DataHack 2021

FM STEREO
FM
AM
88 92 96 100 104 108 MHz
530 600 700 800 1000 1200 1400 1600 kHz
AM/FM 2 BAND STEREO CASSETTE RECORDER
SR2100 SERIES

Agenda

- Problem
- Data
- Models
- Interface
- Up Next

Problem

Scope and Background

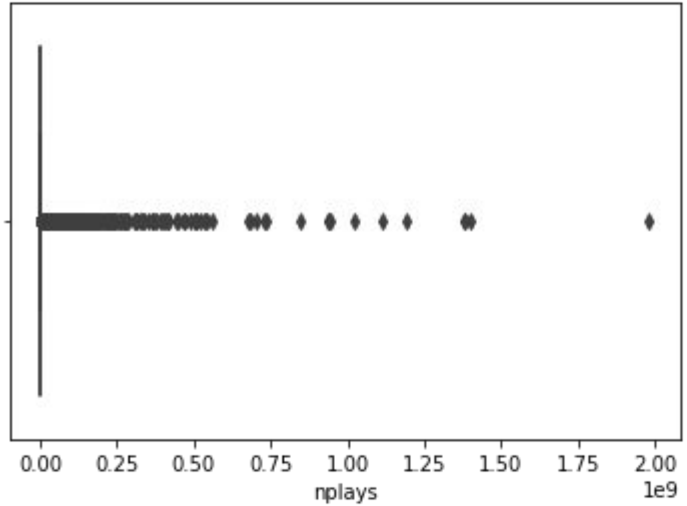
- The goal: Predict success of a song defined by the number of Spotify plays
    - Help small artists improve the likelihood of success based on metrics
- Success is defined this way as music is because the goal of most artists is to share their music with as many people as possible!

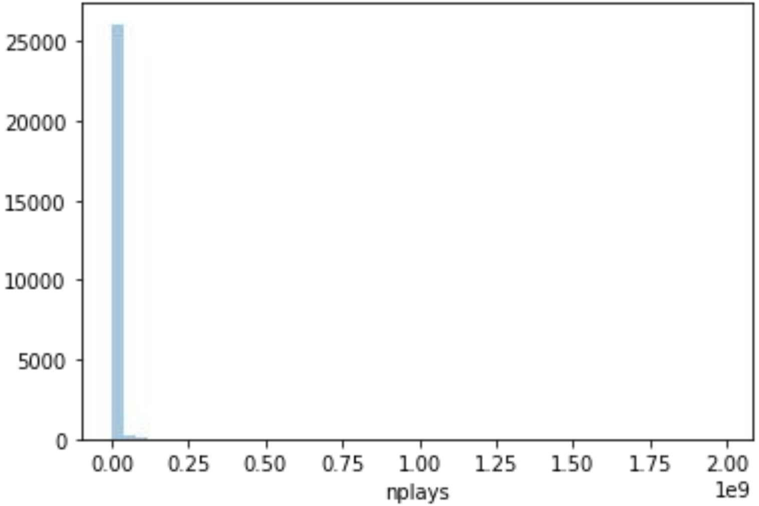
Data

90
IEC II / TYPE II HIGH POSITION
SUPER AVILYN CASSETTE SUPER HIGH RESOLUTION
B:
SUPER PRECISION ANTI-RESONANCE CASSETTE MECHANISM

Data Scope - nplays

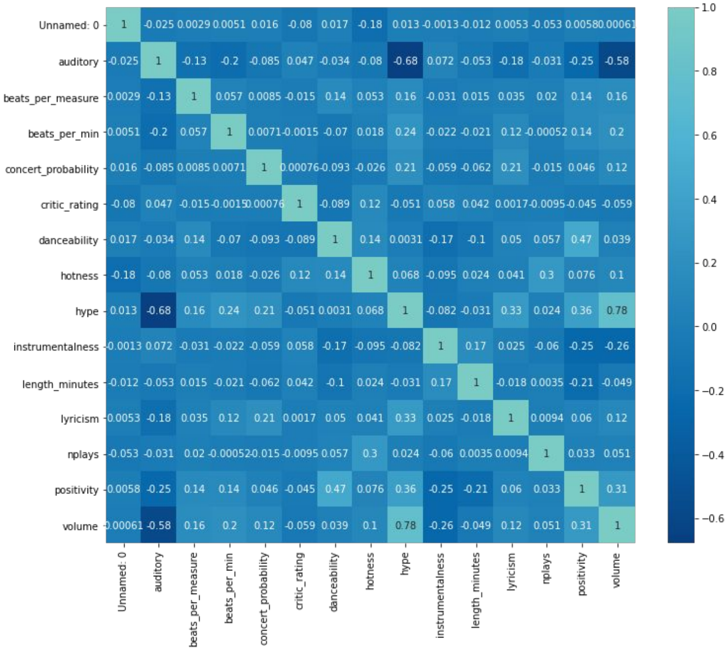
- Nplays has a huge spread
  - Eliminate outliers outside 3 standard deviations
  - Log scale the values





Data Scope - Correlations

- Upon looking at correlations decided to remove all about 0.5



	Unnamed: 0	auditory	beats_per_measure	beats_per_min	concert_probability	critic_rating	danceability	hotness	hype	instrumentalness	length_minutes	lyricism	nplays	positivity	volume
Unnamed: 0	1	-0.025	0.0029	0.0051	0.016	-0.08	0.017	-0.18	0.013	-0.0013	-0.012	0.0053	-0.053	0.0058	0.00061
auditory	-0.025	1	-0.13	-0.2	-0.085	0.047	-0.034	-0.08	-0.68	0.072	-0.053	-0.18	-0.031	-0.25	-0.58
beats_per_measure	0.0029	-0.13	1	0.057	0.0085	-0.015	0.14	0.053	0.16	-0.031	0.015	0.035	0.02	0.14	0.16
beats_per_min	0.0051	-0.2	0.057	1	0.0071	-0.0015	-0.07	0.018	0.24	-0.022	-0.021	0.12	-0.00052	0.14	0.2
concert_probability	0.016	-0.085	0.0085	0.0071	1	0.00076	-0.093	-0.026	0.21	-0.059	-0.062	0.21	-0.015	0.046	0.12
critic_rating	-0.08	0.047	-0.015	-0.0015	0.00076	1	-0.089	0.12	-0.051	0.058	0.042	0.0017	-0.0095	-0.045	-0.059
danceability	0.017	-0.034	0.14	-0.07	-0.093	-0.089	1	0.14	0.0031	-0.17	-0.1	0.05	0.057	0.47	0.039
hotness	-0.18	-0.08	0.053	0.018	-0.026	0.12	0.14	1	0.068	-0.095	0.024	0.041	0.3	0.076	0.1
hype	0.013	-0.68	0.16	0.24	0.21	-0.051	0.0031	0.068	1	-0.082	-0.031	0.33	0.024	0.36	0.78
instrumentalness	-0.0013	0.072	-0.031	-0.022	-0.059	0.058	-0.17	-0.095	-0.082	1	0.17	0.025	-0.06	-0.25	-0.26
length_minutes	-0.012	-0.053	0.015	-0.021	-0.062	0.042	-0.1	0.024	-0.031	0.17	1	-0.018	0.0035	-0.21	-0.049
lyricism	0.0053	-0.18	0.035	0.12	0.21	0.0017	0.05	0.041	0.33	0.025	-0.018	1	0.0094	0.06	0.12
nplays	-0.053	-0.031	0.02	-0.00052	-0.015	-0.0095	0.057	0.3	0.024	-0.06	0.0035	0.0094	1	0.033	0.051
positivity	0.0058	-0.25	0.14	0.14	0.046	-0.045	0.47	0.076	0.36	-0.25	-0.21	0.06	0.033	1	0.31
volume	0.00061	-0.58	0.16	0.2	0.12	-0.059	0.039	0.1	0.78	-0.26	-0.049	0.12	0.051	0.31	1

Data Preprocessing

- **Eliminate metrics that can not be predicted before release(lookahead bias)**
  - Hotness
  - Critic information
- Translated information directly related to individual artist
  - Artist name → length, is uppercase, positive sentiment, negative sentiment
  - Album name → length, is uppercase, positive sentiment, negative sentiment
  - Song name → length, is uppercase, positive sentiment, negative sentiment
- Log scaled nplays
- Replace all NaNs in features with the most common class
- Removed variables with correlation above 0.5
- Min Max scaled the numerical features to [0, 1]

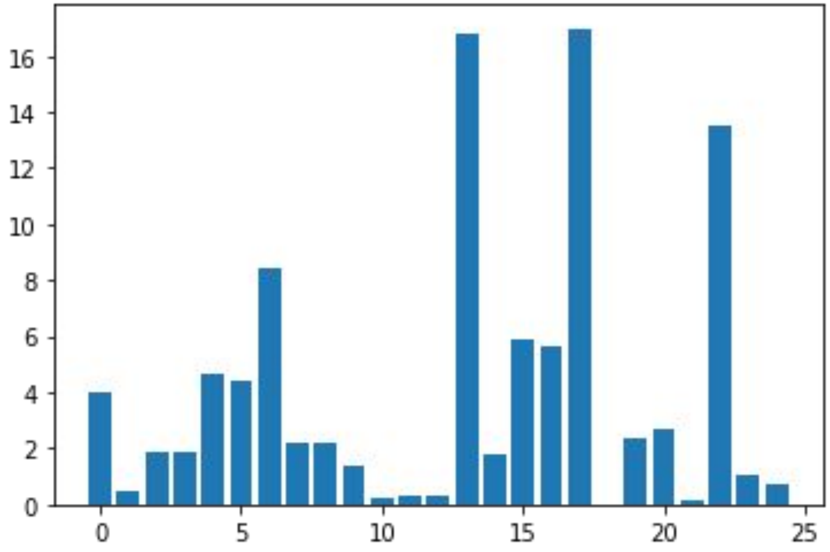
Models

Models

- XGBoost, Catboost
  - Performed similarly overall
  - Some hyper parameter tuning revealed that we didn’t need to deep of a tree
  - Since the styles feature ended up mattering a lot, catboost automatic categorical feature handling could have helped it outperform xgb
  - Achieved final R^2 score of 0.49 with Catboost
- Recurrent neural networks
  - Used textgenrnn to train many different models to generate reviews
  - User input used to select which model to use for review generation (currently based on genre)

Insights

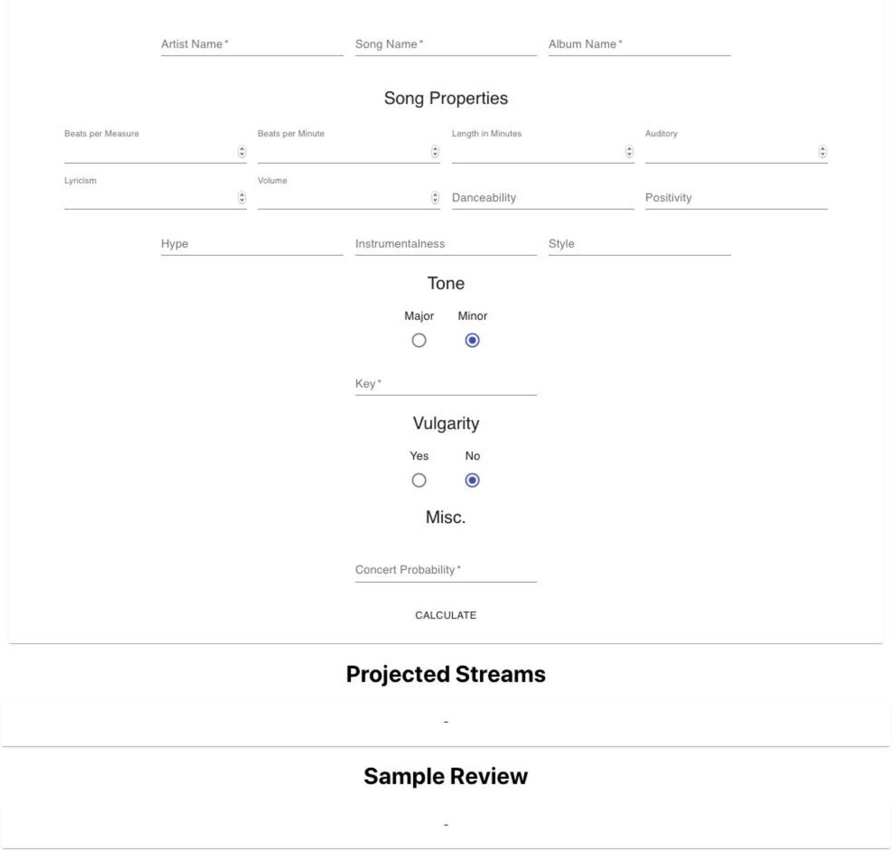
- Top 3 features according to our tuned Catboost model
  - Length of album name
  - Length of artist name
  - Style
- Could be indicative that the visual aspect of the song could influence whether it is streamed or not
  - Also possible that these metrics still too closely associate the song with artist
- Some styles are just more popular than others
  - Styes were also consistently an important feature through our model exploration



Interface

Practical Application

- Flask and React App built on our model, so a user may input info about their song and be returned the expected number of Spotify plays
- The user would not see the complex model underneath



Artist Name * | Song Name * | Album Name *

Song Properties

Beats per Measure | Beats per Minute | Length in Minutes | Auditory

Lyricism | Volume | Danceability | Positivity

Hype | Instrumentalness | Style

Tone

○ Major ● Minor

Key *

Vulgarity

○ Yes ● No

Misc.

Concert Probability *

CALCULATE

Projected Streams

-

Sample Review

-

Up Next

Future Plans

- Upload a finished song and our algorithm would calculate the song statistics
- Displaying what metrics had the most positive and negative influence on the number of plays so artists can adjust this in real time
- Implement grammarly API to help generate reviews that make sense
- Categorize music by genre to improve metrics and success prediction

Questions