Predicting Song Popularity

A linear regression project
by Alison Salerno

Process

Step 1

- Gather 2,000 song data from Kaggle and additional data from Spotify API
- Clean data for modeling

Step 2

Use EDA & Statistical Testing to gain insight into the dataset:

- How my target and explanatory variables relate
- The relationship between x variables

Step 3

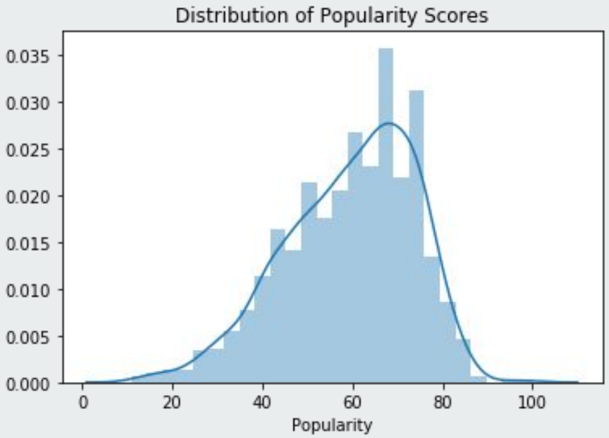
Find the best model for my dataset through:

- Feature transformation
- Feature engineering
- Feature selection
- Compare/contrast model iterations

The Data

Target Variable

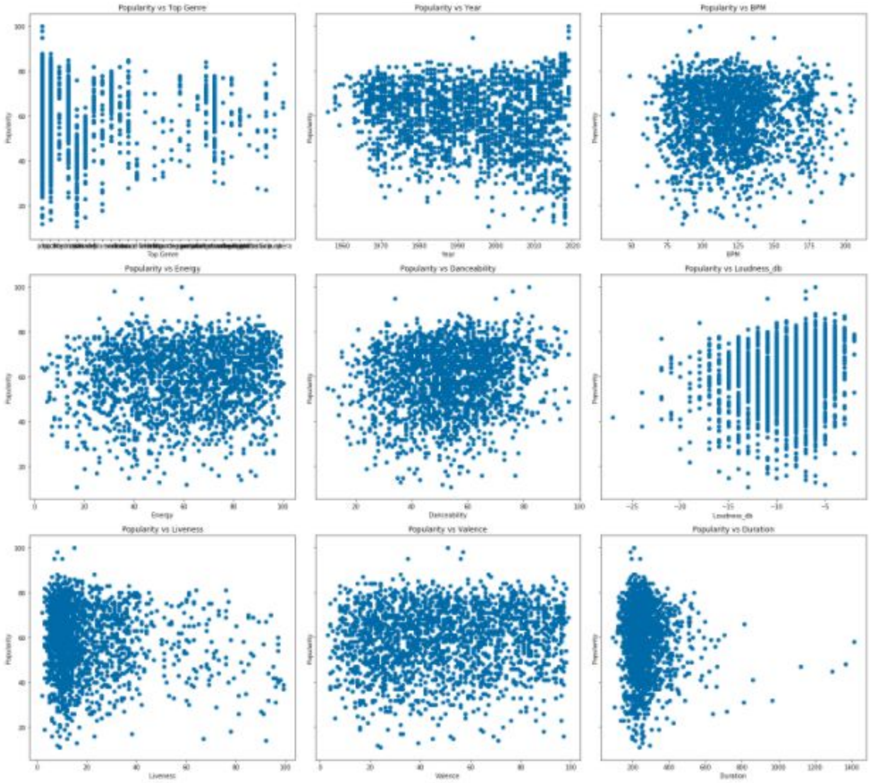
- Popularity Score

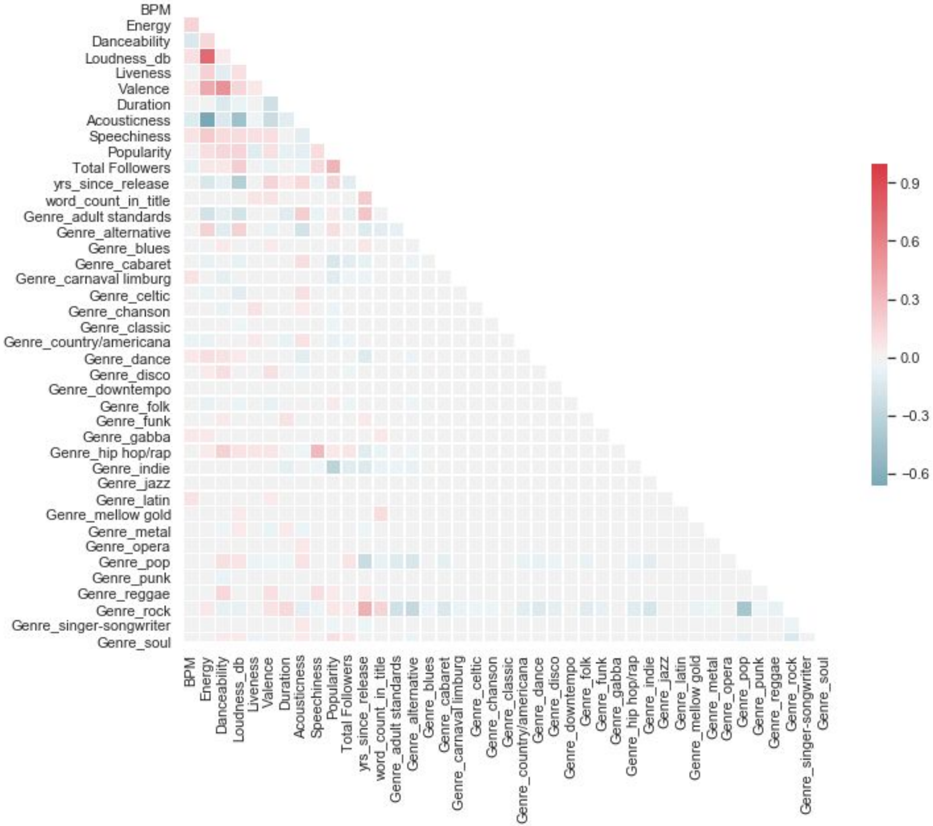


Independent Variables

- Artist follower count
- Top genre
- # of years since released
- # of words in title
- Audio elements:
  - BPM
  - Energy
  - Danceability
  - Loudness
  - Valence
  - Speechiness

EDA Conclusions





Feature Engineering

- Use of Polynomial transformations to create interactions such as:
  - Energy Danceability
  - BPM Loudness
  - Accousticness Speechiness
- Interaction features create a more predictive model

Models & Results

	# of Features	Train RMSE	Test RMSE	R^2
Baseline	1	13.63	13.12	.096
All Features	85	11.67	11.83	.26
Select Features	34	11.86	11.69	.28

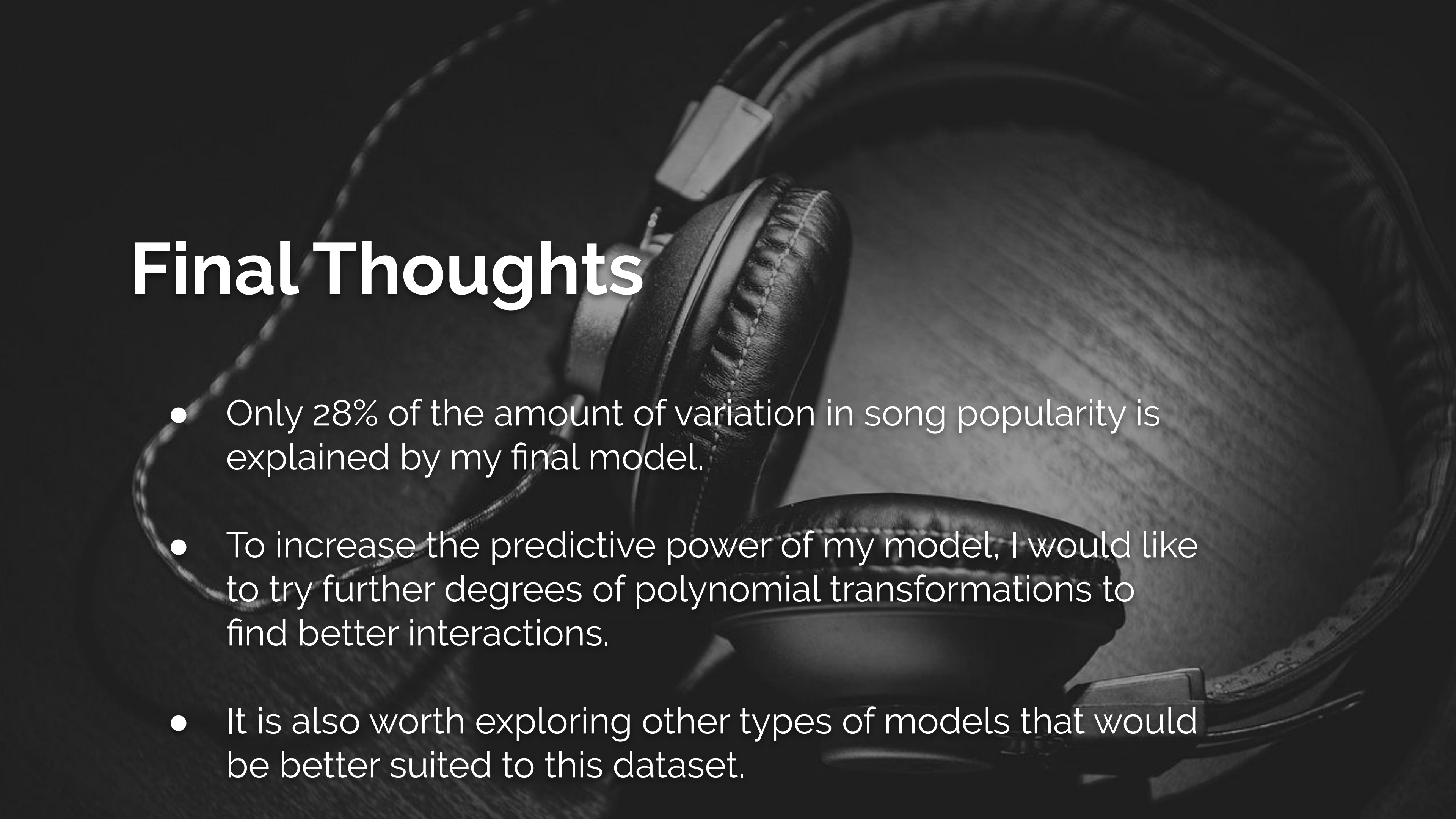
Biggest increasers of song popularity:

- # of years since release
- Artist Followers
- Danceability

Biggest decreasers of song popularity:

- Indie Genre
- Acousticness

Speechiness



Final Thoughts

- Only 28% of the amount of variation in song popularity is explained by my final model.

- To increase the predictive power of my model, I would like to try further degrees of polynomial transformations to find better interactions.

- It is also worth exploring other types of models that would be better suited to this dataset.