*베이스라인 설명 및 사용법

*베이스라인: 성능 비교의 기준이 되는 기본 모델

목표

2014 ~ 2023년의 billboard hot100 데이터를 이용한 top10 여부 예측

사용 데이터

- 파일명: final_data_senti.csv

학습 데이터

독립변수

year, month, week, BPM, duration_sec, R, G, B, compound, genre

종속변수
(target)

isTop10
(0: X, 1: O)

데이터 전처리

[1] Genre: 힙합 -> 랩/힙합

[2] RGB1: R, G, B로 나누기

[3] top10 여부 컬럼 추가

[4] 사용 feature 정의

사용

- Year
- Month
- Week
- Genre
- BPM
- Duration_sec
- R
- G
- B
- Featuring
- compound
- isTop10

제외

- Rank: ‘isTop10’으로 대체
- Title
- Artist
- Lyrics
- color1, color2, color3
  : ‘R’, ‘G’, ‘B’로 대체
- neg, neu, pos: compound로 대체

데이터 전처리

[5] 데이터 타입 변경

	Year	Month	Week	Genre	BPM	Duration_sec	R	G	B	Featuring	compound	isTop10
0	2014	1	2	랩/힙합	110.009	251.246	186	97	156	1.0	0.9904	1
1	2014	1	2	POP	129.992	204.160	18	22	36	1.0	0.9958	1
2	2014	1	2	랩/힙합	122.013	257.840	82	68	48	0.0	-0.9867	1
3	2014	1	2	POP, 록/메탈	84.876	190.185	151	123	112	0.0	0.9887	1
4	2014	1	2	발라드	141.284	229.400	33	39	67	0.0	0.9771	1
...	...	...	...	...	...	...	...	...	...	...	...	...
52095	2023	12	5	재즈, 보컬재즈, 애시드/퓨전/팝	77.810	1037.907	200	215	236	0.0	0.8405	0
52096	2023	12	5	POP	129.918	143.940	31	30	30	0.0	0.9676	0
52097	2023	12	5	월드뮤직, 라틴	125.012	189.426	161	163	191	0.0	-0.7319	0
52098	2023	12	5	POP	156.975	146.752	54	32	13	0.0	-0.9849	0
52099	2023	12	5	R&B/Soul, 국외드라마	84.828	244.685	71	83	80	0.0	-0.9946	0

52084 rows × 12 columns

```
<class 'pandas.core.frame.DataFrame'>
Index: 52084 entries, 0 to 52099
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Year          52084 non-null  int64
 1   Month         52084 non-null  int64
 2   Week          52084 non-null  int64
 3   Genre         52084 non-null  category
 4   BPM           52084 non-null  float64
 5   Duration_sec  52084 non-null  float64
 6   R             52084 non-null  int64
 7   G             52084 non-null  int64
 8   B             52084 non-null  int64
 9   Featuring     52084 non-null  float64
 10  compound      52084 non-null  float64
 11  isTop10       52084 non-null  category
dtypes: category(2), float64(4), int64(6)
memory usage: 4.5 MB
```

전처리가 끝난 데이터

모델링

[1] 데이터 준비

1) X, y로 데이터 나누기

2) 수치형 변수 스케일링
- Standard Scaling: 평균을 0, 표준편차를 1로 하는 데이터로 변환
- MinMax Scaling: 최대값은 1, 최소값은 0으로 하여 0~1 사이의 값으로 변환
- Robust Scaling: 중앙값, IQR 사용

3) 범주형 변수 인코딩

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

4) 2), 3) 데이터 병합

5) train, test 데이터 나누기

모델링

[2] 모델 정의

경사하강법 기반: Logistic Regression

확률 기반: Naive Bayes

거리 기반: SVM

트리 기반: 기본) Decision Tree

앙상블) Random Forest

[3] 모델 학습 & 예측

```
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

모델링

[4] 성능 평가

Accuracy

F1-Score

Classification_report

Classification Report

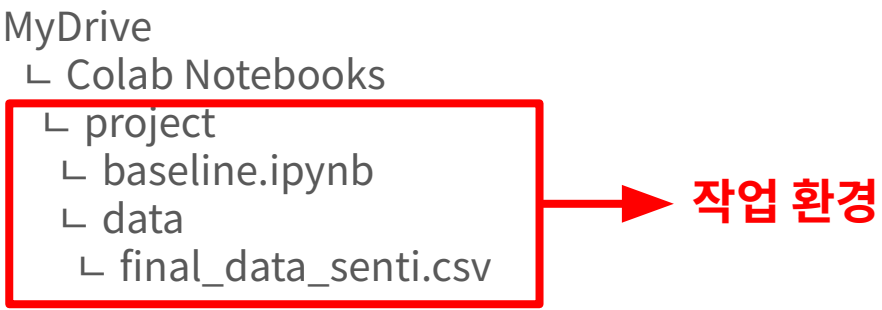
	precision	recall	f1-score	support
0	0.90	1.00	0.95	14072
1	0.00	0.00	0.00	1554
accuracy			0.90	15626
macro avg	0.45	0.50	0.47	15626
weighted avg	0.81	0.90	0.85	15626

Confusion Matrix

```
Confusion matrix
[[14072     0]
 [ 1554     0]]
```

베이스라인 사용법

- 환경: Google Colab
- 준비
  1) 구글 드라이브 ‘Colab Notebookes’ 폴더 내에 ‘project’ 폴더 생성
  2) project 폴더 내에 ‘baseline.ipynb’ 파일 업로드
  3) project 폴더 내에 ‘data’ 폴더 생성
  4) data 폴더 내에 ‘final_data_senti.csv’ 파일 업로드



베이스라인 사용법

[1] 구글 드라이브 마운트

[3] 데이터 로드

[2] 라이브러리 로드

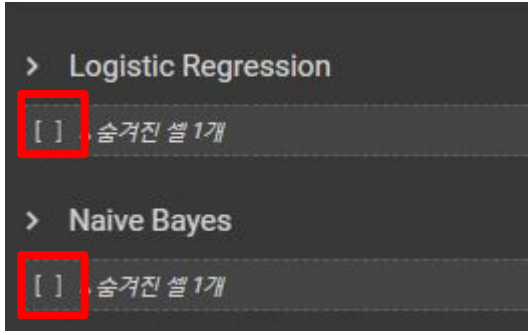
[4] 전처리

베이스라인 사용법

[5] 모델링

[5-1] 데이터 준비

[5-2] 모델 정의

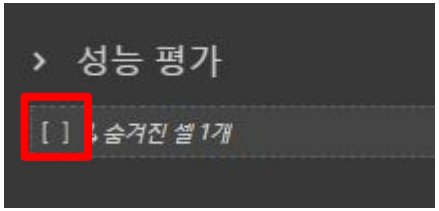


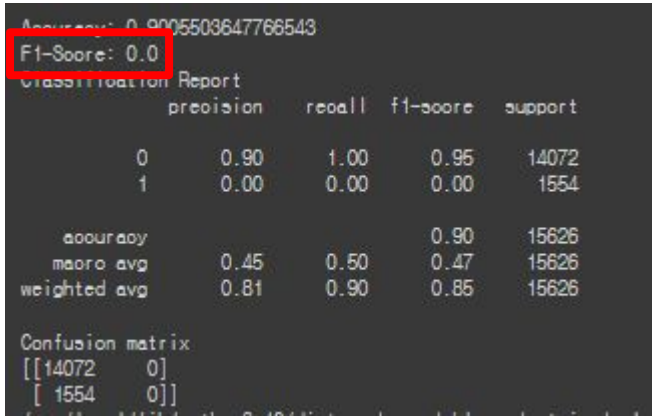
5개 중 하나만 실행!

베이스라인 사용법

[6] 모델 학습 & 예측

[7] 성능 평가





[8] 테스트 결과 정리: 링크