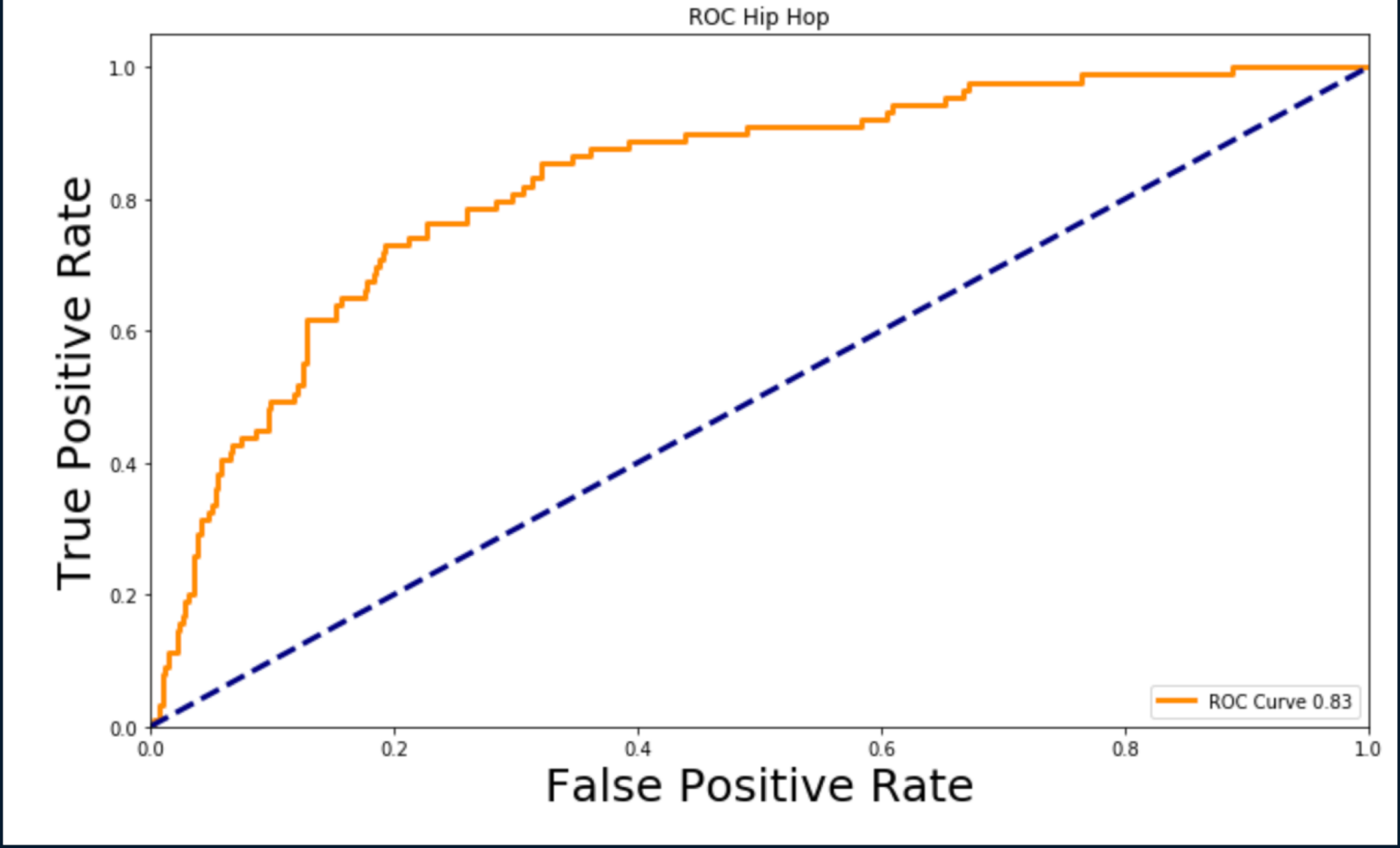
Predicting a Song's Genre-Audio

- Started with the genre "Hip Hop"
- 1 of 10 genres, should predict better
- Modeling:

Model	Score
XGB	0.835932
ADA	0.826259
Logistic Regression	0.821364
Random Forest	0.813674

	XGB Imp
segments_loudness_start_mean	0.127527
segments_pitches_mean	0.115086
segments_loudness_max_time_mean	0.113530
loudness	0.099533
tempo	0.082426
tatums	0.054432
segments_start_mean	0.052877
mode	0.049767
key	0.048212
segments_timbre_mean	0.048212

Predicting a Song's Genre-Audio



Predicting a Song's Genre-Audio

- Use Sensitivity to evaluate model
- Optimized created scores of:
  - Accuracy = .73
  - Sensitivity = .78
  - [301 114]
    [19 70]
  - Still predicting many True Negatives and False Positives

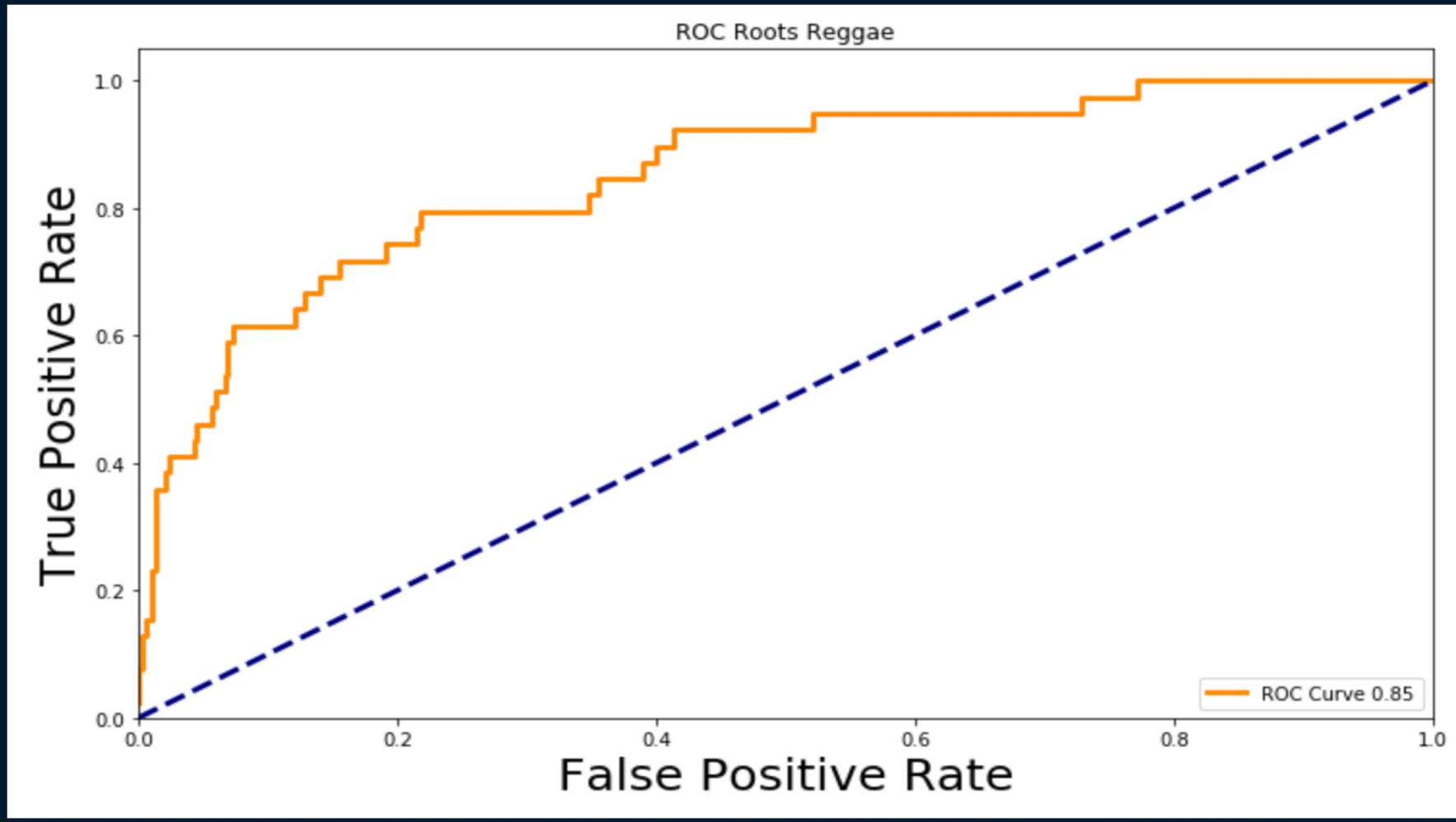
Predicting a Song's Genre-Audio

- Predicted again on lowest of 10 genres "Roots Reggae"
- Modeling:

Model	Score
XGB	0.887895
Random Forest	0.844468
Logistic Regression	0.839470
ADA	0.821948

	XGB Imp
tempo	0.162295
tatums	0.144262
segments_loudness_start_mean	0.108197
mode	0.090164
segments_pitches_mean	0.088525
bars	0.070492
segments_loudness_max_time_mean	0.052459
segments_timbre_mean	0.049180
sections_start_mean	0.037705
segments_loudness_max_mean	0.031148

Predicting a Song's Genre-Audio



Predicting a Song's Genre-Audio

- Use Sensitivity to evaluate model
- Optimized created scores of:
  - Accuracy = .89
  - Sensitivity = .61
  - [425 40]
    [15 24]
  - Models are very bad at predicting just songs in Roots Reggae compared to Hip Hop

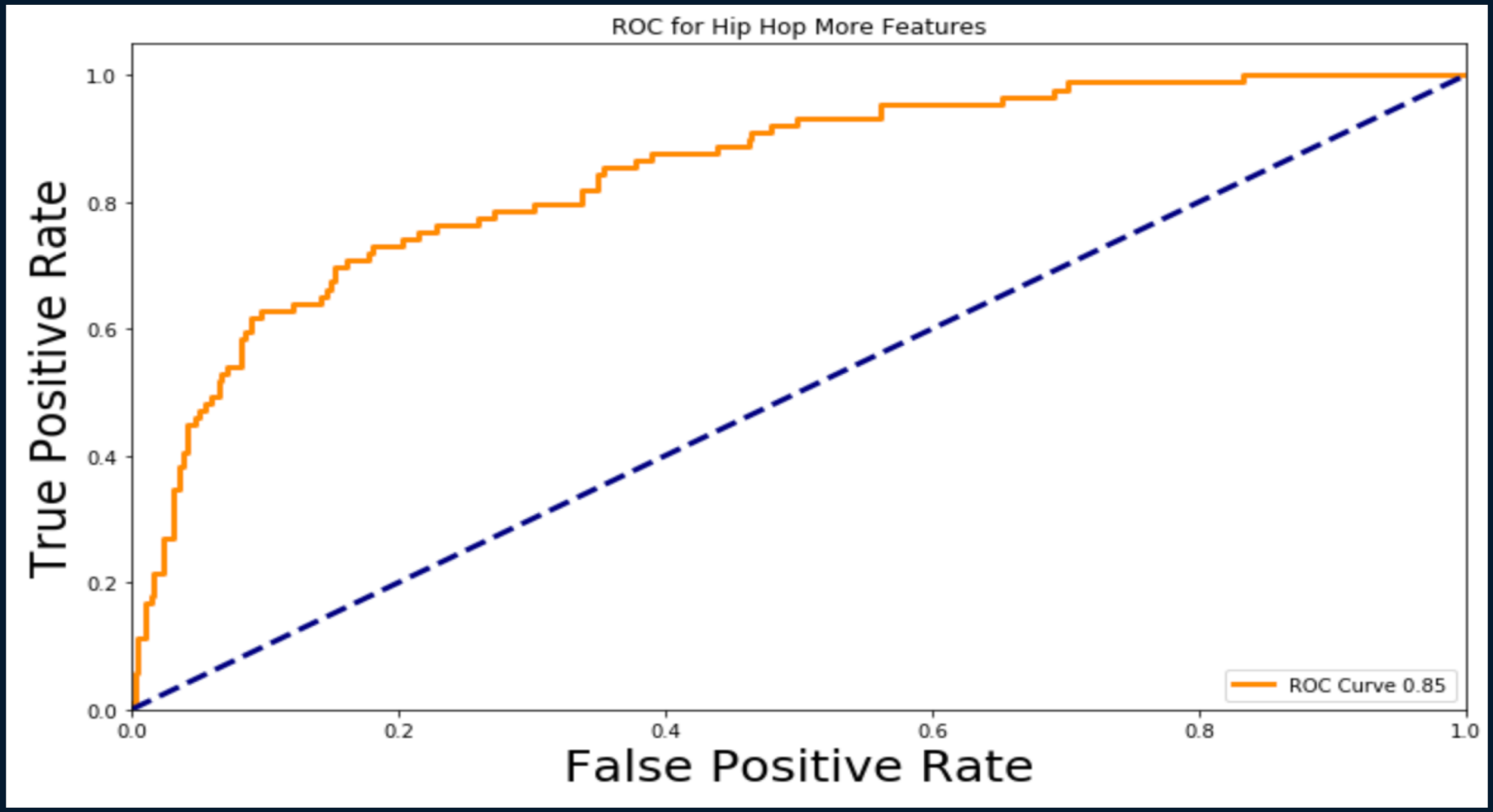
Predicting a Song's Genre-All

- Final attempt taking audio + artist features
- Used Hip Hop genre
- Model:

Model	Score
XGB-Audio +	0.865076
XGB-Audio	0.835932

	XGB Imp
artist.name_new	0.128593
segments_loudness_start_mean	0.110439
location_new	0.098336
segments_pitches_mean	0.093797
segments_loudness_max_time_mean	0.078669
loudness	0.069592
tempo	0.069592
segments_timbre_mean	0.042360
key	0.040847
tatums	0.034796

Predicting a Song's Genre-All



Predicting a Song's Genre-All

- Use Sensitivity to evaluate model
- Optimized created scores of:
  - Accuracy = .76
  - Sensitivity = .76
  - [316 99]
    [21 68]
  - Model improved with adding more than just audio features, but still grabbing many songs not labeled Hip Hop

Conclusions

- Song Hotness
  - Audio and artist features prove less than useful to predict if a song is hot or not
  - Certain consistent audio measurements are found in hot songs
  - Other factors must be involved in determining this; record labels, money, sheer luck

Conclusions

- Genre Identification
  - Audio features are helpful, but too many genres ”bleed” into other genres, which makes it hard to draw a black and white line (think blues vs. rock)
  - Additional artist information improves the predictions
  - Too many songs get misclassified when applying sensitivity metrics

Next Steps

- Sample size greatly reduced because of missing data
  - Use AWS for full million songs
- Isolate just the 90+ audio measurements by themselves
  - Use neural networks
- Find higher level of genre “buckets”
- Fill in missing data through other means

Next Steps

- Identify if social or political trends have any influence on a songs popularity
- Examine lyrics, both for popular songs, and to classify genres
- Use Spotify's playlists to do a similar analysis

Citation

https://labrosa.ee.columbia.edu/millionsong/

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere.

The Million Song Dataset. In Proceedings of the 12th International Society

for Music Information Retrieval Conference (ISMIR 2011), 2011.