The Next Big Sound:

Using Data Science to Predict Hit Songs

By: José Cacho

Can we predict hit songs *before* they chart on the Billboard?

- Record labels spend millions of dollars looking for the next big sound and then promoting their prospective musicians heavily
- Historically done by ‘feel’ - with scouts
- Can we provide a data-based approach to this process?

Data Collection

- 290,000 songs
  - ~285,000 “non-hits”
  - ~5,000 “hits” from the Billboard Top 100 charts
- Used Spotify’s API to collect 13 acoustic features for each song
  - E.g. ‘danceability’, ‘energy’, ‘liveness’
- SMOTE to upsample “hits”

Data Cleanup

- Removals
  - Songs from 1950-1963 and from 2016-2019, as my “hits” dataset did not contain any songs from this time period
  - Karaoke versions
  - Mismatches between the track and Spotify’s API

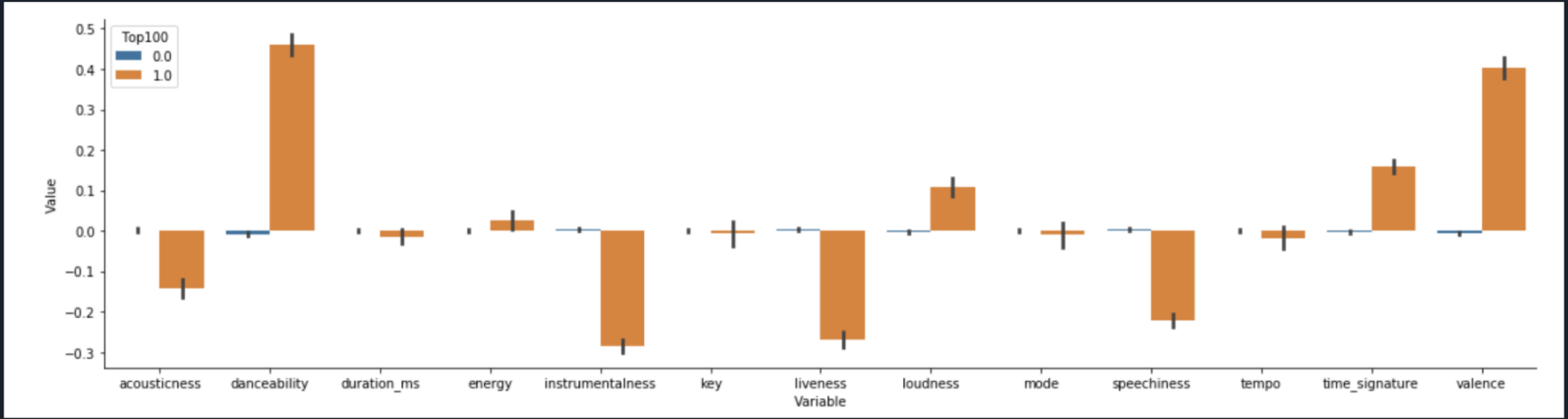
Data Analysis

- To the naked eye, no major differences

	top100	other
acousticness	0.253845	0.299235
danceability	0.619139	0.539385
duration_ms	236487.830790	237971.909518
energy	0.624470	0.617738
instrumentalness	0.029598	0.098727
key	5.223402	5.246971
liveness	0.176781	0.236091
loudness	-8.662994	-9.162115
mode	0.704048	0.708443
speechiness	0.063522	0.092811
tempo	119.354481	119.873555
time_signature	3.960186	3.887181
valence	0.615800	0.511869

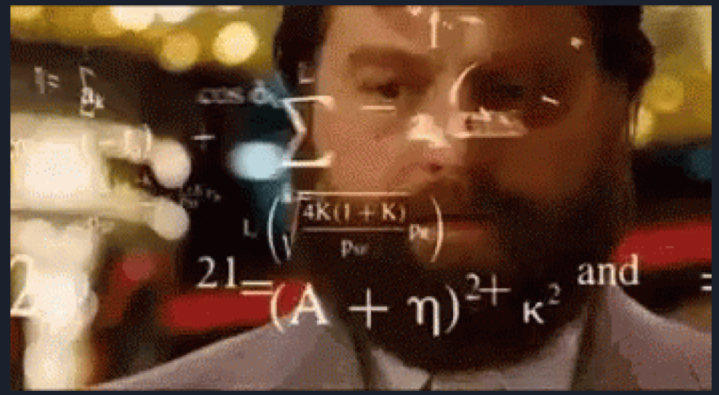
Data Analysis

However, when scaled...



Modeling

- Train/Test split of 75/25
- 8 total models:
  - Logistic Regression
  - KNN
  - Decision Tree
  - Bagging
  - Random Forest
  - AdaBoost
  - Gradient Boosting
  - Voting Classifier



Results

Bagging

98.6%

Test Accuracy

Results

Model	Training Accuracy	Testing Acccuracy
Logistic Regression	66.00%	65.65%
KNN	87.83%	82.27%
Decision Trees	99.99%	97.01%
Bagging	99.82%	98.65%
Random Forest	99.88%	98.18%
AdaBoost	86.04%	85.86%
Gradient Boosting	95.11%	95.11%
Voting Classifier	99.82%	98.43%

Results

Feature	Coefficient
acousticness	0.112218
danceability	0.070404
duration_ms	0.05618
energy	0.05787
instrumentalness	0.043434
key	0.214133
liveness	0.042887
loudness	0.035769
mode	0.170644
speechiness	0.061317
tempo	0.037828
time_signature	0.04598
valence	0.051336

key	int	The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.
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 is most danceable. The distribution of values for this feature look like this: