DACA

NGO 머신러닝 : 3주차

2021.1.8 ~ 2021.1.14

빅데이터 응용학과 석사 1기 양선욱

목차

1. 베르누이 나이브 베이즈

* tripadvisor_hotel_review.csv

긍정/부정	rating	개수	비율
긍정	5	9054	44.1%
	4	6039	29.4%
부정	3	2184	10.6%
	2	1793	8.7%
	1	1421	6.9%
총합		20491	100%

- 4점 이상을 '긍정', 3점 이하를 '부정' 으로 간주한다.
- 리뷰의 긍정과 부정을 분류하는 모델을 생성한다.

* 모델평가&분석 결과

```
CountVectorizer(binary=True, stop_words='english', min_df=3)
```

- train : test = 7:3
- 학습용 데이터 세트 정확도 : 0.853
- 평가용 데이터 세트 정확도 : 0.817

	precition	recall	f1	support
부정	0.67	0.62	0.65	1649
긍정	0.87	0.89	0.88	4499
macro avg	0.77	0.76	0.76	6148
weighted avg	0.81	0.82	0.81	6148

2. 다항분포 나이브 베이즈

* winemag-data

총 데이터 수 : 117,228

와인 지방 카테고리 수 : 50개

- 와인은 생산되는 지방에 맛을 특징이 다르다.
- 와인을 시음한 사람들의 설명이 지방의 특징을 설명하고 있다면 이것만으로 와인 생산 지방을 분류할 수 있다고 예측하였다.
- 와인 지방 390여 곳 중에 데이터의 수가 100이하인 곳을 제거하고 남은 곳이 50곳

* 모델평가&분석 결과

- 학습용 데이터 세트 정확도 : 0.471
- 평가용 데이터 세트 정확도 : 0.455
- 정밀도 평균 : 0.28
- 재현률 평균 : 0.07
- F1 스코어 평균 : 0.08

* 데이터의 불균형 문제

- California 카테고리 수 : 39,816
- 데이터가 100개 미만인 카테고리는 제거 하였지만 역부족

3. 합성곱 신경망

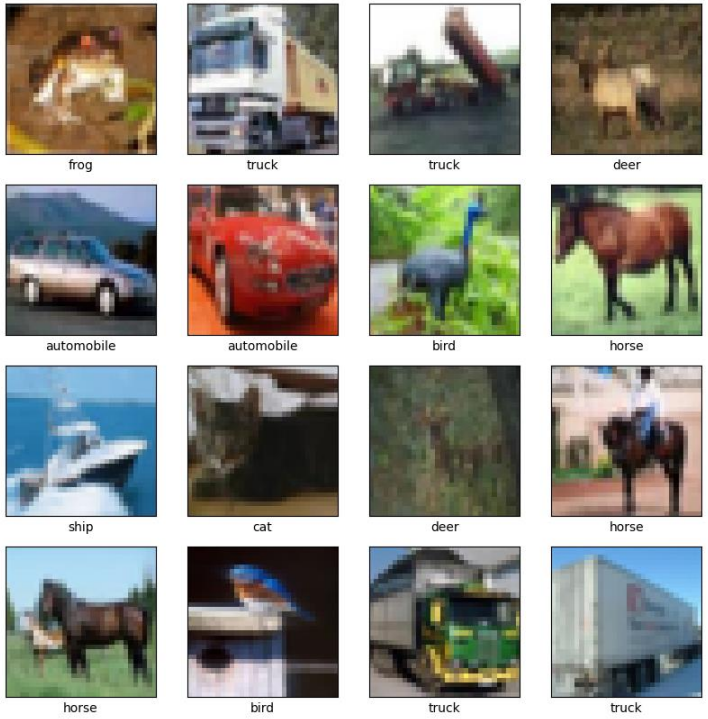
* datasets.cifar10.load_data()

train_date : 50,000 test_data : 10,000

* 10가지의 카테고리

```
['airplane', 'automobile', 'bird', 'cat', 'deer',
'dog', 'frog', 'horse', 'ship', 'truck']
```

* Train Data 이미지



* 모델 학습, 평가

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 30, 30, 32)	896
max_pooling2d (MaxPooling2D)	(None, 15, 15, 32)	0
conv2d_1 (Conv2D)	(None, 13, 13, 64)	18496
max_pooling2d_1 (MaxPooling2	(None, 6, 6, 64)	0
conv2d_2 (Conv2D)	(None, 4, 4, 64)	36928
flatten (Flatten)	(None, 1024)	0
dense (Dense)	(None, 64)	65600
dense_1 (Dense)	(None, 10)	650

Total params: 122,570
Trainable params: 122,570
Non-trainable params: 0

epochs = 20	오차	정확도
학습용 데이터 세트	0.309	0.889
평가용 데이터 세트	1.104	0.704

4. 순환 신경망

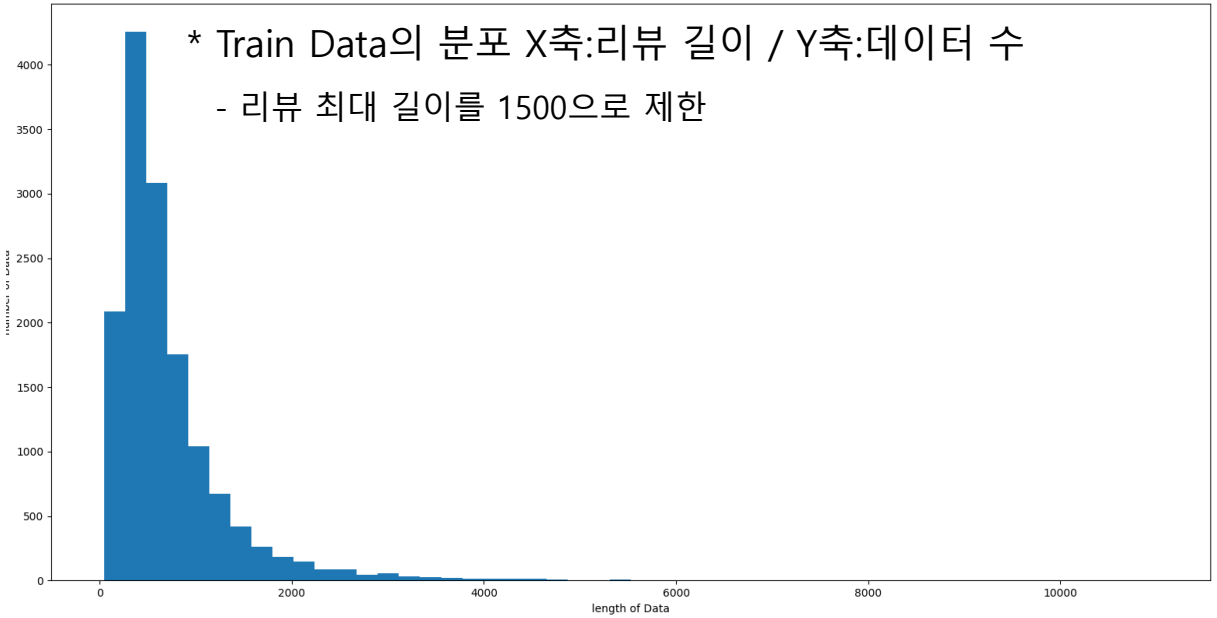
* tripadvisor_hotel_review.csv

train_date : 14,343 test_data : 6,148

카테고리 : [0, 1] {0-부정 : 3749, 1-긍정 : 14343}

* Train Data의 단어별 인덱싱

```
for k in range(0,len(X),1) :
    vocab = sorted(set(X[k]))
    char2idx = {u: i for i, u in enumerate(X[k])}
    idx2char = np.array(vocab)
    X2[k] = np.array([char2idx[c] for c in X[k]])
    # print(X2[k])
```



* Train Data의 분포 X축:리뷰 길이 / Y축:데이터 수

- 리뷰 최대 길이를 1500으로 제한

* 모델 학습, 평가

Model: "sequential"

Layer (type)	Output Shape	Param #
embedding (Embedding)	(None, None, 128)	1920000
simple_rnn (SimpleRNN)	(None, 128)	32896
dense (Dense)	(None, 1)	129

data_length : max 1500

Embedding(15000,128)

SimpleRNN(128, activation='tanh'

epochs : 5

dense : sigmoid

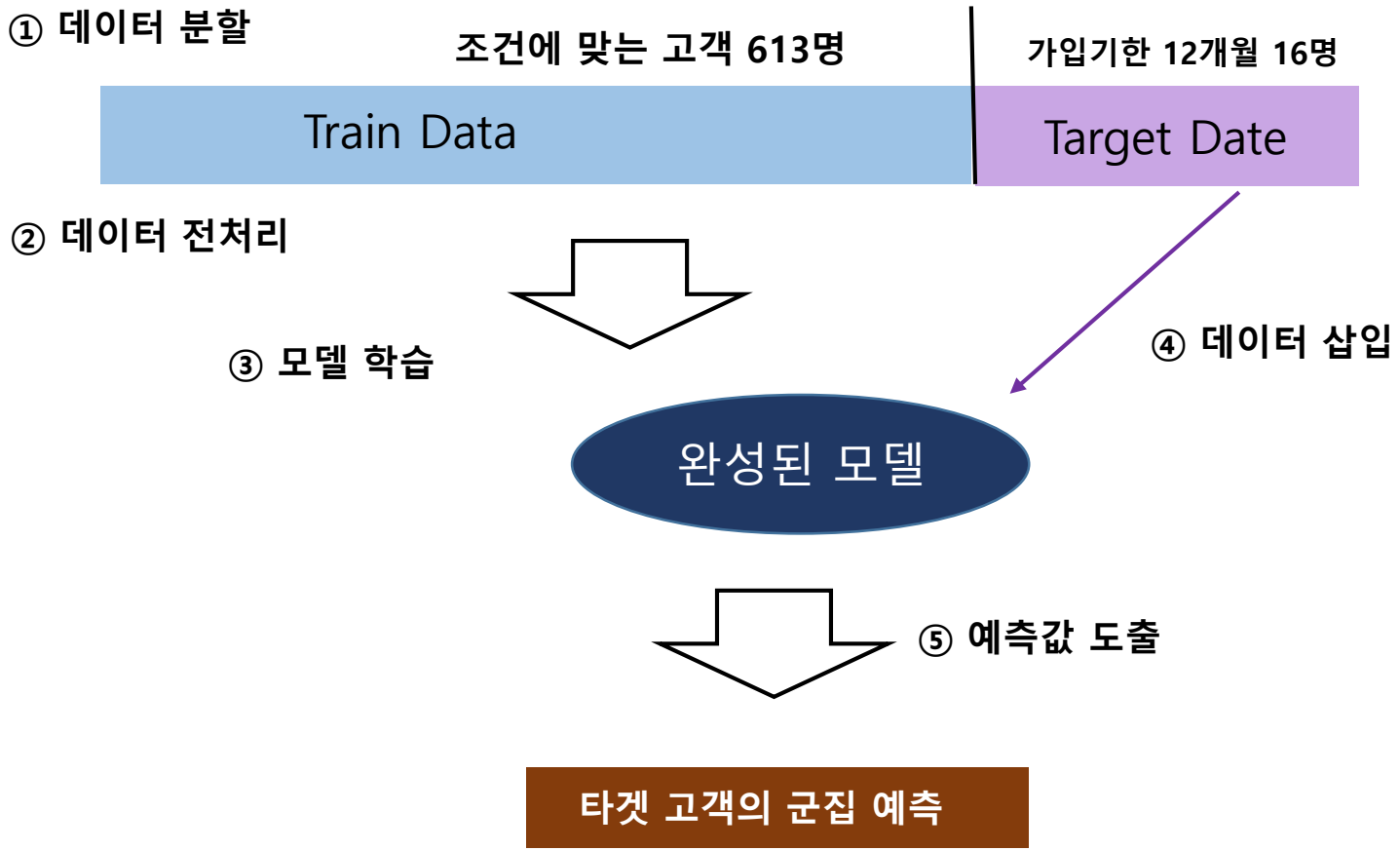
	오차	정확도
학습용 데이터 세트	0.505	0.760
평가용 데이터 세트	0.607	0.719

군집분석 문제 정의

* 분석 목적

- 비이탈, 가입기한 1년 이상의 고객 데이터로 군집 분석을 수행하여 고객을 분류한다.
- 가입기한이 12개월 째인 고객을 분류 모델에 넣어 어느 군집에 속할 지 미리 예측한다.

* 분석 과정



군집분석 데이터 전처리

* 사용 데이터

- 독립변수 : 나이, 가입나이, 가입일수, 총납입금액

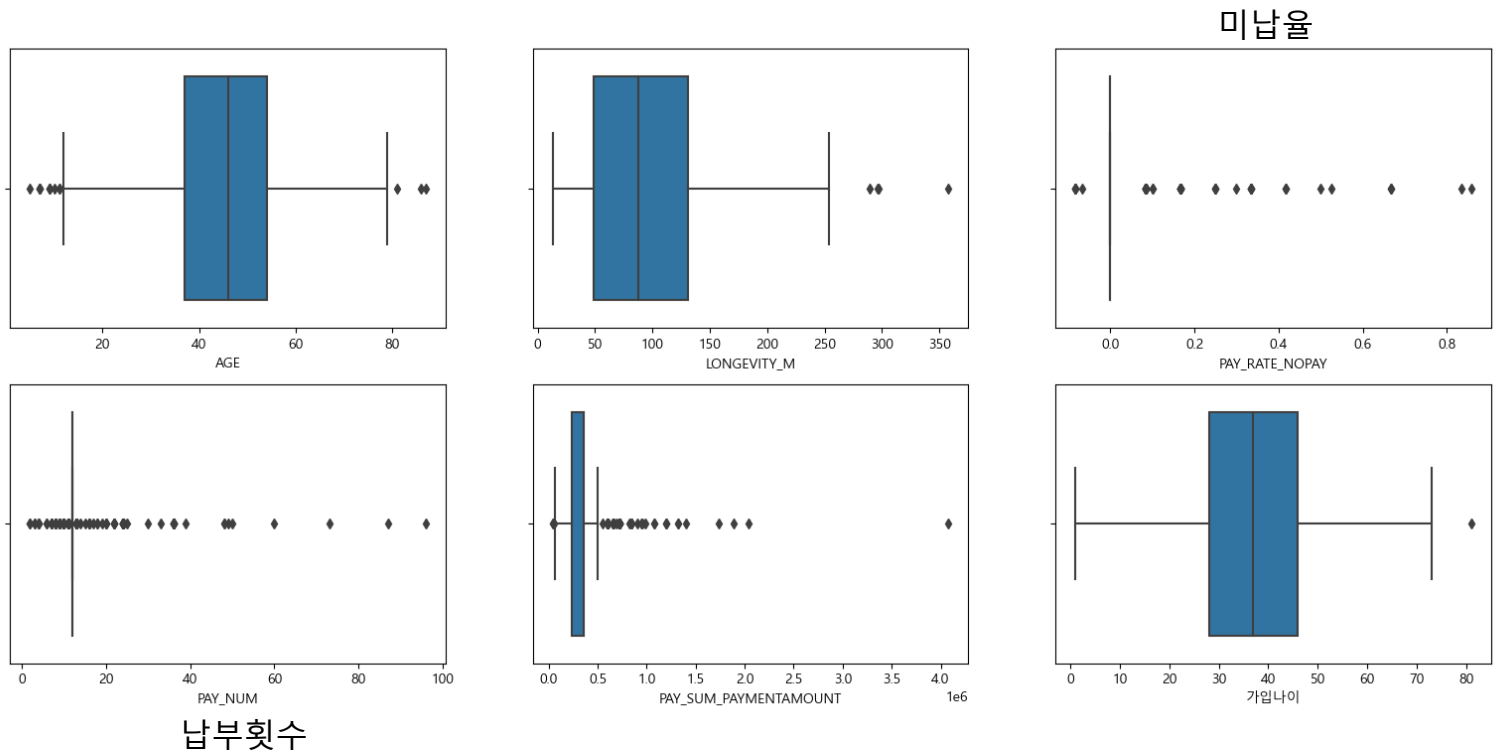
기존 : n=613

전처리 후 : n=530

* 데이터 전처리

- 독립변수 표준화와 명목형에 대한 원핫인코딩
- 이상치 제거 (Q1 – 1.5*IQR이하, Q3+1.5*IQR이상)

* 파이 차트



- 모든 항목에서 이상치의 존재를 확인. -> 이상치 제거
- 미납율은 613명 중 524명이 0%이므로 독립변수에서 배제한다.
- 납부 횟수는 전처리 후 모든 데이터가 12이므로 배제한다.

5. 계층적 군집분석

* 모델 평가

ⓐ 모델 Confusion Matrix

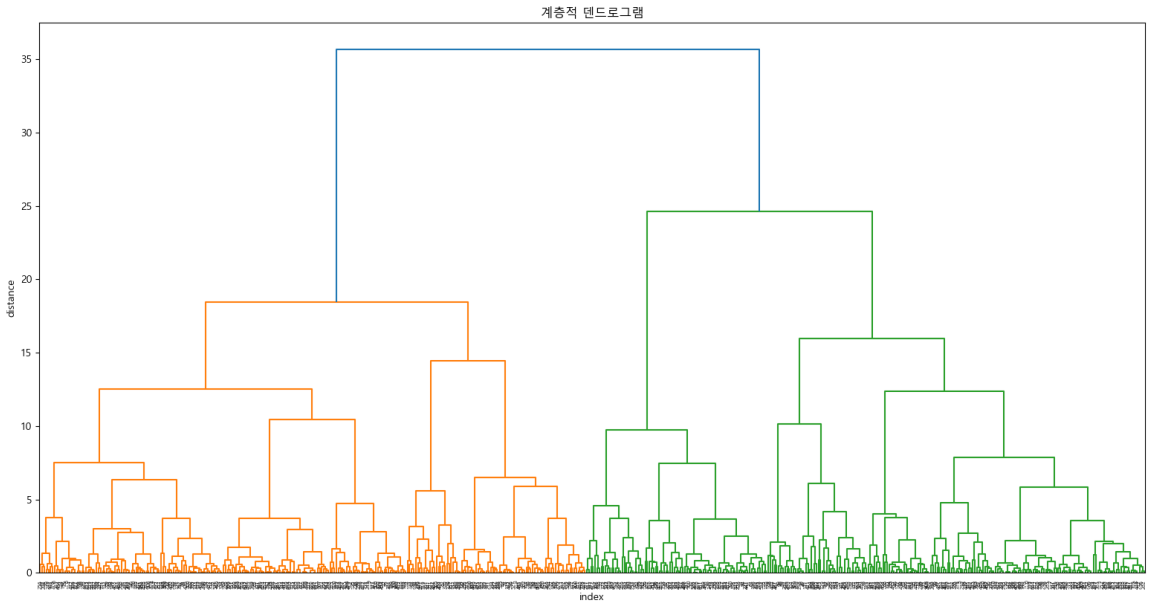
linkage = 'ward'

실루엣 계수 : 0.274

CH 점수 : 226

군집	수량	비율
0	268	50.5
1	262	49.5
총합	530	100%

ⓑ 모델 덴드로그램



* 분석 결과

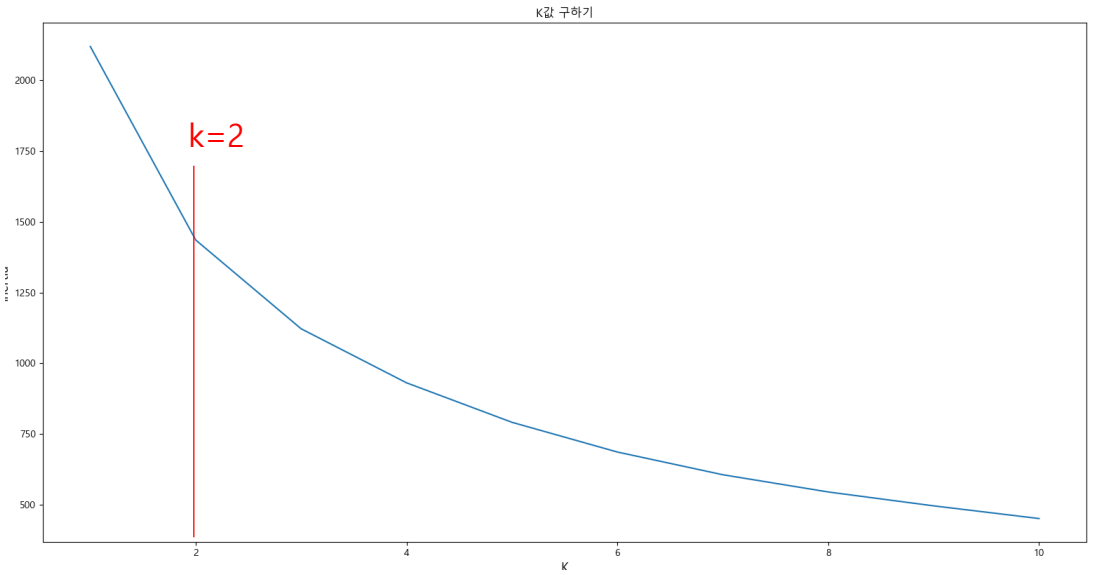
ⓒ 비이탈, 가입기한 12개월인 고객 16명의 분류 예측 결과

AGE	SEX	LONGEVIT	PAY_SUM	가입나이	가입나이	연령대	계층적군집예측
41	1	12	220000	40	40대	40대	0
40	2	12	600000	39	30대	40대	1
26	2	12	110000	25	20대	20대	0
20	2	12	120000	19	10대	20대	0
33	1	12	550000	32	30대	30대	1
38	1	12	240000	37	30대	30대	0
42	1	12	120000	41	40대	40대	0
53	2	12	220000	52	50대	50대	0
31	1	12	240000	30	30대	30대	0
9	2	12	330000	8	유아	유아	0
36	1	12	220000	35	30대	30대	0
45	1	12	240000	44	40대	40대	0
40	1	12	240000	39	30대	40대	0
48	2	12	240000	47	40대	40대	0
21	2	12	360000	20	20대	20대	0
46	1	12	200000	45	40대	40대	0

실루엣 -1~1, 1:good / CH 클 수록 군집 간 분산이 크다

6. K-평균 군집분석

* K값 찾기



* 모델 평가

k = 2

실루엣 계수 : 0.290

CH 점수 : 251

군집	수량	비율
0	271	51.2%
1	259	48.8%
총합	530	100%

* 분석 결과

비이탈, 가입기한 12개월인 고객 16명의 분류 예측 결과

AGE	SEX	LONGEVIT	PAY_SUM_	가입나이	가입나이인	연령대	K-mean군집예측
41	1	12	220000	40	40대	40대	0
40	2	12	600000	39	30대	40대	1
26	2	12	110000	25	20대	20대	0
20	2	12	120000	19	10대	20대	0
33	1	12	550000	32	30대	30대	1
38	1	12	240000	37	30대	30대	0
42	1	12	120000	41	40대	40대	0
53	2	12	220000	52	50대	50대	0
31	1	12	240000	30	30대	30대	0
9	2	12	330000	8	유아	유아	1
36	1	12	220000	35	30대	30대	0
45	1	12	240000	44	40대	40대	0
40	1	12	240000	39	30대	40대	0
48	2	12	240000	47	40대	40대	0
21	2	12	360000	20	20대	20대	1
46	1	12	200000	45	40대	40대	0

실루엣 -1~1, 1:good / CH 클 수록 군집 간 분산이 크다

7. DBSCAN

* 모델 찾기

eps	min_san	실루엣	CP
0.8	18	0.249837711	56.34724035
0.8	20	0.242532782	55.8949508
0.8	19	0.243057747	54.28838131
0.7	16	0.230512742	54.1860207
0.7	15	0.236434224	54.05518893
0.7	17	0.073796006	47.71511227
0.9	20	0.252994802	47.19689356
0.8	15	0.142017459	41.71649638
0.6	20	0.028779929	40.97802124
0.6	13	0.017443716	40.43594926

eps : 0.5~1.0 / 0.1간격

min_samples : 5~20 / 1간격

*채택

CP, 실루엣 최대 기준

eps : 0.8

min_samples : 18

* 모델 평가

eps : 0.7

min_samples : 10

실루엣 계수 : 0.121

CH 점수 : 25

군집	수량	비율
-1	99	18.6%
0	409	77.1%
1	22	4.1
총합	530	100%

* 분석 결과

ⓒ 비이탈, 가입기한 12개월인 고객 16명의 분류 예측 결과

AGE	SEX	LONGEVIT	PAY_SUM_	가입나이	가입나이연	연령대	DBSCAN군집예측
41	1	12	220000	40	40대	40대	-1
40	2	12	600000	39	30대	40대	-1
26	2	12	110000	25	20대	20대	-1
20	2	12	120000	19	10대	20대	-1
33	1	12	550000	32	30대	30대	-1
38	1	12	240000	37	30대	30대	-1
42	1	12	120000	41	40대	40대	-1
53	2	12	220000	52	50대	50대	-1
31	1	12	240000	30	30대	30대	-1
9	2	12	330000	8	유아	유아	-1
36	1	12	220000	35	30대	30대	-1
45	1	12	240000	44	40대	40대	-1
40	1	12	240000	39	30대	40대	-1
48	2	12	240000	47	40대	40대	-1
21	2	12	360000	20	20대	20대	-1
46	1	12	200000	45	40대	40대	-1

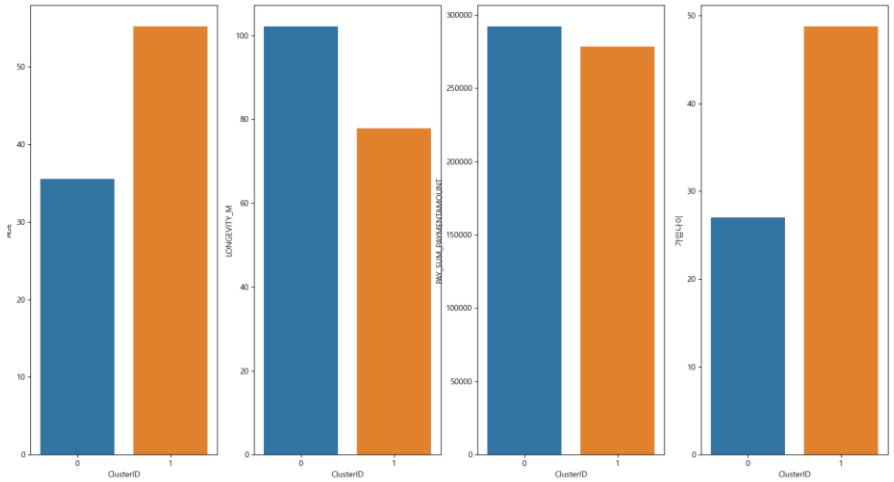
실루엣 -1~1, 1:good / CH 클 수록 군집 간 분산이 크다

모델 비교 - 군집분석

* 군집 품질 비교

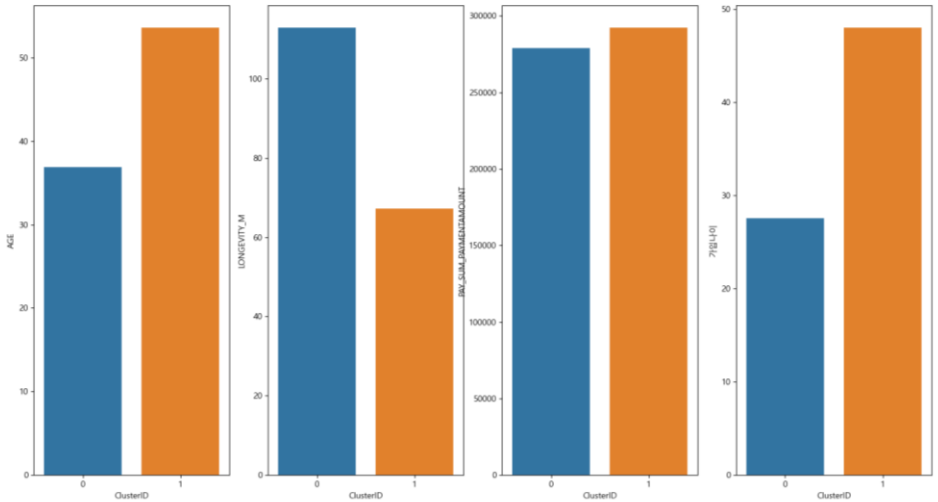
	실루엣	CH
계층적	0.274	226
K-means	0.290	251
DBSCAN	0.121	25

* 군집별 고객 프로파일링



K- means

계층적
AGE – 가입월수 – 총납입금액 - 가입나이



DBSCAN