금융 데이터를 활용한
“나의 금융생활정보 지수” 개발

2018 BIG CONTEST

FRAME

이주영, 김현우, 민은주, 박주연, 이지예

REPORT CONTENTS

- 제공 데이터
- 단계별 수행과제

0. 문제 소개

1. 문제 정의

- 분석 목표
- ‘보통 사람’에 대한 정의

- 변수 형태 정의
- 이상치 제거
- 결측치 처리

2. 변수 처리

3. 분석

- 문제 1번
- 문제 2번
- 문제 3번
- 문제 4번

- 한계점
- 해결 방안

4. 한계 및 보완

00 문제 소개

제공 데이터

- '신한은행 2018 보통사람 금융생활보고서'
- 약 1만 7천 명의 설문조사 DATA SET + 신한카드 평균거래 정보
- 고객정보(8개 - 성별, 나이, 직업, 지역, 소득, 결혼, 맞벌이 여부, 자녀 수), 금융 거래 정보(26개)

금융 거래 정보 (26개)

총 자산

금융 자산
부동산 자산
기타 자산

총 부채

신용 대출
담보 대출
아파트/주택 담보대출
전세 자금 대출

월 총 저축액

적금
펀드
주식
펀드/주식
저축성보험
청약

그 외 변수

청약 보유 여부
은퇴 후 필요자금
금융상품잔액_정기예금
...

00 문제 소개

단계별 수행과제

- 문제1 : 고객기본정보 8개 항목 중 5가지 필수정보(성별, 연령, 지역, 직업, 가구소득)와 3가지 선택정보(결혼여부, 맞벌이여부, 자녀 수)를 모두 조합하면 141,750개 고객 유형을 만들 수 있으나, 결측치가 발생합니다. **141,750개 고객유형의 25개 금융거래정보 항목의 결측치를 추정**하시오.

- 문제2 : 문제1의 금융거래 정보를 이용하여 유사한 집단을 **Peer Group으로 묶으시오.**

- 문제3 : 고객이 8개의 고객기본정보와 본인의 "금융자산", "월 저축 금액", "월 소비 금액"을 입력하면 소속된 **Peer Group**을 찾아 금융점수를 제시할 수 있도록 Peer group의 "금융자산", "월 저축 금액", "월 소비 금액"의 금액 **분포를 백분위**로 표시하시오.

- 문제4 : **고객기본정보 8개 항목 중 고객 정보 수집을 최소화**하여 상담시스템을 만들 수 있다면 필요한 정보는 무엇인지 근거를 설명하시오.

출처 : Innovation 분야 금융문제 설명회 자료

01 문제 정의

분석 목적

- 8가지의 개인 정보 입력 → 비슷한 개인 정보를 가진 사람들의 평균 금융 정보 제공
- 상담 시스템을 구축하여 금융 생활에 도움

분석 대상

‘보통 사람’

- ‘유사한 8가지 개인 정보를 가진 일반적인 사람에게 기대되는 금융 상태를 가진 사람’이라고 정의
- 또한, ‘상담시스템과 창구를 이용할 것이라 예상되는 주요 이용 고객 층’으로 가정

02 변수 처리

변수 형태의 정의

- **명목 변수** : 성별, 직업 구분, 지역 구분, 결혼 여부, 맞벌이 여부, 청약 보유 여부
- **순위 변수** : 연령, 가구 소득 구간
- **연속형 변수** : 그 외의 변수
- **단위 통일** : 월 평균 카드 사용 금액을 10000으로 나눠 단위를 만 원으로 맞춤

이상치

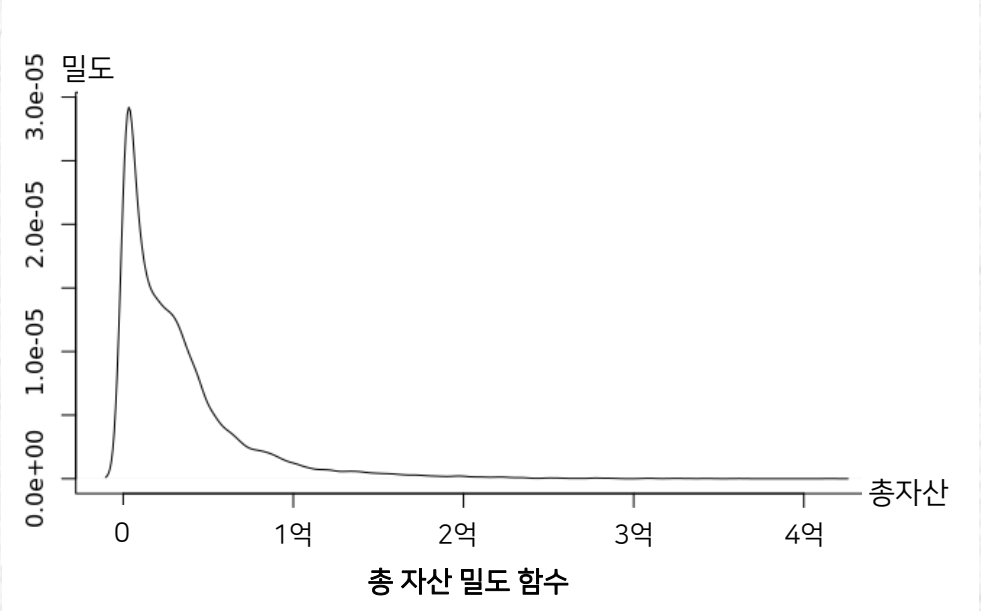
- '보통 사람' 에서 벗어나는 사람
- '연령'과 '가구 소득' 구간별 총 자산의 상위 5%와 하위 1% 내외
- '가구 소득'의 증가에 따른 총 자산의 증가를 만족하지 않는 유형

02 변수 처리

이상치 제거 과정

- 금융정보를 가장 대표하는 변수인 총 자산의 분포가 치우친 것을 확인
- 동일한 기본 정보를 가지고 있어도 총 자산의 편차가 큰 문제 존재
- 대푯값의 왜곡 → 이상치 제거 필요

총 자산의 분포



동일 기본 정보 내에서도 큰 총 자산의 편차 - 예시

인덱스	성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수	총자산
231	1	2	2	2	6	1	-	-	79100
8221	1	2	2	2	6	1	-	-	71200
8447	1	2	2	2	6	1	-	-	47300
22392	1	2	2	2	6	1	-	-	940
23066	1	2	2	2	6	1	-	-	6980

02 변수 처리

이상치 제거 과정

- 제거할 비율 결정 위해 상위, 하위 10%의 총 자산을 확인
- 상위 10% 사람들이 전체 자산에서 차지하는 비중 확인→ 많은 사람들을 대표할 수 있는 백분위 수 설정
- 총 자산만을 기준으로 제거 시 소득 구간이 높은 데이터가 대부분 제거, 이는 이상치가 아니라고 판단

상위 10%에 해당하는 총 자산의 백분위 수

90	91	92	93	94	95	96	97	98	99	100
75625	79916	84350	89500	95850	104000	116550	132055	152005	185993	415500

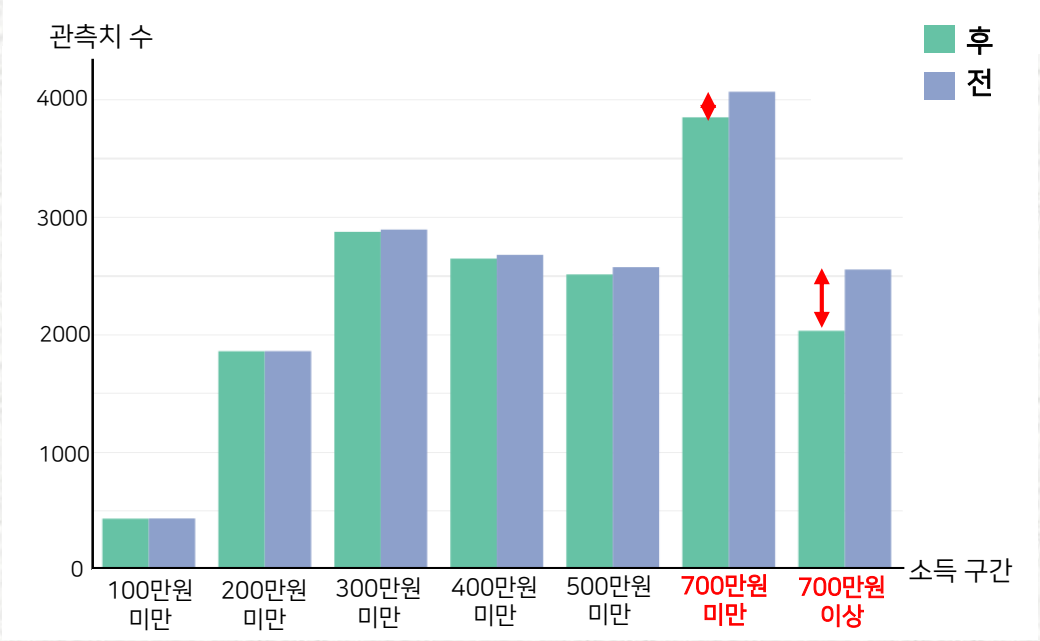
하위 10%에 해당하는 총 자산의 백분위 수

0	1	2	3	4	5	6	7	8	9	10
30	134.5	270	400	553	700	880	1050	1190	1380	1550

상위 10% 사람들이 전체 자산에서 차지하는 비중

0	1	2	3	4	5	6	7	8	9	10
0	7.2	12.4	16.8	20.8	24.5	27.3	30.2	32.9	35.5	37.9

상위 5% 제거 전과 제거 후 소득 구간 별 총 자산 분포

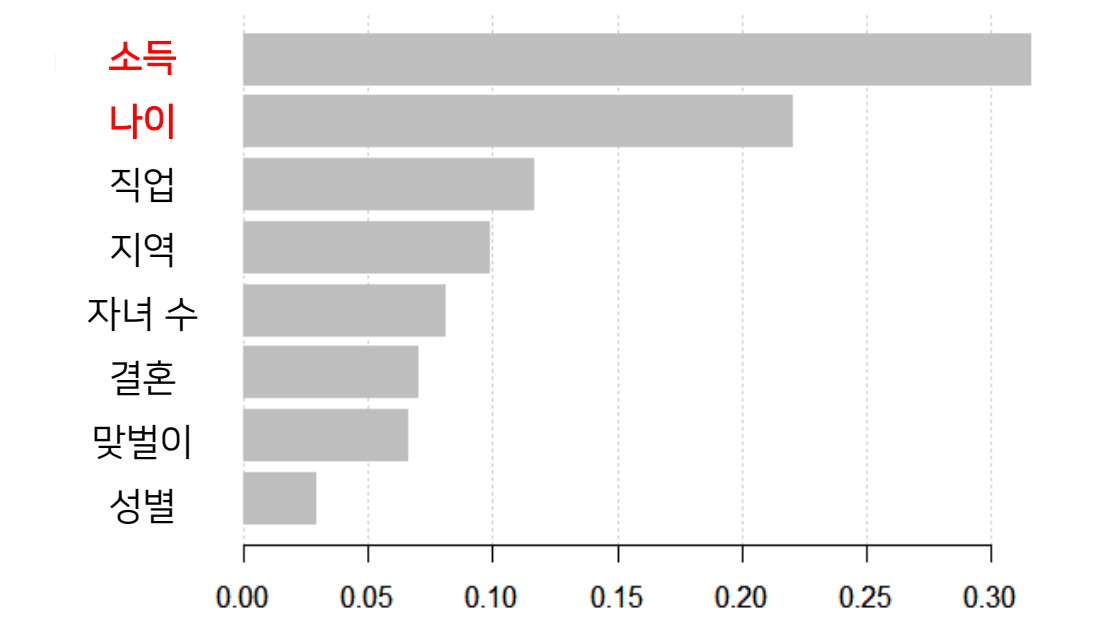


02 변수 처리

이상치 제거 과정

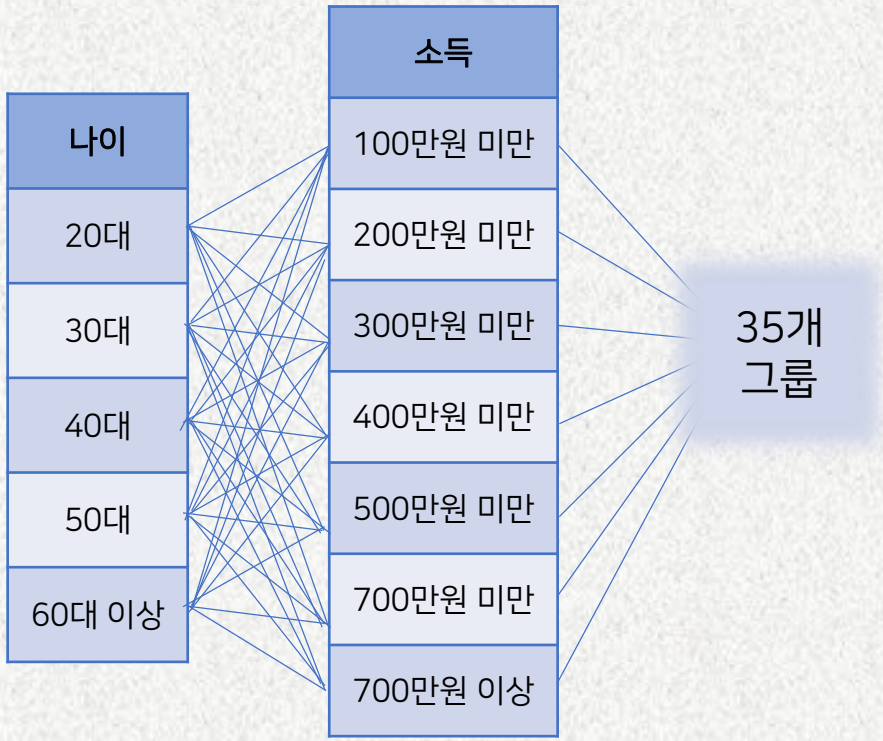
- XGBoost의 Gain importance를 이용한 변수 중요도 결과를 통해 총자산에 대한 기본 변수들의 영향 확인
- 나이(5개 구간), 소득 구간(7개 구간)에 따른 35개 그룹에서의 이상치 제거

총 자산에 영향을 미치는 기본변수 중요도



- 나이와 소득 구간이 총 자산에 가장 영향이 큰 변수인 것을 확인
  → 두 변수를 기준으로 이상치 제거

두 변수에 따른 그룹 생성



02 변수 처리

이상치 제거 과정

- 나이와 소득 구간 별 백분위수 확인, '보통 사람'으로 판단되는 수치를 고려

나이와 소득 구간 별 백분위수 일부

나이	소득	백분위수 1	백분위수 2	백분위수 3	백분위수 90	백분위수 91	백분위수 92	백분위수 93	백분위수 94	백분위수 95	백분위수 96	백분위수 97	백분위수 98	백분위수 99
2	1	40	56	74	10270	10337	10600	11301	12121	14153	17059	18100	23098	36479
2	2	50	100	100	10450	10920	11660	12500	13500	15400	18300	21300	30900	46600
2	3	106	155	238	17433	18171	19712	22044	24460	28090	32784	41500	53064	82037
2	4	190	201	275	34400	37000	40480	43800	47721	52100	59300	62200	80500	106500
2	5	502	812	959	37461	39282	40959	42780	44607	48380	53263	67691	85045	120074
2	6	379	570	612	55620	57901	60061	62885	67296	70616	72852	82283	97458	124760
2	7	350	905	1178	98100	108410	111452	127826	131824	149555	182096	199030	206129	254332

- 총 자산의 편차를 줄이면서 최대한 많은 사람들을 설명하는 비율이자, 창구 이용 예상 고객 가정을 고려한 비율이 상위 5%, 하위 1%라고 판단

02 변수 처리

이상치 제거 과정

- 동일한 나이에서 소득의 증가에 따라 총 자산이 증가하는지 확인
- 증가하지 않을 시 이상치 제거 비율을 조정

30대에 소득 구간이 1, 2, 3에 해당하는 총자산의 상위 5% 수치

나이	소득	총 자산 (분위수 95)
30대	100만원 미만	5억 4165만원
30대	200만원 미만	2억 2383만원
30대	300만원 미만	3억 5850만원

- 일반적으로 소득 구간이 1인 사람의 총 자산이 2인 사람과 3인 사람보다 작은 것이 타당
  → 2 구간에 해당하는 수치 수준이 되도록 비율 조정

30대에 소득 구간이 1, 2, 3에 해당하는 관측치 수와 비율

나이	소득	관측치 수, 비율 (명, %)
30대	100만원 미만	35 (0.82)
30대	200만원 미만	376 (8.84)
30대	300만원 미만	907 (21.33)

- 30대이고 소득이 100만원 미만인 그룹의 관측치 수가 전체 0.82%에 해당
- 실제로 관측치 수가 부족해 극단적인 값이 나온 것으로 추정

02 변수 처리

결측치 처리

- 기본 변수 2개(맞벌이 여부, 자녀 수), 금융 변수 7개(청약 보유 여부, 은퇴 후 필요 자금, 정기 예금, 적금, 청약, 펀드, ELS/DLS/ETF 잔액)로 총 9개의 변수에서 결측치 존재
- 14만 개 유형의 데이터를 추정하기 위해 결측치 처리가 필요하다고 판단

결측치 확인

맞벌이	자녀 수	청약 보유 여부
6076	6700	8885
은퇴 후 필요자금	**금융상품 잔액_ 정기예금**	**금융상품 잔액_적금**
11001	9680	8269
금융상품 잔액_ 청약	**금융상품 잔액_ 펀드**	**금융상품 잔액_ ELS/DLS/ETF 등**
8885	13390	15475

결측치 처리 방법의 순서

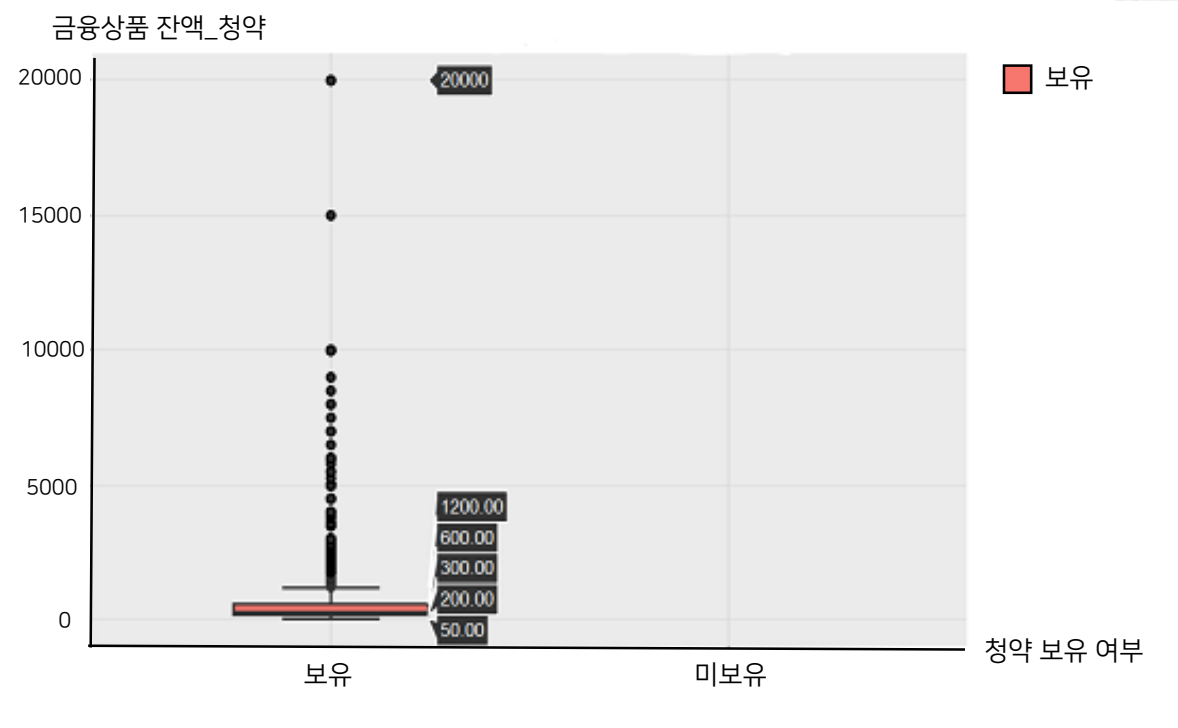
1. 다른 변수와의 의미 파악을 통한 대체
2. 인구통계학적 가정 활용
3. 결측치 대체 패키지 사용
4. 아무런 처리를 하지 않음

02 변수 처리

다른 변수와 의미 파악을 통한 대체

- "청약 보유 여부"와 "금융상품 잔액_청약"에 결측치 존재
  두 변수 간의 관계 파악을 통해 대체
  → "금융상품 잔액_청약" 변수의 결측치를 0으로 대체

"청약 보유 여부"와 "금융상품 잔액_청약" 변수



- "청약 보유 여부"가 NULL일 때 "금융상품 잔액_청약" 변수도 NULL
- "금융상품 잔액_청약" 변수에서 0값 존재하지 않음
  → "금융상품 잔액_청약" 변수의 결측치를 0으로 처리하는 것이 타당

02 변수 처리

다른 변수와 의미 파악을 통한 대체

- "금융상품 잔액_펀드", "금융상품 잔액_적금" : 각 금융상품의 월 저축액 변수와의 의미 파악을 통해 대체
  → 동일한 유형 존재할 때 대푯값으로 처리, 존재하지 않을 때 0으로 처리
- "맞벌이 여부": "결혼 여부" 변수와의 의미 파악을 통해 값을 대체하지 않음

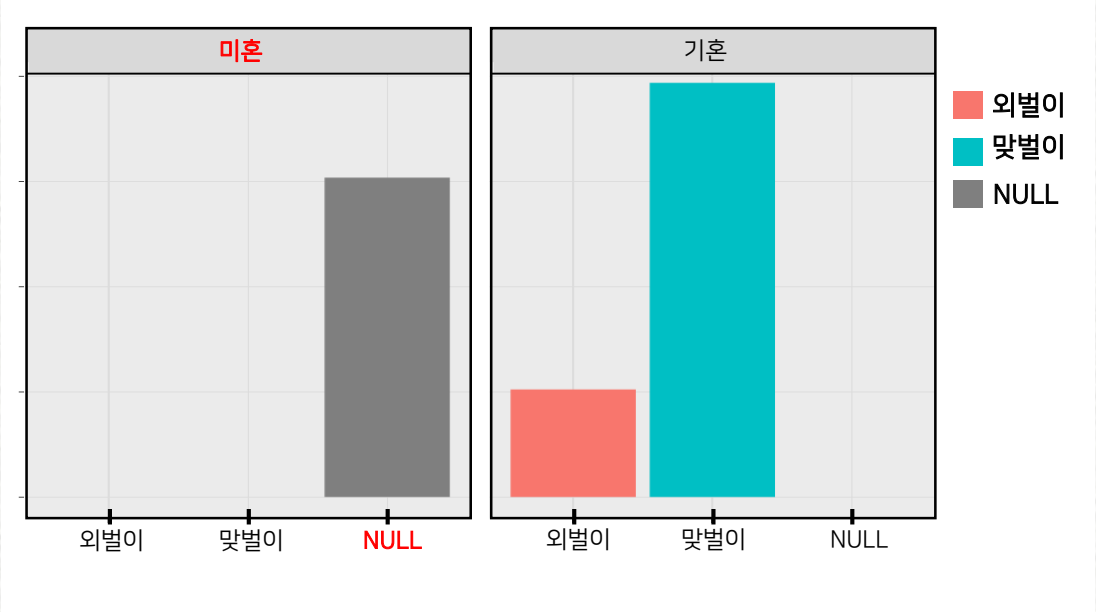
펀드, 적금 잔액와 매월 금액

		금융상품 잔액_펀드	
		결측치	비결측치
월 저축액_펀드	0	13390	1165
	0이외의 값	0	1447

		금융상품 잔액_적금	
		결측치	비결측치
월 저축액_적금	0	8269	706
	0이외의 값	0	7027

- 기본변수 기준 동일한 유형이 존재할 경우 평균으로 결측치 처리,
  동일한 유형이 없을 경우 0으로 처리해 왜곡을 피함

맞벌이 여부 변수와 결혼 여부 변수



- 미혼일 때 맞벌이 여부는 항상 NULL, 기혼일 때 값 항상 존재
  → 맞벌이 여부에서의 결측치는 미혼으로 인한 미응답

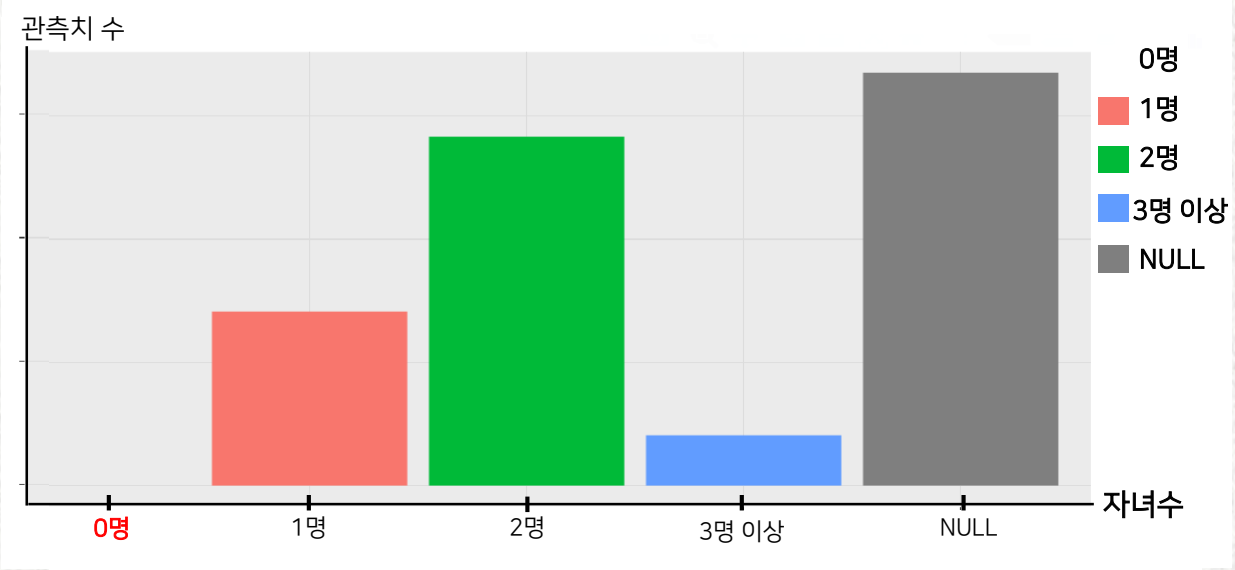
02 변수 처리

인구통계학 가정 사용

- "자녀 수" 변수에 결측치 존재
  "결혼 여부"와 "맞벌이 여부"를 고려 + 통계청의 [2018년 한부모가족 실태조사 사전 연구] 활용
  → 결측치를 0으로 대체

1. 0값은 존재하지 않음. 결측치 유형은 아래 2가지로 판단.

자녀 수 변수



- 자녀가 없는 경우
- 응답을 회피한 경우

2. 다음의 이유로 자녀 수의 결측치는 자녀 수가 없는 경우라고 판단.

자녀 수가 결측치일 때 결혼 여부와 맞벌이 여부

결혼	맞벌이	자녀 수	빈도 수, 비율 (명, %)
미혼	NULL	NULL	5946 (83.06)
기혼	외벌이	NULL	177 (2.47)
기혼	맞벌이	NULL	1035 (14.45)

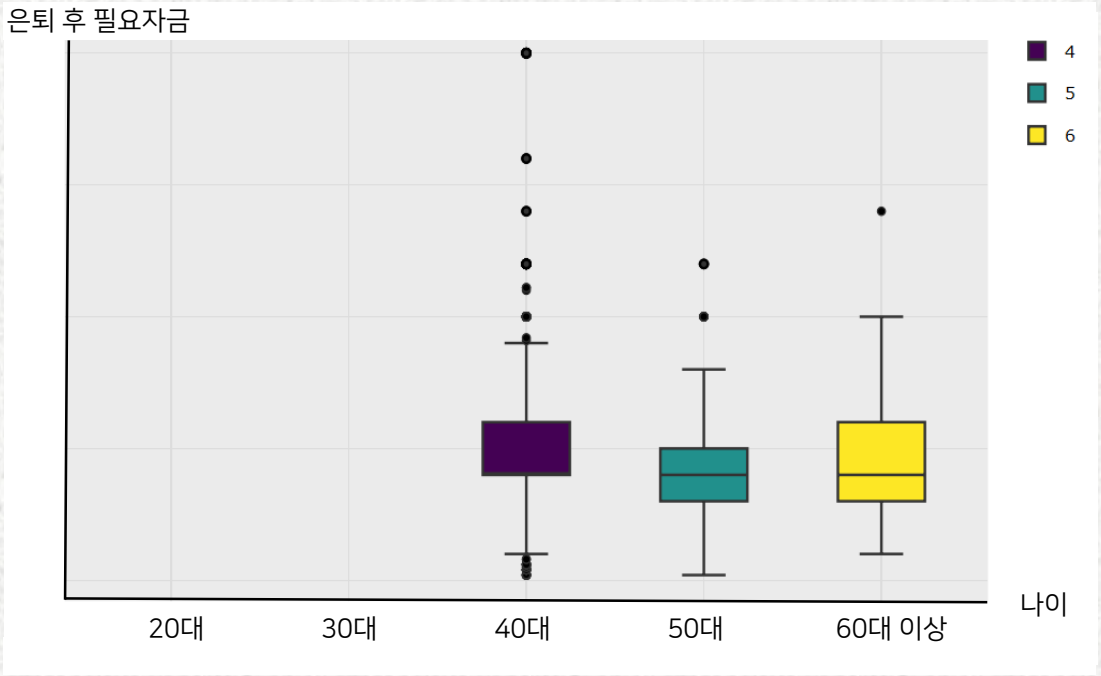
- 결측치의 대부분이 미혼.
- 응답 거부 유형은 미혼인데 자식이 있는 경우, 즉 한부모가정에서 나타난다고 가정. 설문조사 데이터에서 한 부모가정의 비율은 통계청 수치보다 1%p 많아 응답 거부는 극소수일 것으로 판단.

02 변수 처리

결측치 처리 패키지 사용

- "은퇴 후 필요 자금" 변수에 결측치 존재
  : Amelia 패키지를 이용해 결측치 추정

은퇴 후 필요 자금 변수 결측치



- 20대와 30대의 경우 모두 결측치인 것을 확인
  → 14만 개 유형의 추정을 위해 결측치 추정 필요

다양한 결측치 추정 패키지의 RMSE 비교

- 단일 대체법 (Xgboost / Decision tree)
  - 이상치가 많아 이상치에 덜 민감한 Tree 기반 모델을 사용
- 다중 대체법 (Amelia, Mice)
  - Row에 두 개 이상의 결측치가 있어 다중 대체 모델 사용.

	XGboost	Amelia	Mice	Decision tree
RMSE	86	78	94	109

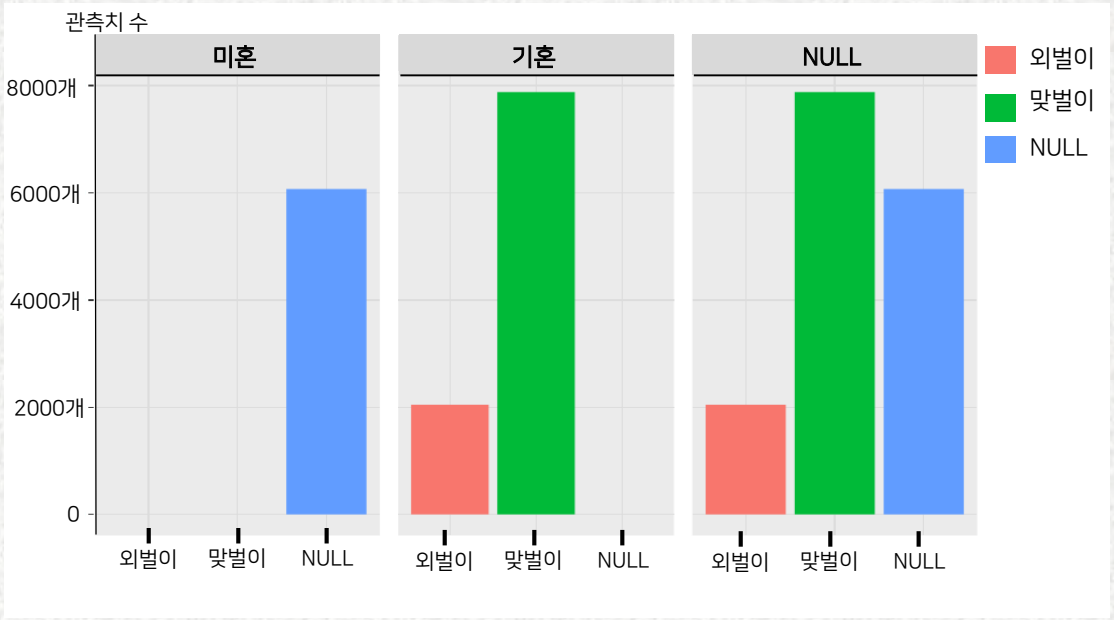
- 이 외 결측치가 있는 타 금융 변수도 위 패키지를 이용해 추정 시도
  → 천 만 단위 이상의 오차 발생으로 적합하지 않다고 판단

02 변수 처리

존재하지 않는 유형 처리

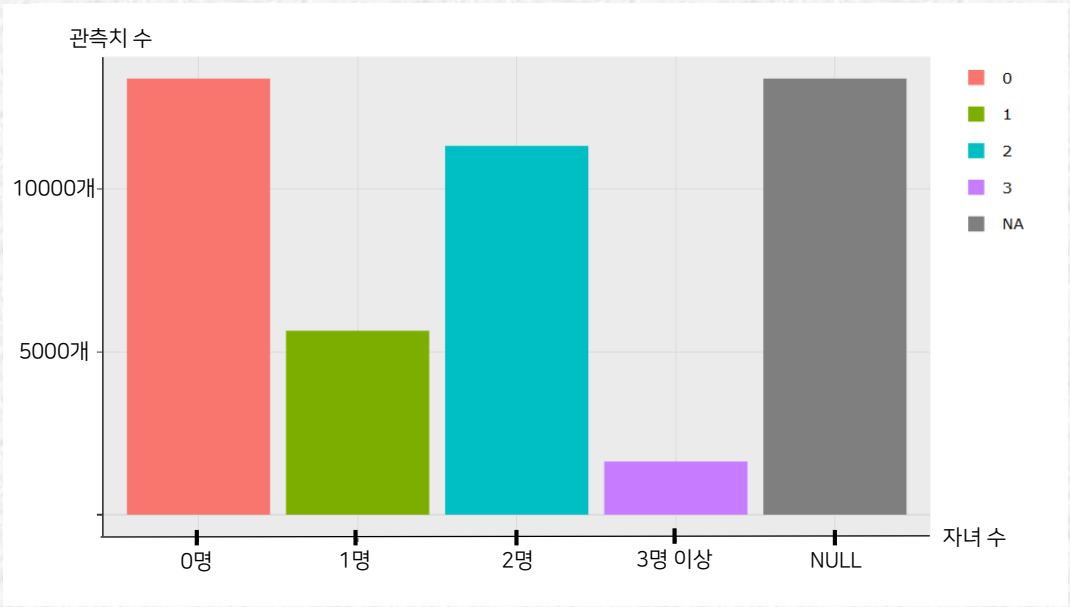
- 기본 변수 – "결혼 여부", "자녀 수"
  추후 예측을 위해 데이터에서 존재하지 않은 선택지를 같은 분포로 가정해 대입

맞벌이 여부, 결혼 여부에 따른 관측치



- 원래 결측치 존재하지 않음
  → 결측치에 대한 데이터를 같은 분포로 가정해 대입

자녀 수의 관측치



- 원래 0 값이 존재하지 않음, 결측치를 0으로 대입
  → 없어진 결측치에 대한 데이터를 같은 분포로 가정해 대입

03 분석 – 문제 1번

141,750개 고객 유형의 25개 금융거래정보 항목의 결측치 추정

추정 방법

- 정보가 제공되지 않은 유형의 값의 추정 필요
- 기본 정보 중 가장 영향력이 적은 변수를 제거하여 상위 유형 생성, 그 대푯값을 사용
- 행 별로 예측함으로써 26가지의 금융 정보들의 상관성을 고려
- 머신러닝 사용은 부적절하다고 판단 . 그 이유는,

  1) 변수 간 상관성이 크고, 2) 정보가 제공되지 않는 유형이 더 많고, 3) 기본 정보에 대한 금융 정보 수치의 편차가 큼

1. 변수 제거 조합 생성

	GROUP1	GROUP2	GROUP3	GROUP4	GROUP5	GROUP6	GROUP7	GROUP8
성별	X	O	O	O	O	O	O	O
나이	O	X	O	O	O	O	O	O
직업	O	O	X	O	O	O	O	O
지역	O	O	O	X	O	O	O	O
소득	O	O	O	O	X	O	O	O
결혼	O	O	O	O	O	X	O	O
맞벌이	O	O	O	O	O	O	X	O
자녀수	O	O	O	O	O	O	O	X

- 변수를 1~4개씩 제거할 때 가능한 모든 조합을 생성
- 표는 변수를 1개 제거했을 경우의 8개 조합

03 분석 – 문제 1번

141,750개 고객 유형의 25개 금융거래정보 항목의 결측치 추정

2. 제거할 변수 선택 - RMSE 구하기

1. 제거하는 변수 기준, 상위 유형들의 대푯값 구함

- 성별을 제거하는 조합 → [표1], [표2]는 같은 유형, 평균 구함

[표1]

성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
1	2	2	2	1	2	-	0

[표2]

성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
2	2	2	2	1	2	-	0

2. 모든 조합에 대한 RMSE를 통해 중요하지 않은 변수를 추출

- RMSE는 상위 유형의 평균과 각 관측치의 오차합
- 가장 중요한 금융 정보인 총 자산과 총 부채의 오차합으로 RMSE 계산
- 아래 표에서는 결혼 여부 – 맞벌이 여부 – 성별 – 자녀 수 – 나이 – 지역 – 직업 – 소득 순으로 중요하지 않은 변수

	GROUP1 (성별 제외)	GROUP2 (나이 제외)	GROUP3 (직업 제외)	GROUP4 (지역 제외)	GROUP5 (소득 제외)	GROUP6 (결혼 제외)	GROUP7 (맞벌이 제외)	GROUP8 (자녀수 제외)
총 자산과 총 부채의 오차합(RMSE)	8917	14344	15722	14532	19290	0	7282	12062

3. 기본변수 2개 제거부터 4개 제거까지 위의 1~2번 과정을 반복

03 분석 – 문제 1번

141,750개 고객 유형의 25개 금융거래정보 항목의 결측치 추정

3. 값 채워 넣기

1. 제공 데이터에 같은 유형이 있는 경우 대푯값 사용
2. 같은 유형이 없는 경우 RMSE에 따른 변수 조합을 이용
   - 조합에 따라 같은 유형을 가정, 그 대푯값을 대입

	성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
'자산+부채'의 RMSE	8917	14344	15722	14532	19290	0	7282	12062

	성별+나이	성별+직업	성별+지역	성별+소득	성별+결혼	…	…	맞벌이+자녀수
자산+부채	19678	20647	19737	24692	10762			16697

인덱스	성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	총자산	금융자산	실물자산
53	1	2	2	2	1	2	-	0	650	250	250
97	1	2	2	3	1	1	-	0	590	540	0
137	1	2	2	4	1	2	1	0	1430	355	1075
729	2	2	3	2	1	0	3	0	807	557	0
1891	1	2	7	3	1	1	1	0	4238	1011	50

03 분석 – 문제 1번

141,750개 고객 유형의 25개 금융거래정보 항목의 결측치 추정

3. 값 채워 넣기

1. 제공 데이터에 같은 유형이 있는 경우 대푯값 사용
2. 같은 유형이 없는 경우 RMSE에 따른 변수 조합을 이용
   - 조합에 따라 같은 유형을 가정, 그 대푯값을 대입

	성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
'자산+부채'의 RMSE	8917	14344	15722	14532	19290	0	7282	12062

	성별+나이	성별+직업	성별+지역	성별+소득	성별+결혼	…	…	맞벌이+자녀수
'자산+부채'의 RMSE	19678	20647	19737	24692	10762			16697

인덱스	성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	총자산	금융자산	실물자산
53	1	2	2	2	1	2	-	0	650	250	250
97	1	2	2	3	1	1	-	0			
137	1	2	2	4	1	2	1	0			
729	2	2	3	2	1	0	3	0			
1891	1	2	7	3	1	1	1	0			

03 분석 – 문제 1번

141,750개 고객 유형의 25개 금융거래정보 항목의 결측치 추정

3. 값 채워 넣기

1. 제공 데이터에 같은 유형이 있는 경우 대푯값 사용
2. 같은 유형이 없는 경우 RMSE에 따른 변수 조합을 이용
   - 조합에 따라 같은 유형을 가정, 그 대푯값을 대입

	성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
'자산+부채'의 RMSE	8917	14344	15722	14532	19290	0	7282	12062

	성별+나이	성별+직업	성별+지역	성별+소득	성별+결혼	…	…	맞벌이+자녀수
'자산+부채'의 RMSE	19678	20647	19737	24692	10762			16697

인덱스	성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	총자산	금융자산	실물자산
53	1	2	2	2	1	2	-	0	650	250	250
97	1	2	2	3	1	1	-	0	**590**	**540**	**0**
137	1	2	2	4	1	2	1	0			
729	2	2	3	2	1	0	3	0			
1891	1	2	7	3	1	1	1	0			

03 분석 – 문제 1번

141,750개 고객 유형의 25개 금융거래정보 항목의 결측치 추정

3. 값 채워 넣기

1. 제공 데이터에 같은 유형이 있는 경우 대푯값 사용
2. 같은 유형이 없는 경우 RMSE에 따른 변수 조합을 이용
   - 조합에 따라 같은 유형을 가정, 그 대푯값을 대입

	성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
'자산+부채'의 RMSE	8917	14344	15722	14532	19290	0	7282	12062

	성별+나이	성별+직업	성별+지역	성별+소득	성별+결혼	…	…	맞벌이+자녀수
'자산+부채'의 RMSE	19678	20647	19737	24692	10762			16697

인덱스	성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	총자산	금융자산	실물자산
53	1	2	2	2	1	2	-	0	650	250	250
97	1	2	2	3	1	1	-	0	590	540	0
137	1	2	2	4	1	2	1	0	1430	355	1075
729	2	2	3	2	1	0	3	0			
1891	1	2	7	3	1	1	1	0			

03 분석 – 문제 1번

141,750개 고객 유형의 25개 금융거래정보 항목의 결측치 추정

3. 값 채워 넣기

1. 제공 데이터에 같은 유형이 있는 경우 대푯값 사용
2. 같은 유형이 없는 경우 RMSE에 따른 변수 조합을 이용
   - 조합에 따라 같은 유형을 가정, 그 대푯값을 대입

	성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
'자산+부채'의 RMSE	8917	14344	15722	14532	19290	0	7282	12062

	성별+나이	성별+직업	성별+지역	성별+소득	성별+결혼	…	…	맞벌이+자녀수
'자산+부채'의 RMSE	19678	20647	19737	24692	10762			16697

인덱스	성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	총자산	금융자산	실물자산
53	1	2	2	2	1	2	-	0	650	250	250
97	1	2	2	3	1	1	-	0	590	540	0
137	1	2	2	4	1	2	1	0	1430	355	1075
729	2	2	3	2	1	0	3	0	**807**	**557**	**0**
1891	1	2	7	3	1	1	1	0			

03 분석 – 문제 1번

141,750개 고객 유형의 25개 금융거래정보 항목의 결측치 추정

3. 값 채워 넣기

1. 제공 데이터에 같은 유형이 있는 경우 대푯값 사용
2. 같은 유형이 없는 경우 RMSE에 따른 변수 조합을 이용
   - 조합에 따라 같은 유형을 가정, 그 대푯값을 대입

	성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
'자산+부채'의 RMSE	8917	14344	15722	14532	19290	0	7282	12062

	성별+나이	성별+직업	성별+지역	성별+소득	성별+결혼	…	…	맞벌이+자녀수
'자산+부채'의 RMSE	19678	20647	19737	24692	10762			16697

인덱스	성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	총자산	금융자산	실물자산
53	1	2	2	2	1	2	-	0	650	250	250
97	1	2	2	3	1	1	-	0	590	540	0
137	1	2	2	4	1	2	1	0	1430	355	1075
729	2	2	3	2	1	0	3	0	807	557	0
1891	1	2	7	3	1	1	1	0	**4238**	**1011**	**50**

03 분석 – 문제 2번

금융 거래 정보를 이용한 Peer Group 도출

클러스터링 기법

- K-prototype Clustering
- 연속형과 명목형 변수를 모두 사용할 수 있는 기법

k-prototype 선택 이유

- **기본변수**만 사용한다면 금융 변수의 편차를 고려해주지 못함
  - 같은 기본정보 유형 내에서도 8억에 가까운 편차가 존재.
- **금융변수**만을 사용한다면 기본 정보에 따른 차이를 고려해주지 못함
  - 20대와 60대의 경우 현재와 미래의 다른 소비패턴을 반영해 줄 수 없음.

→ 연속형과 명목형 변수를 모두 고려할 수 있는 K-prototype 클러스터링 선택.

기준 변수 및 선택 이유

기본 변수

가구 소득 구간
나이
자녀 수

- 가구 소득 구간과 나이는 자산에 가장 큰 영향을 미치는 변수
- 자녀 수는 개인의 미래 소비를 예측할 수 있는 변수로 판단, factor로 바꿔 사용

금융 변수

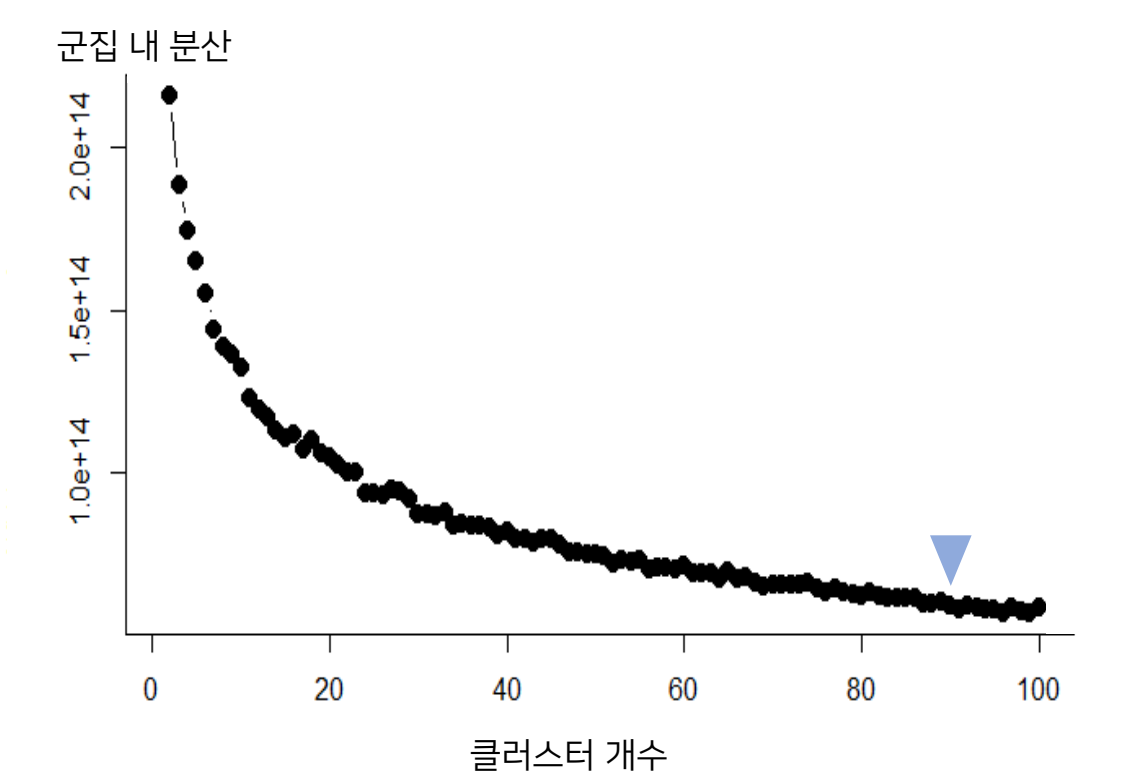
총 자산
총 부채

- 개인의 금융정보를 담고 있는 가장 상위의 변수, 나머지 변수들은 분배의 문제
- 해당 고객과 금융 생활이 가장 비슷한 사람들의 정보를 얻을 수 있을 것이라 판단

03 분석 – 문제 2번

금융 거래 정보를 이용한 Peer Group 도출

Elbow Method로 최적 군집 개수 도출



<Elbow Method를 활용한 최적 군집 개수에 따른 군집 내 분산>

- 군집의 개수가 늘어남에 따라 분산 내 거리가 크게 작아지지 않는 지점인 90을 최적 군집 개수로 선택
- 왼쪽의 그래프만 보면 20~30개 정도가 최적 군집 수로 보여지나 분산 내 거리의 수치가 90개인 경우 20개인 경우에 비해 23%로 크게 감소
- 클러스터 개수가 50 ~ 100개인 부분의 분산 내 거리만 봤을 때, 90개가 Local Minimum

03 분석 – 문제 3번

Peer Group의 '금융자산', '월 저축 금액', '월 소비 금액'의 금액 분포

분포 표시

- 도출된 Peer group으로 묶은 후, 제시된 세 변수에 대한 분포를 백분위로 확인

데이터 예시

Peer Group No.	비교대상칼럼	백분위수1	백분위수2	백분위수3	백분위수4	백분위수5	백분위수6	백분위수7	백분위수8	…	백분위수96	백분위수97	백분위수98	백분위수99
1	금융자산	100	200	300	300	300	300	305	330	…	5200	5789	6750	6750
1	월저축금액	3	3	10	10	15	15	20	20	…	210	248.3333	300	300
1	월소비금액	32.5	50	50	65.35714	75	76.66667	76.66667	80.8	…	400	400	450	450
2	금융자산	100	160	200	500	500	500	655.9	680	…	29247	29475	37590	37590
2	월저축금액	10	10	10	10	10	10	10	10	…	300	300	300	300
2	월소비금액	10	20	50	67.5	70	70	70	70	…	300	320	340.75	437.5
90	금융자산	80	150	200	200	200	300	350	350.4	…	900	900	900	933
90	월저축금액	1	1	2	5	5.666667	9	9	12	…	20000	20000	21800	21872.8
90	월소비금액	9	20	37.14286	40	70	75	80	80	…	350	400	415	415

03 분석 – 문제 4번

고객 기본 정보 수집 최소화 시 필요한 정보

변수 중요도

- All subset, GAIN, RMSE 사용해 중요도 점수와 변수 제거 순위 도출

All subset

- Adj.R과 AIC을 통해 유의미함 판단

All subset을 통한 총 자산에 대한 변수 중요도

설명변수의 수	설명변수	Adj.R	AIC
1	소득	0.305	402325
2	소득, 지역	0.333	401662
3	소득, 지역, 나이	0.360	400936
4	소득, 지역, 나이, 맞벌이	0.368	400714
5	소득, 지역, 나이, 맞벌이, 직업	0.371	400652
6	소득, 지역, 나이, 맞벌이, 직업, 자녀 수	0.3735	400590
7	소득, 지역, 나이, 맞벌이, 직업, 자녀 수, 성별	0.3737	400585

All subset을 통한 총 부채에 대한 변수 중요도

설명변수의 수	설명변수	Adj.R	AIC
1	맞벌이	0.068	344291
2	맞벌이, 소득	0.083	344036
3	맞벌이, 소득, 지역	0.089	343928
4	맞벌이, 소득, 지역, 자녀 수	0.091	343879
5	맞벌이, 소득, 지역, 자녀 수, 직업 or 나이	0.093	343854
6	맞벌이, 소득, 지역, 자녀 수, 직업, 나이	0.095	343830
7	맞벌이, 소득, 지역, 자녀 수, 직업, 나이, 성별	0.095	343816

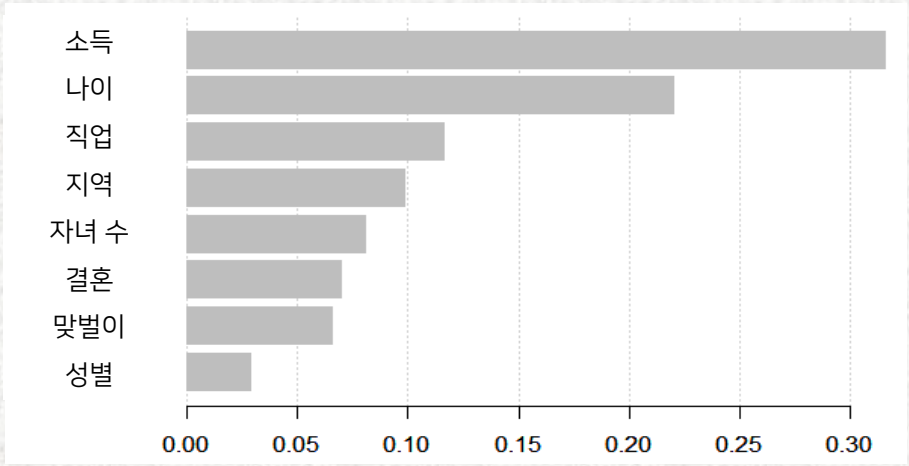
03 분석 – 문제 4번

고객 기본 정보 수집 최소화 시 필요한 정보

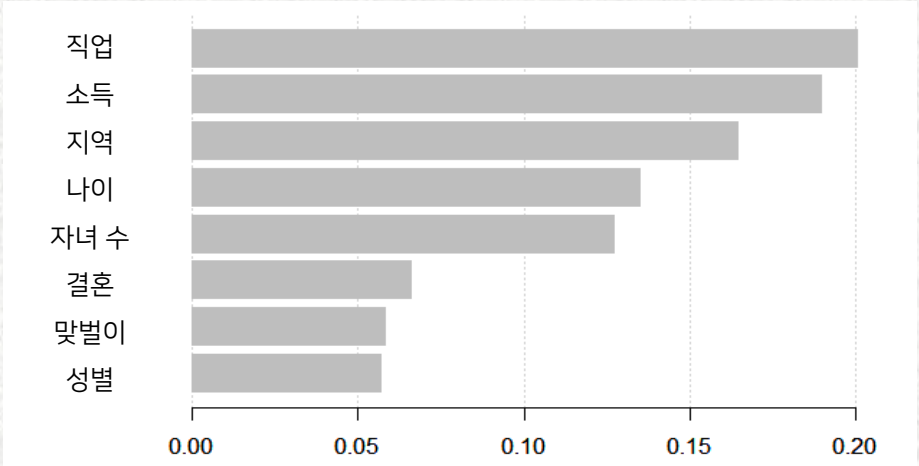
Gain

- XGBoost 모델에서 제공하는 변수의 중요도 기준. 평균 교육 손실을 점수화

총 자산의 Information Gain



총 부채의 Information Gain



RMSE

- 실제 y 값과 예측한 y 값 사이의 차이로 에러를 계산

	성별	나이	직업	지역	소득	결혼	맞벌이	자녀 수
총 자산	8551	13852	15075	13988	18731	0	7001	11538
총 부채	2210	3395	3809	3469	3560	0	1784	3134

총 자산, 총 부채에 대한 RMSE

- 기본 변수를 하나씩 제거하는 8가지 그룹의 RMSE
- **RMSE가 높을수록 영향이 큰 변수**

03 분석 – 문제 4번

고객 기본 정보 수집 최소화 시 필요한 정보

중요도 점수

- 가장 중요한 변수를 8점, 가장 중요하지 않은 변수를 1점으로 중요한 순서대로 점수를 매김
- SCORE = 0.5 * All Subset 점수 + GAIN 점수 + 0.5 * RMSE 점수

전체 중요도 점수와 순위

	All subset	GAIN	RMSE	SCORE	RANK
성별	4	2	6	7	8
나이	9	12	10	21.5	4
직업	7	14	15	25	2
지역	13	11	12	23.5	3
소득	15	15	15	30	1
결혼	2	6	2	8	7
맞벌이	13	4	4	12.5	6
자녀 수	8	8	8	16	5

- 순위는 성별 – 결혼 – 맞벌이 – 자녀 수 – 나이 – 지역- 직업 – 소득 순서로 중요하지 않다고 판단 가능.
- 순위를 바탕으로 성별, 결혼 변수를 제외 가능
  - 6위인 맞벌이 부터는 2배 가까이의 차이가 존재

04 한계 및 보완

한계점 1

- 현업에서 활용하는 파생변수를 고려하지 않아 실제 활용 및 풍부한 해석의 어려움

해결 방안 1

- 자산의 비율, 투자 성향, 저축 비율 , 소비 비율 등 여러가지 파생변수를 만들어서 클러스터링.

  Ex) **1. 은행 이용 목적(대출 vs 자산관리) = 부채/금융자산**

  2. 리스크 성향(리스크 선호 vs 리스크 회피) = 안정적 금융자산/위험 금융상품

  안정적 금융자산 : 적금, 정기예금, 청약, 보험금 납입액 등

  위험 금융상품 : 주식, 펀드, ELS 등

04 한계 및 보완

한계점 2

Peer Group에 대한 해석이 빠져 있음. 그로 인해 아래와 같은 문제점들이 발생

- 설명력 측면 - 클러스터링이 잘 되었는지 확인이 되지 않음
- 활용 측면 - 클러스터링 결과를 활용해 활용 측면(마케팅 등)에 사용하기가 힘듦

해결 방안 2

- 설명력 측면 - 클러스터링 설명력 확인 프로세스 도입

  Ex) Peer Group 내의 기본 정보 조합과 시나리오 분석 결과 비교 – 실제 상담 상황을 기반으로 시나리오 도출
- 활용 측면

  Ex) 대표적인 금융변수(총자산, 소비, 부채)와 파생변수들이 어떠한 분포를 띄고 있는지 그래프를 통해 확인하고

  어떤 상품을 추천할 수 있을지 고려

감사합니다

FRAME