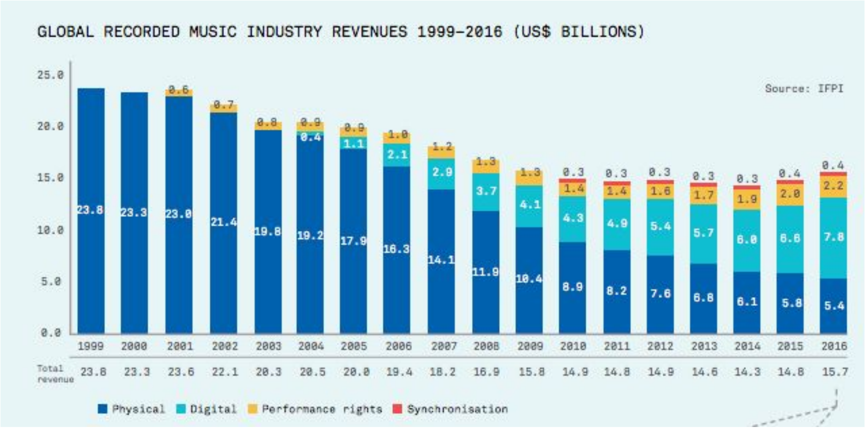
Exploration of Music Popularity with Data Science

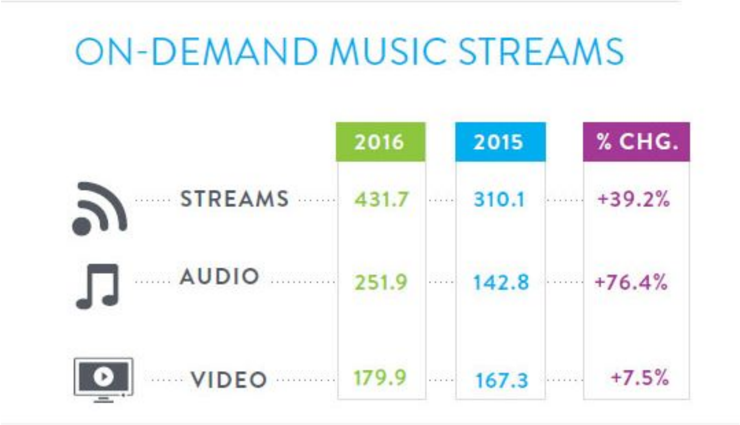
Author: Matthew Oakes
https://github.com/moakes010/ML-PopMusic

Motivation

Music Industry Revenue Boosted By Audio Streaming Services



The Decline of Music Industry Revenue



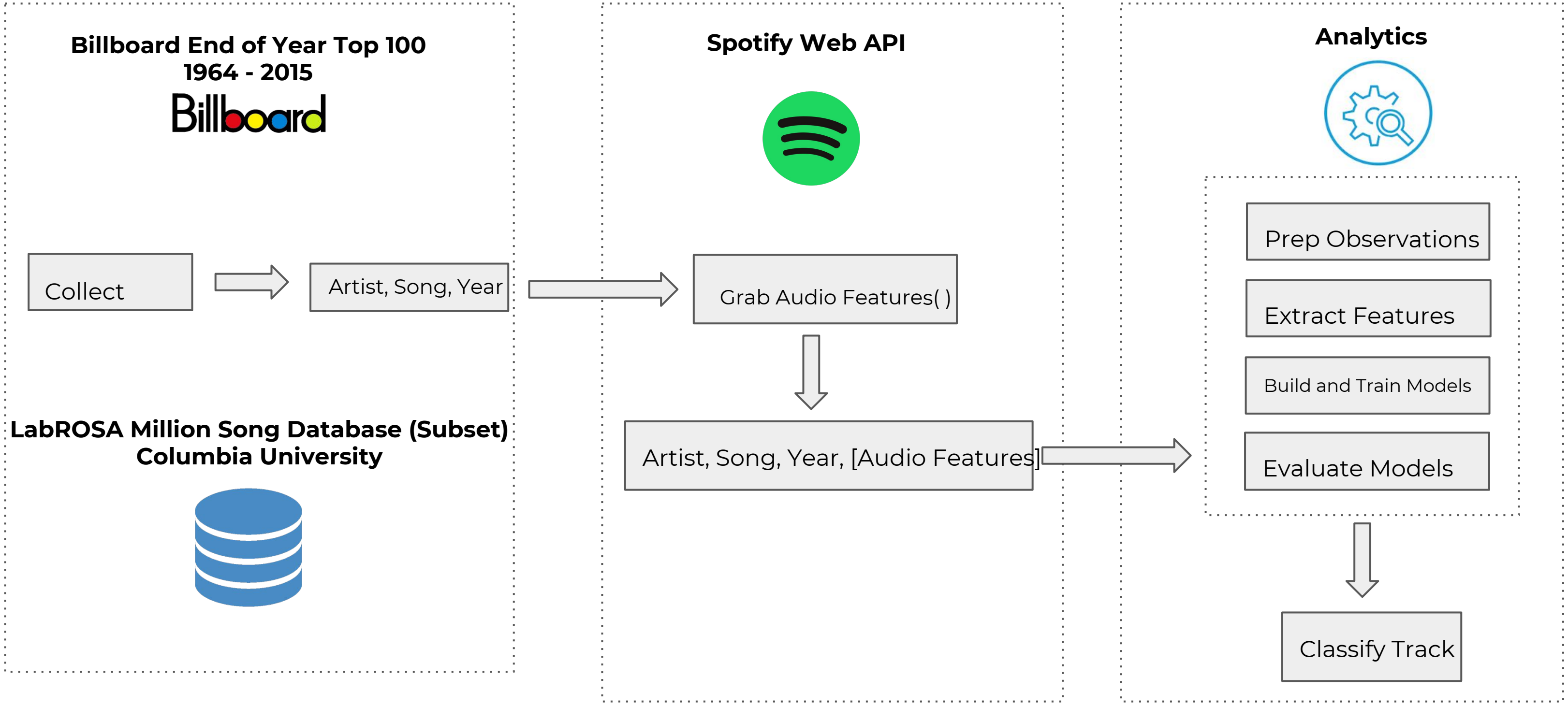
On-Demand Audio to the Rescue

The Question

Due to the change in the way we consume music and the flood of user data as a result of switching to a subscription model, is it possible to build a model to predict commercial success of a track?

Choosing the Data to Build Models

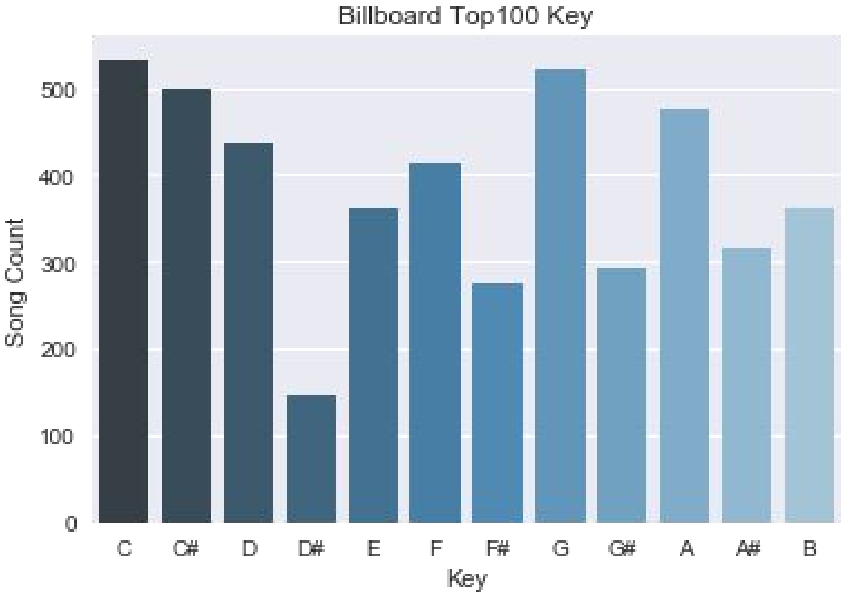
Gather, Localize , Analyze Data, Generate and Apply Model

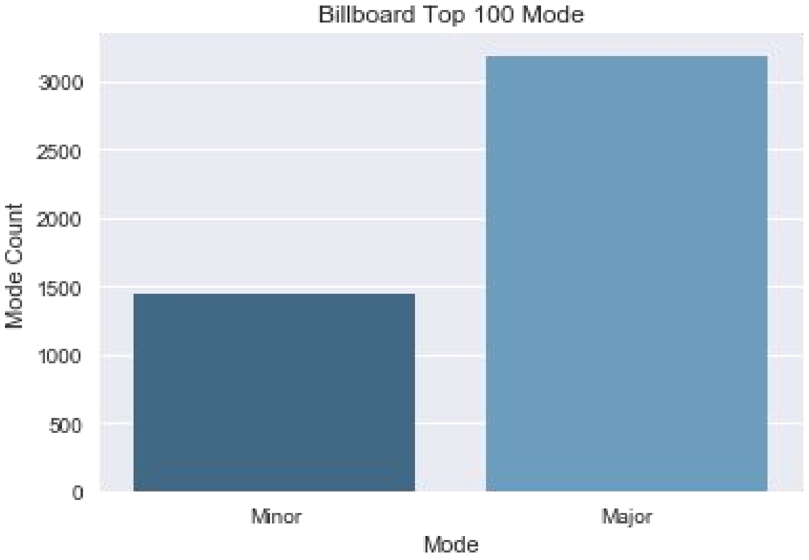


Key Value	Type	Description
acousticness	float	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic
danceability	float	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0
duration_ms	int	The duration of the track in milliseconds.
energy	float	Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate and general entropy.
instrumentalness	float	Predicts whether a track contains no vocals. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal". The closer the instrumentalness value is to 1.0, the greater likelihood the track contains no vocal content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
key	int	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.
liveness	float	Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 0.8 provides strong likelihood that the track is live.
loudness	float	The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 db.
mode	int	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
speechiness	float	Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
tempo	float	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.
valence	float	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
popularity	integer	The popularity of a track is a value between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Duplicate tracks (e.g. the same track from a single and an album) are rated independently. Artist and album popularity is derived mathematically from track popularity. Note that the popularity value may lag actual popularity by a few days: the value is not updated in real time.

Explore Dataset

Lots of interesting Visualizations! Example of Key and Modality.



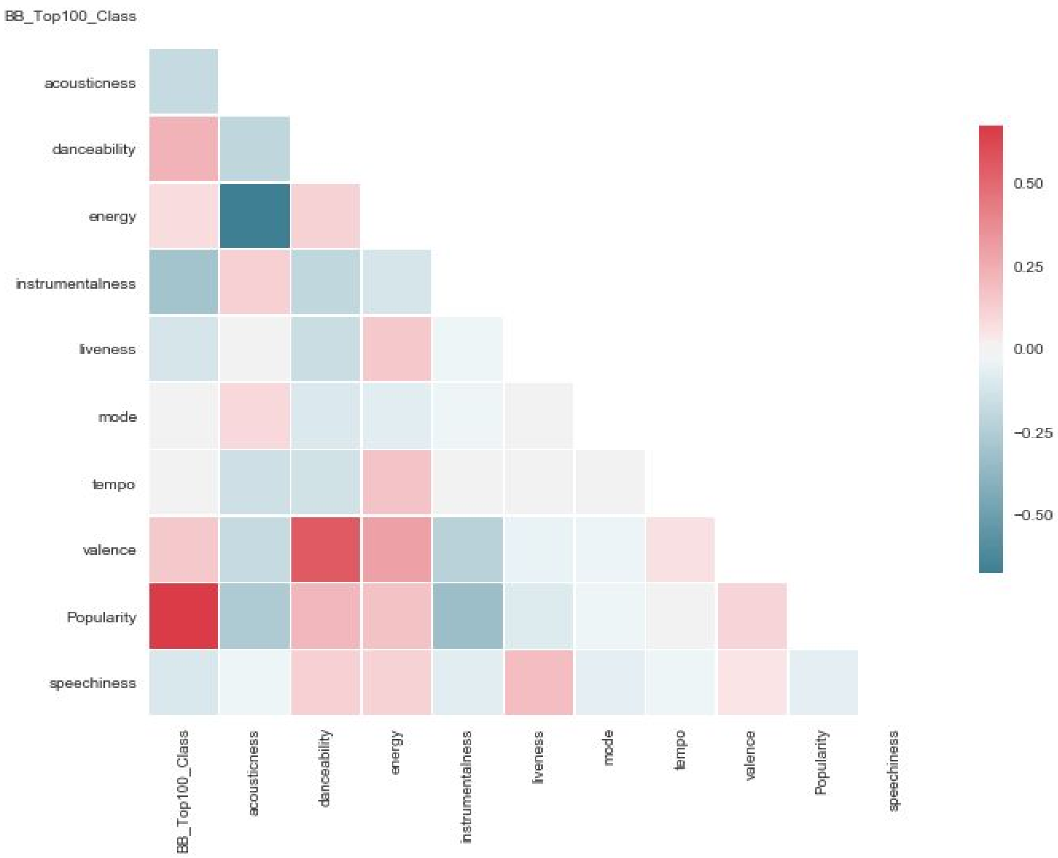


Explore Dataset

Correlations b/t classifier* and selected features:

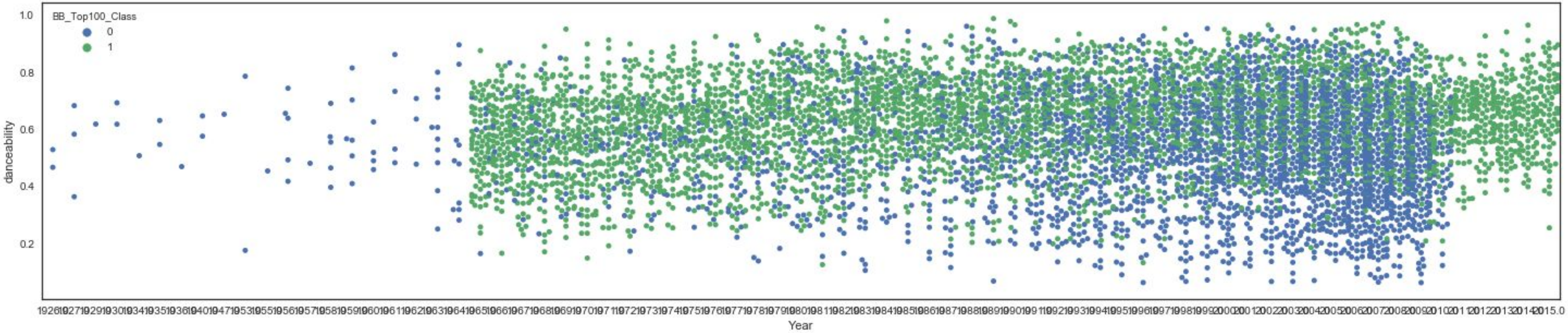
Energy
Danceability
Valence
Popularity

*Classifier Track in Top100 list 1, Not in List 0

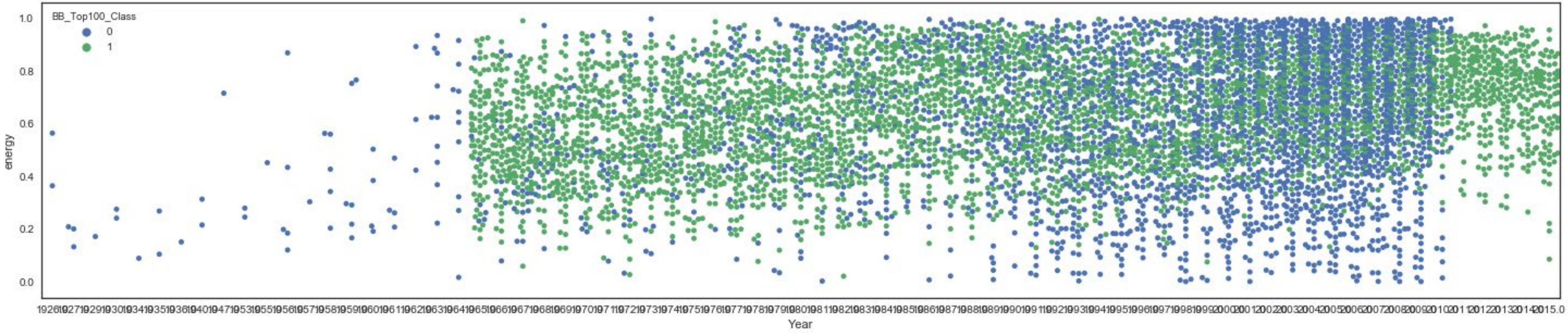


Explore Dataset

Danceability

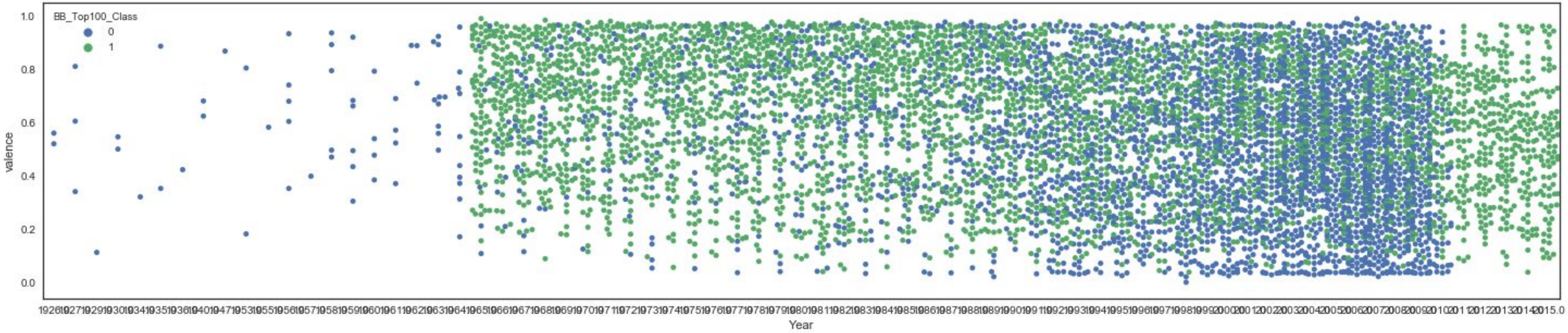


Energy

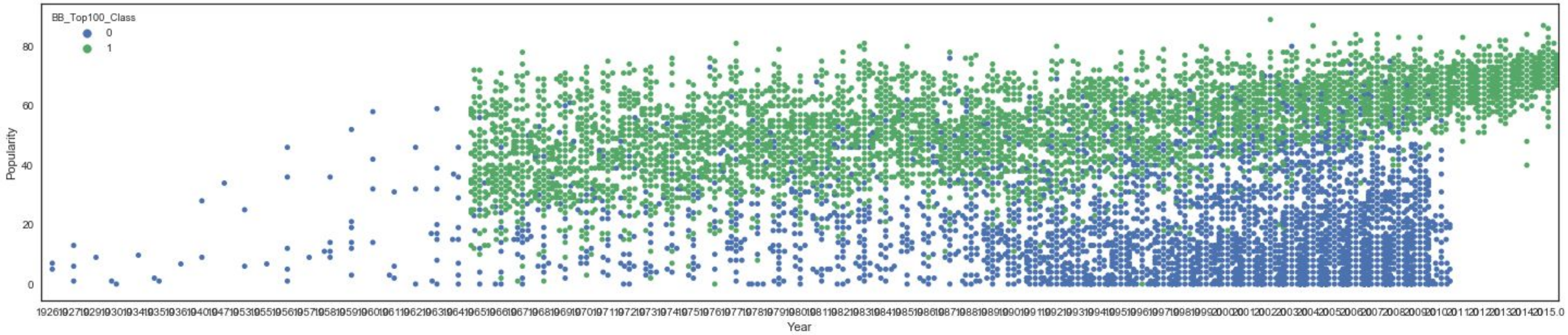


Explore Dataset

Valence



Popularity



Model Evaluation

Selected Features

acousticness
danceability
energy
instrumentalness
liveness
mode
tempo
valence
Popularity
speechiness

Observations

Total : 10192

Top 100 Tracks : 4631

Response

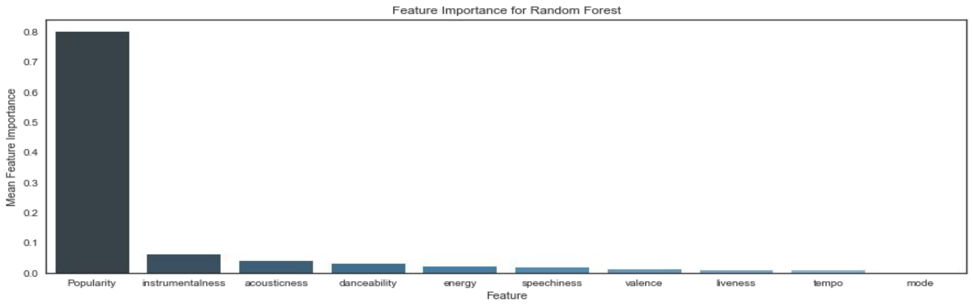
0 - Not Top 100

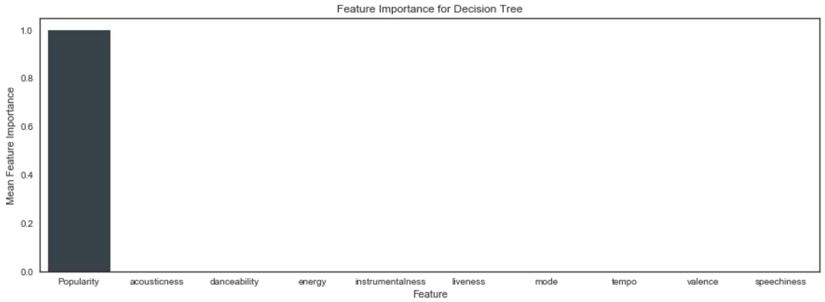
1 - Top 100

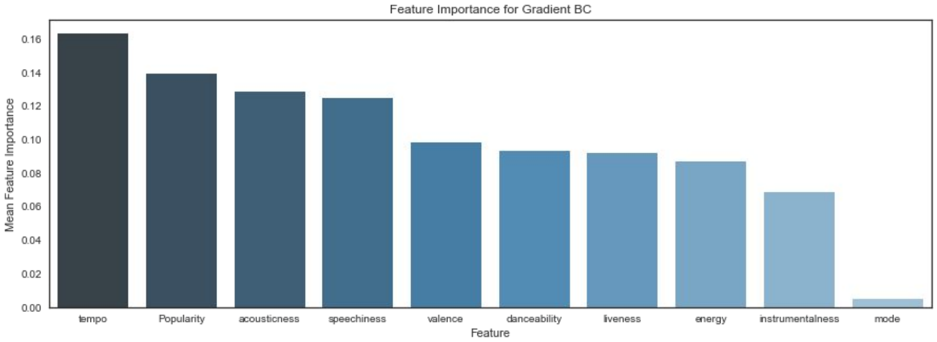
Models

Decision Tree
Random Forest
Logistic Regression
Gradient Boosting Classifier

Model Evaluation

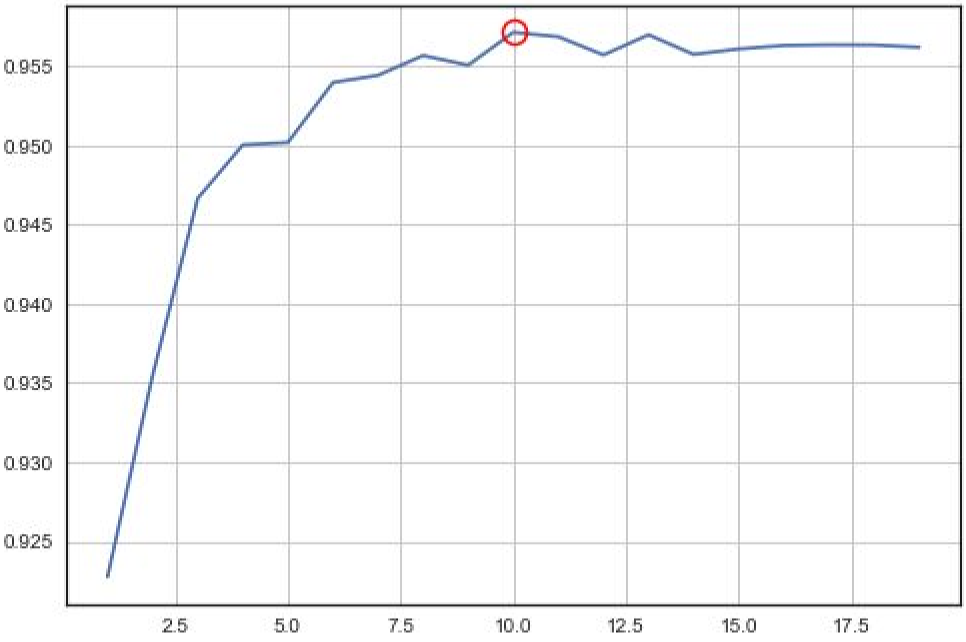




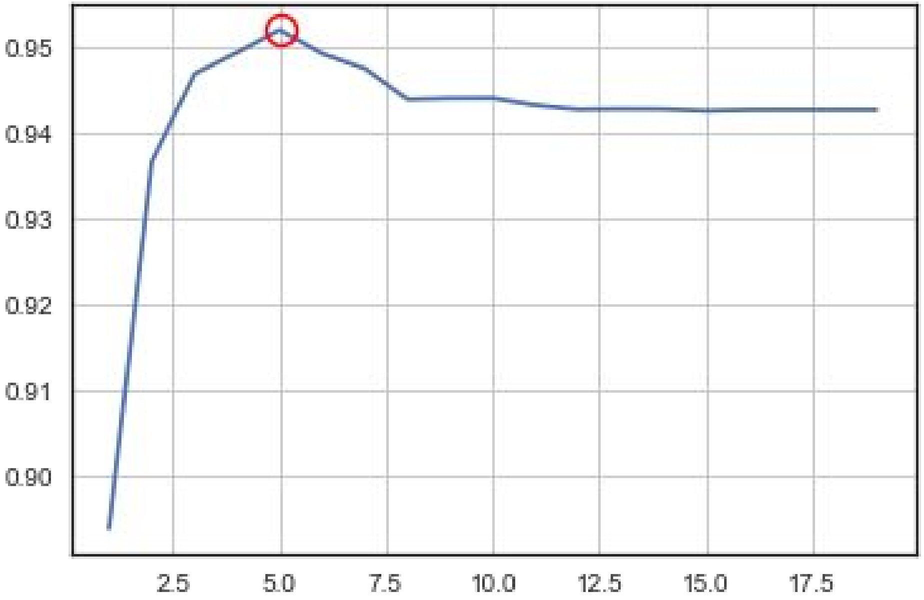


Model Evaluation

Decision Tree Best Parameters



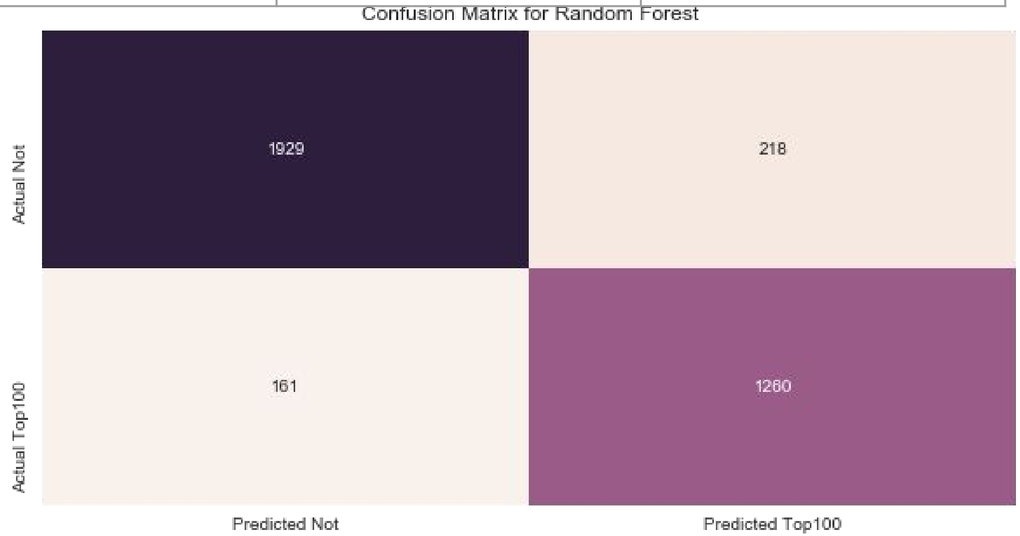
Random Forest Best Parameters



Model Evaluation

Model	AUC	Accuracy	Precision	Recall
Random Forest	0.931374875569	0.893744407989 25359	0.85250338295	0.923997185081

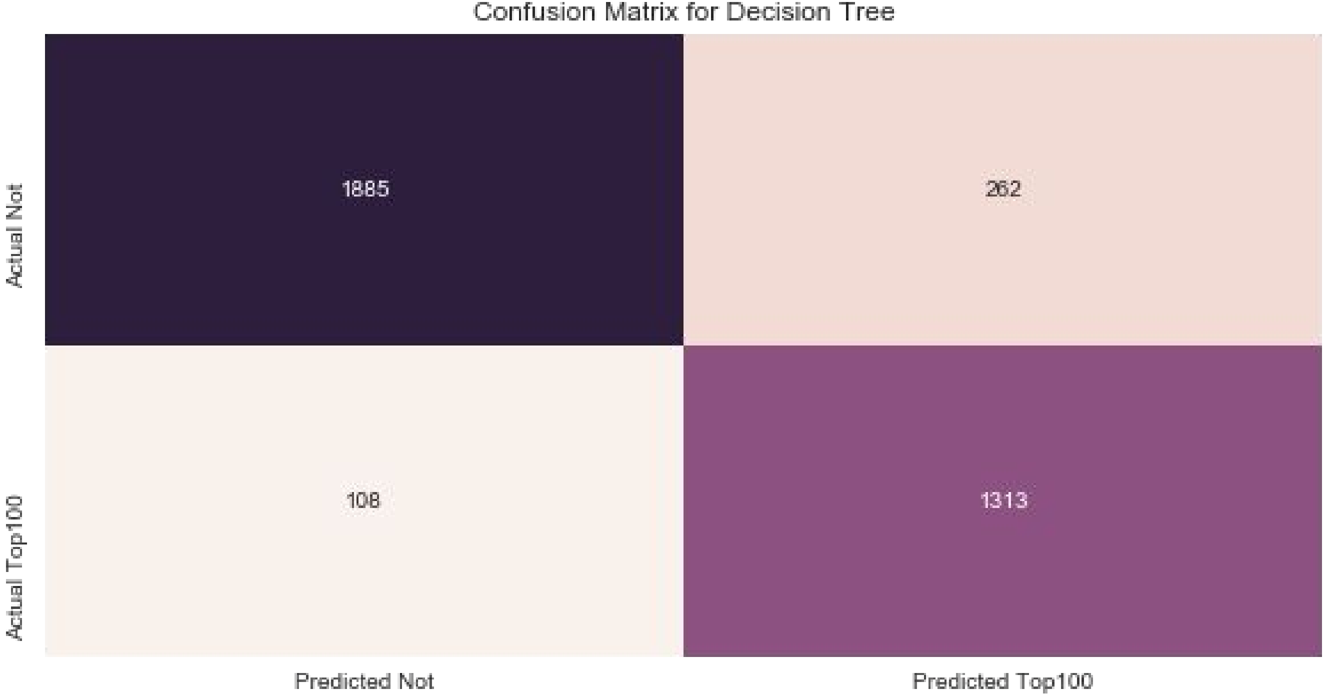
n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	



Accuracy: (TP + TN)/TOTAL
True Positive Rate (Recall): When it's actually yes, how often does it predict yes?
TP/(FN+TP)
Precision: When it predicts yes, how often is it correct?
TP/(FP+TP)

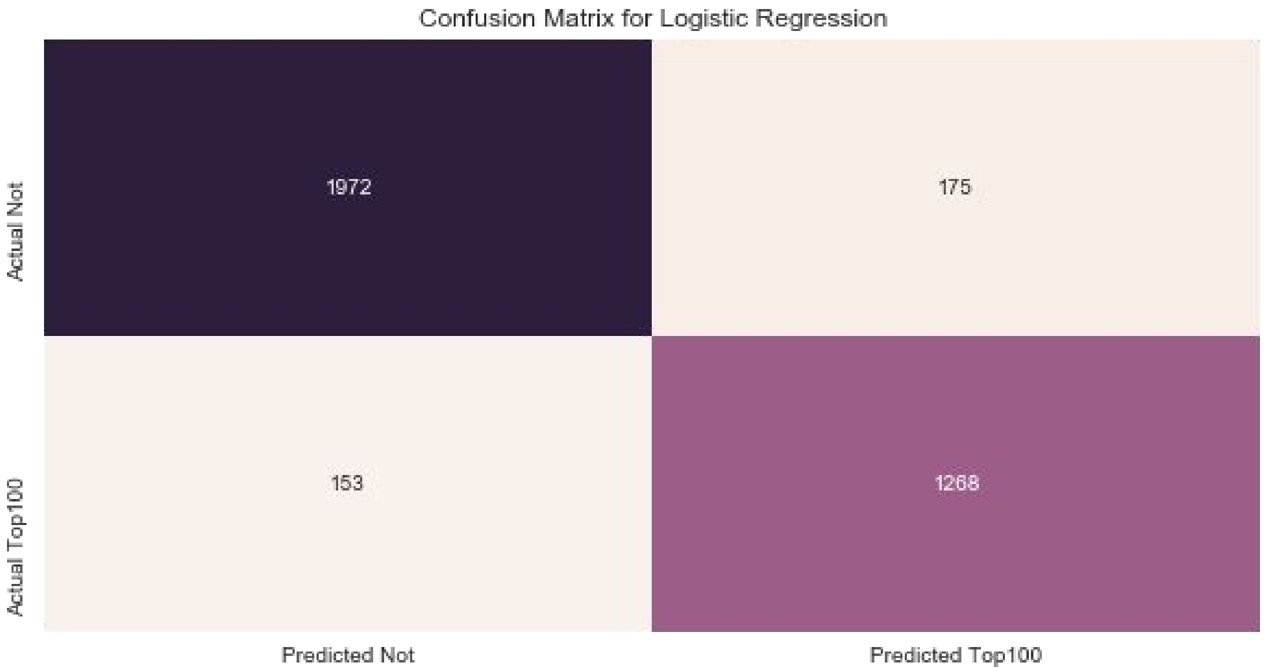
Model Evaluation

Model	AUC	Accuracy	Precision	Recall
Decision Tree	0.941631892627	0.90051589683218458	0.833650793651	0.886699507389



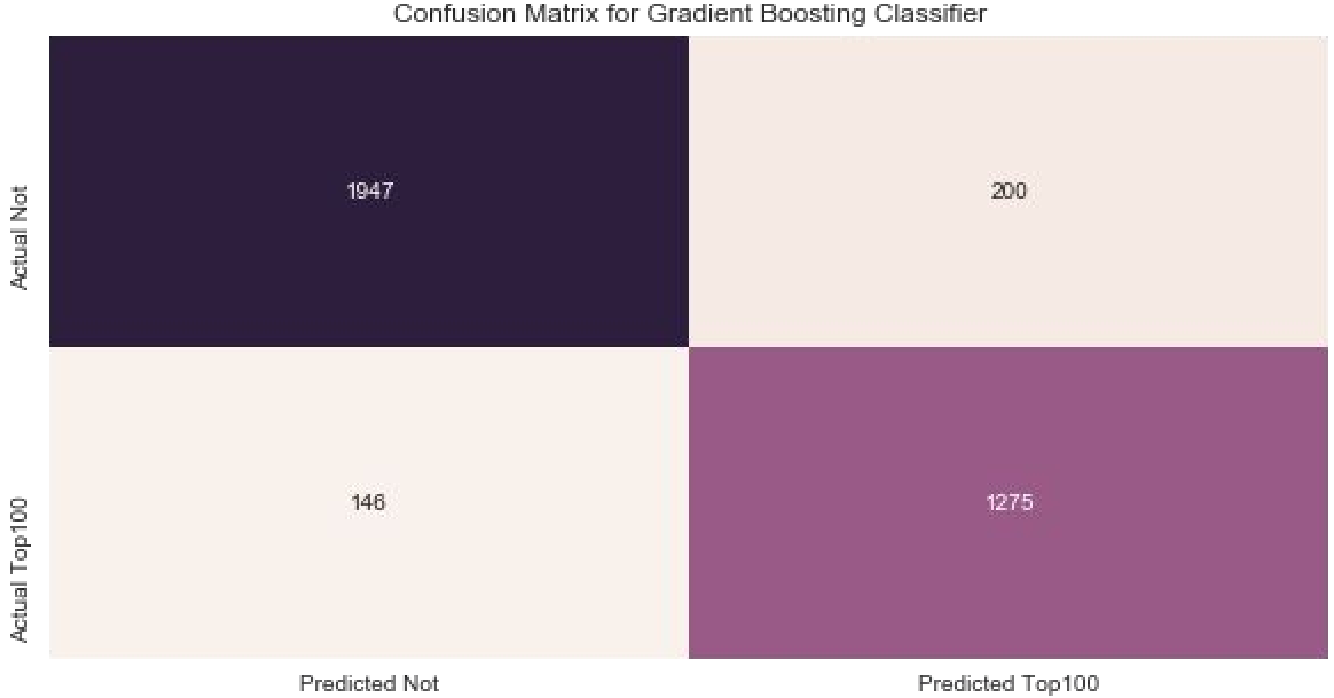
Model Evaluation

Model	**AUC**	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.964765165016	0.89570961924513	0.878724878725	0.892329345531



Model Evaluation

Model	AUC	Accuracy	Precision	Recall
Gradient Boosting Classifier	0.959321010578	0.88707288348259006	0.864406779661	0.897255453906



Model Evaluation

Results Summary

Model	AUC	Accuracy	Precision	Recall
Decision Tree	0.941631892627	0.90051589683218458	0.833650793651	0.886699507389
Random Forest	0.931374875569	0.89374440798925359	0.85250338295	0.923997185081
Gradient Boosting Classifier	0.959321010578	0.88707288348259006	0.864406779661	0.897255453906
Logistic Regression	0.964765165016	0.89570961924513	0.878724878725	0.892329345531

2016 End of Year Top 100

Initial Dataset missing did not have 2016 Billboard Data. Scrap the data from:

http://www.billboard.com/charts/year-end/2016/hot-100-songs

Spotify API returned 95 of 100 entries.

Model	**Accuarcy**
Decision Tree	0.947368421053
Random Forest	0.947368421053
Gradient Boosting Classifier	0.978947368421
Logistic Regression	1.0

Closing Thoughts

Based on the features provided by Spotify, one can predict whether a song could end up in Billboard's End of Year Top 100 list; however, I believe there are other factors that would need to be accounted for like genre, etc.

Spotify audio feature results are limited to artists that have can release music on Spotify and have been processed by the Echo Nest software. Websites like Soundcloud and Purevolume where user's upload audio would never be discovered.

Would like to include lyrics in future iterations of the project to see if lyric composition has any impact on music popularity.

Would like to build model from audio analysis rather than audio features (features is just averages features rom audio analysis).