K-Pop 장르 분류

Spotify 데이터를 이용한
세부 장르 예측

목 차

조원 소개

데이터 수집

이은서 [조장]

- 데이터 수집
- 데이터 전처리
- PPT 제작

이종혁

- 데이터 수집
- 데이터 전처리
- 모델링

모델링

김원진

- 데이터 수집
- 모델링

류가연

- 데이터 수집
- 모델링
- PPT 제작

평가지표

김사무엘

- 데이터 수집
- 평가지표

이정우

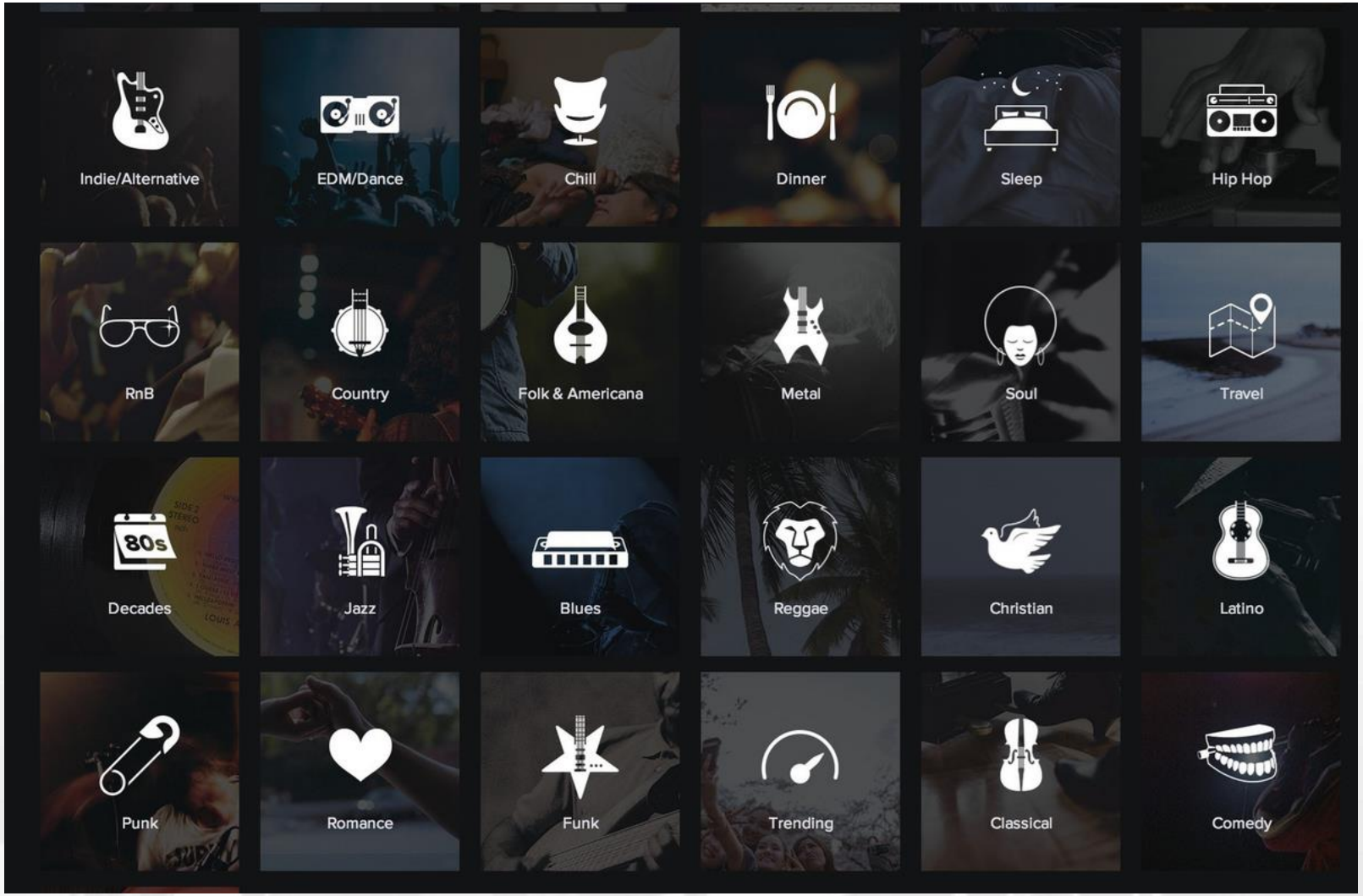
- 데이터 수집
- 평가지표

주제 선정 배경

종합	장르종합	국내종합	해외종합						
한국대중음악	발라드	댄스	랩/힙합	R&B/Soul	인디음악	록/메탈	트로트	포크/블루스	
해외POP음악	POP	록/메탈	일렉트로니카	랩/힙합	R&B/Soul	포크/블루스/컨트리			
그외인기장르	OST	재즈	뉴에이지	J-pop	월드뮤직	CCM	어린이/태교	종교음악	국악

▲ 멜론의 장르 분류 체계

주제 선정 배경



▲ 스포티파이의 장르 분류 체계

주제 선정 배경

음원 사이트마다
음악 장르 분류 체계
모호

Spotify 장르 분류를 이용하여
K-Pop 장르 분류

데이터 수집

- **Spotify API** 를 이용하여 데이터 수집
- 추천 장르 목록 **122개 중 64개 장르**를 선택
- **각 장르별 1,000곡씩 수집**하여 훈련용 데이터셋 생성

	id	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	genre	genre2
0	5ydPIlHBSV0w2Mo2ichlmR	0.681	0.500	10	-8.854	1	0.0294	0.566	0.000000	0.1120	0.3060	119.993	257333	4	acoustic	acoustic
1	3cnRYoW4nYzluVPCp7k5iG	0.773	0.394	0	-9.252	1	0.0516	0.627	0.000014	0.0953	0.4430	74.977	212707	4	acoustic	acoustic
2	0HLWvLKQWpFdPhgk6ym58n	0.584	0.509	2	-13.109	1	0.0287	0.267	0.002520	0.3600	0.3670	90.457	175480	4	acoustic	acoustic
3	6EGlya6vCQzo3PdutOJEJP	0.640	0.355	9	-8.579	1	0.0292	0.637	0.000120	0.1080	0.2080	127.874	245840	4	acoustic	acoustic
4	2Oehrcv4Kov0SuIgWyQY9e	0.327	0.710	3	-2.928	1	0.0547	0.202	0.000090	0.2800	0.4160	179.561	175200	4	acoustic	acoustic
5	78wvHiOarfOut310PnUnlr	0.479	0.305	11	-9.328	1	0.0509	0.914	0.000000	0.1480	0.0953	140.086	238505	4	acoustic	acoustic
6	4RL77hMWUq35NYnPLXBpih	0.379	0.290	4	-8.485	1	0.0510	0.952	0.001060	0.1180	0.1690	166.467	201080	4	acoustic	acoustic
7	2QaaXYzEzLio3POCFCI19r	0.644	0.337	7	-10.906	1	0.0295	0.606	0.016900	0.3010	0.3100	108.041	236080	4	acoustic	acoustic
8	3eOYmWvienElgHJFAWE3ZD	0.671	0.426	1	-9.745	0	0.0293	0.359	0.000091	0.1250	0.6130	112.030	215906	4	acoustic	acoustic
9	3xEP8kzYIjbkUrjfsgn8Hh	0.655	0.660	0	-6.544	1	0.0274	0.227	0.000000	0.0549	0.4500	116.959	248920	4	acoustic	acoustic

▲ 생성된 훈련용 데이터셋

데이터 수집

122개 장르에서
64개 장르를
선정한 기준은 ?

번호	장르	선택/삭제(O/X)	솎아내기
100	rock	O	
101	rock-n-roll	O	
102	rockabilly	O	
103	romance	X	2
104	sad	X	2
105	salsa	X	1
106	samba	X	1
107	sertanejo	X	1
108	show-tunes	X	3
109	singer-songwriter	O	
110	ska	O	
111	sleep	X	2
112	songwriter	O	
113	soul	O	
114	soundtracks	X	6
115	spanish	X	2
116	study	X	2
117	summer	X	2
118	swedish	X	1
119	synth-pop	O	
120	tango	O	
121	techno	O	
122	trance	O	

[제외할 장르의 기준]

1. 지역 특성이 짙은 장르
2. 국내 음악과 상관이 없는 느낌의 장르
3. 분위기 장르
4. 시대적 장르
5. 기타

선정한 64개 장르를
다시 28개의 장르로
재그룹화

번호	장르_대분류	소분류_번호	장르_소분류
1	acoustic	1	acoustic
		2	guitar
2	alternative	3	alternative
		4	grunge
3	dance	5	dance
		6	dancehall
4	electronic	7	deep-house
		8	dubstep
		9	edm
		10	electro
		11	electronic
		12	house
		13	idm
		14	industrial
		15	progressive-house
		16	trance
		17	chicago-house
5	hip-hop	18	hip-hop
		19	trip-hop

[그룹화할 장르의 기준]

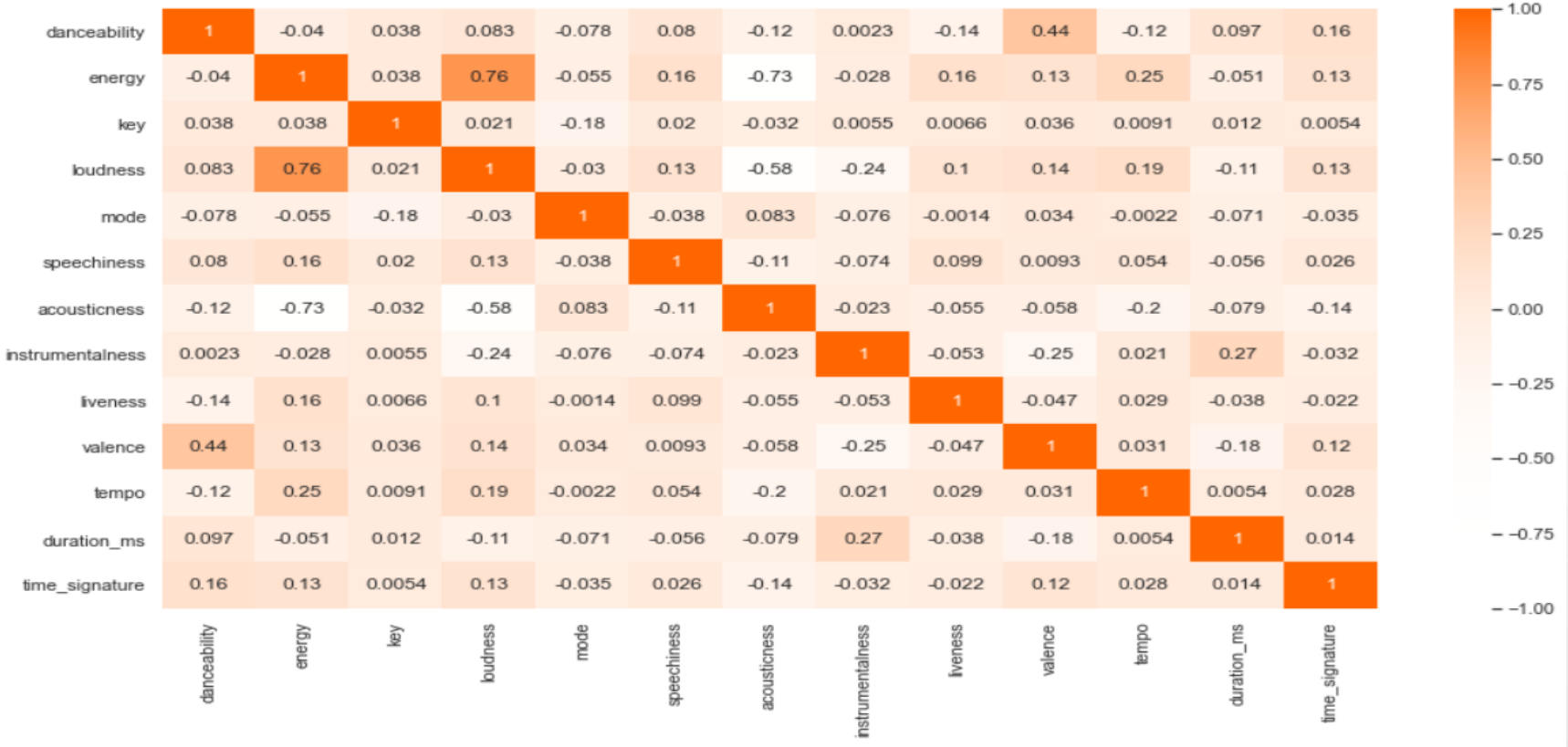
1. 대분류로 그룹화할 장르는
   보편적으로 하위 장르의 특성을
   포함하고 있어야 함
2. 상식적으로 누구나 인정할 수 있는
   대표장르를 대분류로 지정

변수 설명

변수명	설명	변수명	설명
id	곡의 이름	**instrumentalness**	악기의 연주길이
danceability	노래가 춤추기에 좋은가	**liveness**	1에 가까울수록 라이브음원
energy	1에 가까울수록 활기참	**valence**	1에 가까울수록 밝음
key	노래의 Key	**tempo**	BPM
loudness	0에 가까울수록 시끄러움	**duration_ms**	음악의 길이
mode	1장조음악,0단조음악	**time_signature**	박자
speechiness	보컬(나레이션)의 비중	**genre**	곡 장르
acousticness	노래의 어쿠스틱함 정도	**genre2**	대분류 장르[타겟 변수]

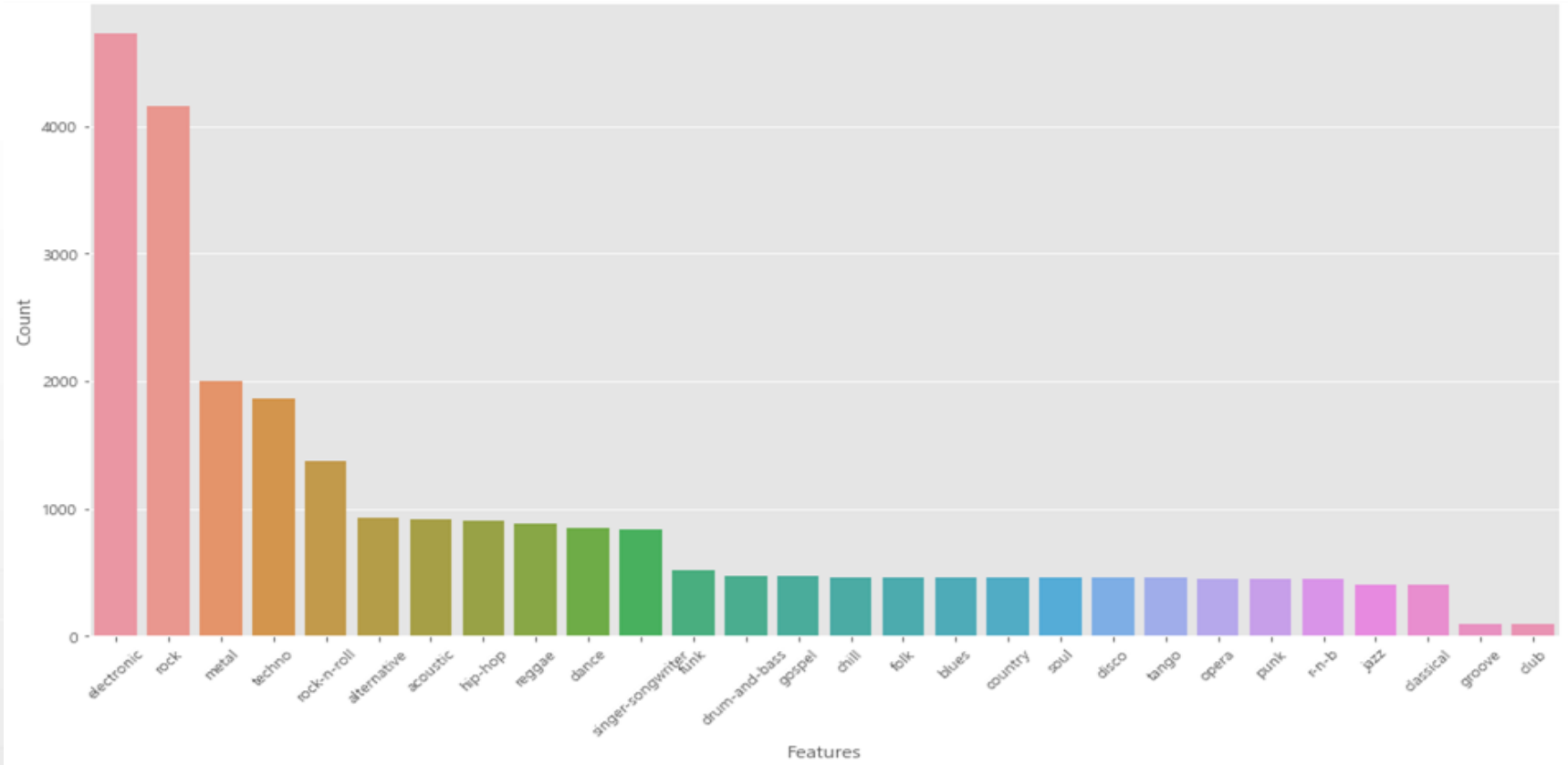
출처 : https://developer.spotify.com/documentation/web-api/reference/#/operations/get-audio-features

EDA



▲ Feature 상관 계수

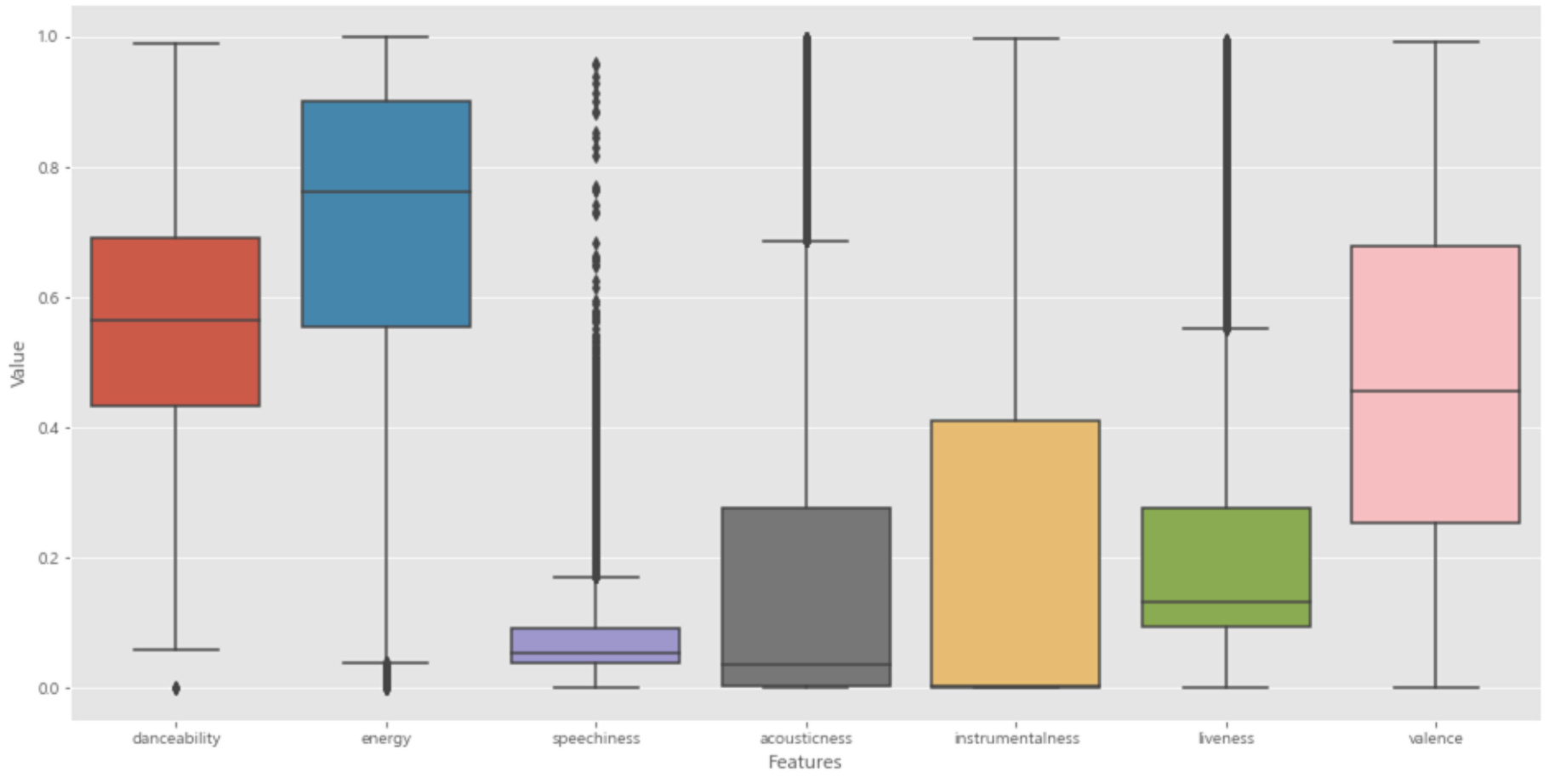
EDA



▲ 장르별 개수

	genre2_name	genre2_count
0	acoustic	920
1	alternative	929
2	blues	459
3	chill	462
4	classical	405
5	club	100
6	country	457
7	dance	841
8	disco	456
9	drum-and-bass	469
10	electronic	4725
11	folk	462
12	funk	517
13	gospel	468

EDA



▲ 오디오 Feature Box Plot

모델링

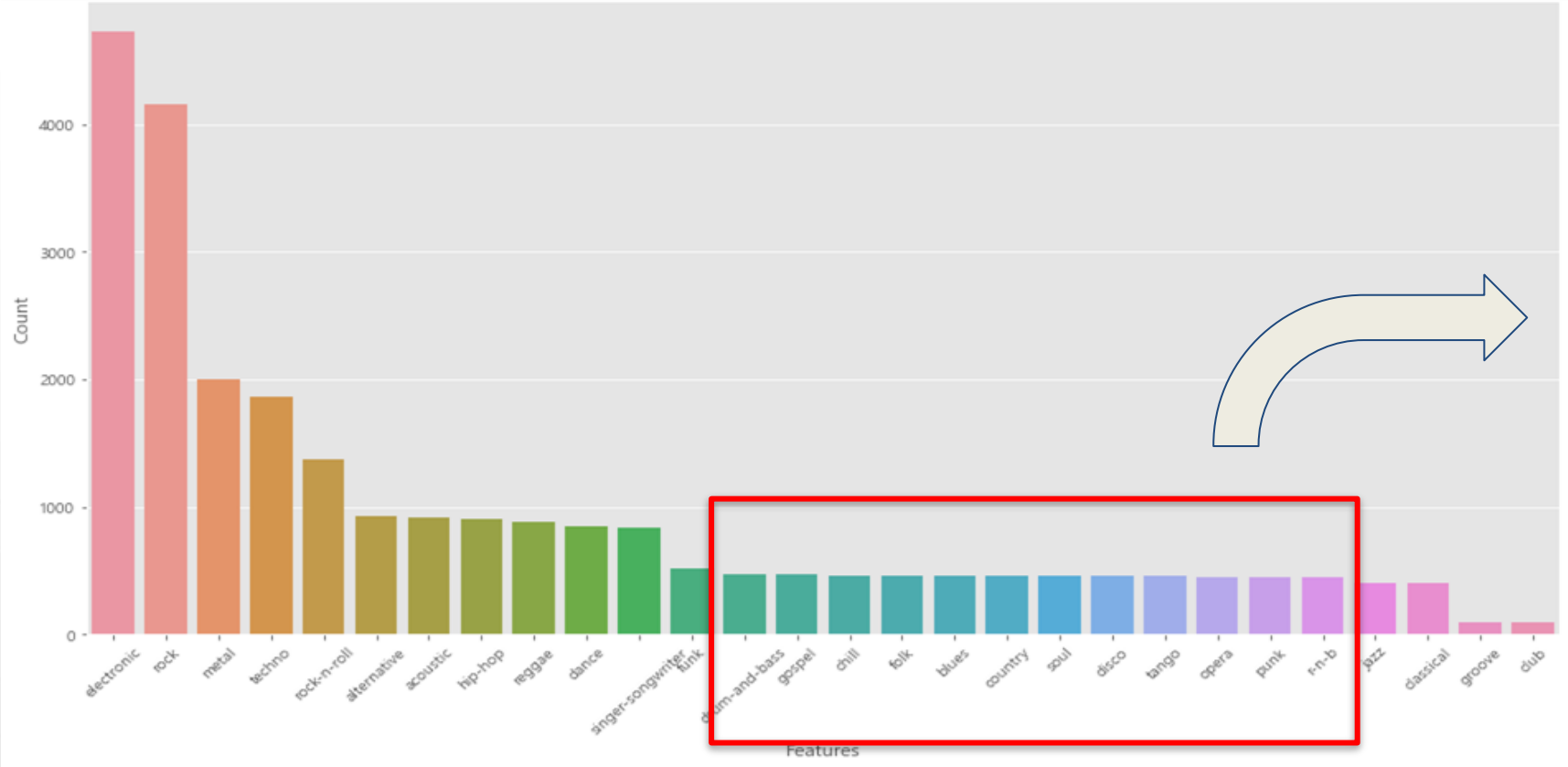
1) 기본 모델링

사용 모델	LightGBM	LightGBM	RandomForest
CV	K-Fold	Stratified K-Fold	K-Fold
샘플링	-	-	-
Train Best Score	**0.4386**	**0.43756**	**0.2826**
Validation Best Score	**0.431**	**0.4345**	**0.2824**

→ 스코어가 낮음

모델링

빈도수가 균일한 장르로만 모델링을 하면?



▲ 장르별 개수

사용 모델	LightGBM
CV	K-Fold
Train Best Score	0.71868
Validation Best Score	0.73967

2) Split → Sampling (Over/Under)

사용 모델	LightGBM	LightGBM	LightGBM
CV	K-Fold	K-Fold	K-Fold
샘플링	SMOTE (Oversampling)	ENN (Undersampling)	TOMEK LINKS (Undersampling)
Train Best Score	**0.7234**	**0.8455**	0.4343
Validation Best Score	**0.3546**	**0.7500**	0.4446

→ 과적합

3) Sampling → Split

복합 샘플링 (Combining Under- and Over-sampling) : 오버 샘플링과 언더 샘플링을 섞어 사용하는 기법

사용 모델	LightGBM	LightGBM	LightGBM	LightGBM	LightGBM
CV	K-Fold	K-Fold	Repeated K-Fold	Stratified Repeated K-Fold	K-Fold
샘플링	SMOTE	SMOTE & TOMEK LINKS (Over- + Under-sampling)	SMOTE & TOMEK LINKS (Over- + Under-sampling)	SMOTE & TOMEK LINKS (Over- + Under-sampling)	SMOTE & ENN
Train Best Score	**0.68227**	**0.66903**	0.669	0.6699	**0.881188**
Validation Best Score	**0.693644**	**0.68254**	0.6825	0.6825	**0.89464**

- *SMOTE 적용 후 레이블 값 분포*

0 4726
1 4726
26 4726
25 4726
24 4726
.
.
.
.
.
3 4726
10 4726
2 4726
22 4726

- *SMOTE & TOME LINKS 적용 후 레이블 값 분포*

0 4726
9 3889
4 3727
18 3626
6 3330
19 3325
23 3297
10 3228
.
.
.
.
25 3052
13 3048
8 3041
22 3033
12 2909

- *SMOTE & ENN 적용 후 레이블 값 분포*

0 4726
9 717
4 212
26 207
6 200
.
.
.
1 57
24 54
16 48
10 40
3 28
22 6

3-1) 최종 선정 모델

사용 모델	LightGBM	XGBoost
CV	K-Fold	K-Fold
샘플링	SMOTE & TOMEK LINKS	SMOTE & TOMET LIKNS
Train Best Score	0.6690	0.6555
Validation Best Score	0.6825	0.6720

Score : LightGBM > XGBoost

3-2) 최종 모델 CV 비교

사용 모델	LightGBM	LightGBM
CV	**K-Fold**	**Stratified K-Fold/ Repeated K-Fold**
샘플링	SMOTE & TOMEK LINKS	SMOTE & TOMEK LINKS
Train Best Score	0.6690	0.6690
Validation Best Score	0.6825	0.6825

→ 점수가 동일

평가지표 - Multi Class

		PREDICTED		
	Classes	Positive (1)	Negative (0)	Total
ACTUAL	Positive (1)	TP = 20	FN = 5	25
	Negative (0)	FP = 10	TN = 15	25
	Total	30	20	**50**

▲ 이중분류 평가지표

		PREDICTED classification				
	Classes	a	b	c	d	Total
ACTUAL classification	a	5	23	17	17	62
	b	10	540	21	14	585
	c	166	96	436	110	808
	d	1	2	5	87	95
	Total	182	661	479	228	**1550**

▲ 다중분류 평가지표

평가지표 - 오차 행렬

		PREDICTED classification			
	Classes	a	b	c	d
ACTUAL classification	a	TN	FP	TN	TN
	b	FN	TP	FN	FN
	c	TN	FP	TN	TN
	d	TN	FP	TN	TN

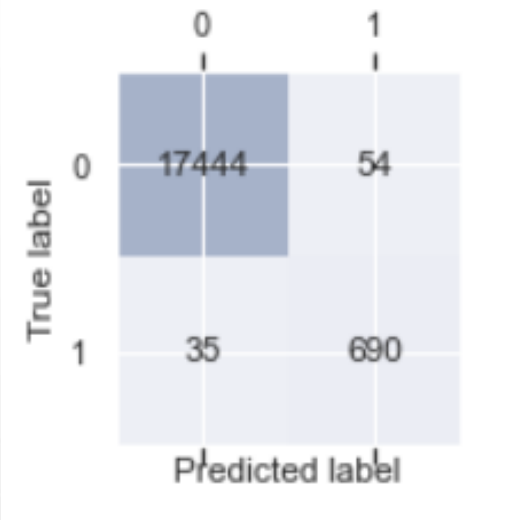
```
from sklearn.metrics import confusion_matrix

predict = model.predict(x_val)
confmat = confusion_matrix(y_true=y_val, y_pred=predict)
print(confmat)
```

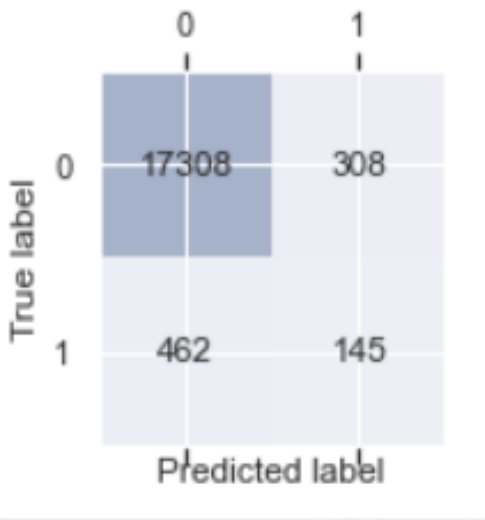
```
[[276  95  28  67   3   4  43   1  15   1   4  69   9   9   1   4  31  24
    2  24  18   3  26  41 108  20  12   7]
 [ 26 453  11  46   0  14  24   5  18   7   5  44   5  12   1  15   2  29
    2  64   5   7  67  50  20   5   5   3]
 [ 15  18 660   5   0   4   7   1   6   0   1  18  21   3   4   1  27   1
    1   2   6   1   9  33  26  42  32   1]
 [ 48  44  19 381   3   8  42  16  34   1   8  51   8  12   7  46   7   0
    1   1  43  21  12   9  89  16  15   3]
 [  0   0   1   0 908   0   0   0   0   0   0   1   0   0   0   1  14   0
   16   0   0   0   0   0   0   0   4   0]
 [  0   0   0   2   0 901   0   9   6   1  12   0   1   0   0   3   0   2
    0   0   2   0   1   0   0   0   0   5]
 [  6   4   4  21   0   7 808   2  13   0   2  11   3   3   3   1   0   0
    0   3   6   2  10  15  16   4   0   1]
 [  1   6   2  11   0  41  13 533  52   1  38   0  22   6   2  43   0   2
    0   1  21 107   2  12   0   1   3  25]
```

▲ 오차행렬 (Confusion Matrix)

결과 - 평가지표 - 개별 Class Score

< LGBM >

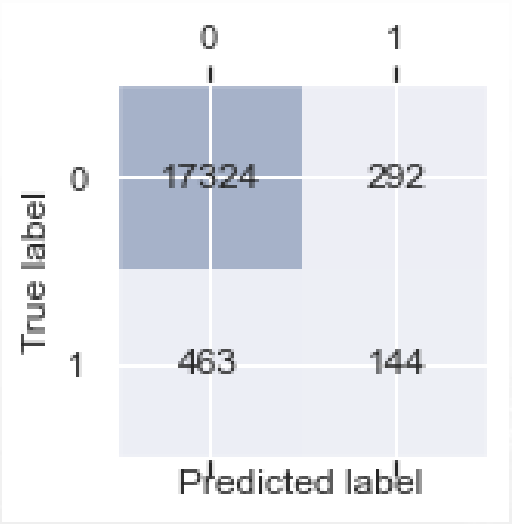


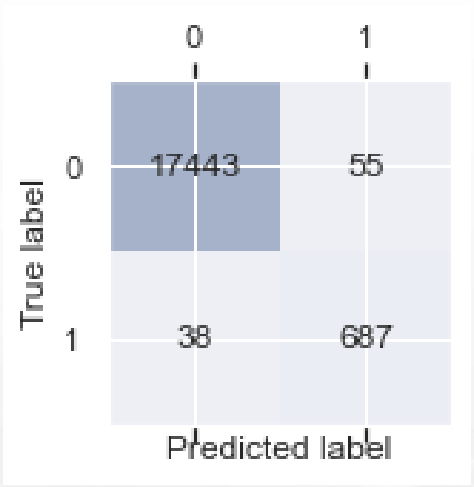


장르	Rock	Opera
Accuracy	0.9957	0.9951
Precision	0.3200	0.9274
Recall	0.2388	0.9517

결과 - 평가지표 - 개별 Class Score

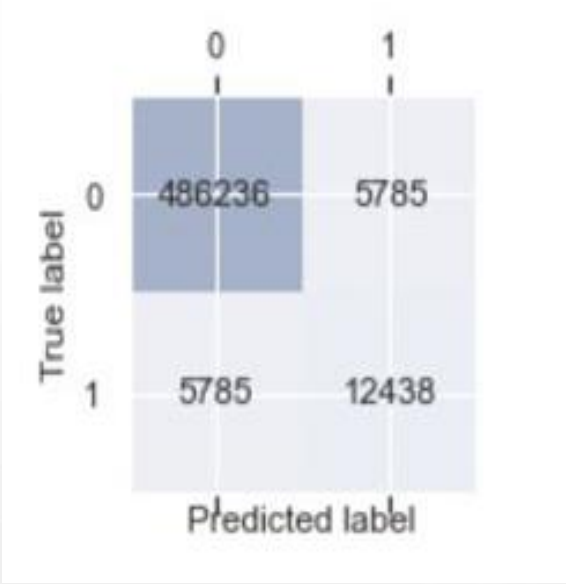
< XGBoost>

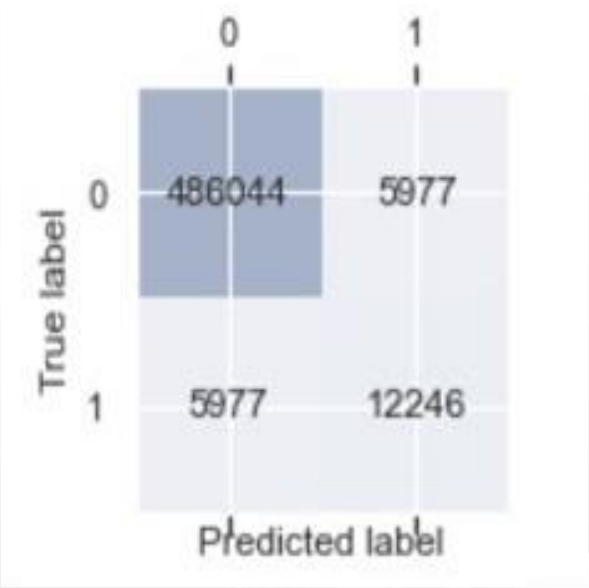




장르	Rock	Opera
Accuracy	0.9585	0.9948
Precision	0.2372	0.9475
Recall	0.3302	0.9258

결과 - 평가지표 - Score (micro average)





	LGBM	XGBoost
Accuracy	0.9773	0.9766
Precision	0.6825	0.6720
Recall	0.6825	0.6720

Melon Top 100 트랙별 장르 예측

멜론 Top 100의 곡들의 장르 분류

	song	artist	track_id	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature
0	strawberry moon	IU	2g0LdZQce9xlcHb1mBJyuz	0.475	0.712	6	-3.306	1	0.0431	0.1380	0.000000	0.0936	0.443	169.796	205333	4
1	Savage	aespa	3dbLT62Cvs46Ju7a8gpr36	0.727	0.879	9	-1.167	1	0.1290	0.1240	0.000012	0.2440	0.671	146.959	238144	4
2	Traffic light	Lee Mujin	03qu1u4hDyepQQi2INxCka	0.609	0.717	6	-3.547	0	0.0597	0.0965	0.000000	0.0694	0.545	97.026	231587	4
3	Next Level	aespa	2zrhoHIFKxFTRF5aMyxMoQ	0.820	0.852	11	-2.567	0	0.1660	0.4880	0.000006	0.0907	0.820	109.036	221573	4
4	If you lovingly call my name	GyeongseoYeji	0tgxvf4rqBBeEB54h0nnRD	0.382	0.672	7	-2.836	1	0.0320	0.3510	0.000000	0.1440	0.236	143.643	231339	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
94	Sticker	NCT 127	4bEa9VAnyVJWBxOUyVvzie	0.527	0.769	2	-0.419	1	0.2800	0.3110	0.000000	0.7420	0.719	156.032	227773	4
95	Face ID (Feat. GIRIBOY, Sik-K, JUSTHIS)	Epik High	6YcJycTGtvfTt5ACVH29q7	0.830	0.889	7	-4.703	1	0.1950	0.0215	0.000000	0.0651	0.763	115.025	216503	4
96	Let's forget it - Drama Version	Mido and Falasol	7kU5GMeSxtOTtl9l1ipgYa	0.751	0.575	9	-6.874	1	0.0282	0.4850	0.000214	0.0771	0.366	97.011	275504	4
97	I can't run away	SEVENTEEN	0bFiLUmBmFX0aWyhVqlwiN	0.640	0.509	5	-6.223	1	0.0312	0.1010	0.000000	0.1080	0.420	140.074	210680	4
98	How can we become friends who we loved	Ha Yea Song	6m1vhC6VXa06TSTpKmAQKr	0.555	0.617	9	-3.562	1	0.0348	0.6600	0.000000	0.1050	0.342	131.918	226920	4

99 rows × 17 columns

▲ 멜론 일간 Top 100 테스트 셋

Melon Top 100 트랙별 장르 예측

모델의 정확도를 조금 더 향상시키기 위해 기존 훈련용 데이터셋에서 멜론 Top 100 안 곡들이 가지는 특성을 고려하여 데이터셋을 수정

	danceability	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms	time_signature	genre2
0	0.681	-8.854	1	0.0294	0.5660	0.000000	0.1120	0.306	119.993	257333	4	0
1	0.773	-9.252	1	0.0516	0.6270	0.000014	0.0953	0.443	74.977	212707	4	0
3	0.640	-8.579	1	0.0292	0.6370	0.000120	0.1080	0.208	127.874	245840	4	0
4	0.327	-2.928	1	0.0547	0.2020	0.000090	0.2800	0.416	179.561	175200	4	0
9	0.655	-6.544	1	0.0274	0.2270	0.000000	0.0549	0.450	116.959	248920	4	0
...	...	...	...	...	...	...	...	...	...	...	...	...
26446	0.539	-6.113	1	0.0597	0.0160	0.000000	0.1660	0.790	99.261	210387	4	11
26458	0.670	-3.679	0	0.0818	0.0351	0.000055	0.2910	0.423	130.555	248254	4	11
26463	0.696	-5.843	1	0.0402	0.0491	0.000068	0.1240	0.758	122.989	292907	4	11
26469	0.784	-5.153	0	0.0923	0.3000	0.000016	0.2720	0.819	118.738	216733	4	11
26471	0.747	-7.700	1	0.0453	0.1270	0.000000	0.0840	0.870	112.964	220867	4	11

5410 rows × 12 columns

▲ 멜론 Top 100 곡들이 가지는 특징만을 반영한 수정 훈련용 셋

Melon Top 100 트랙별 장르 예측

사용 모델	LightGBM	XGBoost
Train Best Score	0.7052	0.6823
Validation Best Score	0.7205	0.6936

→ 최종 선정 모델 2가지의 성능 향상

Melon Top 100 트랙별 장르 예측

	song	artist	track_id	predict1	predict2	predict3	voting
0	strawberry moon	IU	2g0LdZQce9xlcHb1mBJyuz	country	dance	dance	dance
1	Savage	aespa	3dbLT62Cvs46Ju7a8gpr36	dance	rock	disco	dance
2	Traffic light	Lee Mujin	03qu1u4hDyepQQi2INxCka	disco	reggae	rock-n-roll	disco
3	Next Level	aespa	2zrhoHIFKxFTRF5aMyxMoQ	dance	disco	disco	disco
4	If you lovingly call my name	GyeongseoYeji	0tgxvf4rqBBeEB54h0nnRD	rock	dance	dance	dance
5	OHAYO MY NIGHT	D-Hack	4iJprGt1rt5iy0sxXXaRWn	reggae	r-n-b	reggae	reggae
6	NAKKA (with IU)	AKMU	4t2FIqZJORKZGSKg30SShr	hip-hop	reggae	reggae	reggae
7	Foolish Love	MSG WANNABE	7I7TTfKcDDAeSf6HPgbdPT	rock	rock-n-roll	rock-n-roll	rock-n-roll
8	Permission to Dance	BTS	0LThjFY2iTtNdd4wviwVV2	disco	drum-and-bass	drum-and-bass	drum-and-bass
9	Weekend	TAEYEON	6cqH1q7g5GeRVQVMK1Vc7f	funk	rock	rock	rock

- predict1 : 수정된 LGBM 최종 모델 예측값
- predict2 : 원 LGBM 최종 모델 예측값
- predict3 : 원 XGBoost 최종 모델 예측값
- voting : predict1 * 1.5 + predict2 * 1 + predict3 * 1 보팅 모델

Melon Top 100 트랙별 장르 예측

	song	artist	track_id	predict1	predict2	predict3	voting
20	Rock with you	SEVENTEEN	6LnEoRQKMcaFTR5UvaKuBy	rock	rock	rock	rock
24	Space	BOL4	2AAYL6JwiPKHn33buQqo4P	alternative	alternative	alternative	alternative
53	eight(Prod.& Feat. SUGA of BTS)	IU	0pYacDCZuRhcrwGUA5nTBe	rock	rock	folk	rock

Rock With You - 세븐틴
https://www.youtube.com/watch?v=WpuatuzSDK4
Space - 볼빨간사춘기
https://www.youtube.com/watch?v=mo9zv9P-anI
에잇 - 아이유
https://www.youtube.com/watch?v=TgOu00Mf3kI

추후 과제 제안

1. 데이터 불균형이 심한 데이터를 수집하여 분석에는 샘플링을 적용해 모델링을 실시하였지만 데이터 품질과 모델의 성능을 향상하는 것에는 어려움 존재

2. 추후 정확한 기준을 선정하여(타겟의 개수, 장르 선정) 데이터 수집, 데이터 품질 개선 필요하다고 판단

3. 학습한 분류모델을 더 다양한 음원 사이트 장르 분류에 이용 (가온, 빌보드 등)

4. 타겟이 너무 많은 점과 적합한 샘플링기법을 적용하지 못한점을 개선하여 이후 다양한 기법을 활용하여 모델링을 실시하고 후에 스태킹과 보팅까지 진행 제안


감사합니다