Spotify

Song Hit Prediction

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About us

- About Us
- Motivation
- Dataset
- EDA
- Models
- Results
- Future Work

AC 209

Spotify Song Hit Prediction

Which song would be the next hit?

Jinglun Gao
Chao Wang
Sophy Huang
Hongwen Song

12/11/2022 Team33

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

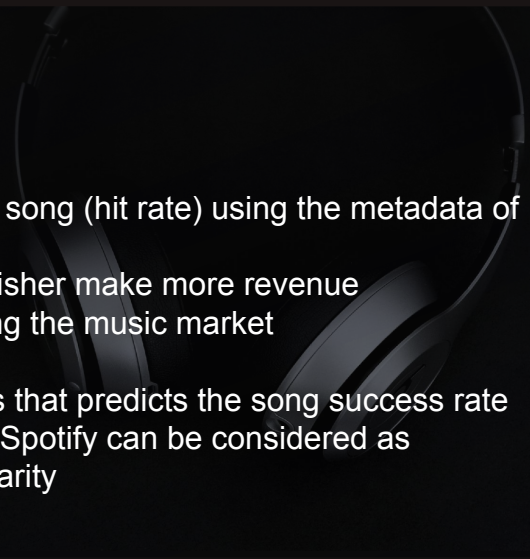
EDA

Models

Results

Future Work

Motivation



- Spotify
  - Most famous music app
  - 30 million tracks
- Hit Song Science
  - Predicting the success of a given song (hit rate) using the metadata of the song
  - Helping music producer and publisher make more revenue
  - AI-assisted tools for understanding the music market
- Our Contribution
  - develop machine learning models that predicts the song success rate
  - investigate if audio features from Spotify can be considered as determinants of the stream popularity

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

EDA

Models

Results

Future Work

Dataset

- Spotify Song Hit Prediction Dataset
- 41106 rows and 19 feature
- 13 numerical variables and 6 categorical variables
- Target variable: Song hit (0 or 1)
- Balanced class
- All columns do not contain missing values

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

Exploratory Data Analysis (EDA)

About Us

Motivation

Dataset

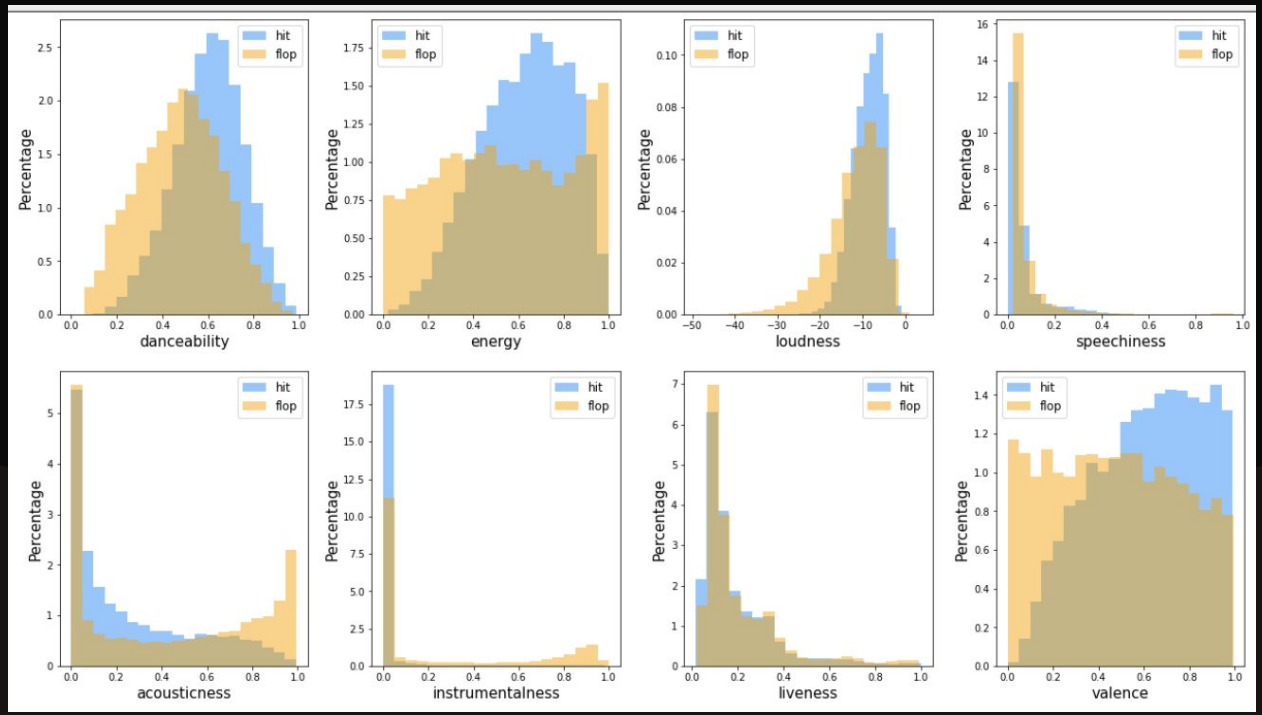
EDA

Models

Results

Future Work

Distributions of the predictors with respect to the target variable



https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

EDA

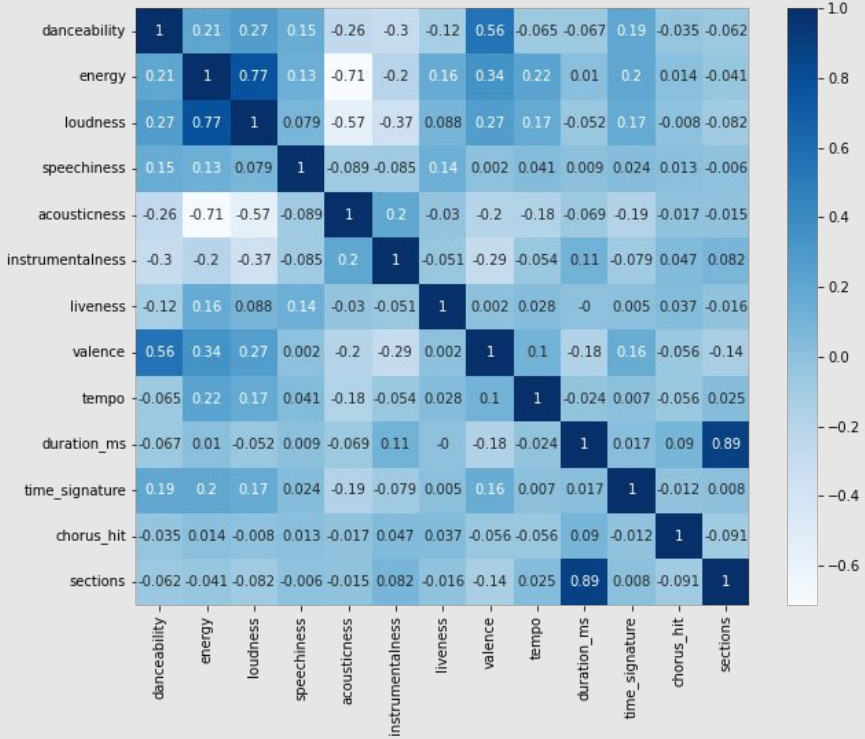
Models

Results

Future Work

Exploratory Data Analysis (EDA)

Heatmap of the correlation matrix



https://github.com/miaowu128/Spotify_Song_Hit_Prediction

Exploratory Data Analysis (EDA)

About Us

Motivation

Dataset

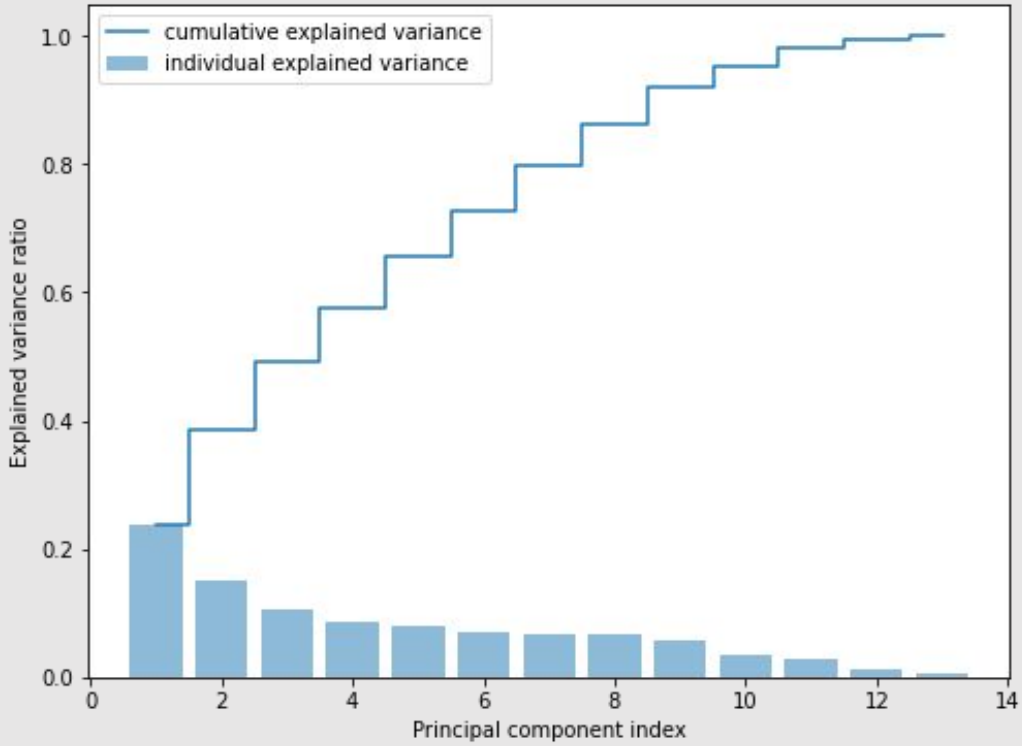
EDA

Models

Results

Future Work

Principal Component Analysis



https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

EDA

Models

Results

Future Work

Models

- Linear Classifier
  - Logistic Regression
- Non-Linear Classifier
  - KNN
  - Decision Tree
  - Random Forest
  - AdaBoost
  - Stacking
  - MLP

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

EDA

Models

Results

Future Work

Logistic Regression

- One of the simplest classifiers
- Extremely fast to train
- Most importantly, high interpretability to study feature importance

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

EDA

Models

Results

Future Work

KNN

- Also a simple predictive model based on k-nearest data points
- This dataset has balanced labels

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

EDA

Models

Results

Future Work

Tree Family

- Decision Tree
  - Less work regarding data preprocessing as it only cuts b/w points
  - Highly interpretable
- Random Forest
  - DT tends to overfit, and RF can reduce it
  - Better than bagging in reducing correlation b/w trees
- AdaBoost
  - Adaptively weights the difficult-to-classify samples more heavily
  - Focus on reducing the bias thus very quick to fit

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

EDA

Models

Results

Future Work

Stacking & MLP

Stacking

- An advanced way to ensemble heterogeneous models
- Second-level meta-classifier to possibly make better decisions

MultiLayer Perceptron (MLP)

Cope with large data where underlying nonlinear relationship might be more likely

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

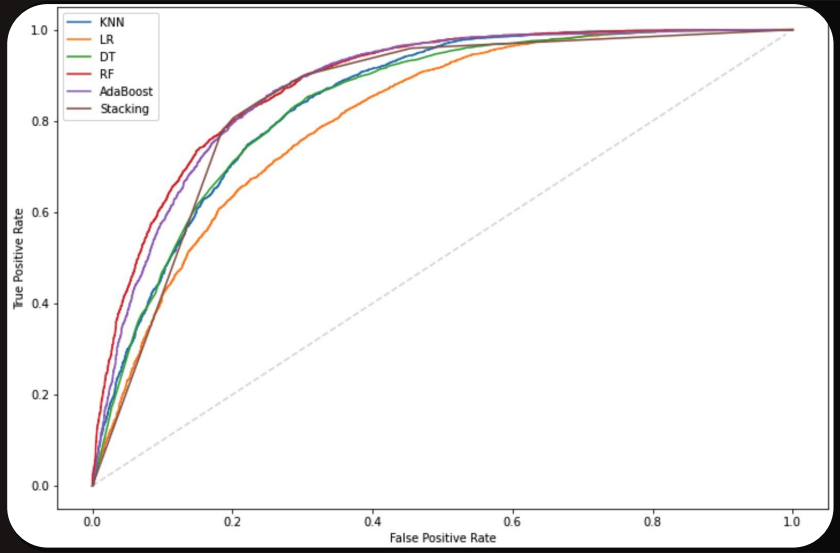
EDA

Models

Results

Future Work

Results



ROC Curves

Trade-offs between precision and recall

Random Forest

Better than LR, KNN, and DT

AdaBoost

Slightly better precision score

Stacking

Same performance as AdaBoost

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us
Motivation
Dataset
EDA
Models
Results
Future Work

Results

Model	Training Time	Precision	Recall	F1-Score	Accuracy
LR	0.23s	0.74	0.73	0.73	0.73
KNN	0.01s	0.79	0.75	0.75	0.76
DT	0.22s	0.78	0.77	0.77	0.77
RF	25.1s	0.80	0.80	0.80	0.80
AdaBoost	14.26s	0.81	0.80	0.80	0.80
Stacking	55.43s	0.81	0.80	0.80	0.80
MLP	365s	0.80	0.80	0.80	0.80

Random Forest & MLP

Better than LR, KNN, and DT

AdaBoost

Slightly better precision score

Stacking

Same performance as AdaBoost

https://github.com/miaowu128/Spotify_Song_Hit_Prediction

About Us

Motivation

Dataset

EDA

Models

Results

Future Work

Future Work

More features: Adding more predictors such as textual lyrics may provide more information for song hit predictions

Support Vector Machine: Utilizing the power of kernel-based framework to improve flexibility

Deep learning methods: Using neural networks to make predictions with more complex data

Broader project scope: Expanding problem statements to provide a more user-friendly tool for music investors

Thank You