Predicting Music Genre & Popularity

Yucong Chen, Tianhui Zhao

Introduction

Music Information Retrieval (MIR)

(the research of studying information in music)

Applications

Music Classification
(genre, mood, artist, etc.)

Algorithmic Composition

With the goal of

Protecting Intellectual Property

Music Cognition, therapy

Music Development

This project aims to…

Classify Classical music

Predict Music Popularity

II

Description of Data

1. Data variables

Data set is from Kaggle (generated from Spotify API); ~40,000 observations with 11 variables.

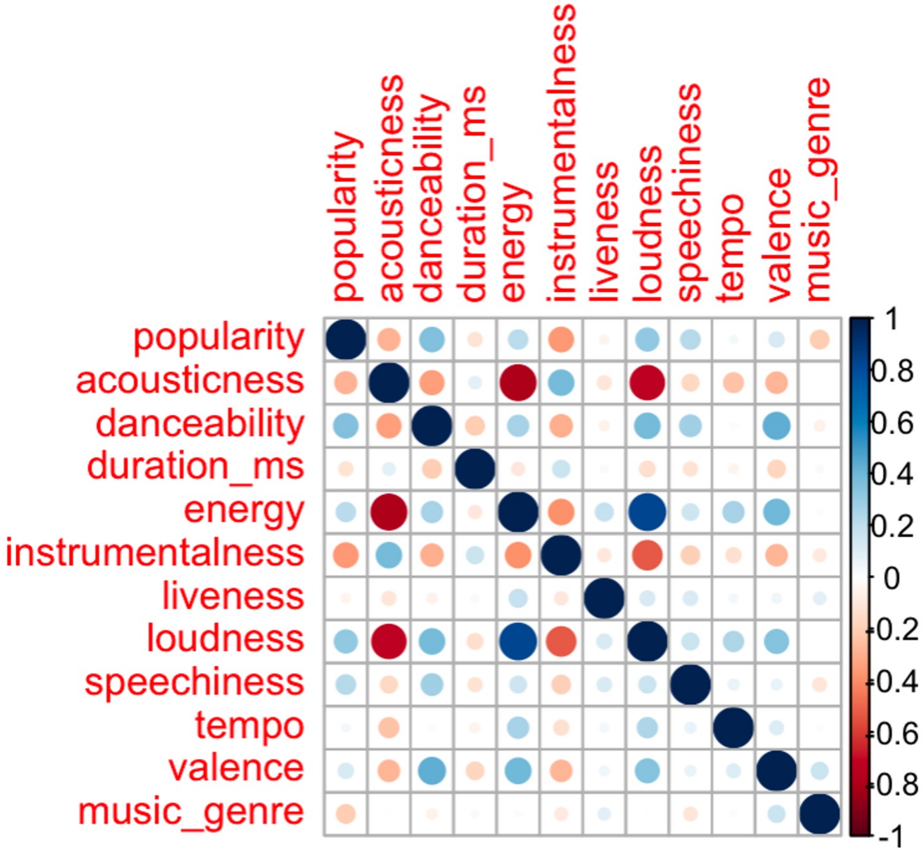
- **Popularity**: a number ranged from 0 to 100 representing how popular the song is; larger value means more popular.
- **Danceability**: a number ranged from 0 to 1 representing how likely can we dance through the music; larger value means more possible to dance.
- **Acousticness**: a number ranged from 0 to 1 measuring if the song uses instruments and no electronic components; larger value means more instruments.
- **Energy**: a number ranged from 0 to 1 measuring if the song makes you want to move forward; larger value means more energy.
- **Instrumental**: a number between 0 to 1 representing if the music is instrumental versus vocal; the higher the more instrumental the music is.
- **Liveness**: A number interpreting if the music is “live”; higher means more possibility.

1. Data variables

- **Valence**: measures positiveness; higher value means more positive.
- **Loudness**: records loudness in dB.
- **Speechiness**: Measures the presence of spoken words in a song.
- **Tempo**: How fast the song is in BPM (beats per minute).
- **Music Genre**:
  - We encode this data three times:
    - When we do the music classification models, we will have variable = 1 for classical music, 0 otherwise.
    - When we do K-means model for predicting music popularity (spoiler alert!), we will encode 1 for blues, 2 for anime, 3 for alternative, 4 for rock, 5 for rap, 6 for jazz, 7 for hip-hop, 8 for electronic, 9 for country, and 10 for classical
    - When we do other models for predicting music popularity, we treat it more like “time,” so that we can see how new and past music genres affect popularity. So 1 for classical, 2 for blues, rock, jazz, country, and 3 for anime, alternative, rap, hip-hop, electronic.

2. Correlation plot

(for music_genre takes value of 0 & 1)



- Music genre does not show significant relationship between any variables - not determined by one variable.
- Potential models: logit, kNN, classification trees, and more.
- Strong relationship between loudness and instrumentalness.
- Instrumentalness and valence seem to have strong relationship with other variables.
- Interesting: valence and music genre has a positive relationship (maybe subject to particular songs)

III.1

Music genre classification

Logit model - analysis

```
Coefficients:
                     Estimate     Std. Error z value                Pr(>|z|)
(Intercept)       1.1203048813  0.2639513331   4.244              0.00002192  ***
popularity       -0.0515586490  0.0018316034 -28.149 < 0.0000000000000002  ***
acousticness     -0.5871963587  0.1293019729  -4.541              0.00000559  ***
danceability     -2.3810369042  0.2036140037 -11.694 < 0.0000000000000002  ***
duration_ms       0.0000010078  0.0000002168   4.649              0.00000334  ***
energy           -1.9516031187  0.2350871971  -8.302 < 0.0000000000000002  ***
instrumentalness -2.1209237924  0.1182677417 -17.933 < 0.0000000000000002  ***
liveness          1.1959095785  0.1383227257   8.646 < 0.0000000000000002  ***
loudness          0.0312907082  0.0091158989   3.433                0.000598  ***
speechiness      -4.6914480005  0.4387415952 -10.693 < 0.0000000000000002  ***
valence           3.5800010113  0.1383572598  25.875 < 0.0000000000000002  ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13163  on 20279  degrees of freedom
Residual deviance: 10651  on 20269  degrees of freedom
AIC: 10673

Number of Fisher Scoring iterations: 6
```

−

Popularity
Acousticness
Danceability
Energy
instrumentalness
speechiness

+

Duration
Liveness
Loudness
valence

Results of logit model & LDA

logit

```
logit.pred     0     1
         0 18065  1921
         1   192   102
The training error is 0.104191321499014.
The correct up rate is 0.0504201680672269.
The correct down rate is 0.989483485786274
The Overall fraction of correct predictions is 0.895808678500986
```

LDA

```
LDA:
       0     1
  0 18036  1909
  1   221   114
The training error is 0.105029585798817.
The correct up rate is 0.0563519525457242.
The correct down rate is 0.987895053951909
The Overall fraction of correct predictions is 0.894970414201183
```

Results of QDA & kNN

QDA

```
qda.class     0     1
        0 16964  1032
        1  1293   991
The training error is
0.114644970414201.
The correct up rate is
0.489866534849234.
The correct down rate is
0.929177849591937
The Overall fraction of correct
predictions is 0.885355029585799
```

KNN

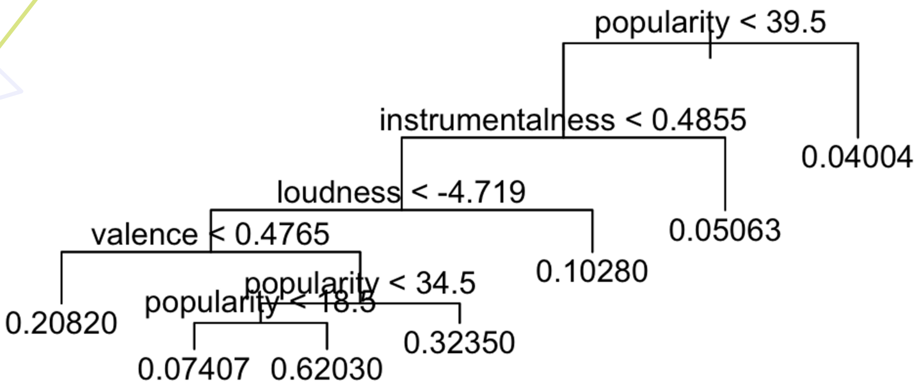
```
K = 1
knn.pred     0     1
       0 16506  1687
       1  1751   336
[1] 0.8304734

K = 10
knn.pred     0     1
       0 18241  2021
       1    16     2
[1] 0.8995562

K = 100
knn.pred     0     1
       0 18257  2023
       1     0     0
[1] 0.9002465
```

Results of logit model, LDA, QDA, kNN

	Correction Rate in Predicting Classical Songs
Logit	0.050
LDA	0.056
QDA	0.490
KNN (k=1)	0.166
KNN (k=10)	0.001
KNN (k=100)	0

Classification Tree

(Popularity plays an important role!)




```
tree.pred      0      1
        0 17364   1140
        1   893    883
The training error is
0.100246548323471.
The accuracy is
0.0997534516765286.
The correct up rate is
0.436480474542758.
The correct down rate is
0.951087254203867
```

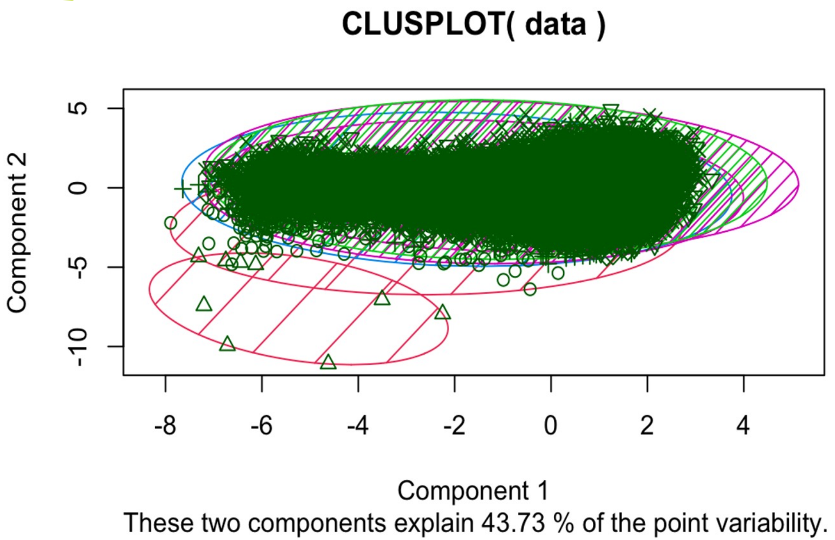
Conclusion:

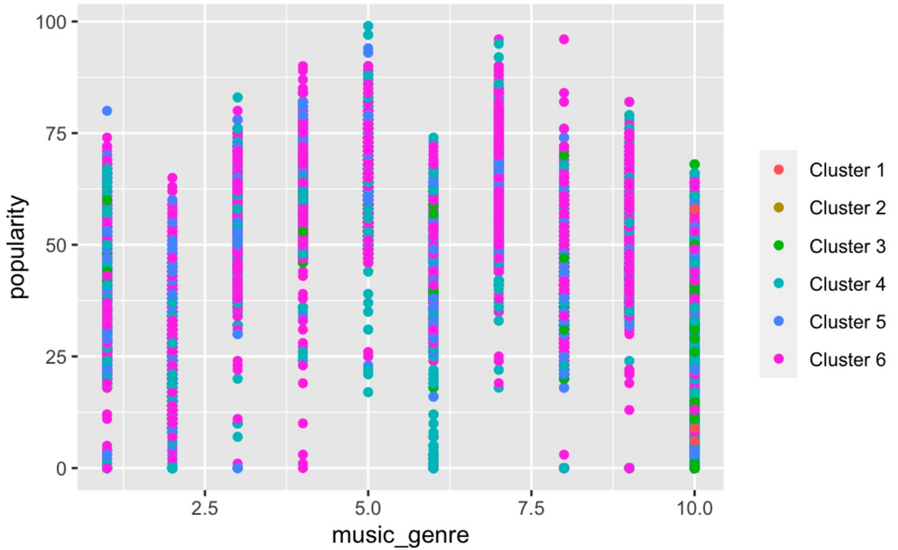
- Tree Model has best training error rate
- Best performing model is LDA, which has highest correct up rate in predicting classical songs.
- Popularity is essential in classifying → move next

III.2

Predicting popularity

K-means

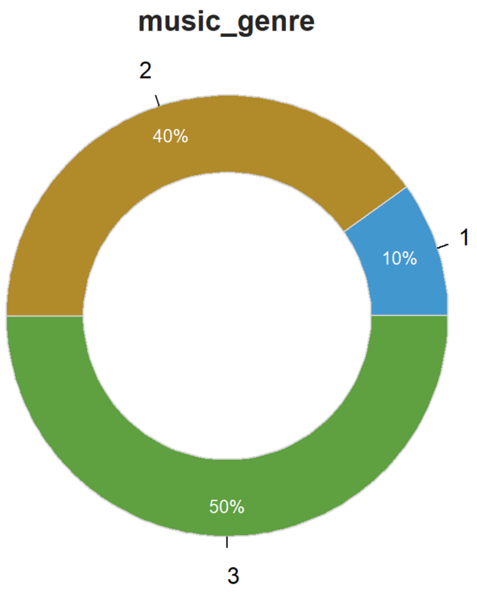




Blues	Anime	Alternative	Rock	**Rap**	Jazz	**Hip-Hop**	Electronic	Country	Classical
1	2	3	4	**5**	6	**7**	8	9	10

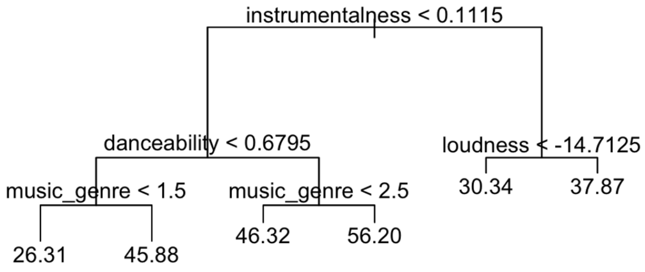
Rearrange Music Genre Signal

Blues	Anime	Alternative	Rock	Rap	Jazz	Hip-Hop	Electronic	Country	Classical
2	3	3	2	3	2	3	3	2	1

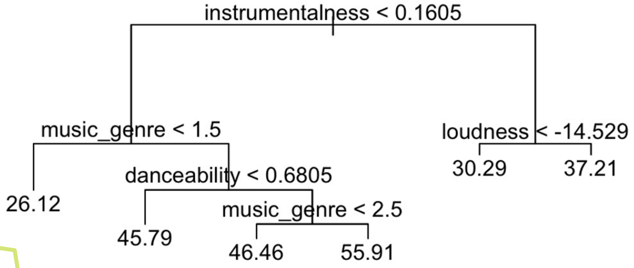


Regression Tree

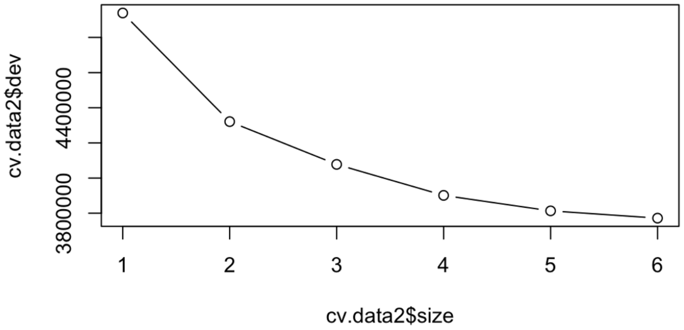
Regression Tree



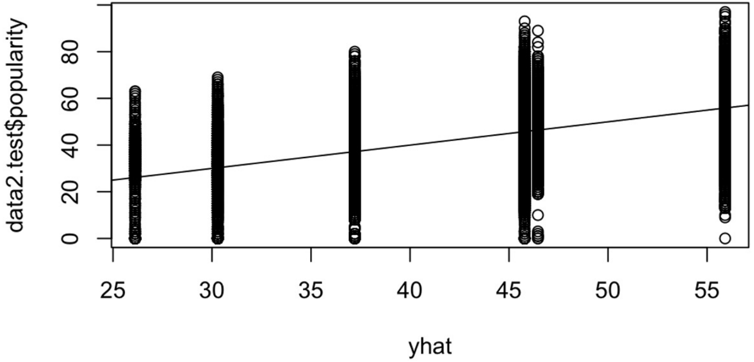
Pruned



Cross Validation



Test MSE: 185.4912



Random Forest

Call:
randomForest(formula = popularity ~ ., data = data2, mtry = 10, importance = TRUE, subset = train)
Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 10

Mean of squared residuals: 143.3584
% Var explained: 40.13

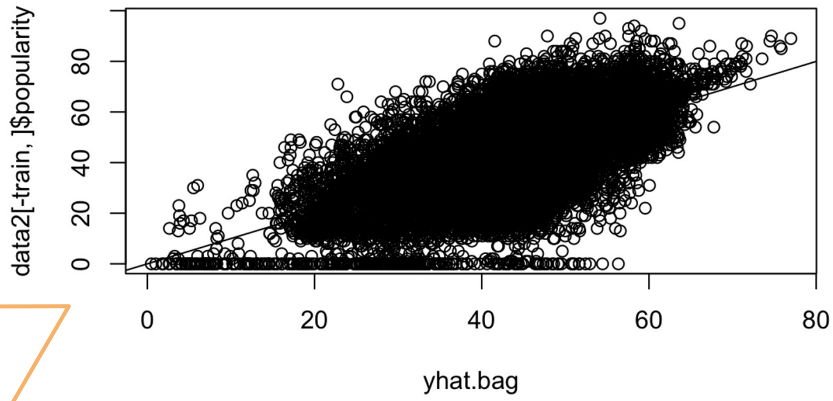
Comment:
- Does not fit data well - only 40% var explained;
- However, MSE is lower than regression tree
- In general, yields similar results with regression tree in terms of importance of variables

mse.rf10 = 142.3174
mse.rf6 = 141.7925

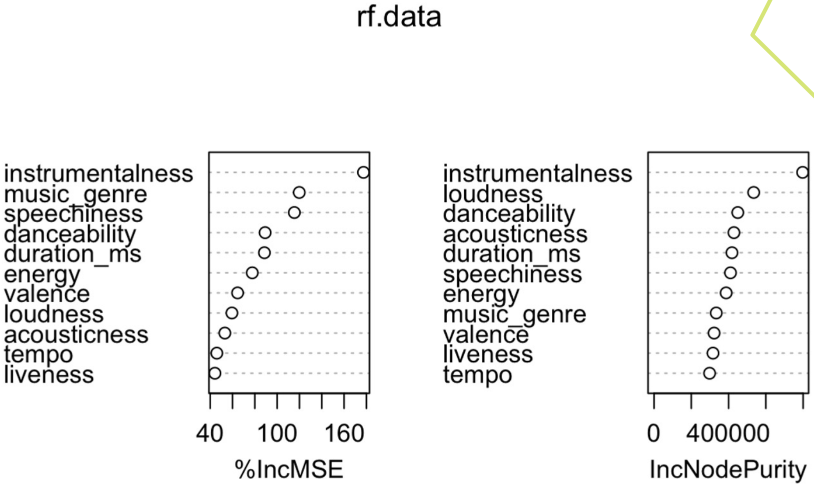
	%IncMSE	IncNodePurity
acousticness	53.12549	428671.5
danceability	89.21148	449387.1
duration_ms	88.58486	418156.8
energy	77.70694	386965.2
instrumentalness	177.41350	797470.0
liveness	44.20859	316378.6
loudness	59.45598	534698.1
speechiness	115.41905	410276.7
tempo	45.94167	298115.9
valence	64.56217	322327.5
music_genre	119.83352	333873.3

Random Forest

Fitting



Importance



GAM

```
Analysis of Deviance Table

Model 1: popularity ~ s(speechiness, 2) + instrumentalness
Model 2: popularity ~ s(speechiness, 2) + instrumentalness + music_genre
Model 3: popularity ~ s(speechiness, 10) + instrumentalness + music_genre

  Resid. Df Resid. Dev     Df Deviance       F              Pr(>F)
1     20276    4097281
2     20275    4046752 1.0000    50529 253.8285 < 0.0000000000000000022 ***
3     20267    4034474 7.9996    12278   7.7103       0.000000002265 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model	***Test RMSE***
Model 1	**16.60855**
Model 2	16.69624
Model 3	16.70890

IV

Conclusion + Further work

Conclusion and Further work

- For predicting music genre:
  - LDA performs best.
  - Random forecast gives desirable prediction in popularity.
  - Even though tree models have straightforward presentation, their accuracy is not well enough to be considered.
- For predicting popularity:
  - Random forest is the best in terms of MSE.
- Need more data! Perhaps data on years of release.
- May cause more confusion if one song has more than one genre.

Thank you!

Data set from Kaggle: https://www.kaggle.com/vicsuperman/prediction-of-music-genre