Spotify Audio Features

Hamid Niki
Springboard
Capstone Project 1

Objective: Predict popularity of a music track based on its audio features

- Being able to predict a track/album’s popularity has applications in the music business
- Helps label companies base their decisions of whether or not they want to sponsor an artist
- This is a purely data oriented decision. The sponsors do not need to even know the artist or like their music. In the past the company had to intimately know the artists and their work to sponsor them

Dataset Description

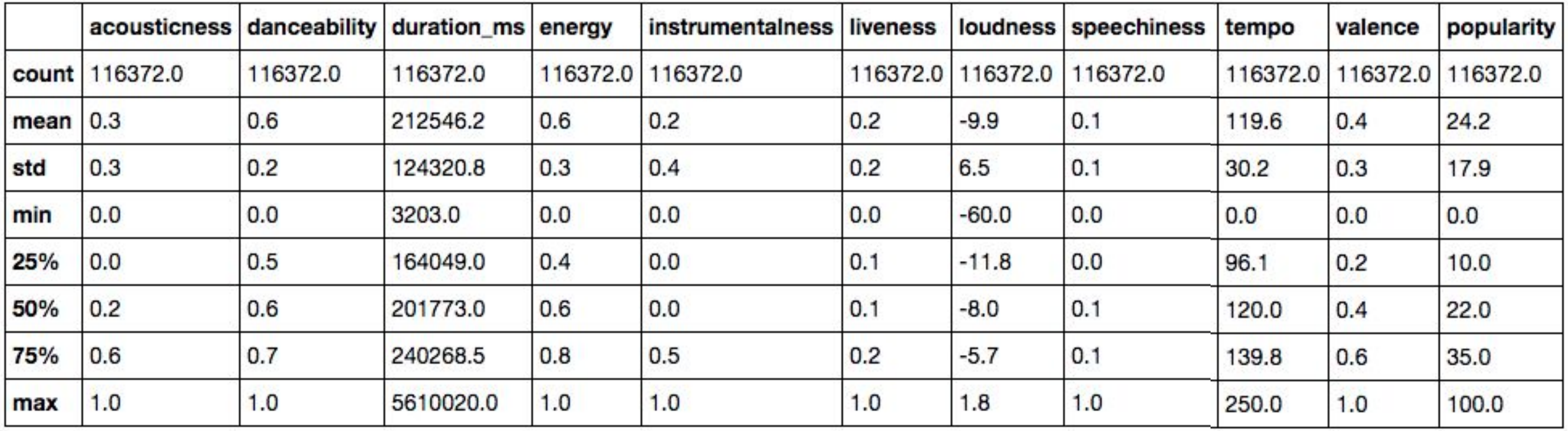
- 116,372 music tracks from spotify (one row per track) and 17 variables
- **Outcome variable:** Popularity (between 0 and 100, calculated based on the number of times a trak was played)
- **Numeric predictors:** Acousticness, danceability, duration_ms, energy, instrumentalness, liveness, loudness, speechiness, tempo and valence (more info: https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/)
- **Categorical predictors:** mode (major or minor), key, time signature

Data Cleaning and Wrangling:

- **Tidiness:**
  - Each track attribute is measured in a separate variable
  - Each observation (i.e. track) is in a separate row
  - There are no duplicates in the data
- **Missing Ratio:** There were no missing values in the data
- **Outliers:** There were some outliers in the duration_ms variables (tracks with 80-90 minutes length). Tracks were looked up in spotify the duration values were confirmed.
  - Artist name: John, Tracks: Whatever, 5,040,048 ms (~84 mins) and Ever, 5,610,020 ms (~93.5 mins)
  - Artist name: Gentle Whispering, Track: Fluffy Sleepy Whispers, 2,097,245 (~35 mins)

Descriptive Statistics

Numerical variables:

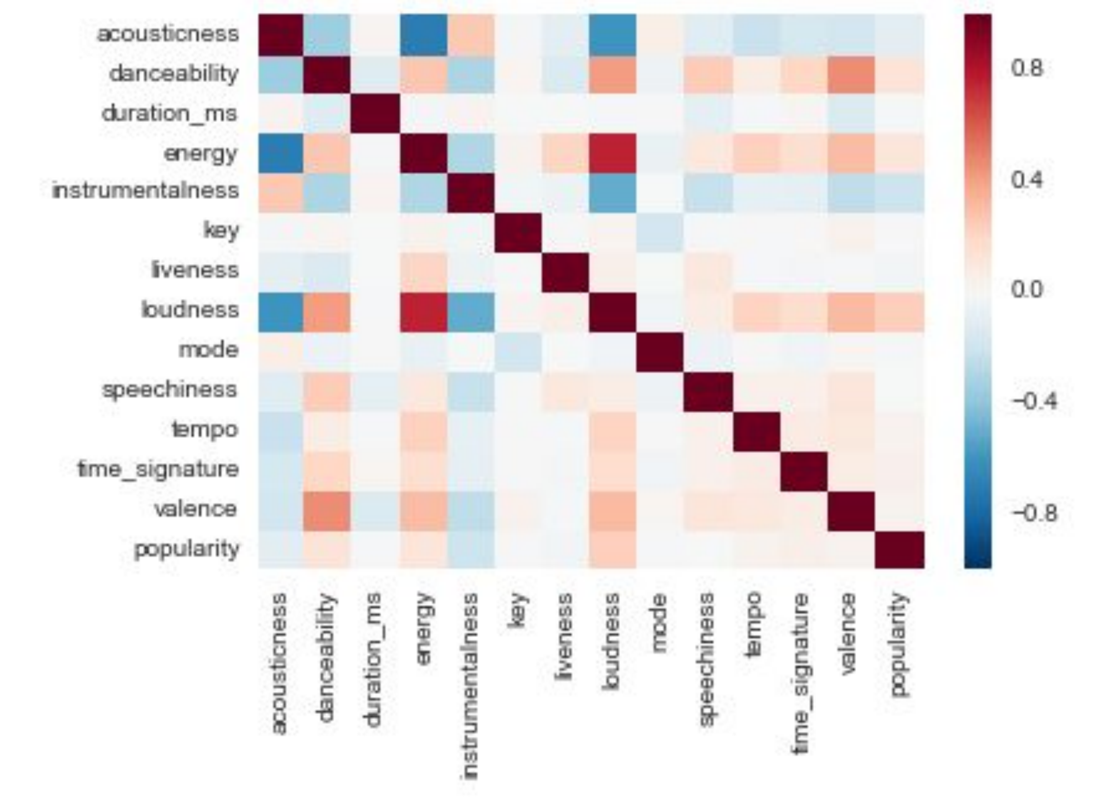


	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
count	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0
mean	0.3	0.6	212546.2	0.6	0.2	0.2	-9.9	0.1	119.6	0.4	24.2
std	0.3	0.2	124320.8	0.3	0.4	0.2	6.5	0.1	30.2	0.3	17.9
min	0.0	0.0	3203.0	0.0	0.0	0.0	-60.0	0.0	0.0	0.0	0.0
25%	0.0	0.5	164049.0	0.4	0.0	0.1	-11.8	0.0	96.1	0.2	10.0
50%	0.2	0.6	201773.0	0.6	0.0	0.1	-8.0	0.1	120.0	0.4	22.0
75%	0.6	0.7	240268.5	0.8	0.5	0.2	-5.7	0.1	139.8	0.6	35.0
max	1.0	1.0	5610020.0	1.0	1.0	1.0	1.8	1.0	250.0	1.0	100.0

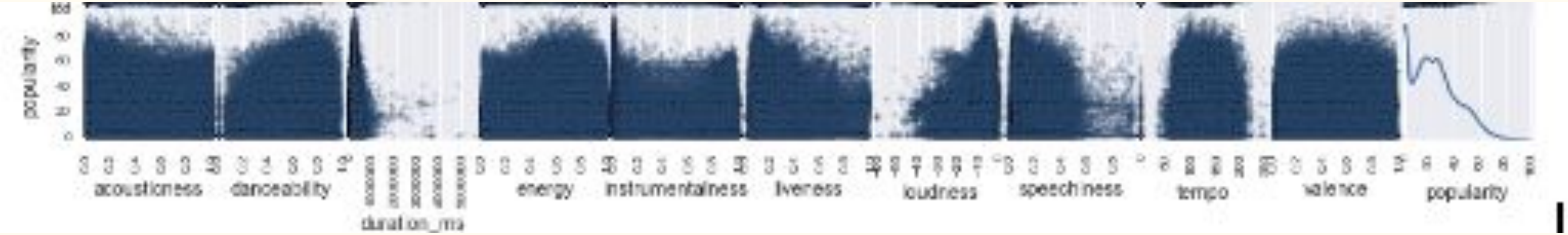
Correlation Plot(matrix):

- Loudness, energy and danceability are positively correlated with popularity

- Instrumentalness and acousticness are negatively correlated with popularity



Scatter Matrix:



- Tracks with very long length (duration_ms) are less popular
- Live tracks (liveness) are less popular
- Tracks with extreme tempo levels (< 50 BPM or > 250 BMP) are not popular

Formal Tests of Correlation - Popularity vs Numerical Variables

All distributions under the null hypothesis were generated using bootstrap method:

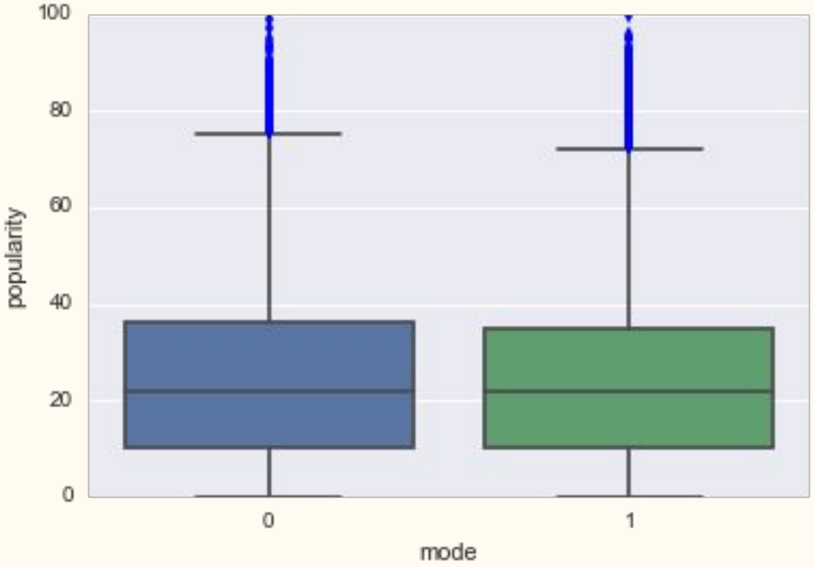
	Hypothesis to Test	**Observed correlation**	**P_value**	**Conclusion**
Loudness vs popularity	H0: No Correlation Ha: Positive correlation	0.235	0.0	Rejecting the null
Danceability vs popularity	H0: No Correlation Ha: Positive correlation	0.134	0.0	Rejecting the null
Instrumentalness vs popularity	H0: No Correlation Ha: Negative correlation	-0.21	0.0	Rejecting the null
Track duration vs popularity	H0: No Correlation Ha: Negative correlation	-0.009	0.001	Rejecting the null
Liveness vs popularity	H0: No Correlation Ha: Negative correlation	-0.028	0.001	Rejecting the null

Categorical Variables: Mode, Time Signature, Key

These variables, though presented as integers in the dataset, should be used as categorical variables in the analysis because they do not have an ordinal nature

Popularity vs Mode (major or minor):

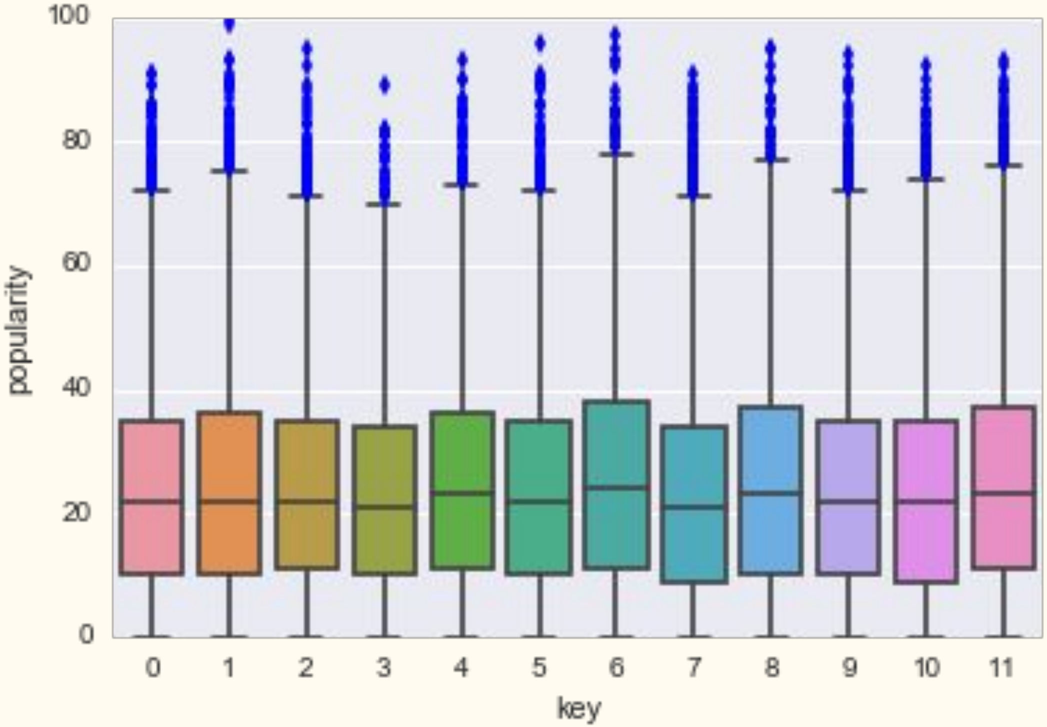
No visible difference in popularity between modes



Popularity vs Key(C, Db, D, Eb, E, etc)

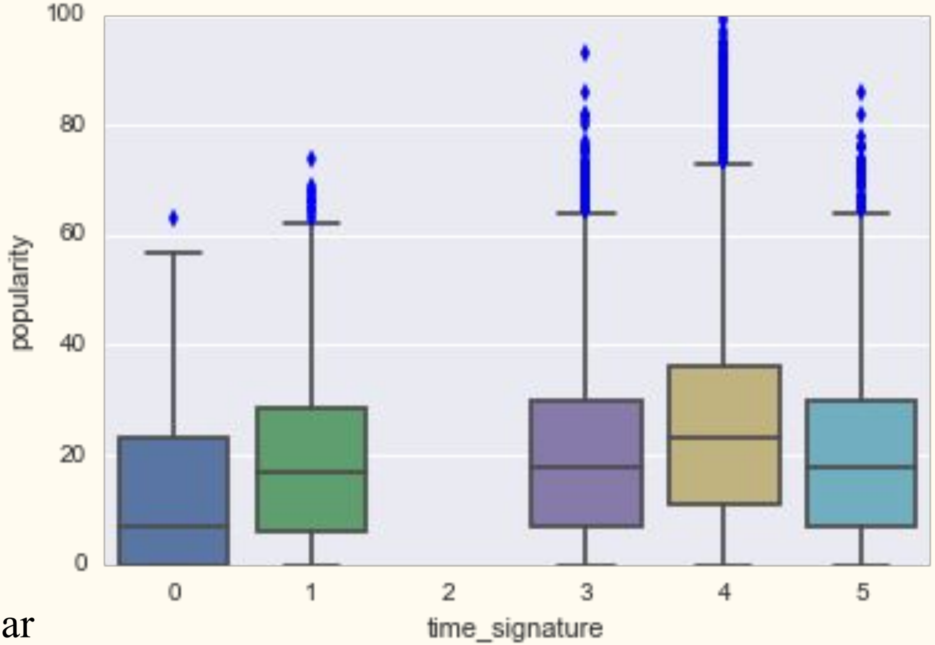
- There are 12 possible keys

- There is no visible difference in popularity between keys



Popularity vs Time Signature:

- 4-beat time signatures seem to be more popular
- 4-beat rhythm is more common than other time signatures (86% of the tracks in the dataset)



- 4-beat is the most familiar beat to human ear
- 2-sample bootstrap hypothesis test between 4-beat and ‘other’

Observed difference in means popularity: **4.56**, p_value close to 0.0 -> Reject the null

Multivariate Analysis

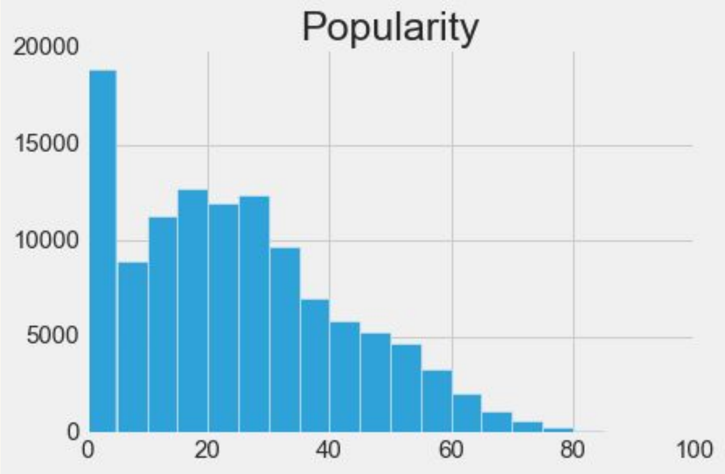
- Supervised learning since data is labeled
- Regression problem since target variable, popularity, is numerical

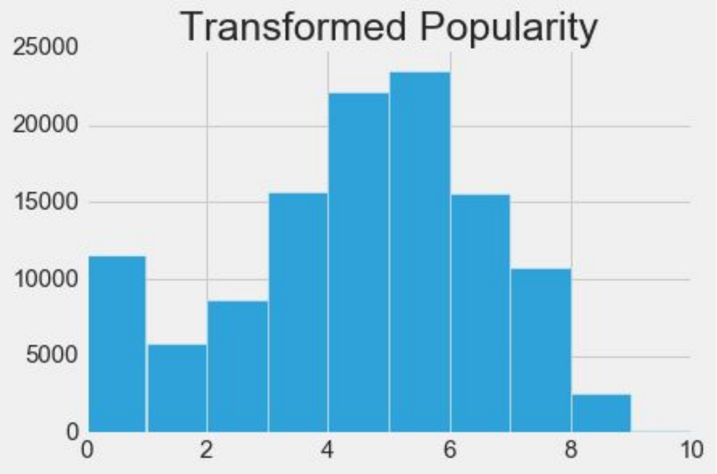
Modeling Approaches Used:

- LASSO
- Gradient Boosting Regression

Preprocessing:

- Turned the 3 categorical variables (mode, key, time signature) to dummy variables. Total number of features: 27
- No missing values in the data
- Square root transformation on target variable





LASSO Analysis:

- 70-30 train-test split
- Baseline using OLS regression:

  $R^2 = 0.06$, MSE = 4.5

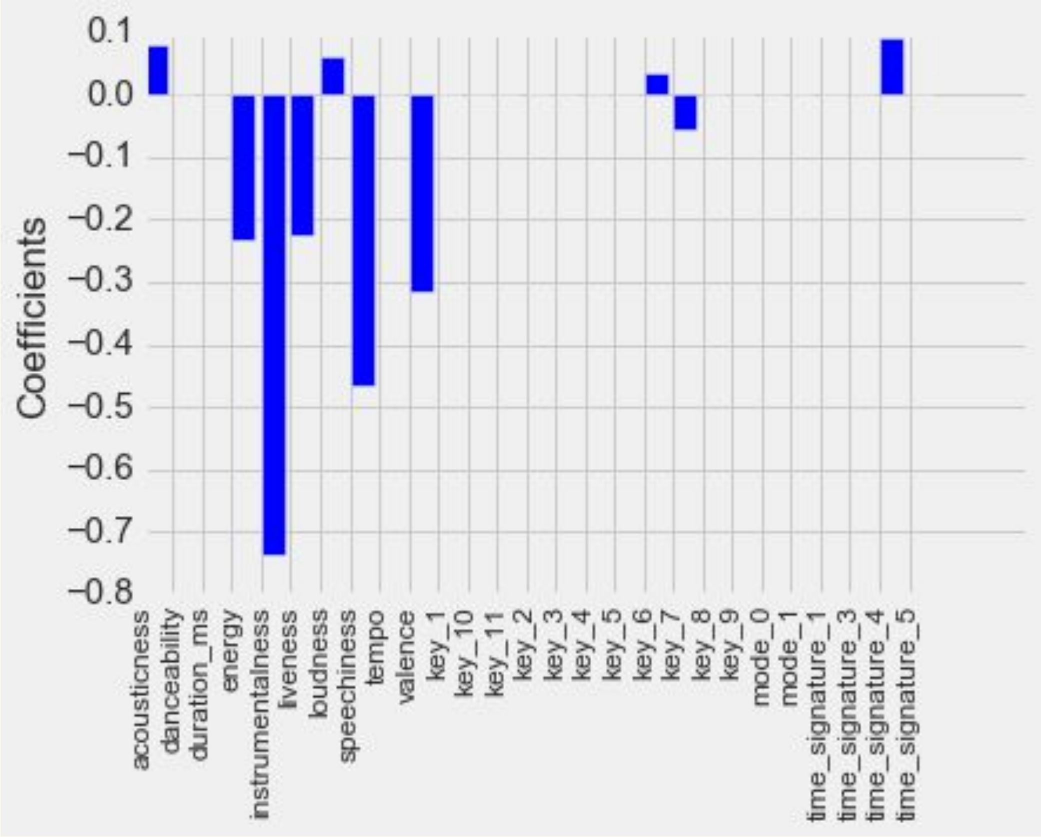
  (same values on train and test data)

- LASSO with hyperparameter tuning:

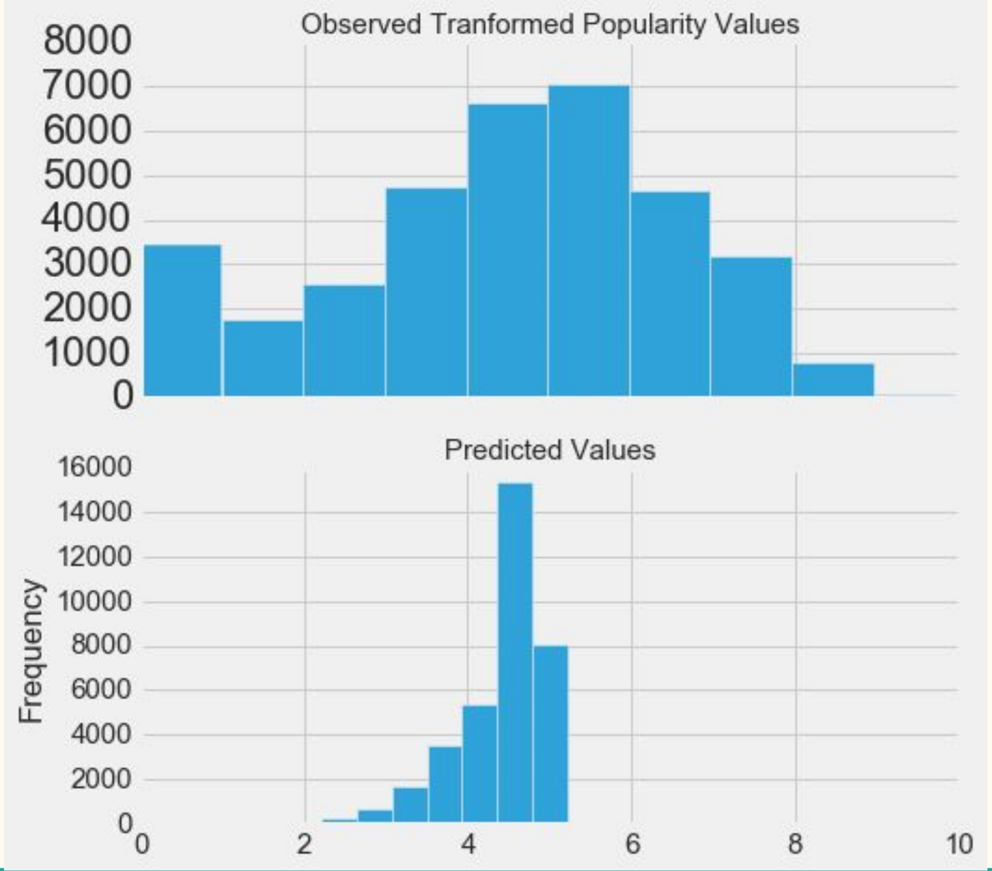
  $R^2 = 0.07$, MSE =4.5 (same values on train and test data)

- Model under fits, but it is generalizable to unseen data

LASSO Coefficients:



Actual vs Predicted Popularity:



Gradient Boosting

- 70-30 train-test split
- Baseline GB with default parameters:

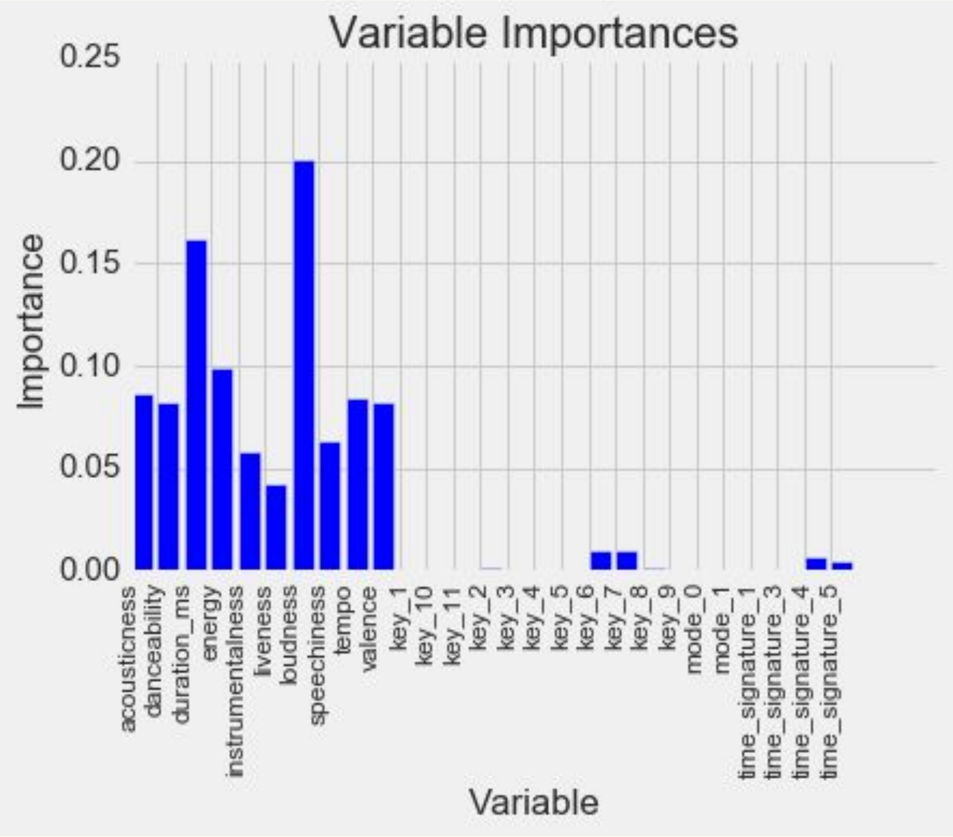
  R^2 train= 0.14, R^2 test = 0.12 , MSE test = 4.2

- Gradient boosting with hyperparameter tuning:

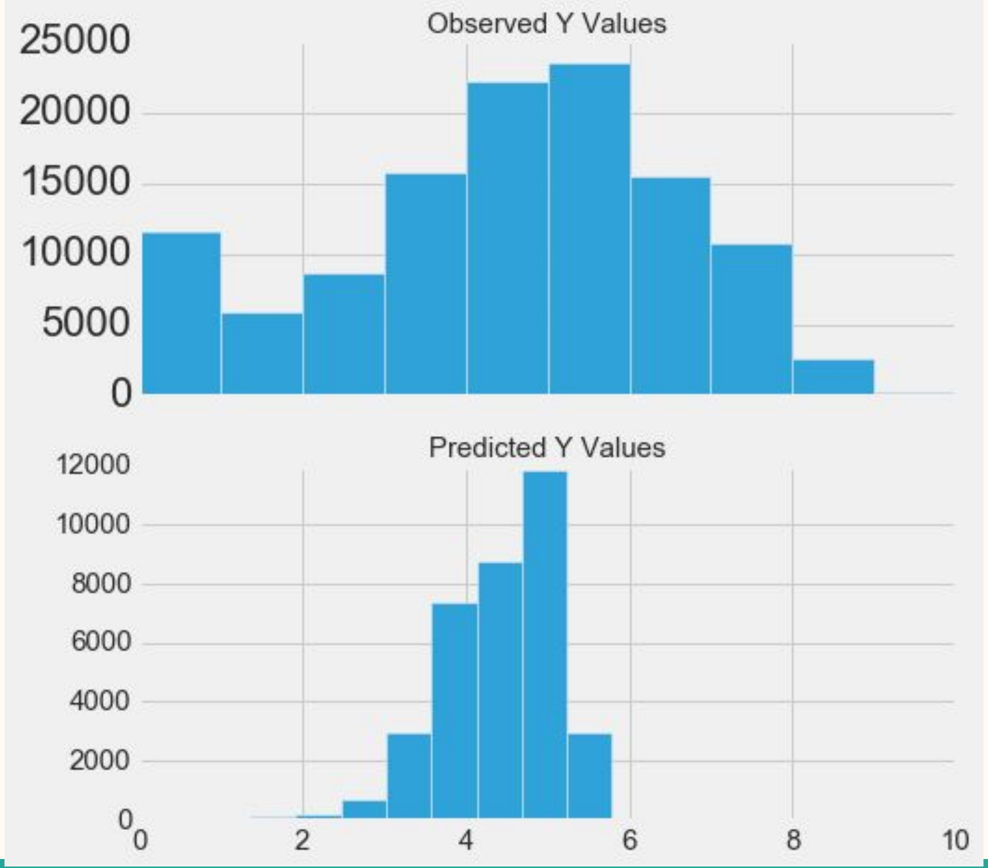
  R^2 train= 0.14, R^2 test = 0.12 , MSE test = 4.2

- We are perhaps at near max. performance
- Model performs better than LASSO and is generalizable to unseen data

Variable Importance in GB Model:



Actual vs Predicted Popularity:



Conclusion and Next Steps:

- Additions to the features in near future might improve the model performance
- Adding artist name as a predictor is likely to improve performance, though artist name would not be an audio feature