

Spotify Billboard Classifier

Sayan

Reca

Databa

Data Analysis

Cross

Validation and Mode Selection

Cross Validation Model

Model Selection

Random

XGBoo

Spotify Billboard Classifier

Course Project: DATA 1030 - Hands On Machine Learning

Sayan Samanta



Instructor: Dr. Andras Zsom TA Advisor: Natalie Delworth GitHub Repo: shorturl.at/BEN08



Recap: Looking at the Database

Spotify Billboard Classifier

Sayan Samant

Recap Database

Explorat Data Analysis

Validationand Moo

Cross Validatio Model

Results
Random
Forests
XGBoost
AdaBoos

album id artist id billboard acousticness 0.00784 2Auw0pTT6EcQdvHNimhLQI 04gDigrS5kc9YWfZHwBETP 1.0 0 0.05910 2Auw0pTT6Ec0dvHNimhL0I 04qDigrS5kc9YWfZHwBETP 1.0 0.28400 2Auw0pTT6Ec0dvHNimhL0I 04gDigrS5kc9YWfZHwBETP 1.0 duration ms danceability date energy id 0.532 2015-05-15 238200.0 0.599 1m39XApxHXb2UkGItyyIU0 0 2015-05-15 0.788 2iuZJX9X9P0GKaE93xcPik 0.748 235493.0 13YQutqrAhT5iX9H0ctlu5 0.501 2015-05-15 203453.0 0.704 instrumentalness key lenath liveness loudness mode speechiness 0.0 6.0 12.0 0.1400 -6.543 1.0 0.0333 0 1.0 0.0 12.0 0.0863 -7.055 1.0 0.0334 0.0 3.0 12.0 0.1550 -5.640 1.0 0.0326 time signature track length valence tempo 130.019 4.0 225984.25 0.113 120.076 4.0 225984.25 0.884 97.039 4.0 225984.25 0.270

Figure: Sample data with Features



Recap: Exploratory Data Analysis

Spotify Billboard Classifier

Sayar Saman

Reca

Database Exploratory Data

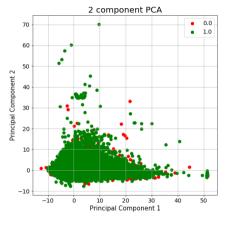
Data Analysis Cross

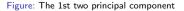
and Mod Selection Cross Validation

Model Selection

Random Forests XGBoost

Outloo





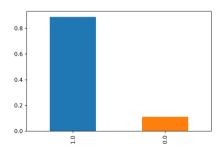


Figure: Balance of the dataset



Recap: Exploratory Data Analysis

Spotify Billboard Classifier

Sayan Samant

Recap

rtecaj

Exploratory Data

Analysis

Validation

Cross Validation

Model Selection

Random Forests

, (0000)

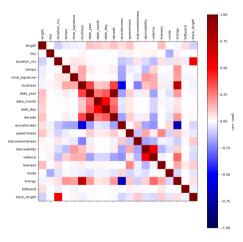


Figure: Correlation of features



Recap: Exploratory Data Analysis

Spotify Billboard Classifier

Sayan

Reca

Exploratory Data

Data Analysis

Validatio and Mod Selection

Validatio Model

Results
Random
Forests

AdaBoo

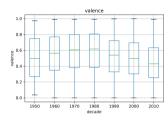
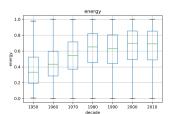


Figure: Valence evolution with time



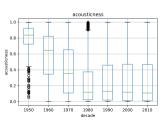


Figure: Energy evolution with time



Cross Validation

Spotify Billboard Classifier

Sayan

Reca

Databa Explora

Data Analys

Cross Validatio

and Mode Selection Cross

Validation Model Selection

Results
Randon
Forests

, (0000)

■ Since the dataset is highly imbalanced, we use stratified KFold split.

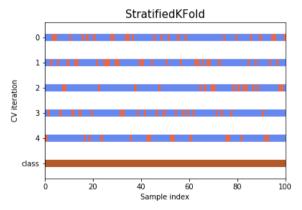


Figure: Different KFold Splits



Classification Models

Spotify Billboard Classifier

Sam

Datab Explor

Analys Cross Valida

and M Selecti

Cross Validation Model Selection

Results
Random
Forests
XGBoos
AdaBoo

Outloo

Random Forest

Parameters tuned:

max_depth 1,210 10 min sample split 1,210 3	Parameter Name	Parameter Range	Optimal Parameter
min sample split 1.210 3	max_depth	$1, 2 \dots 10$	10
1,21110	min_sample_split	$1, 2 \dots 10$	3

XGBoost

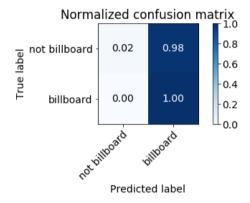
Parameters tuned:

Parameter Name	Parameter Range	Optimal Parameter
max_depth	3, 4, 5, 6, 8, 10, 12, 15	15
min_child_weight	1, 3, 5, 7	3
gamma	0.0, 0.1, 0.2 , 0.3, 0.4	0.4

AdaBoost

Parameters tuned:

Parameter Name	Parameter Range	Optimal Parameter
learning_rate	$10^{-3},\ldots,10^4$	0.1
algorithm	SAMME.R and SAMME	SAMME.R





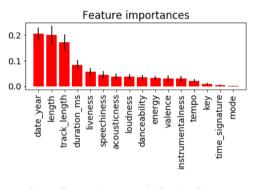


Figure: Feature Importance for Random Forest

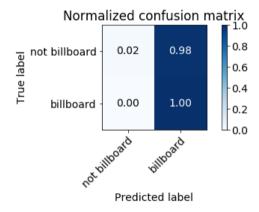


Figure: Confusion Matrix for XGBoost

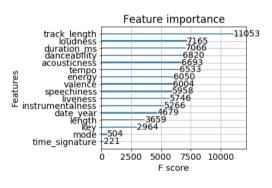


Figure: Feature Importance for XGBoost



Spotify Billboard Classifier

Sayan Samant

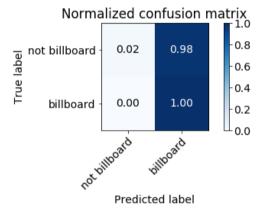
Recap Databas Explorat Data

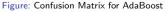
Cross
Validation
and Mode
Selection

Cross Validation Model Selection

Results

AdaBoost





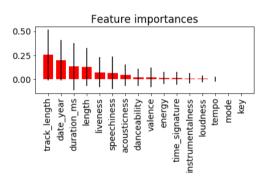


Figure: Feature Importance for AdaBoost

Sayan Samant

Reca

Datal

Explora Data

Anai

Validatio and Mod Selection

Cross Validation

Model Selection

Randoi Forests

AdaBoost

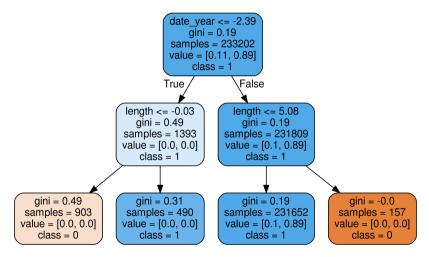


Figure: Estimator with least error

Outlook

Spotify Billboard Classifier

Sayaı

Databa Explora Data Analysi

Validati and Mo Selection

Cross Validation Model Selection

Results
Random
Forests
XGBoost
AdaBoos

Outlook

The dataset was highly imbalanced. Having a much more evenly divided set among the classes would improve the results.

- The acoustic features of songs in both the classes overlap highly. Different data mining algorithm might do better
- Due to lack of time, certain methods such as support vector machine classification or K-nearest neighbours couldnt be implemented
- Advanced deep-learning methods could improve the result.



Spotify Billboard Classifier

Outlook

Thank You. Question?