





Spotify Predictive Model

Data Mining for the Next Big Hit

Chelsea Alford, Marlon Do Couto, Yujia Huo, Richard Locarni

DATA UNDERSTANDING

- Data by genre
- Focus on variables reflecting musical qualities
- Making a subjective question more scientific



DATA PREPARATION

- Missing or Incomplete records
- improperly formatted or structured data
- inconsistent values
- “spotify.csv” VS “data_w_genres.csv” → Rock, Pop, Jazz, Rap



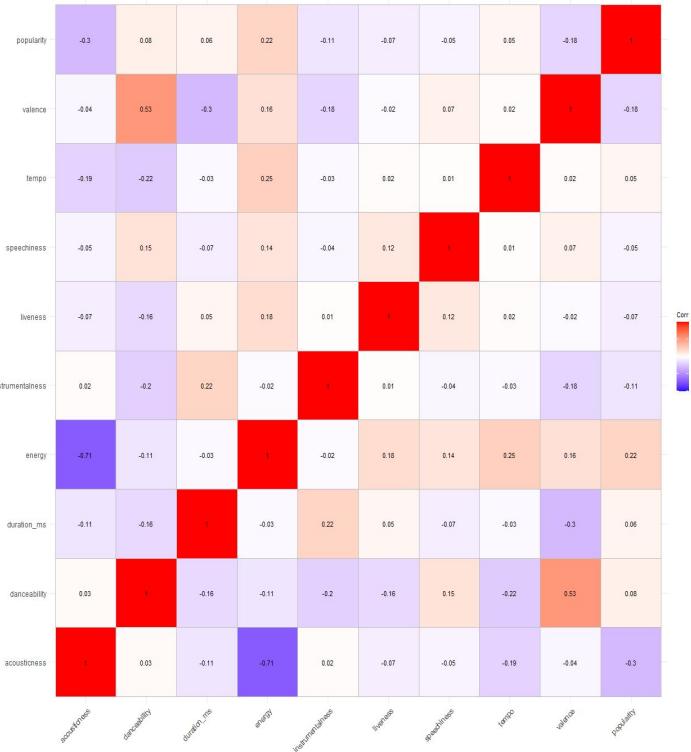
MODELING

Preliminary R² findings:

- Rock: 0.1908
- Jazz: 0.2243
- Pop: 0.2353
- Rap: 0.2774

R² for a complete model:

- Rock: 0.2906
- Jazz: 0.3087
- Pop: 0.3565
- Rap: 0.2617



The Data

Modeling

Evaluation

Deployment

MODELING

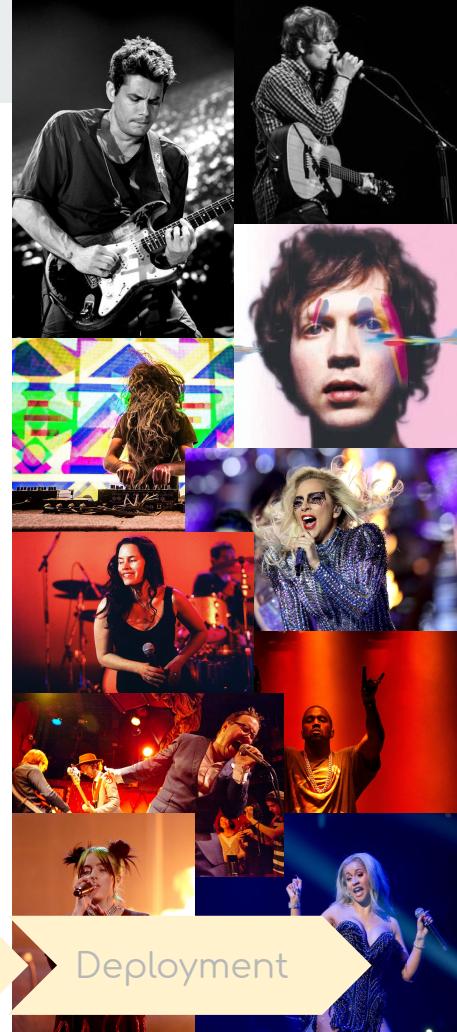
- Rock model:
 - *Duration*
 - Final R²: 0.2275
- Jazz model:
 - *Speechiness, tempo, and instrumentalness*
 - However, AIC for first model was 8,211 vs 8,215
 - Final R²: 0.387
- Rap model:
 - *Speechiness*
 - Final R²: 0.2745
- Pop model:
 - *No variable eliminated*
 - Final R²: 0.3565

The Data

Modeling

Evaluation

Deployment



EVALUATION

Rock music:

- Less valence
- More danceability
- Fewer spoken words

Rap music:

- Less valence
- More danceability
- Less instrumentation

Jazz music:

- Less valence
- More danceability
- Less liveness

Pop music:

- Less valence
- More danceability
- More energetic



The Data

Modeling

Evaluation

Deployment

DEPLOYMENT

- Multi-level implementation
- Future development
 - Machine learning
 - Artist-specific implementation
- Risks
 - Immediate impact
 - Ethics



The Data

Modeling

Evaluation

Deployment

CONCLUSION

- Making sense of unpredictability
- Tailor songs by genre
- Improve efficiency of record companies

