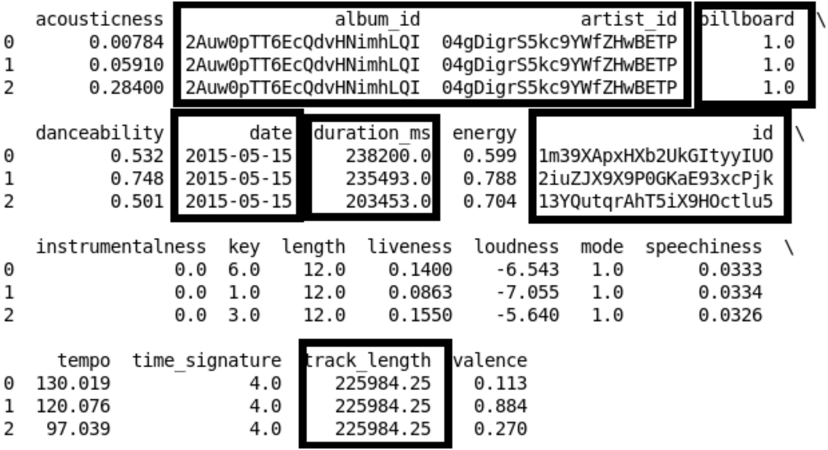
Spotify Billboard Classifier

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



```
   acousticness                album_id               artist_id  billboard  \
0       0.00784  2Auw0pTT6EcQdvHNimhLQI  04gDigrS5kc9YWfZHwBETP        1.0
1       0.05910  2Auw0pTT6EcQdvHNimhLQI  04gDigrS5kc9YWfZHwBETP        1.0
2       0.28400  2Auw0pTT6EcQdvHNimhLQI  04gDigrS5kc9YWfZHwBETP        1.0

   danceability        date  duration_ms  energy                      id  \
0         0.532  2015-05-15     238200.0   0.599  1m39XApxHXb2UkGItyyIUO
1         0.748  2015-05-15     235493.0   0.788  2iuZJX9X9P0GKaE93xcPjk
2         0.501  2015-05-15     203453.0   0.704  13YQutqrAhT5iX9HOctlu5

   instrumentalness  key  length  liveness  loudness  mode  speechiness  \
0               0.0  6.0    12.0    0.1400    -6.543   1.0       0.0333
1               0.0  1.0    12.0    0.0863    -7.055   1.0       0.0334
2               0.0  3.0    12.0    0.1550    -5.640   1.0       0.0326

     tempo  time_signature  track_length  valence
0  130.019             4.0     225984.25    0.113
1  120.076             4.0     225984.25    0.884
2   97.039             4.0     225984.25    0.270
```

Figure: Sample data with Features

Recap: Exploratory Data Analysis

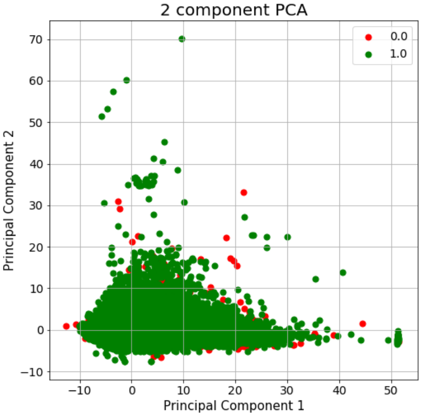


Figure: The 1st two principal component

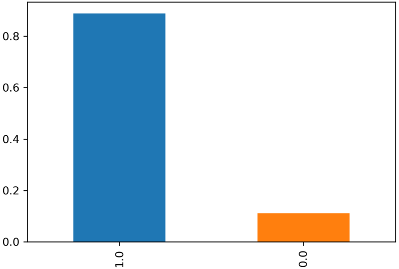


Figure: Balance of the dataset

Recap: Exploratory Data Analysis

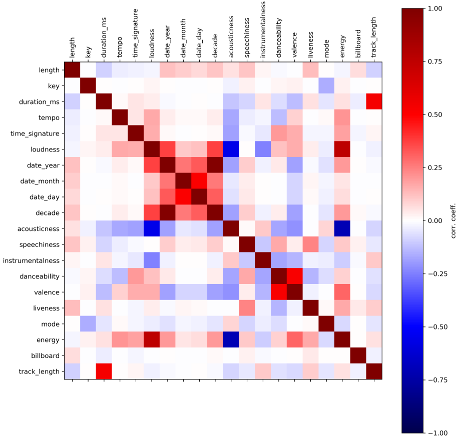


Figure: Correlation of features

Recap: Exploratory Data Analysis

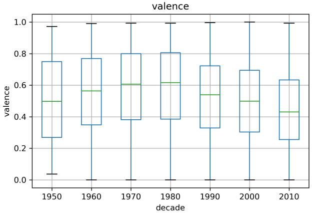


Figure: Valence evolution with time

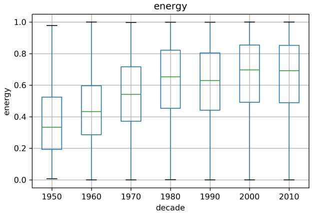


Figure: Energy evolution with time

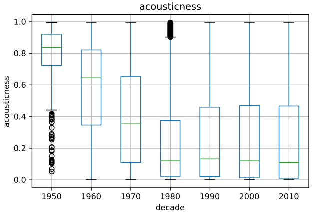


Figure: Acousticness evolution with time

Cross Validation

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

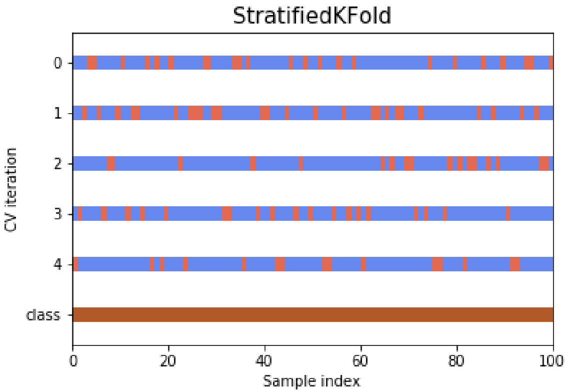


Figure: Different KFold Splits

Classification Models

Random Forest

Parameters tuned:

Parameter Name	Parameter Range	Optimal Parameter
max_depth	$1, 2 \ldots 10$	10
min_sample_split	$1, 2 \ldots 10$	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

Results: RandomForest

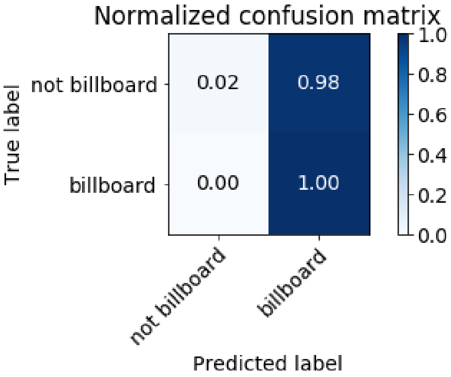


Figure: Confusion Matrix for Random Forest

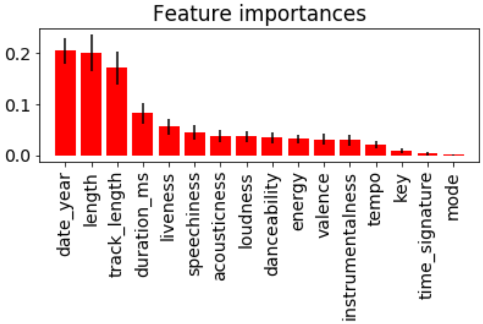


Figure: Feature Importance for Random Forest

Results: XGBoost

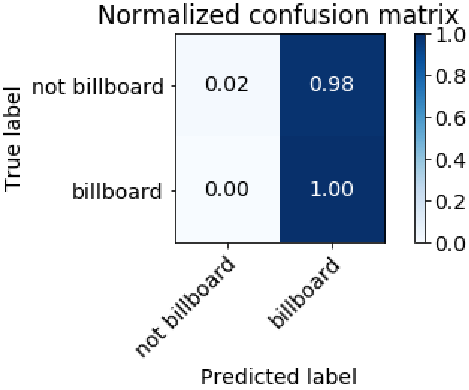


Figure: Confusion Matrix for XGBoost

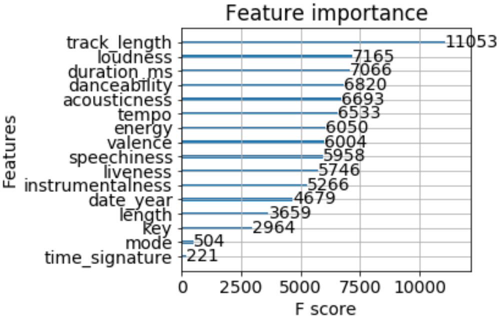


Figure: Feature Importance for XGBoost

Results: AdaBoost

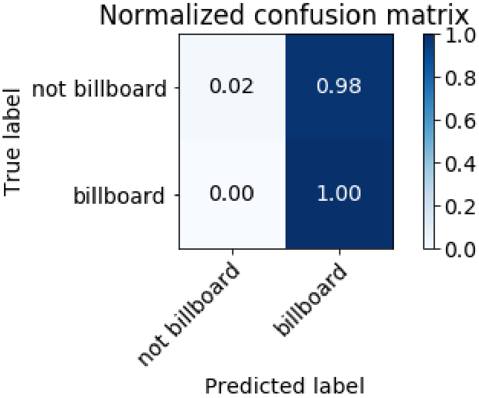


Figure: Confusion Matrix for AdaBoost

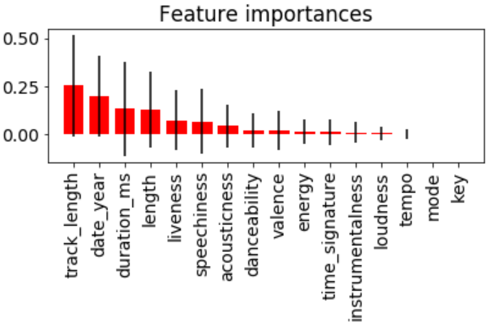


Figure: Feature Importance for AdaBoost

Results: AdaBoost

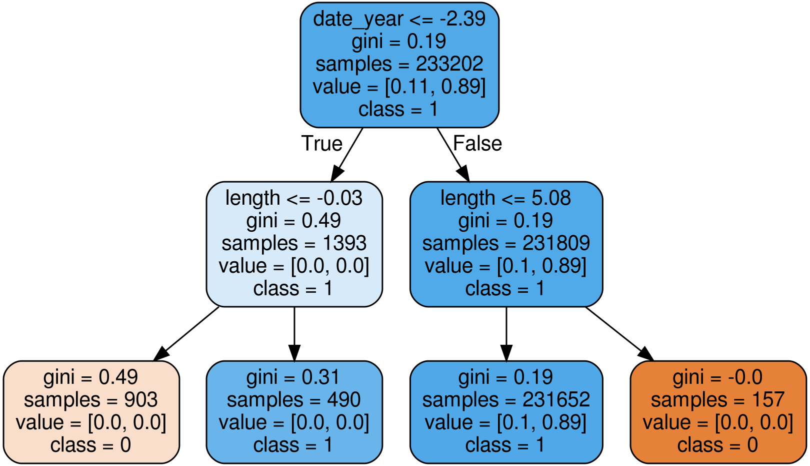


Figure: Estimator with least error

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.

Thank You. Question?