



Spotify Song Attributes

An attempt to build a classifier
that can predict whether or not I like a song

**노래를
골라
판단해조**

김도윤 김동훈 최혜린 이재호



Home



Browse



Radio

YOUR LIBRARY



Q Search



목차

Play

0. 주제 선정 배경

1. 데이터 셋 소개

2. 모델 수립과 종류

3. 모델 성능 평가 및 선택

4. 최종 모델 결과 해석

5. Test data set 성능 평가

6. Follow up action



New Playlist



주제 선정 배경

Play

언어적 표현 yes/no

음악을 즐김 : 감정, 기분, 느낌의 움직임을 즐김

음악의 선호도 판별

장르, 아티스트 영향

음악의 선호를 Linguistic 한 방법이 아닌

수치적 속성들을 통해 판별할 수 있을까?



Home



Browse



Radio

YOUR LIBRARY

주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action



Q Search



데이터 셋 소개

Play

해외 유명 음악 스트리밍 사이트
'Spotify'

음량, 길이, 가사와 음이 있는 소리의 비율, 악기 소리의 비율 등
수치로 나타낼 수 있는 속성들에 대한 정보들을 공개 중



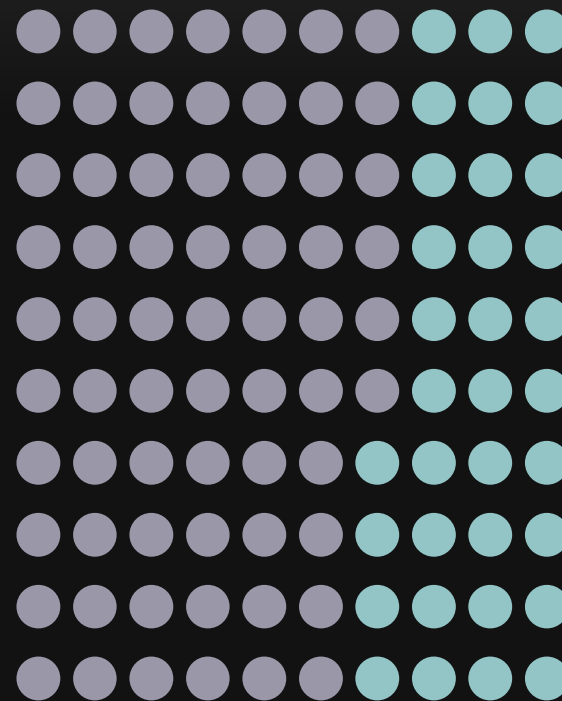
New Playlist



데이터 셋 소개

Play

Instance



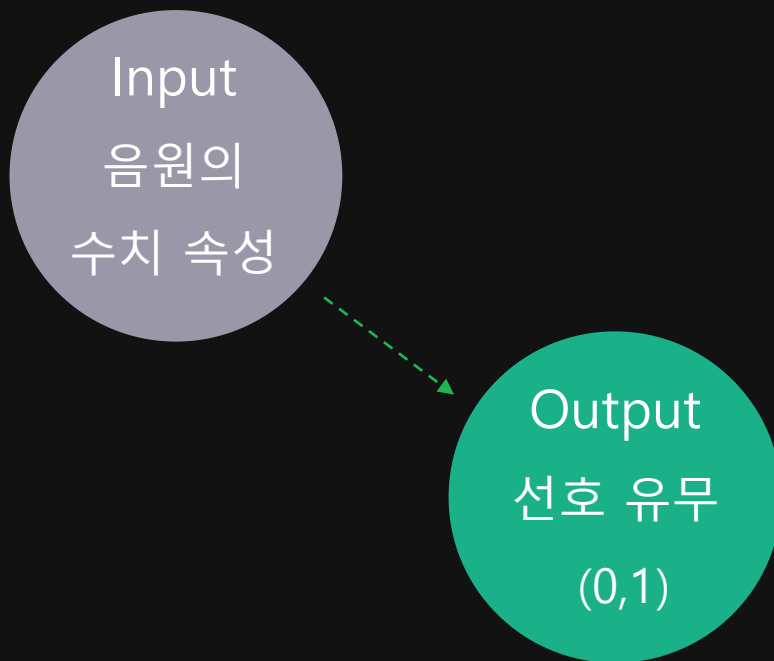
1020개

선호 음원

997개

비선호 음원

George McIntire의 데이터 셋



주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action



데이터 셋 소개

Acousticness	float	음원이 악기 소리로 이루어져 있는지
Danceability	float	춤추기에 얼마나 적절한지 (템포, 리듬의 일정성, 비트의 강도)
Duration_ms	int	음원의 길이
Energy	float	빠르고, 강하고, 시끄러운 정도
Instrumentalness	float	가사가 아닌 음가가 있는 소리
Key	int	음원의 조성
Liveness	float	공연 현장의 녹음으로 이루어져있는 정도



데이터 셋 소개

loudness	float	음원의 크기 (단위 :dB)
mode	int	장조 =1 / 단조 = 0
speechiness	float	가사의 비율 / 0.66< : 내레이션 / (0.33,0.66) : 배경음과 가사가 같이 공존 / 0.33> : 가사가 없음
tempo	float	음원의 빠르기 (BPM)
time_signature	int	how many beats are in each bar / 4beat, 8beat, 16beat 등
valence	float	음원의 분위기 / 긍정적일수록 높음 / 어두울수록 낮음



데이터 셋 소개

다음 변수들에서 특정한 패턴을 확인할 수 있었음

주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action

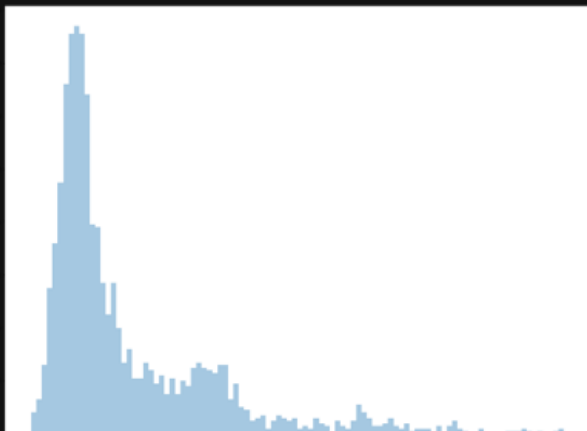
Intsrumentalness

0~0.01사이의 관측치 개수가
1,569개로 절반 이상



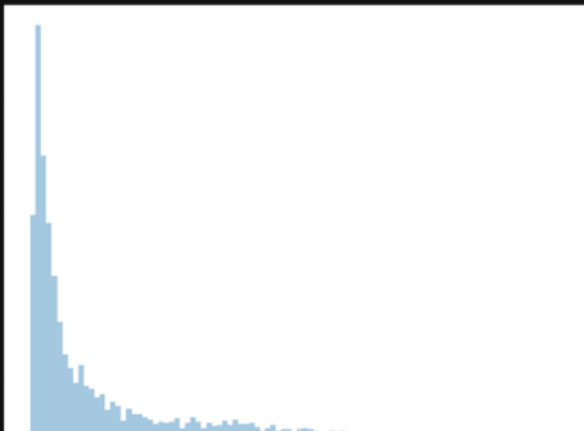
Liveness

0.1 주변의 관측치 개수가
전체 분포에 많은 비중을 차지



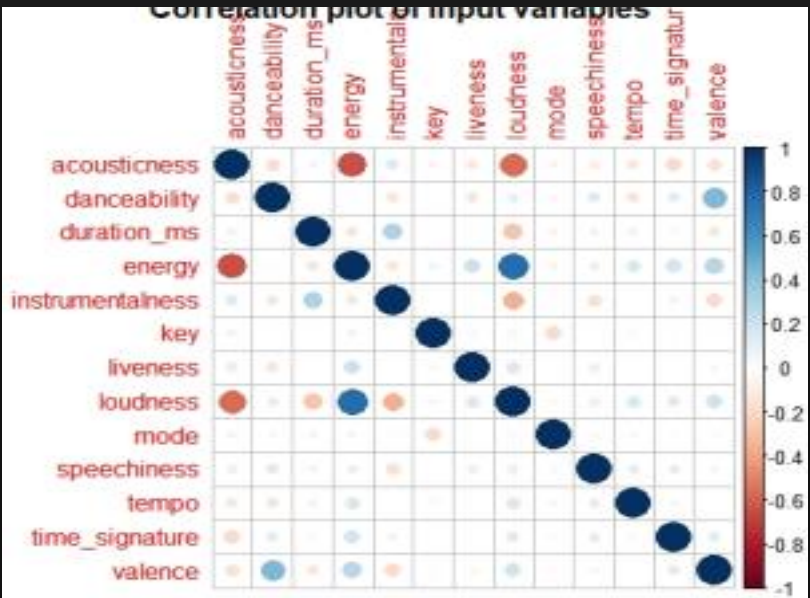
Speechiness

right-skewed 되어 있음





데이터 셋 소개



상관 관계가 높은 변수들이 일부 존재

> >vif(LR.trn)

acousticness

1.890721

danceability

1.465879

duration_ms

1.143158

instrumentalness

1.315264

key

1.054382

liveness

1.083151

mode

1.054841

speechiness

1.083285

tempo

1.115848

valence

1.434047

time_signature

1.059941

loudness

3.292487

energy

3.513963

VIF test 결과 그 수치가 10을 넘는 변수가 없음을 확인
즉 다중공선성을 나타내는 변수가 없음.
특히 변수의 제거를 실시하지 않음



모델 수립과 종류

절대적으로 많은 것은 아니나 한 개인에 대한 데이터로는 충분하다고 판단



■ Training

■ Validation

■ Test

→ 이 중 Training data를 통해 모델을 수립

Logistic Regression

Shrinkage : Ridge, LASSO, Elastic Net

Decision Tree : Gini Index, Deviance

Random Forest

ANN : Full, 5-fold validation



주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action



Home



Browse



Radio

YOUR LIBRARY

주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action



Q Search



모델 성능 평가 및 선택

Validation Data Set 모델 성능을 평가

Confusion Matrix 4가지 평가지표 : Recall, Precision, Accuracy, F1-Measure

시각적 해석 지표 : AUC

“이 중 가장 중요하다고 판단한 지표는 F1으로 선정.”



New Playlist



모델 성능 평가 및 선택

주제 선정 배경

데이터 셋 소개

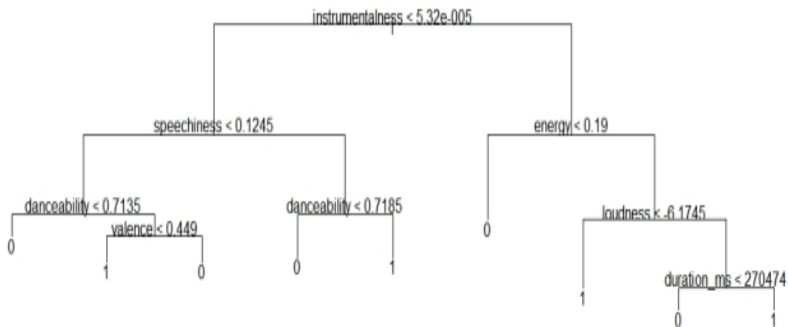
모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

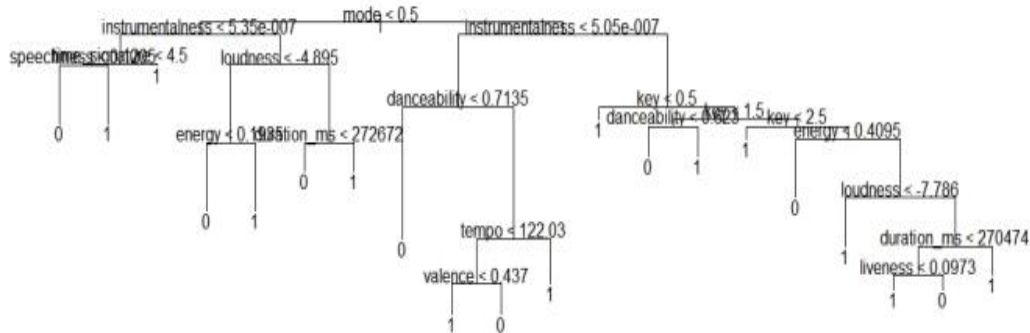
Test data set 성능 평가

Follow up action



[DT-Deviance]

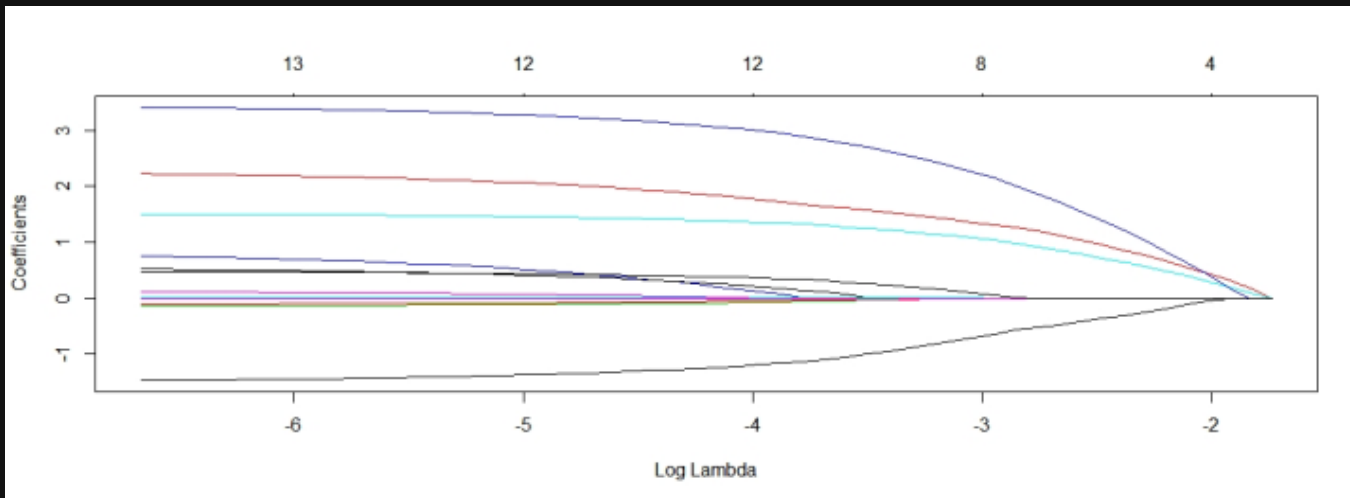
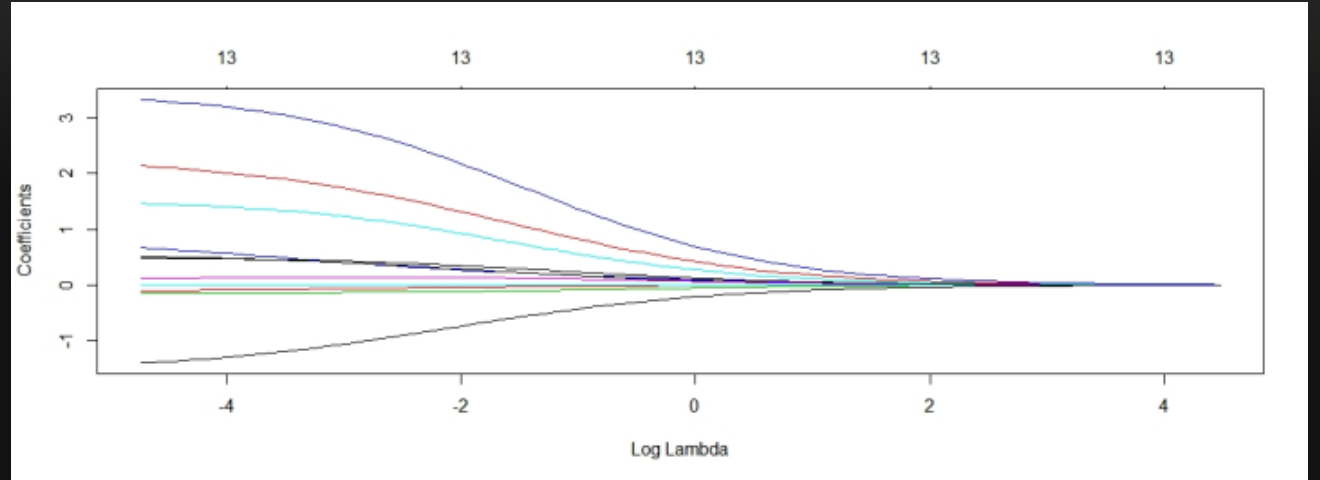
[DT-Gini Index]





모델 성능 평가 및 선택

[Ridge]



[Lasso]

주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action



모델 성능 평가 및 선택

주제 선정 배경

데이터 셋 소개

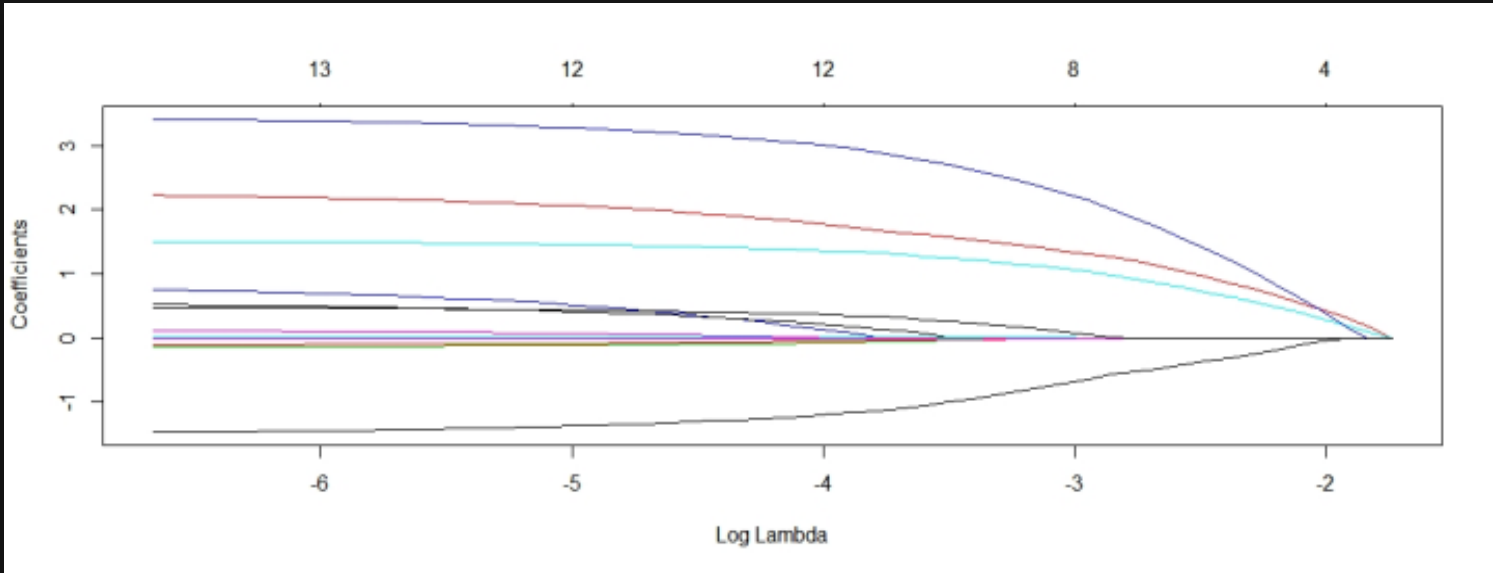
모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

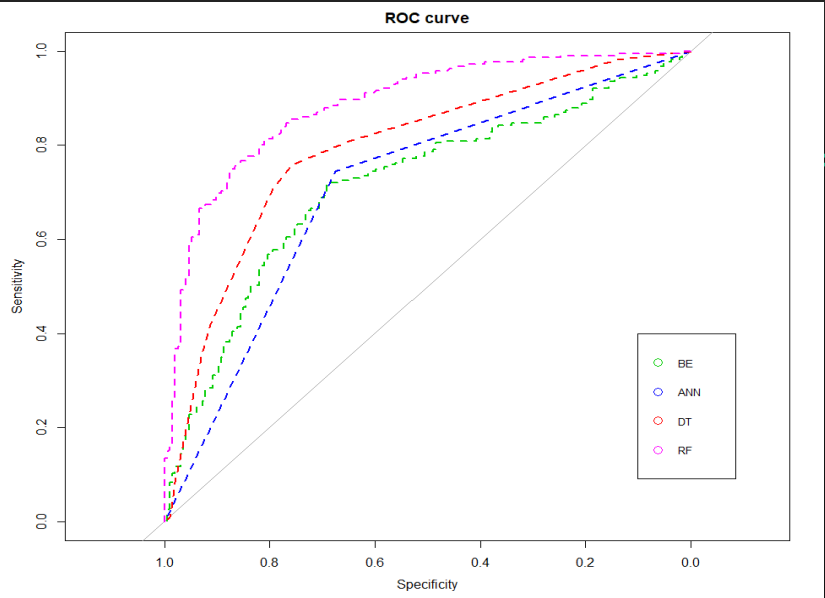
Test data set 성능 평가

follow up action



[Elastic Net]

모델 성능 평가 및 선택



← (pink : RF, red : DT, blue : ANN, green : BE)

Random Forest 모델이 가장 좋은 성능

	Recall	Precision	Accuracy	F1-Measure	AUC
Random Forest	0.7377778	0.8217822	0.7721823	0.7775176	0.8461458
Decision Tree(Gini)	0.6755556	0.7342995	0.6930456	0.7037037	0.7361111
ANN	0.6800000	0.7285714	0.6906475	0.7034483	0.6915625
LASSO	0.6355556	0.7447917	0.6858513	0.6858513	0.6901736
Elastic Net	0.6355556	0.7409326	0.6834532	0.6842105	0.6875694
ANN (5-fold)	0.6581197	0.6363636	0.6514523	0.6470588	0.6516405
Decision Tree	0.6133333	0.7709497	0.6930456	0.6831683	0.7482523
Ridge	0.6266667	0.7382199	0.6786571	0.6778846	0.6831250
Logistic Regression	0.6311111	0.7319588	0.6762590	0.6778043	0.6910880



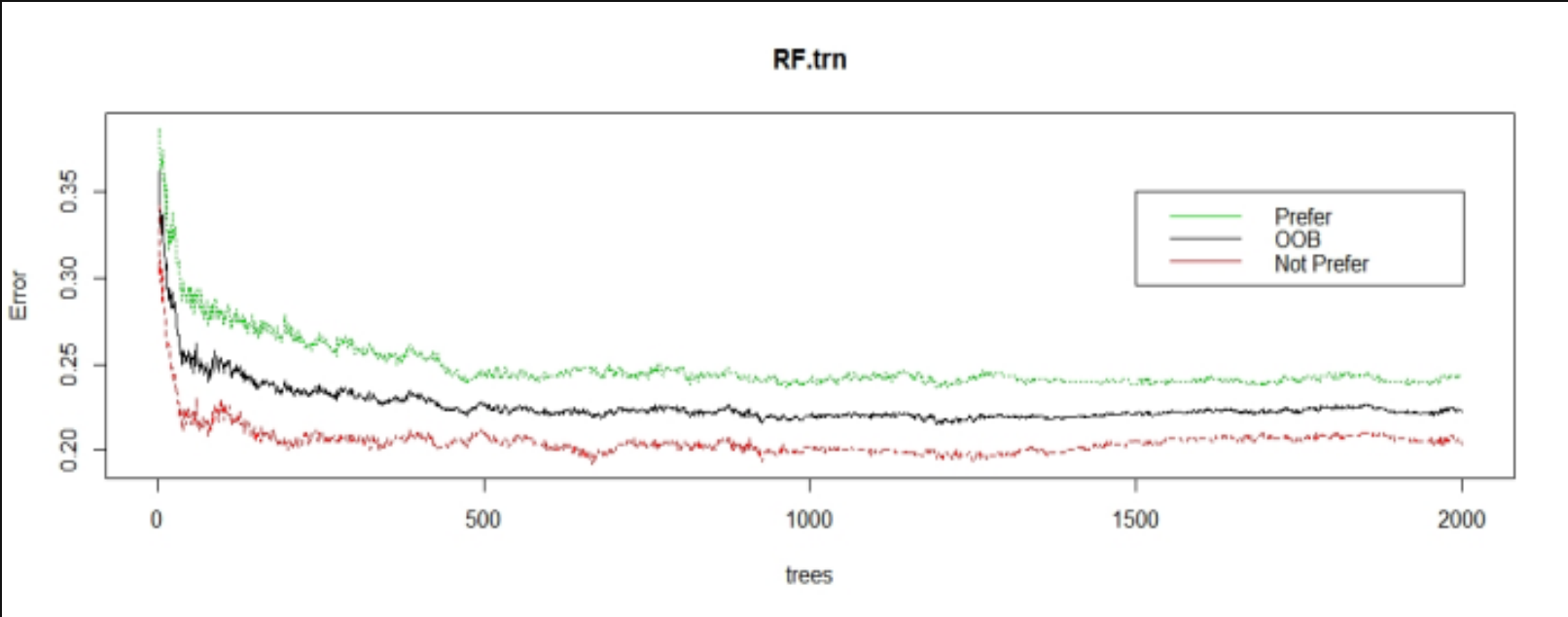
최종 모델 결과 해석

randomForest
(formula = target ~ ., data = df.trn, ntree = 2000, mtry = 3)
Type of random forest: classification
Number of trees: 2000
No. of variables tried at each split: 3

OOB estimate of error rate: 22.19%

Confusion matrix:

	0	1	class.error
0	492	123	0.2000000
1	145	448	0.2445194



주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action



최종 모델 결과 해석

주제 선정 배경

데이터 셋 소개

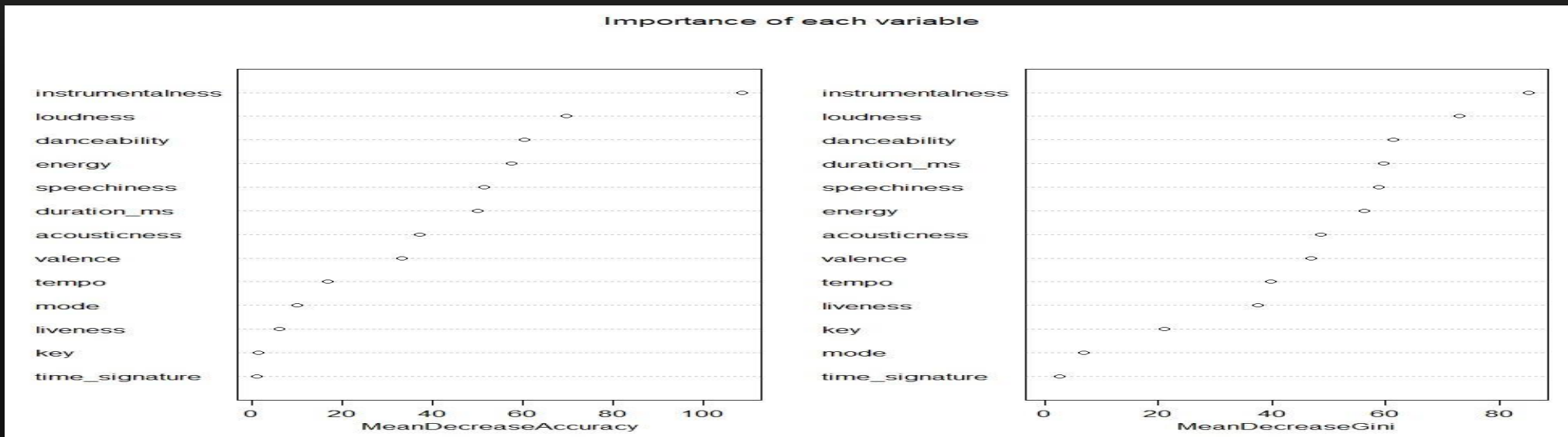
모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action



	MeanDecreaseAccuracy
Instrumentalness	111.6863905
Loudness	71.7887672
Danceability	58.1483537
Energy	56.9498823
...	...

	MeanDecreaseGini
Instrumentalness	75.889884
Loudness	70.133236
Speechiness	61.461251
Danceability	61.357943
...	...

🔊 Instrumentalness, Loudness, Danceability 가장 중요한 변수



최종 모델 결과 해석

주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action

[Coefficients of LASSO Model]

(Intercept)	-4.129548e+00
Acousticness	-1.410578e+00
Danceability	2.092996e+00
Duration_ms	2.321664e-06
Energy	5.278189e-01
Instrumentalness	1.476618e+00
Key	.
Liveness	4.108626e-01
Loudness	-8.860980e-02
Mode	-1.114590e-01
Speechiness	3.318261e+00
Tempo	4.486710e-03
Time_signature	6.278235e-02
Valence	4.349544e-01

Rythmical 할수록

배경음악 소리가 많을수록,

소리가 크지 않을 수록

해당 음악을 선호할 확률이 높아짐.



Home



Browse



Radio

YOUR LIBRARY

주제 선정 배경

데이터 셋 소개

모델 수립과 종류

모델 성능 평가 및 선택

최종 모델 결과 해석

Test data 성능 평가

Follow up action



New Playlist



Q Search



Test data Set 성능 평가

	Recall	Precision	Accuracy	F1-Measure	AUC
Random Forest	0.8069307	0.815	0.8061224	0.8109453	0.8684211

80% 이상의
높은 예측률

따라서 우리는 음악의 수치적 속성으로도
개인의 음악의 선호를 충분히 판별할 수 있음을 확인 !



Home



Browse



Radio

YOUR LIBRARY



Search



Follow Up Action

음악의 수치적 속성으로도 개인의 음악의 선호를 충분히 판별할 수 있음을 확인

Instance를 개개인에 맞추어 변경한 뒤 모델을 학습

-> 나만의 음악 선호 판별 모델을 만들 수 있음

아티스트, 장르 등의 음원 Background 정보 없이 수치적 속성으로 만든 모델

-> 음악에 대한 편견과 선입견 없이 즐길 수 있음

나아가 음악 추천 시스템의 핵심 모델로도 발전시킬 수 있음

주제 선정 배경

데이터 셋 소개

모델 수립과 종류

성능 평가 및 최종 선택

최종 모델 결과 해석

Test data set 성능 평가

Follow up action



New Playlist