2022년 빅데이터 붐업 워크숍 데이터톤

인터넷 사용자의 MBTI 예측하기

team10 - 김영준, 정성문, 최윤서

빅데이터 혁신공유대학

데이터셋 설명

INTP	say process model list like subscriber channel
INFJ	upon much manipulate retail finish like sacrifi
INFJ	fit yes certain bff social feel goal go know nor
INTJ	complete love within someone ideal joke solv
ENTJ	public strictly thing person x question persona
INFJ	opinion mean dim nuclear one like upset deci
INFJ	thing zebra fool aw co word take day talk giv
INTP	hard lazy yet technique convince many hard l
INTP	incite use market especially market late try ne
ENTP	east relate probability thinker live know simple
INFJ	understand seem hard one use environment b
INTJ	let true need nice u nice even like fi life fi mal
INFJ	rot discipline live get overall house black te di
INTP	build beyond polynomial perceive unexpectec
ENTP	idea piece moral usually try f shit enfp value a

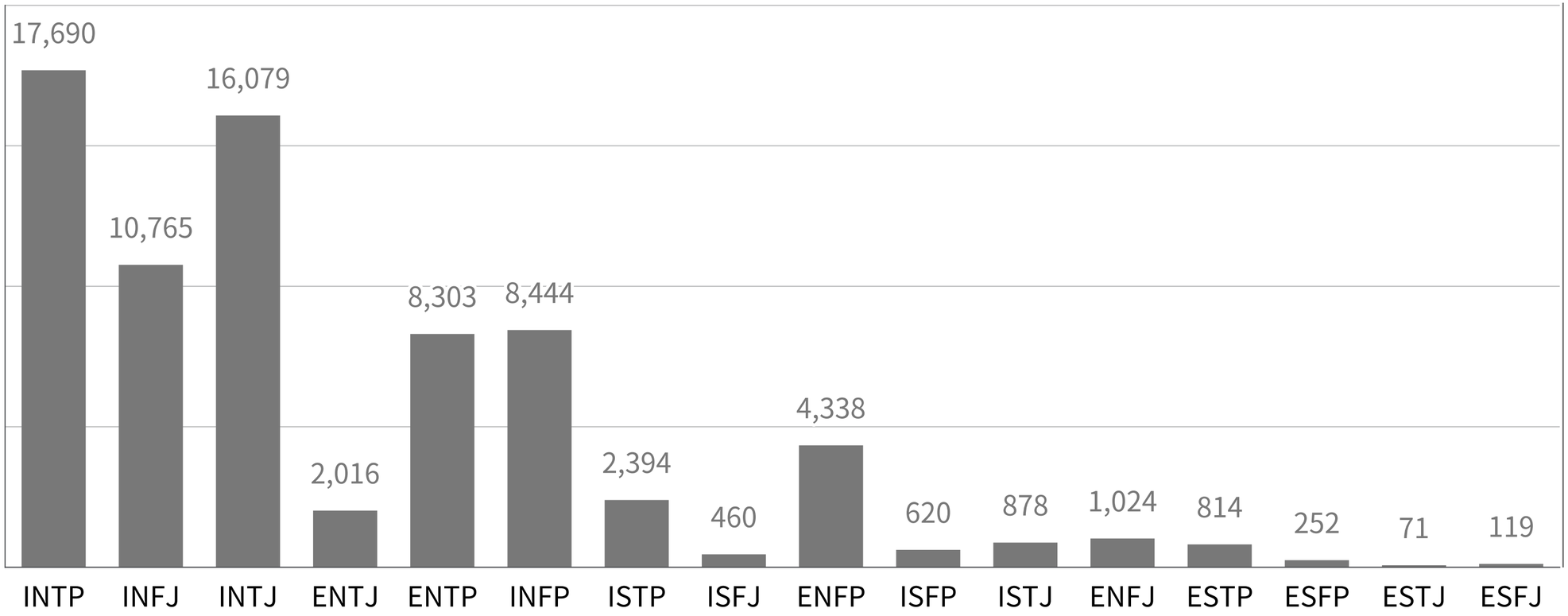
- ▶ 데이터 : 인터넷 사용자의 게시글을 가공한 데이터
- ▶ 종속변수: 인터넷 사용자의 MBTI
- ▶ 설명변수: 게시글에서 랜덤 추출한 400개의 단어
- ▶ 평가지표: 4가지 지표 중 맞춘 지표 수에 따라 정확도 판별

ex)

정답 : ESTJ / 결과 : ISTJ => 3개 정답 인정

정답 : ESFP / 결과 : ISTP => 2개 정답 인정

종속변수 - MBTI


17,690
10,765
16,079
2,016
8,303
8,444
2,394
460
4,338
620
878
1,024
814
252
71
119
INTP
INFJ
INTJ
ENTJ
ENTP
INFP
ISTP
ISFJ
ENFP
ISFP
ISTJ
ENFJ
ESTP
ESFP
ESTJ
ESFJ

데이터 전처리

1. 16개의 MBTI 단어 포함 여부

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

0.8634

2. OneHotEncoding(1000 words)

1	2	3	...	995	996	997	998	999
1.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0
1.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0
1.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0
1.0	1.0	1.0	...	0.0	1.0	0.0	0.0	0.0
1.0	1.0	1.0	...	0.0	0.0	0.0	0.0	0.0

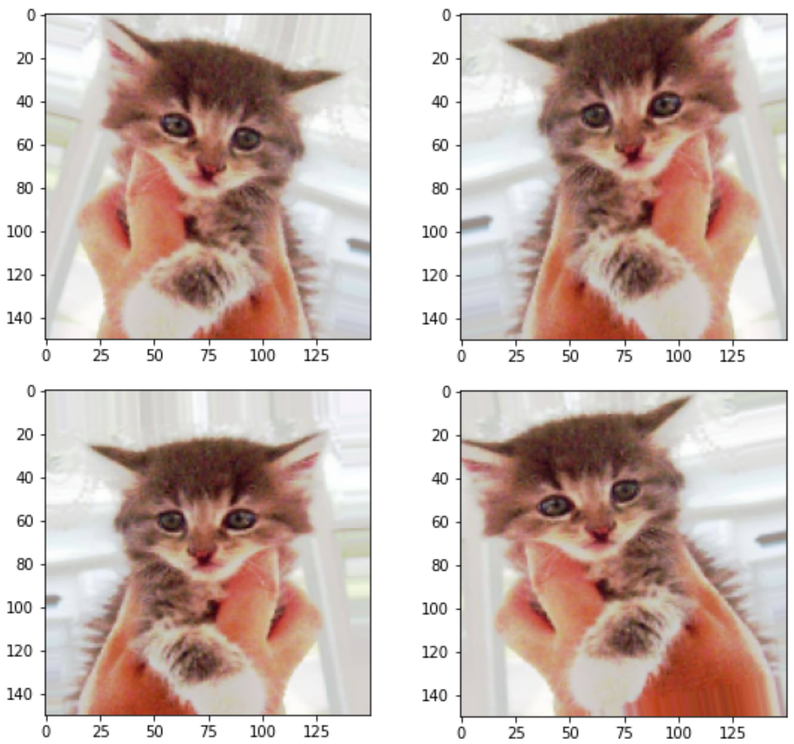
0.8796

3. tf-idf vectors

$$tf(t,d) = freq$$
$$idf(t, D) = num\ of\ document$$
$$tf\text{-}idf = tf * idf$$

0.9193

데이터 증식



INTP	say process model list like subscriber channel...
INFJ	upon much manipulate retail finish like sacrif...
INFJ	fit yes certain bff social feel goal go know n...
INTJ	complete love within someone ideal joke solvea...
ENTJ	public strictly thing person x question person...

x

n

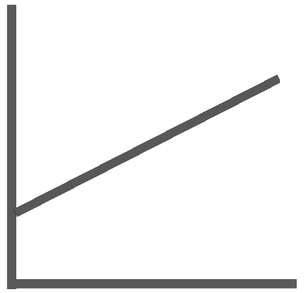
데이터 증식 - 단어 간 교호작용 중요성 확인

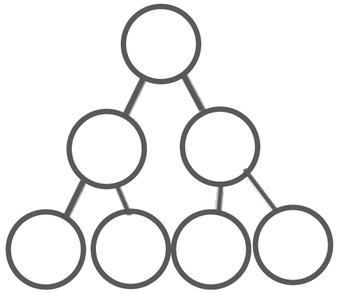
1. 문장이 아닌 단어의 조합

'upon much manipulate retail finish like sacrifice d
eep like alleviate odd one think strategy damage g
rind like u fire without one latter physically never st
upid time perfect good pretty fix generalize explain
must laugh terrible quietly different experience wo
uld couple two leave job thryomanes stuff world lo
ok panic imperative along brace lot one instead ex
troverted lie font live shortage competent treat ti
me go use personality entire get hear try hard tell
employee choice miserably end end level dad alon
e tantrum make emotional sacrifice life put psych
otic flea believe come get really member robot pre
vious job enhance get u thing lot fail emotional rela
tionship image thank employee dayvancowgirl ha
ppen pretend forget peace team weird break drive
introvert extremely thing people surprisingly clarity
gonna really common habitually work ...'

<실제 데이터>

2. logistic regression vs random forest





0.9159

0.4398

데이터 증식 상세 과정

1. MBTI별 단어 빈도 수 계산

	words	frequency
0	like	1454
1	think	1012
2	people	885
3	get	813
4	type	669
...	...	...
8181	snarky	1
8182	plaything	1
8183	starve	1
8184	stumble	1

<ESFP의 단어 빈도수>

2. 빈도수를 확률로 단어 샘플링

```
279           sound
5269      management
1277           intps
732          however
80              time
             ...
1426      friendship
1510     comfortable
172         everyone
375            first
475             meet
Name: index, Length: 40000,
```

<ESFP에서 추출된 단어 40000개>

3. 데이터 생성

y	x
ESFP	sound management intps however time time littl...
ESFP	person band friend whenever think nt thank mak...
ESFP	hate time hard bullshit stereotype opposite fr...
ESFP	confuse go e think remind recommend start sinc...
ESFP	would think fuck say u introvert idea wavelenh...
...	...
ESFP	everyone uncharacteristic talk ne next awkward...
ESFP	take light good location downvoted negative co...
ESFP	disc power ideal throne well put say trust lik...
ESFP	achiever degree big ask change every type men ...
ESFP	esfps relate silly like thing hear life inform...

<ESFP에서 생성된 단어>

데이터 증식 효과

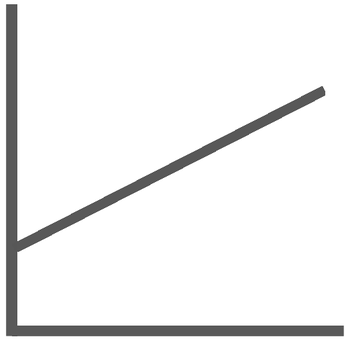
Logistic Regression	Logistic Regression
0.9159	0.9181
LinearSVC	LinearSVC
0.9193	0.9213

<증식 전>

<증식 후>

모델링

1. Logistic Regression

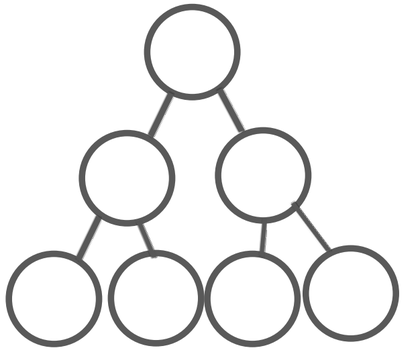


0.9181

2. LinearSVC

0.9213

3. Random Forest



0.8045

활용방안 - 고객 segmentation

1. SNS 게시물 추천 서비스

- 수익 구조: 노출 광고 수익

2. 상품 추천 서비스

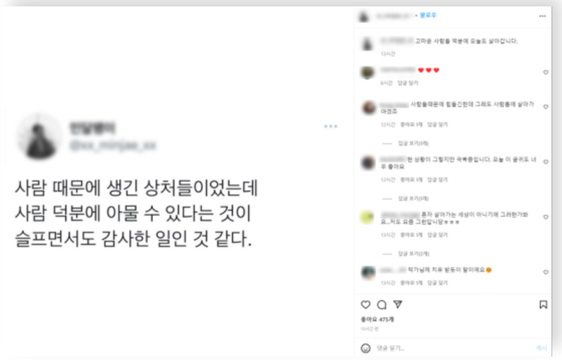
- 수익 구조: 상품 판매 수익

→ 고객이 서비스에 오래 머물수록 수익이 늘어나는 구조

활용방안

1. SNS 게시물 추천 서비스

- 고객이 직접 남긴 댓글, 스토리, 게시물
- 같은 MBTI를 가진 고객의 게시물을 노출



<인스타그램 게시물 및 댓글>

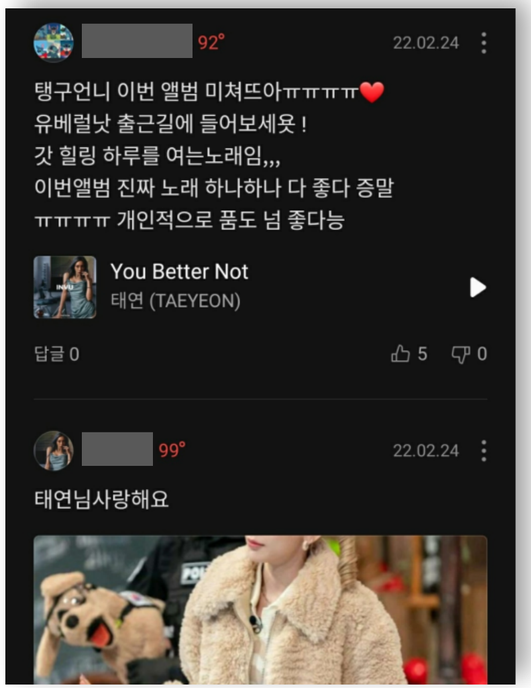
<페이스북 게시물>

활용방안

2. 상품 추천 서비스

ABLY

- 고객이 직접 남긴 리뷰를 바탕으로 데이터 추출
- 같은 MBTI로 예측된 고객이 구매/사용한 아이템 추천
- 컨텐츠 게시물 추천 시 사용



<멜론 리뷰>

<에이블리 스타일 게시물>

개선점

01 다양한 모델 적합 및 하이퍼 파라미터 튜닝

- 본 프로젝트에서는 3개의 모델만 고려
- 더 다양한 분류 모델을 적합하거나, 하이퍼 파라미터 튜닝을 통해 성능 향상의 여지가 있음

02 클래스 불균형 처리

- 본 데이터의 종속변수(MBTI)는 클래스 불균형이 심함
- 클래스 불균형 처리를 적용 후 모델을 돌렸을 때 성능 향상을 기대해볼 수 있음
- 또한, 데이터의 MBTI 비율이 실제 모집단의 비율을 반영하지 않기 때문에 실생활 적용 시 더 타당한 방식으로 생각됨.