디지털 신기술 혁신공유대학 X 신한금융그룹 빅데이터 해커톤 대회

신.빅.해

증권 성향 고객 예측모형 개발

트랙2. 데이터 분석 - 신한투자증권

팀명 SS501

팀원 조민영 (팀장)
고나경
정다운
정재원
최호경

MIRICOMPANY

CONTENTS

목차

STEP 1.

팀 소개

Shinhan(신한) X Sookmyung(숙명) 5명이 모여 0으로부터 하나를 이룬다!

SS501

조민영(팀장)

숙명여자대학교 통계학 전공

고나경

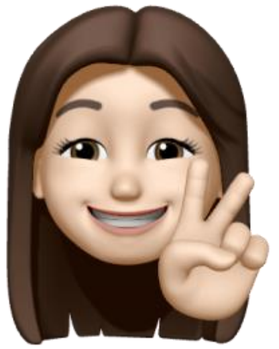
숙명여자대학교 통계학 전공

정다운

숙명여자대학교 통계학 전공

정재원

숙명여자대학교 통계학 전공



최호경

숙명여자대학교 통계학 전공

Preview

[분석 목표]

1. 고객의 소비 데이터 중 증권 성향에 영향을 미치는 유의미한 변수 찾기
2. 성향이 비슷한 고객을 군집화 하여 각 그룹의 특징 정의하기
3. 예측된 모형을 근거로 합리적인 활용 방안 도출하기

[분석 방법] 분석 도구 R과 파이썬 사용

Rank-Sum Test
표본이 서로 독립일때의 비모수 검정 방법
두 모집단의 표본의 갯수가 다를 때 사용 가능

K-Means
분류의 기준을 평균으로 하여 K개의 집단을 분류

K-Prototypes
거리와 비유사도를 가중치를 통해 조절하여 K개의 집단을 분류
연속형과 범주형 모두 사용 가능

Desicion Tree
불순도 양에 따라 feature importance 계산

STEP 2.

데이터 전처리

사용 데이터: 신청 정보, 결제 정보, 기타 정보 등 총 181개의 칼럼으로 이루어진 고객 데이터

- (P변수)
- (B변수)
- (E변수)

STEP 2.

데이터 전처리

총 168개의 방대한 양의 결제 정보 데이터 중 유의미한 변수 추출의 필요성

목 표

증권 성향 예측의 주요 변수인 금융 투자 여부에 따라 두 집단으로 나누어 분포의 형태를 비교해보자

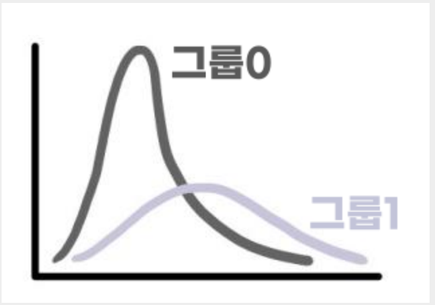
가 설

두 집단 간 분포의 차이가 있다면, 그 변수는 금투 여부에 영향을 끼치는 유의미한 변수이다

과 정

- 그룹0: 금투 경험 없음, 그룹1: 금투 경험 있음
- Rank-Sum Test를 통해 두 집단의 차이 비교

예시)



두 집단 간 차이가 있다!

결제 데이터 중 하나의 변수에 대한 두 집단의 확률밀도함수

STEP 2.

데이터 전처리

사용 통계 기법: Rank-Sum Test를 통한 비모수 검정

* 자료의 왜도가 심해 자료 변환을 통해 정규 근사를 하기 어려운 두 집단의 분포를 비교할 때 사용
* 자료가 어떤 분포를 따르는지 가정하지 않고 자료 값들의 순위를 이용하여 두 그룹 간에 차이가 있는지 확인
* 평균이 아닌 중앙값이 동일한지 여부를 확인

귀무가설	그룹0과 그룹1은 중앙값이 동일한 모집단으로부터 추출된 표본이다
대립가설	그룹0과 그룹1은 동일하지 않은 모집단으로부터 추출된 표본이다

STEP 2.

데이터 전처리

분석 결과 추출된 데이터

```r
#윌콕슨 순위 합 검정 수행
# 귀무가설(H0) : yes_sample의 중앙값 == no_sample의 중앙값
wilcox.test(yes_sample,no_sample, alternative='two.sided')

# 모든 col에 대해서 실행 / 1:다르다 0:같다
result = numeric()
for (i in 2:168) {
  yes_sample = sample(yes[,i],size=10000,replace=T)
  no_sample = sample(no[,i],size=10000,replace=T)

  if (i==112) next
  test = wilcox.test(yes_sample, no_sample)
  if (test$p.value < 0.05) a <- 1 else a <- 0
  result[i-1] <- a

}

length(which(result == 0)) # 두 집단이 같다 : 99개
length(which(result == 1)) # 두 집단이 다르다 : 67개

list(which(result == 1))
```

원본 결제 데이터 칼럼 수

166개

두 집단간 분포가 다른 변수 추출

→

검정 후 추출된 결제 데이터

67개

귀무가설을 기각 당한 결제데이터(B변수) 칼럼:

결론

위 변수들이 금투 여부에 영향을 미친다 할 수 있음

STEP 3.

중요 변수 추출

군집화에 사용할 총 변수 갯수: 74개 (신청정보 7개 + 결제정보 67개)
→ 군집화 결과 해석에 있어, 모든 변수에 대한 해석이 어려움

✓ 목 표	랜덤포레스트 permutation importance를 통해 유의미한 결과 해석을 위한 중요 변수 추출
✓ 과 정	금투여부(P5)를 y변수로 둔 뒤 나머지 73개의 변수 중 금투 여부에 가장 큰 영향을 주는 중요한 특성들을 뽑아냄

랜덤포레스트 permutation feature importance

* 특성과 실제 결과값 간의 관계(연결고리)를 끊어내도록 특성들의 값을 랜덤하게 섞은 후 모델 예측치의 오류 증가량을 측정하는 방법

하나의 특성을 무작위로 섞었을 때,

모델 오류 증가 → 모델이 예측 시 해당 특성에 의존한다는 것을 의미하기 때문에 "중요한" 특성이라 할 수 있음

반대로 오류에 차이 없음 → 그 특성은 "중요하지 않은" 특성이라고 할 수 있음

STEP 3.

중요 변수 추출

1) 언더샘플링

```
금투o 비율: 0.10079475264618187
금투x 비율: 0.8992052473538181
```

금투를 하는 사람의 비율이 매우 낮아 클래스가 불균형하기 때문에 불균형 데이터를 조정하기 위해 언더샘플링을 진행

```
X_resampled, y_resampled = RandomUnderSampler(random_state=0).fit_resample(X, y)
```

2) 사용 변수

범주형 데이터 P변수 라벨링

* P1(
* P2(

* P7(

X

: 라벨링 한 신청정보(P변수) 와 결제정보(B1~B167)

y

: P5(금투여부)

3) 데이터 스케일링

$$\frac{x - x_{min}}{x_{max} - x_{min}}$$

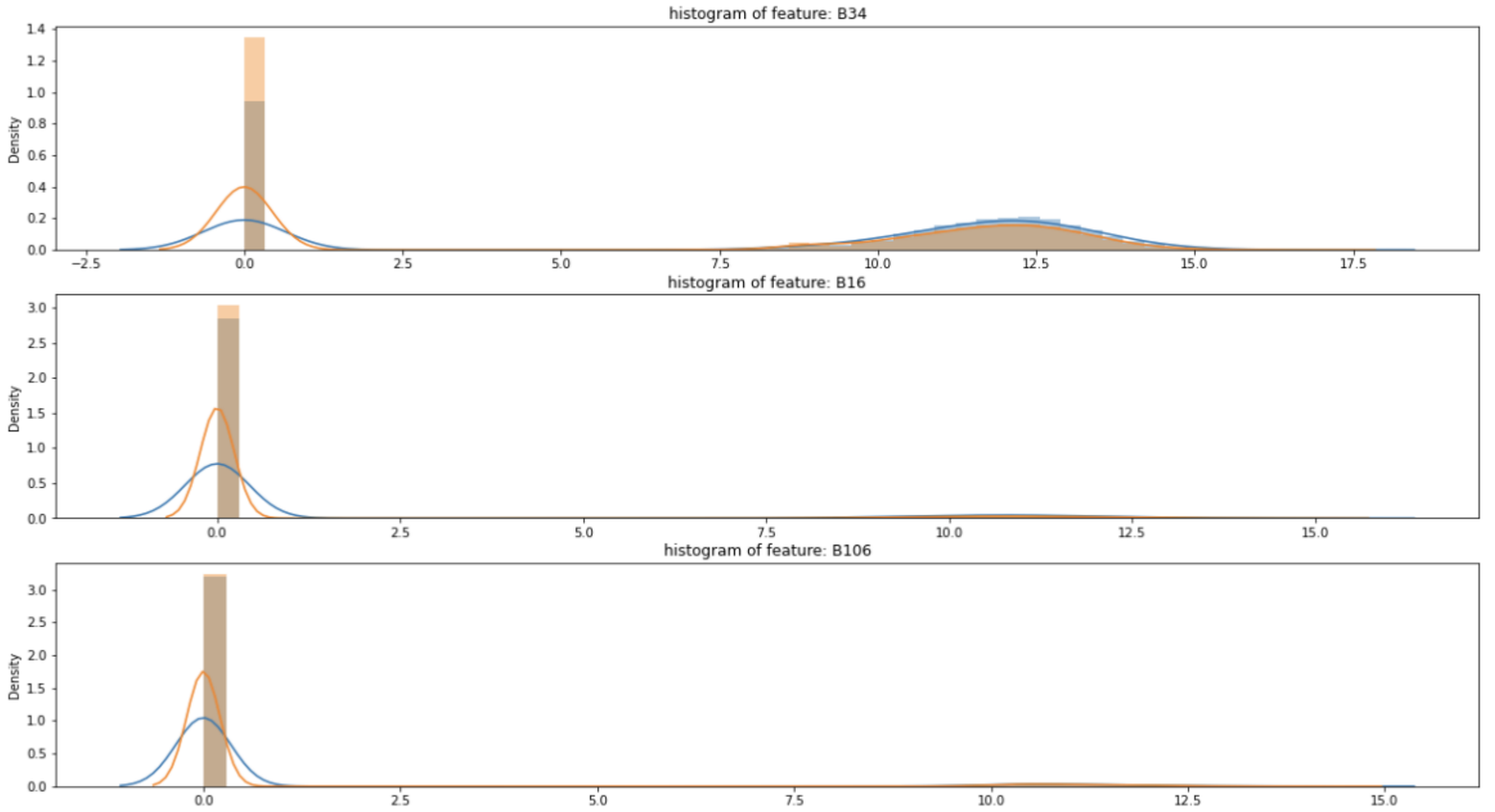
* 데이터를 일반적으로 0~1 사이의 값으로 변환
* 분포가 0의 값에 몰려있는 왜도의 값이 큰 결제 변수들의 스케일을 조정하기 위해 사용.

STEP 3.

중요 변수 추출

4) 중요변수 추출 결과

* P7
* B34
* P4
* P3
* P1
* B16
* B106



-> 중요 변수 중 연속형 데이터인 결제 데이터 (B변수)에 대한 시각화

STEP 4.

군집화

군집화

데이터에 대한 label이 주어지지 않은 상태에서

유사성을 기반으로 데이터를 분할, 그룹화 뒤 그룹별 성격을

진단해 전체 구조를 이해하는 **탐색적 분석 기법**

선택이유

증권 성향고객(Y)에 대해 k개의 군집을 나누어 각 군집별

주요 변수와의 관계를 파악한 후 **금투활동고객의 비율**로

정의하는 것이 옳다고 판단

STEP 4.

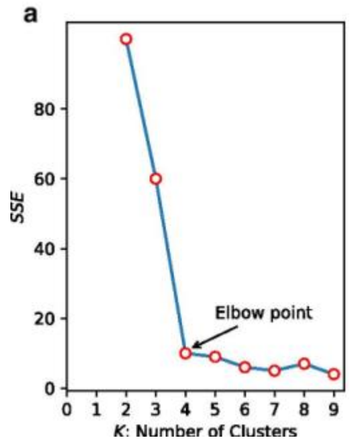
군집화

군집 방법(K-prototype clustering)

1__ K 후보 선정

```
kprototype = KPrototypes(n_jobs = -1, init = 'Huang', random_state = 0)
kprototype.fit_predict(DATA)
```

2__ ELBOW POINT 파악 후 K 선택

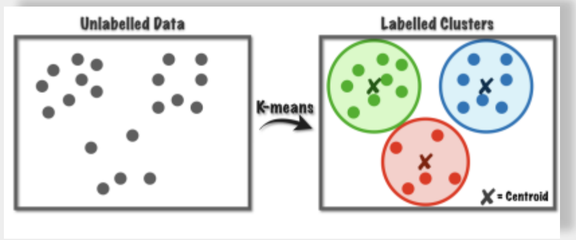


3__ LABEL 생성

```
#최적 k로 다시 군집화
kprototype = KPrototypes(n_jobs = -1, n_clusters = 3, init = 'Huang', random_state = 0)
kprototype.fit_predict(DATA, categorical = DATA Columns)

DATA['cluster_id'] = kprototype.labels_
```

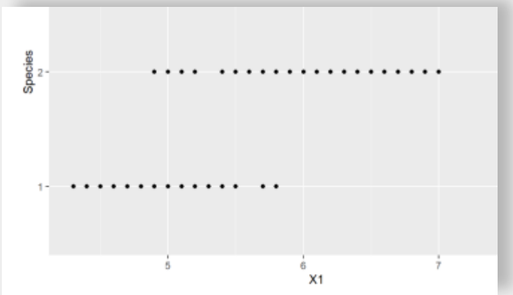
연속형



K개 집단의 분류 기준을 **평균**으로 한다

$$\sum_{j=1}^{q}(x_j - y_j)^2$$

범주형



K개 집단의 분류 기준을 **최빈값**으로 한다

$$\gamma \sum_{j=q+1}^{p} \delta(x_j y_j)$$

K-PROTOTYPE COST

중심점 위주의 모델 K-MEANS와 K-MODES의 개념을 동시에 활용하는 모델.
즉, **연속형**과 **범주형** 자료를 **동시에** 활용할 수 있는 클러스터링 방식

$$d(X,Y) = \sum_{j=1}^{q}(x_j - y_j)^2 + \gamma \sum_{j=q+1}^{p} \delta(x_j y_j)$$

STEP 4.

군집화

[군집화 과정 Preview] **세 종류의 데이터**로 군집화를 시행하였음

데이터 1 전체데이터

군집화가 뚜렷하게 이루어지지 않았음

데이터 2 * 0 값이 많은 데이터

0이 많은 고객은 군집화에 악영향을 끼치는 것을 확인

0이 적은 고객들을 통해 군집화 시행하여

고객의 소비패턴과 금투의 "뚜렷한" 관계를 알아보자 !

* 0 값이 적은 데이터 **데이터 3**

* 0 값이 많은/적은 데이터 ?

: 값이 0인 열의 개수가 많은/적은 행들로 이루어진 데이터

STEP 4.

군집화

: 데이터 1

[데이터 1]

데이터 1 = **전체 데이터**

비모수검정을 통해 추출된 중요 변수들의 전체 데이터로 군집화 시행

[데이터 전처리]

1) 비모수 검정을 통해 유의미한 변수 추출 후 데이터 생성

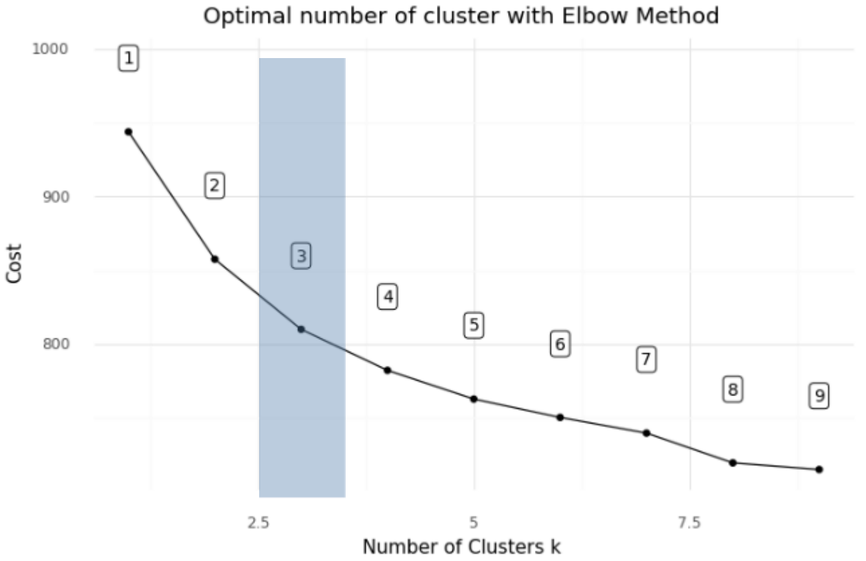
2) MinMaxScaler로 데이터 스케일링

3) 데이터 타입 변경

4) 금융투자 활동고객 (P5) 기준으로 전체 데이터 분할
   - 금투 활동 집단과 그렇지 않은 집단에서 각각 5000개씩 데이터 추출

[k 결정 - Elbow Method]

k = 3 결정



STEP 4.

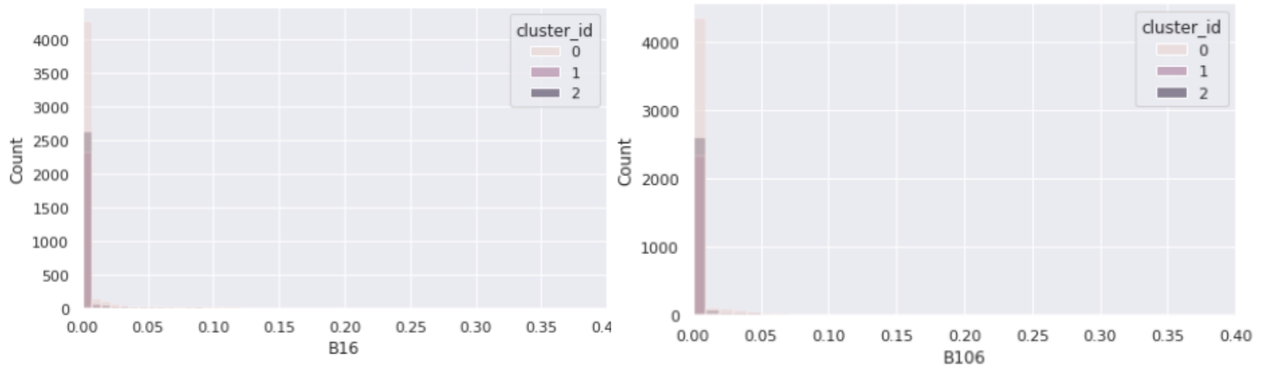
군집화

: 데이터 1

각 군집별 P5(금투여부) 비율

P5	군집 0	군집 1	군집 2
0	3783	1924	2106
1	970	549	668

랜덤포레스트 중요도 변수 (B16, B106)에 대한 군집의 비율



* 군집 개수 별 차이가 존재함
  ⇒ 군집이 균등히 나뉘지 않았음

* 군집1과 2에서 금융투자 고객 비율이 유사함
  ⇒ 집단 별 금투 특성이 뚜렷하게 나타나지 않음

* 모든 군집의 값이 다 0에 몰림
  ⇒ 군집을 나누는 변수가 뚜렷하게 나타나지 않으며
  값이 다 0에 몰려있어 의미를 찾기가 힘들다

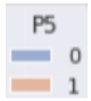
* 군집2의 경우 중요도 변수에 대한 비율이 적음

STEP 4.

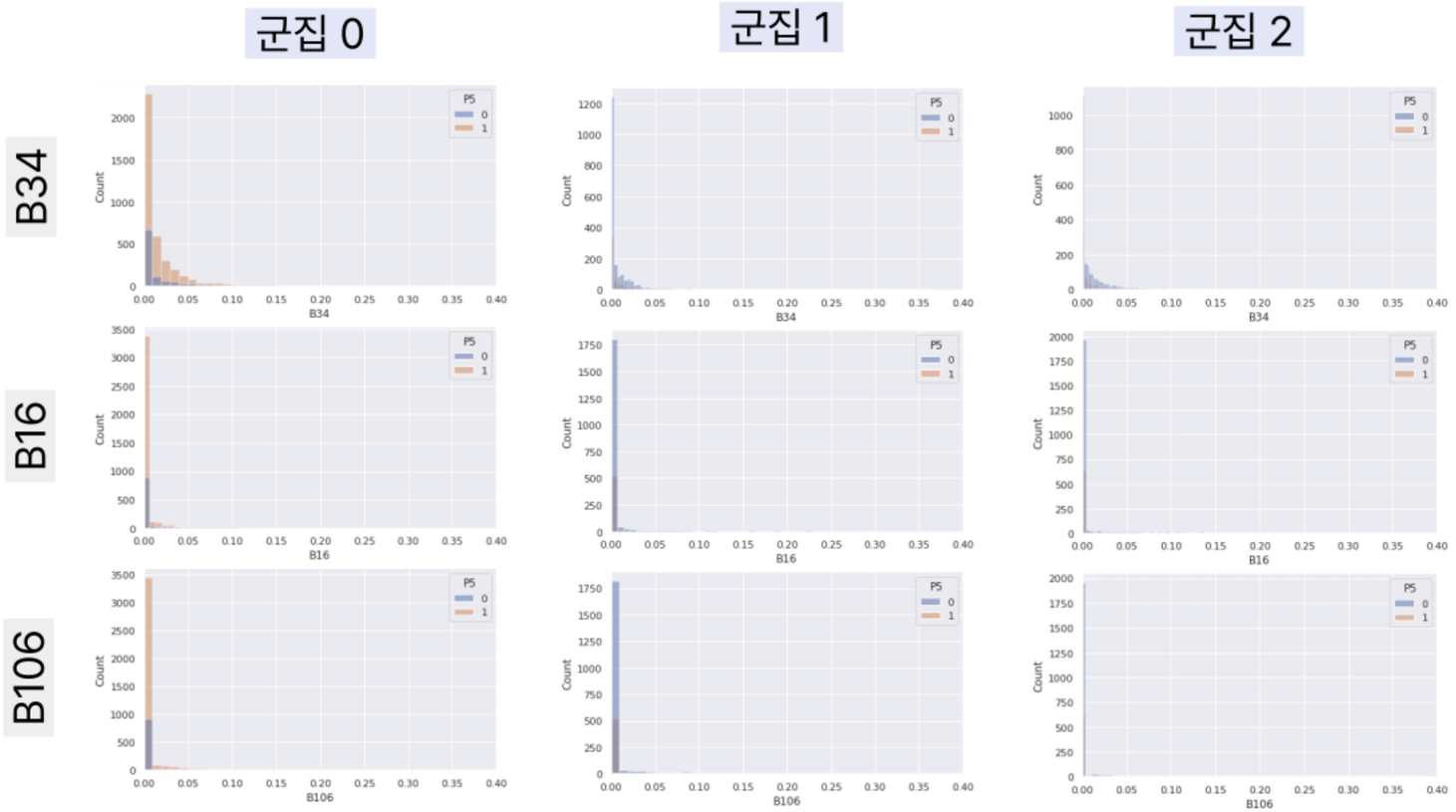
군집화

: 데이터 1

랜덤포레스트 중요도 변수 (B34, B16, B106)에 대한 군집별 시각화



금융투자여부(P5) 변수에 따라 색을 달리함



* 군집1과 군집2에서 "P5==1"과 "P5==0"의 비율이 겹침

⇒ 금융투자에 유의미한 군집화가 아님

뚜렷한 군집의 특성 나타나지 않음

STEP 4.

군집화

[문제점]

데이터 1 군집화 결과

뚜렷한 군집의 특성이 나타나지 않음

[상황]

고객들이 모든 품목(B1-B166)에 대해서 소비를 고르게 하지 않음

즉 (0,0,0,0,0,...) 형태의 데이터들이 많은 상황

0.000000	0.000000	0.0	0.000000	0.000000	0.00015	0.000000	0.000000	0.000000	0.000000	0.000000	0.0

한 고객의 B1-B12 데이터. 아예 소비를 하지 않아 0으로 차있다.

[가설]

군집화 시 "0에 몰린 군집 / 0이랑 먼 군집" 으로만 나뉠 것이다.

→ 소비와 금융 투자 사이의 관계를 담지 못 한다.

[Idea]

값이 0인 열의 개수가 많은 행들로 구성된 데이터프레임으로 군집화 시행.

군집이 잘 나눠지지 않으면, 0이 많은 고객은

군집화에 악영향을 끼친다는 가설에 타당성을 부여할 수 있다.

STEP 4.

군집화

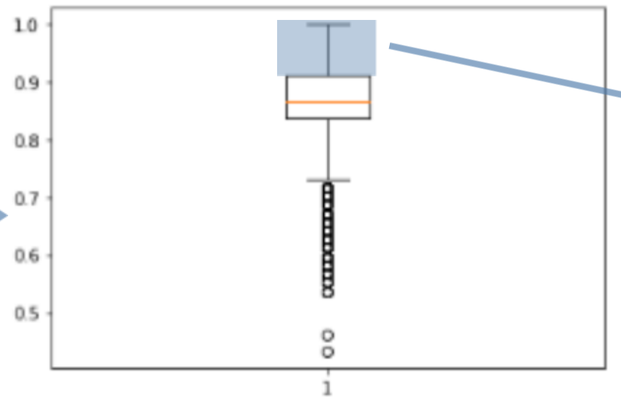
: 데이터 2

데이터 2 = **0이 많은 데이터**

비모수 검정으로 뽑힌 변수들에 대해서,
각 행(고객)마다 값이 0인 열(변수)의 개수와
비율을 담은 새로운 데이터프레임 생성

"0의 비율"(= 각 행의 '0인 변수의 개수/모든 변수의 개수')
을 기준으로 제 4분위수의 데이터들을 사용함

	count_0	count_non0	ratio_0
31520	67	0	1.000000
397609	67	0	1.000000
459173	67	0	1.000000
59201	67	0	1.000000
42784	67	0	1.000000
...	...	...	...
121091	36	31	0.537313
106403	36	31	0.537313
29347	36	31	0.537313
282839	31	36	0.462687
344398	29	38	0.432836



0의 비율에 대한 boxplot
파란색 박스의 부분이 제 4분위수이며 0의 비율이 가장 높은 곳이다.

	P1	P2	P3	P4	P5	P6	P7	B1	B4	B7	B9
0	F	40대_초	0	1	0	0	B은행	0.000000	0.000000	0.0	0.000000
1	F	30대_초	0	0	0	0	B은행	0.000000	0.000000	0.0	0.000000
2	M	40대_후	0	1	0	0	B은행	0.000000	0.000000	0.0	0.050919
3	M	40대_초	0	1	0	0	B은행	0.000000	0.203877	0.0	0.000000
4	M	40대_후	0	1	0	1	B증권사	0.000000	0.000000	0.0	0.000000

STEP 4.

군집화

: 데이터 2

[데이터 전처리]

1) 비모수 검정을 통해 유의미한 변수 추출 후 데이터 생성

2) MinMaxScaler로 데이터 스케일링

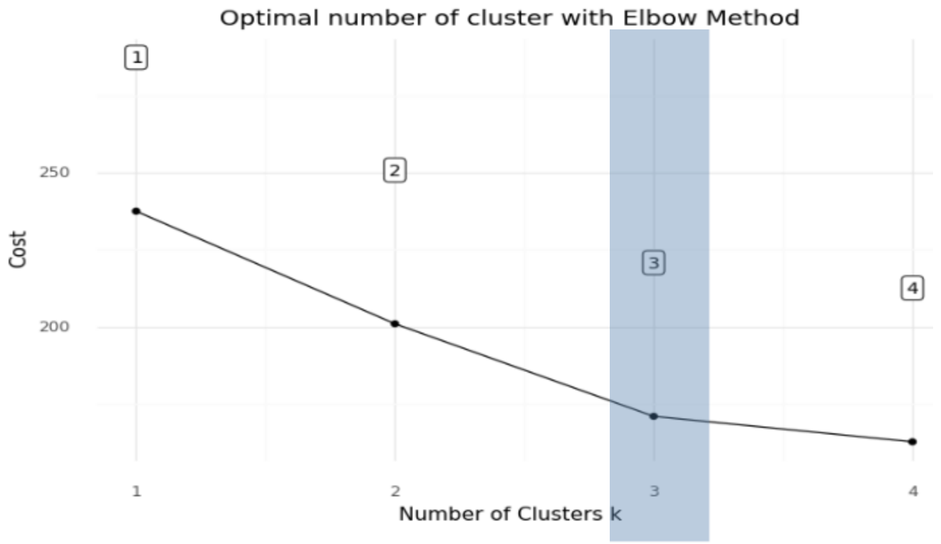
3) 데이터 타입 변경

4) 무작위 표본을 5000개 추출

모든 데이터를 이용하기엔 하드웨어 메모리 상의 문제가 생겨 표본을 추출하였음.

[k 결정 - Elbow Method]

k = 3 결정



STEP 4.

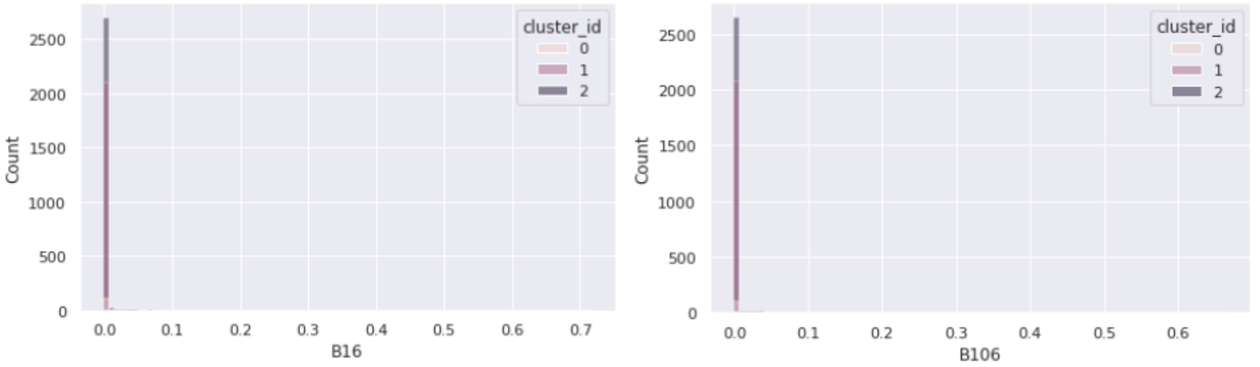
군집화

: 데이터 2

각 군집별 P5(금투여부) 비율

P5	군집 0	군집 1	군집 2
0	102	2060	2510
1	2	78	248

랜덤포레스트 중요도 변수 (B16, B106)에 대한 군집의 비율



* 군집0의 개수가 군집1,2 에 비해 현저히 적음
  ⇒ 군집이 균등히 나뉘지 않았음

* 군집0에서 "P5==1"개수가 2로 굉장히 적음
  ⇒ 군집0에대한 금투 분석이 어려움

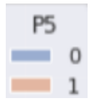
* 모든 군집의 값이 다 0에 몰려있음
  ⇒ 각 변수에서 군집의 분포의 차이를 확인 할 수 없다

STEP 4.

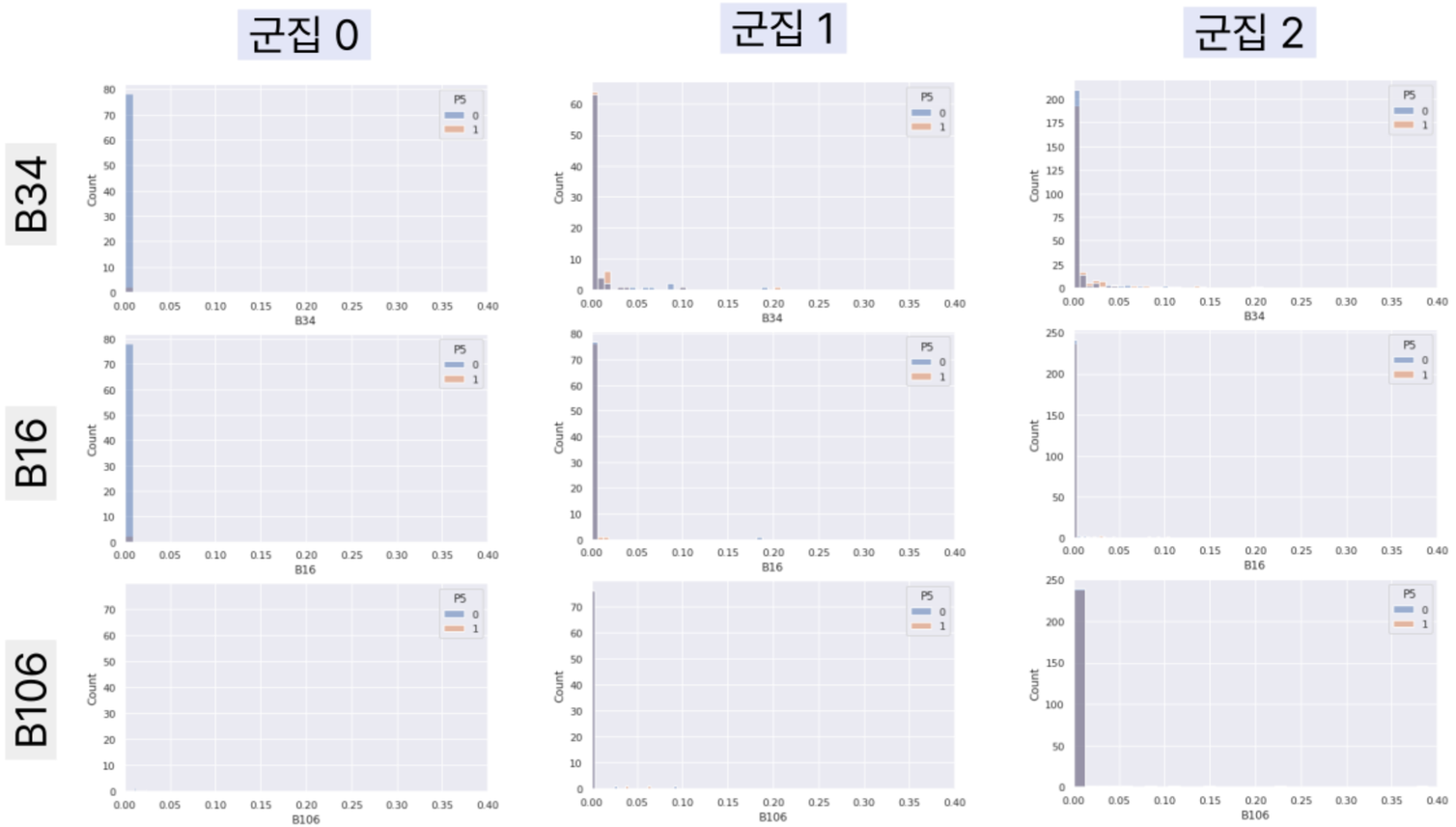
군집화

: 데이터 2

랜덤포레스트 중요도 변수 (B34, B16, B106)에 대한 군집별 시각화



금융투자여부(P5) 변수에 따라 색을 달리함



* 값이 다 0에 몰려 분포가 뚜렷히 나타나지 않음

⇒ 군집을 나누는 변수가 뚜렷하게 나타나지 않으며 군집의 의미를 찾기가 힘들다

* 군집1과 군집2에서 "P5==1"과 "P5==0"의 비율이 겹침

⇒ 금융투자에 유의미한 군집화가 아님

군집이 적절히 이루어지지 않음

0이 많은 고객은 군집화에 악영향을 끼칠 수 있다 !

STEP 4.

군집화

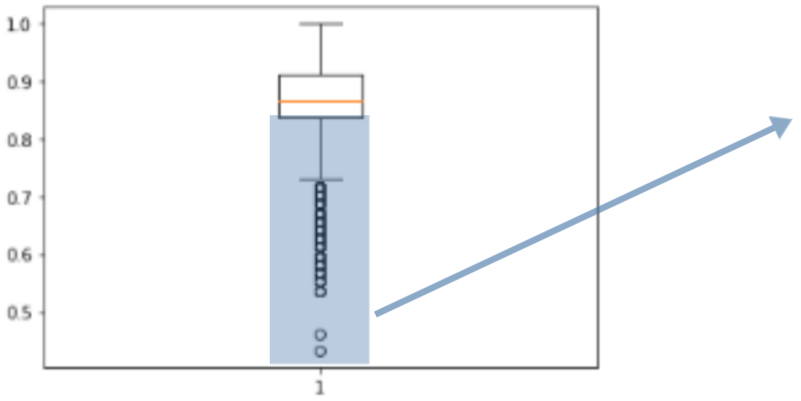
: 데이터 3

데이터 3 = **0이 적은 데이터**

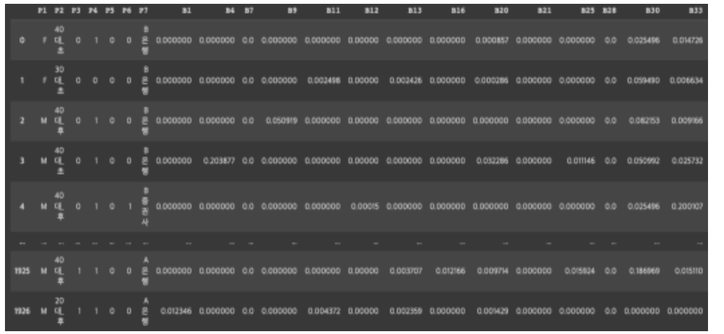
비모수 검정으로 뽑힌 변수들에 대해서,
각 행(고객)마다 값이 0인 열(변수)의 개수와
비율을 담은 새로운 데이터프레임 생성

	count_0	count_non0	ratio_0
31520	67	0	1.000000
397609	67	0	1.000000
459173	67	0	1.000000
59201	67	0	1.000000
42784	67	0	1.000000
...	...	...	...
121091	36	31	0.537313
106403	36	31	0.537313
29347	36	31	0.537313
282839	31	36	0.462687
344398	29	38	0.432836

"0의 비율"(= 각 행의 '0인 변수의 개수/모든 변수의 개수')
을 기준으로 제 1분위수의 데이터들을 사용함



0의 비율에 대한 boxplot
파란색 박스의 부분이 제 1분위수이며 0의 비율이 가장 낮은 곳이다.



STEP 4.

군집화

: 데이터 3

[데이터 전처리]

1) 비모수 검정을 통해 유의미한 변수 추출 후 데이터 생성

2) MinMaxScaler로 데이터 스케일링

3) 데이터 타입 변경

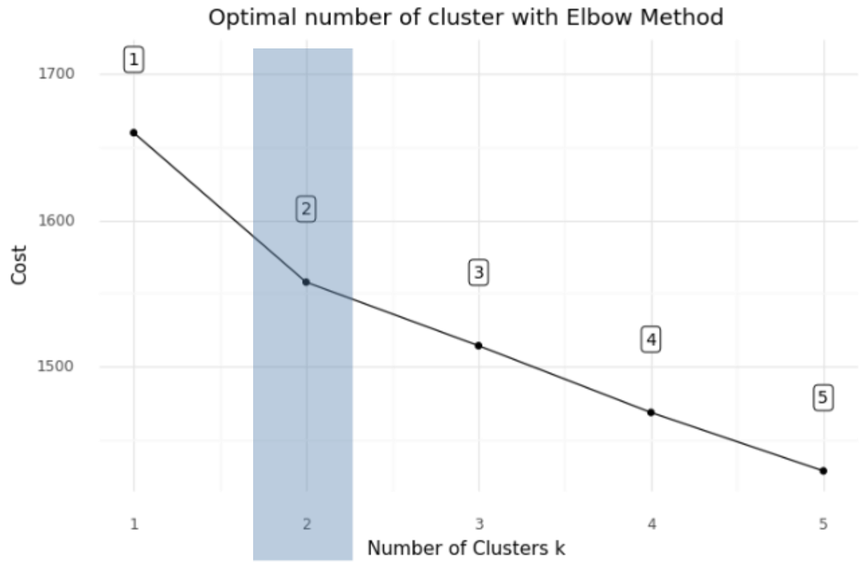
4) 금융투자 활동고객 (P5) 기준으로 전체 데이터 분할

- 금투 활동 집단과 그렇지 않은 집단에서 각각 5000개씩 데이터 추출

모든 데이터를 이용하기엔 하드웨어 메모리 상의 문제가 생겨 표본을 추출하였음.

[k 결정 - Elbow Method]

k = 2 결정



STEP 4.

군집화

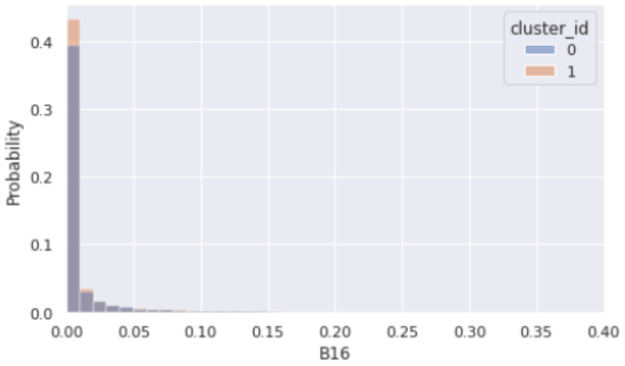
: 데이터 3

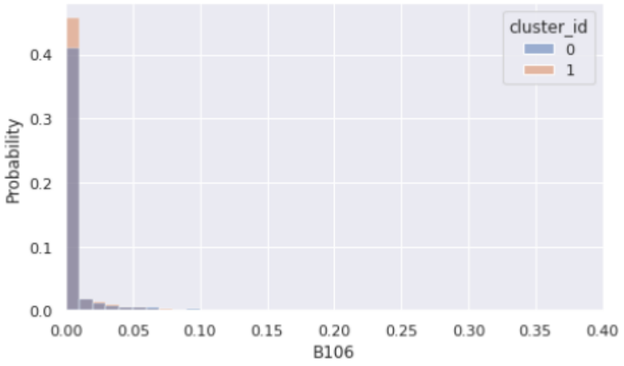
각 군집별 P5(금투여부) 비율

P5	군집 0	군집 1
0	2906	3104
1	896	1094

* 군집0의 개수와 군집1의 개수가 큰 차이를 보이지 않음
  ⇒ 군집이 균등히 나눠짐

랜덤포레스트 중요도 변수 (B16, B106)에 대한 군집의 비율





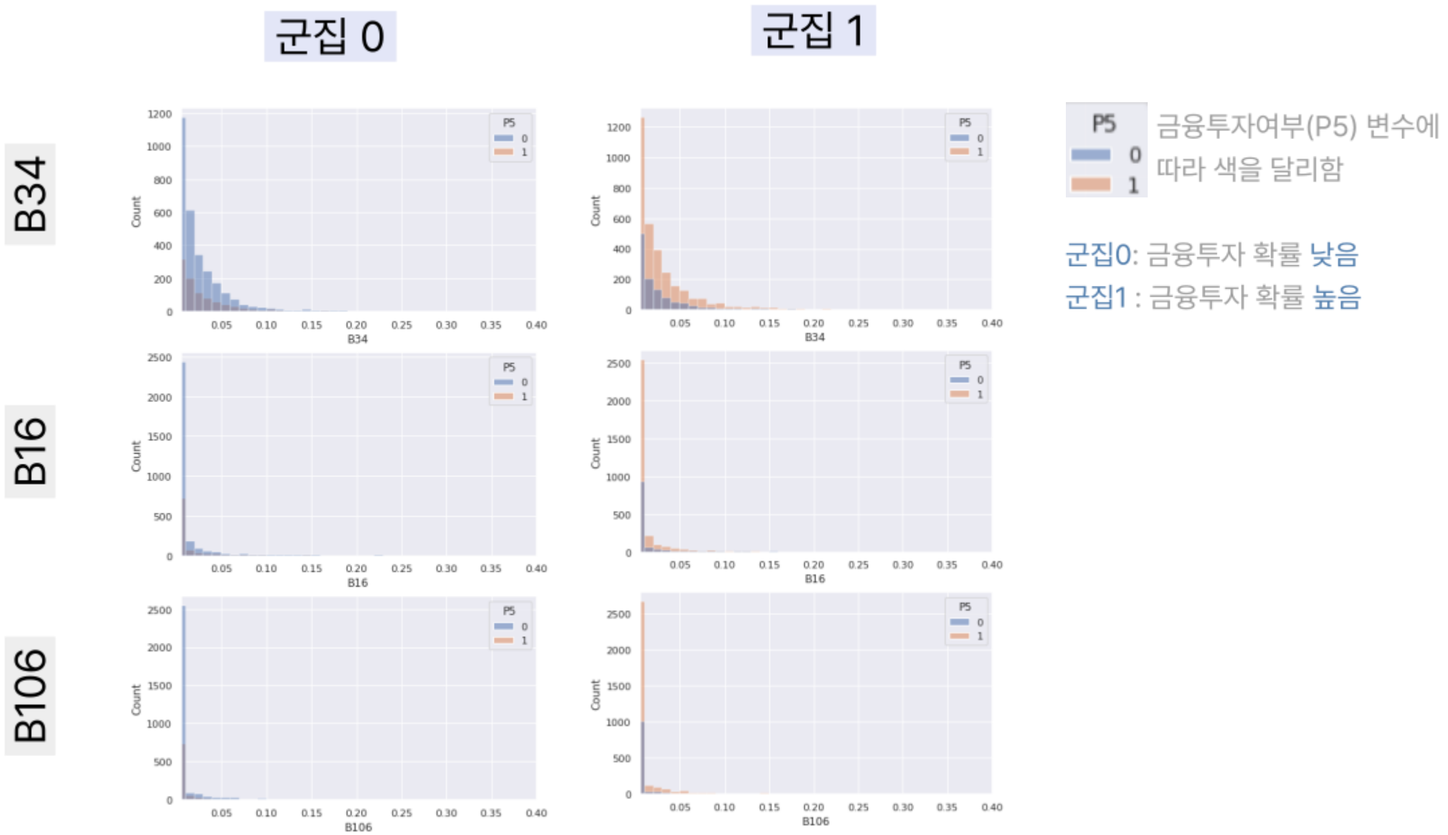
* 변수마다 군집의 비율이 비슷하며 값이 다 0에 몰림
  ⇒ 0에 몰린 데이터가 많긴 하지만 데이터 1,2에 비해 심하게 몰려있지는 않음

STEP 4.

군집화

: 데이터 3

랜덤포레스트 중요도 변수 (B34, B16, B106)에 대한 군집별 시각화



P5 금융투자여부(P5) 변수에 따라 색을 달리함

군집0: 금융투자 확률 낮음
군집1 : 금융투자 확률 높음

* 군집1과 군집2에서 "P5==1"과 "P5==0"의 비율이 뚜렷히 차이남
⇒ 각 군집이 금융투자 여부를 반영해 나뉘어졌음을 알 수 있음

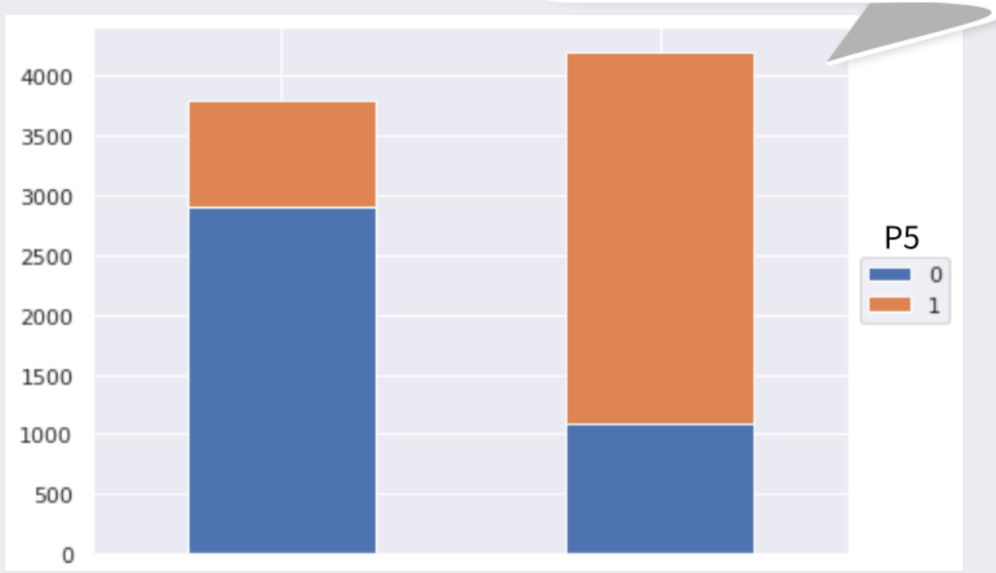
최종 군집으로 선택함

STEP 5.

결론

: 각 군집에 대한 특징

상대적으로 군집0에 비해
금융투자 비율이 더 높다.



상대적으로 군집1에 비해
금융투자 비율이 더 낮다.

금투 비율이 더 높은

군집1의 고객들을

증권성향 고객

으로 정의

+) 군집0은 비증권성향의 고객으로 정의

STEP 5.

결론

: 랜포 중요도 변수 (B변수)

증권성향의 고객은 비증권성향의 고객에 비해

[B16 :]

군집 0		군집 1
0.0092	<	0.0102

[B34 :]

군집 0		군집 1
0.0257	<	0.0278

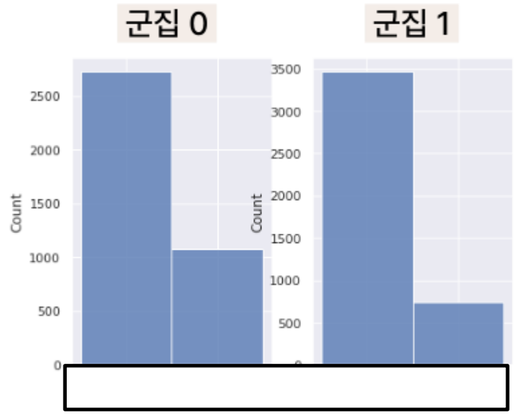
[B106 :]

군집 0		군집 1
0.0093	>	0.0085

STEP 5.

결론 : 랜포 중요도 변수 (P변수)

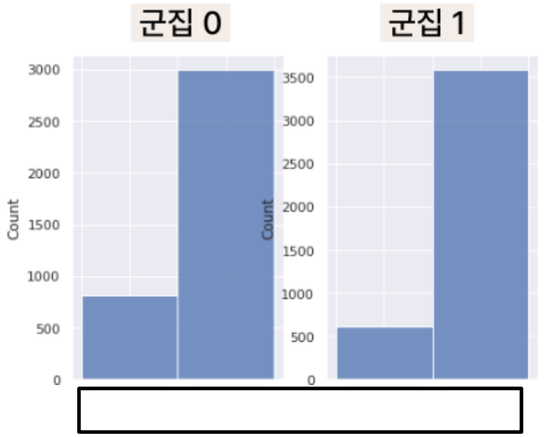
P1 :



군집1에서 의
비율이 높다

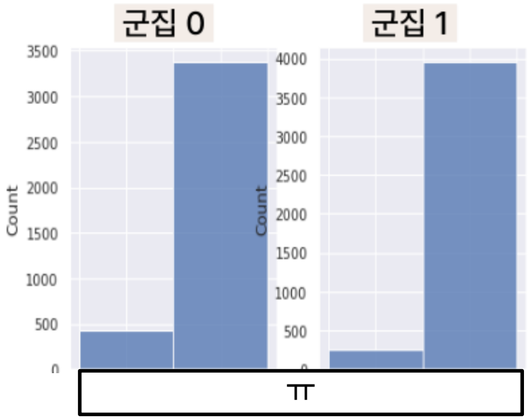
의 증권 성향이 높다

P3 :



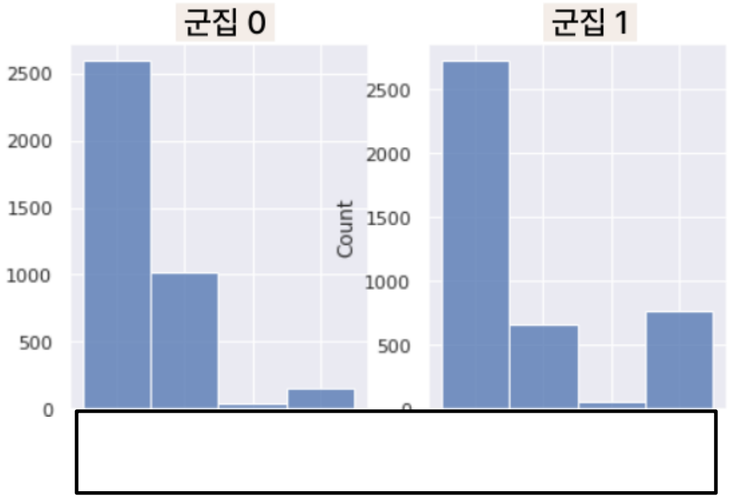
는
증권 성향에
유의미한 영향을
미치지 않는다

P4 :



는
증권 성향에
유의미한 영향을
미치지 않는다

P7



군집1에서 상대적으로
비율이 높고,
비율이 낮다

을 덜 이용하고
을 더 사용하는
고객의 증권 성향이 높다

STEP 6.

활용방안

군집 0 : 비증권 성향 고객

군집 1 : 증권 성향 고객