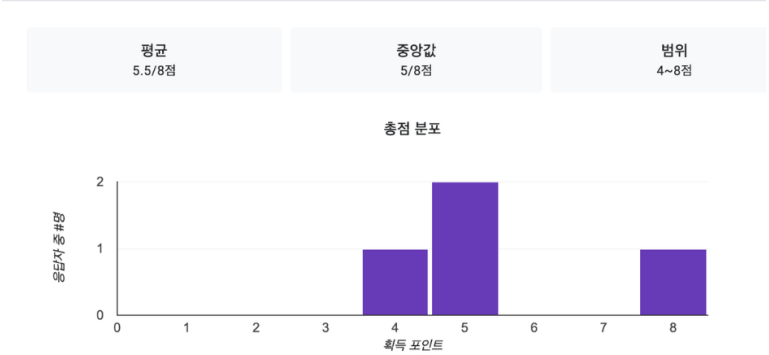
방학 프로젝트 최종 발표

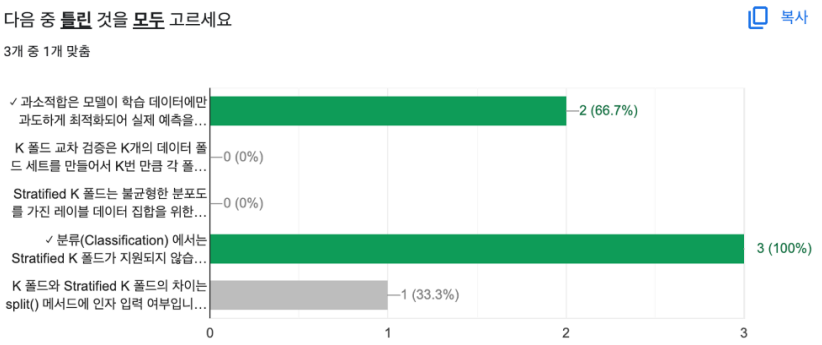
초급 도하연 팀 - 김선향, 김정은, 도하연, 문원정

EWHA EURON

1학기 세션 복습

파이썬 머신러닝 완벽 가이드 (1~9장)

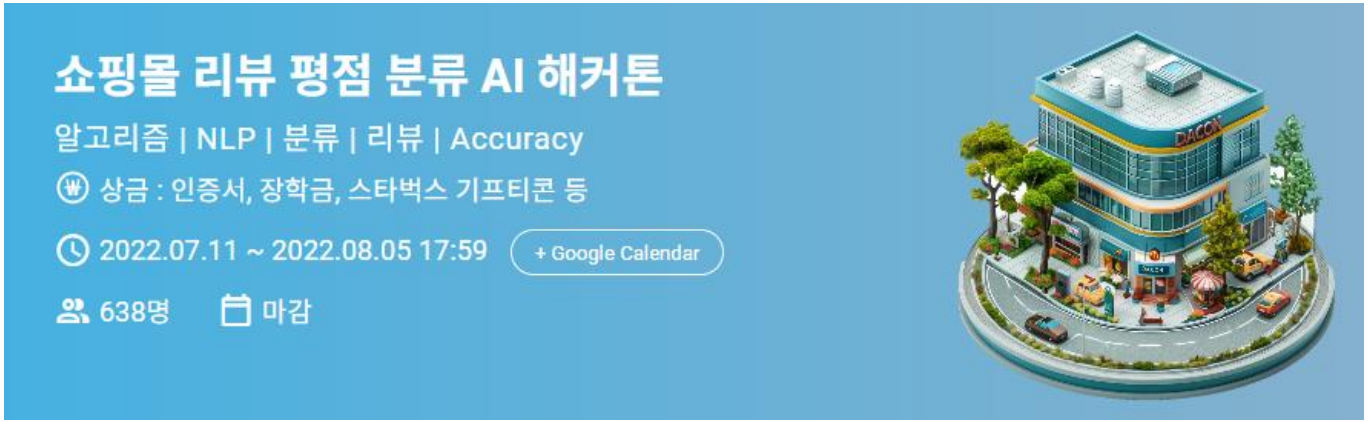




✅ 복습 진행 규칙

- 이틀에 한 단원 복습 진행
- 4명이 돌아가면서 퀴즈 출제 담당
- 비대면으로 주어진 시간동안 풀이 후 상호 피드백

프로젝트 개요



- ✅ 상품 리뷰 텍스트(x)로 평점(y)을 예측
- ✅ 평점 : 1, 2, 4, 5점
- ✅ 자연어 처리 필요

0주차 – 데이터 전처리

```
[4]  train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 3 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   id       25000 non-null  int64
 1   reviews  25000 non-null  object
 2   target   25000 non-null  int64
dtypes: int64(2), object(1)
memory usage: 586.1+ KB
```

✅ 결측치 없음

-> 결측치 전처리 과정 필요X

```
train.head()
```

	id	reviews	target
0	0	조아요 처음구입 싸게햇어요	2
1	1	생각보다 잘 안돼요 매지 바른지 하루밖에 안됐는데ㅠㅠ 25천원가량 주고 사기 너무 ...	1
2	2	디자인은괜찮은데 상품이 금이가서 교환했는데 두번째받은상품도 까져있고 안쪽에 금이가져...	2
3	3	기전에 이 제품말고 이마트 트레이더스에서만 팔던 프리미엄 제품을 사용했었습니다. 샘...	2
4	4	튼튼하고 손목을 잘 받쳐주네요~	5

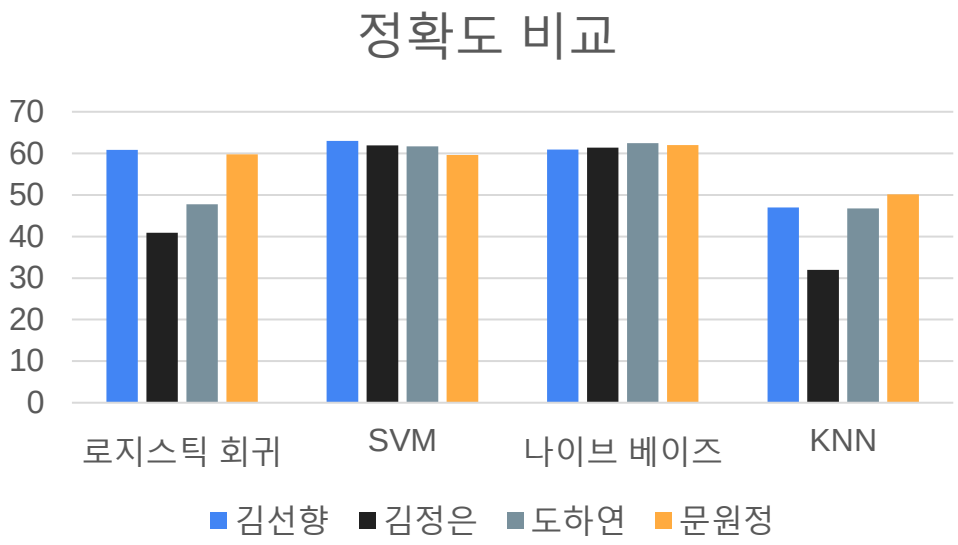
✅ 한국어 텍스트 전처리

- KoNLPy의 Okt 사용
- 토큰화 , 품사 태깅

EWHA EURON

1,2주차 – 다양한 모델 시도

첫번째 시도 : 로지스틱 회귀, SVM, 나이브 베이즈, KNN



✓문제점

각자 다른 자연어 처리, 벡터화 방식

-> 정확도 차이

Ex) 로지스틱 회귀 최고정확도 : 60.8%

최저정확도 : 40.9%

✓ 정확도가 현저히 낮은 KNN, Why?

텍스트 데이터 벡터화 시, 데이터 차원이 매우 높아짐.

KNN은 거리 기반 분류 알고리즘이므로

고차원 데이터에서 성능이 좋지 않음

1,2주차 – 다양한 모델 시도

두번째 시도 : 자연어 처리, 벡터화 베이스라인 역할 분담

	품사 태깅 O	품사 태깅 X
CounterVectorizer	김정은	문원정
TF-IDF	불용어 제거 O : 김선향	도하연
	불용어 제거 X : 김정은	

KNN을 제외한 최대한 많은 모델 시도

모델	앙상블	정규화
로지스틱 회귀 나이브 베이즈 SVM	배깅 – 랜덤 포레스트	L1 L2 복합
	부스팅 - GBM, XGBoost, LightGBM, CatBoost	
	하드 보팅, 소프트 보팅	

3주차 – 베이스라인 확정

	품사 태깅 O	품사 태깅 X
CounterVectorizer	김정은	문원정
TF-IDF	불용어 제거 O : 김선향	도하연
	불용어 제거 X : 김정은	

	김선향	김정은	도하연	문원정
최고 정확도	63.0%	64.1%	62.2%	61.97%
최고 정확도 모델	SVM {'C': 1, 'gamma': 1, 'kernel': 'rbf'}	로지스틱 회귀 (TF-IDF 적용)	소프트 보팅 (로지스틱 회귀, KNN, 나이브 베이즈, SVM)	나이브 베이즈
평균 정확도 (나이브 베이즈, SVM, 로지스틱 회귀)	61.69%	62.47%	60.98%	60.43%

평균 정확도가 가장 높은 품사태깅 + TF- IDF(불용어제거X) -> 베이스라인 확정

중간 결론

✅ 더 높은 정확도를 위해서는?

품사 태깅	품사 태깅 유무	O
	불용어 제거	X
벡터화	CounterVectorizer OR TF-IDF	TF-IDF
사용 모델		나이브 베이즈, SVM, 로지스틱 회귀
하트 보팅 OR 소프트 보팅		소프트보팅

4주차 – 데이터 증강, 교차 검증 fold

세번째 시도 : 베이스 라인 고정 후, 정확도 목표치(70%) 까지 올리기

	테스트 데이터를 다양하게 (CV는 3으로 고정)	교차 검증 fold 수를 다양하게 (test data는 25% 고정)
데이터 증강 X	문원정	김선향
데이터 증강 O	김정은	도하연

- 데이터 증강 방법 : Random_deletion

1. 데이터 증강이 정확도 향상에 기여
   ex) 로지스틱 회귀 : 61.45% -> 64%

2. 랜덤한 데이터 증강 방식
   -> 매번 달라지는 성능

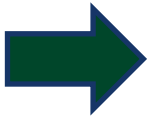
3. 과도한 데이터 증강
   -> 과적합 오류 발생

4. 데이터 증강 시, 94.3%에 달하는 정확도
   -> 과적합 여부 분석 필요! (해커톤 1위 정확도 : 71.3%)

EWHA EURON

5주차 – 데이터 증강 코드 점검

1. Train 성능은 94.3% 이지만 리더보드 결과는 0.32335
2. 훈련 데이터 (94%) 와 검증 데이터 (32%) 정확도 차이가 10% 이상



과적합

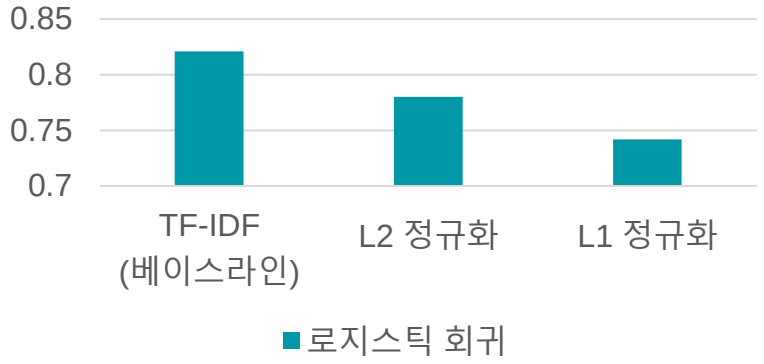
5주차 – 데이터 증강 코드 점검

1. Train 성능은 94.3% 이지만 리더보드 결과는 0.32335
2. 훈련 데이터 (94%) 와 검증 데이터 (32%) 정확도 차이가 10% 이상

→ 과적합

해결 방안

정규화 기법 사용 : L1, L2 규제 사용



5주차 – 데이터 증강 코드 점검

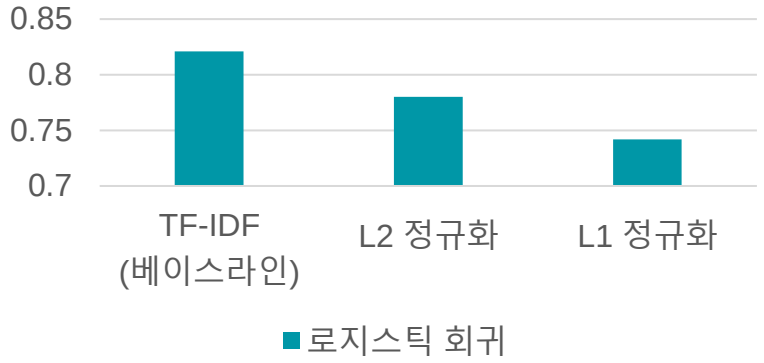
1. Train 성능은 94.3% 이지만 리더보드 결과는 0.32335

2. 훈련 데이터 (94%) 와 검증 데이터 (32%) 정확도 차이가 10% 이상

➡ 과적합

해결 방안

정규화 기법 사용 : L1, L2 규제 사용



결론

- L1 정규화가 과적합 방지에 효과적
  👉가중치 벡터의 일부 요소를 0으로 만들어 불필요한 특성을 제거

- 텍스트 데이터 (고차원 데이터) 에서 유용

6주차 – 데이터 특성 고려

그렇다면 상품 리뷰 텍스트와 평점 사이에는 어떤 관계가 있을까요?

상품 리뷰 텍스트만으로 상품의 평점을 예측할 수 있을까요?

주어진 쇼핑몰 리뷰 데이터셋을 이용하여

상품의 평점 (1점, 2점, 4점, 5점)을 분류해주세요!

▲ 프로젝트 설명

과적합 X

정확도 : 64.3 %

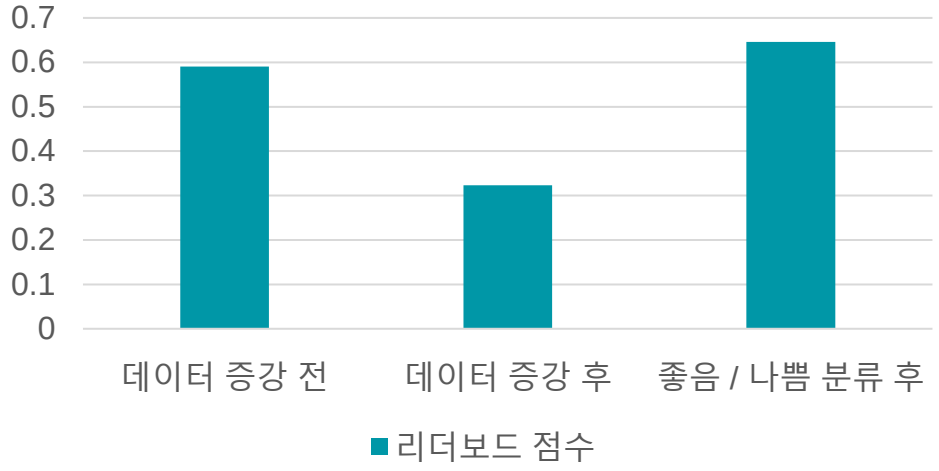
리더보드 점수: 0.64592

- 3점이 존재하지 않는다는 데이터 특성을 고려

- 1,2점 (나쁨) 과 3.4 점 (좋음) 을 초기에 분류하여
  각각 예측하는 방법을 시도

EWHA EURON

최종 결론



1. 데이터 증강 시에도 과적합에 대한 주의 필요
2. 딥러닝 도입의 필요성

소감

⭐<파이썬 머신러닝 완벽 가이드> 교재에서 배운 내용을 대부분 적용해 보았다.

⭐우승자 코드를 먼저 분석하지 않고, 처음부터 팀원들끼리 고민하면서 문제 해결 능력을 길렀다.

⭐모두가 같은 코드를 작성하는 것이 아니라 독립 변수를 조절하면서 다양한 시도를 했다.

⭐'나이브 베이즈' 같이 학기 중에 배우지 않은 내용도 주도적으로 공부하여 적용하였다.

⭐대면으로 만나 더 활발한 토의를 할 수 있었다.

	품사 태깅 O	품사 태깅 X
CounterVectorizer	김정은	문원정
TF-IDF	불용어 제거 O : 김선향	도하연
	불용어 제거 X : 김정은	

감사합니다