금융 데이터를 활용한
“나의 금융생활정보 지수” 개발

2018 BIG CONTEST

FRAME

김현우, 민은주, 박주연, 이지예, 이주영

REPORT CONTENTS

- 변수 형태 정의
- 이상치 제거
- 결측치 채우기

▼

1. 문제 정의

▲

- 서비스 목적
- 분석 목표
- ‘보통 사람’에 대한 정의

2. 변수 처리

3. 분석

▲

- 문제 1번
- 문제 2번
- 문제 3번
- 문제 4번

01. 문제 정의

서비스 목적

- 고객이 창구를 방문하여 8가지의 개인 정보 입력 시,
  '보통 사람' (=고객과 비슷한 개인정보 가진 사람)의 평균 금융 정보 제공
- 고객의 금융 상태 파악 가능 및 재무 상담, 상품 추천 등에 활용 가능

분석 목표

- 금융 거래 정보를 이용한 Peer Group 도출
- Peer Group의 금융자산, 월 저축 금액 등 자산 금액 분포 추정
- 고객 기본 정보 수집 최소화 시 필요한 정보

분석 대상

'보통 사람'

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

02. 변수 처리

타입에 따른 변수

- 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

➡ ‘보통 사람’에 대한 분석을 위해선 이상치 제거가 필수적!

02. 변수 처리

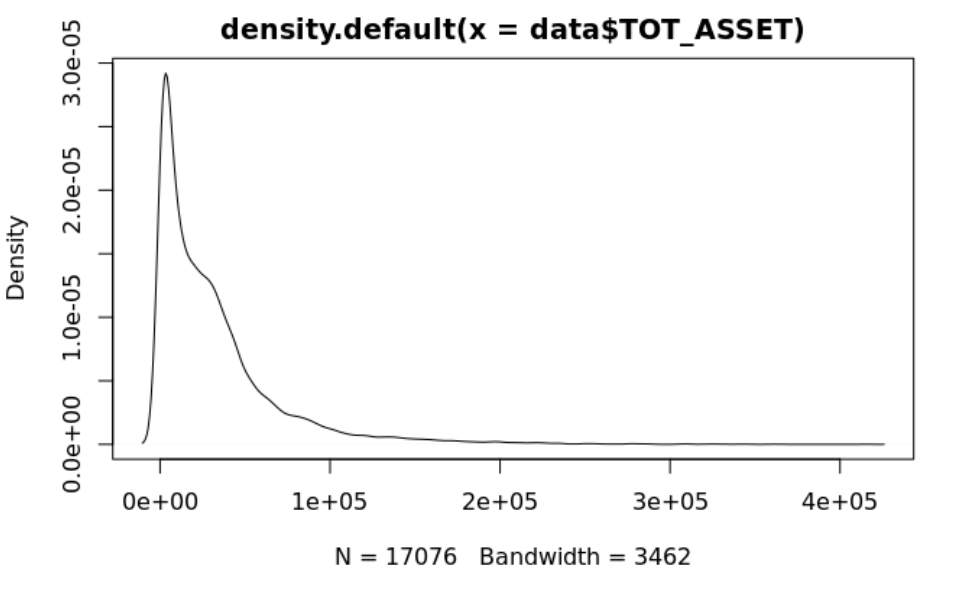
이상치 처리

이상치 정의

- 금융정보를 가장 대표할 수 있는 변수로 '총 자산'이라고 판단
- '총자산' 분포가 skewed 되어 있고 동일한 기본 정보의 사람이지만 '총 자산'의 편차가 크다는 문제점
- 서비스 목적*을 고려했을 때, 이상치 제거 필요

*고객이 창구에서 기본 정보(인구통계정보)를 입력 시 기대 자산 보여주는 서비스

총 자산의 분포



동일 기본 정보 내에서도 큰 총 자산의 편차

인덱스	성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	**총자산**
231	1	2	2	2	6	1	-	-	**79,100**
8221	1	2	2	2	6	1	-	-	**71,200**
8447	1	2	2	2	6	1	-	-	**47,300**
22392	1	2	2	2	6	1	-	-	**940**
23066	1	2	2	2	6	1	-	-	**6,980**

02. 변수 처리

이상치 처리

이상치 정의

- 제거할 비율 결정 위해 상위, 하위 10%의 총 자산을 확인
- 상위 10% 사람들이 전체 자산에서 차지하는 비중 확인→ 많은 사람들을 대표할 수 있게 백분위 수 설정
- 단순히 총 자산만을 이상치 기준으로 삼는다면 소득이 높은 대부분의 사람이 제거되는 문제점 발생

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

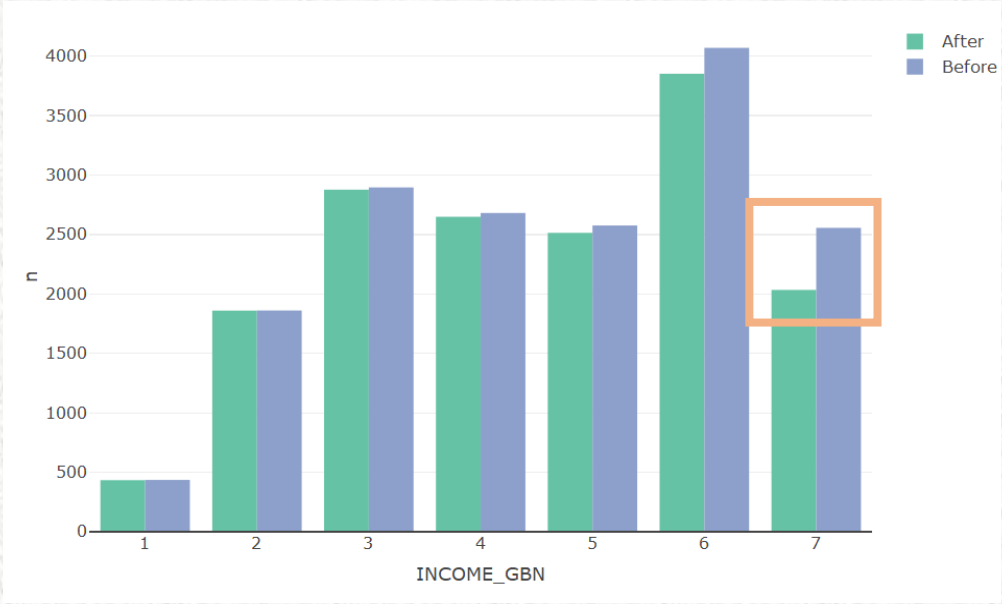
하위 10%에 해당하는 총 자산의 백분위 수

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



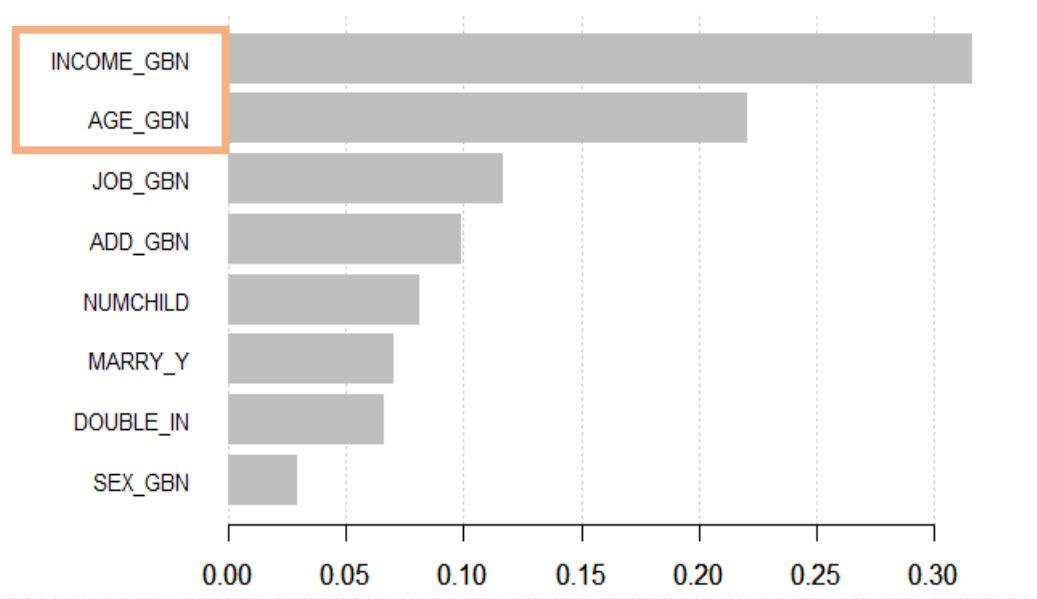
02. 변수 처리

이상치 처리

그룹 정의

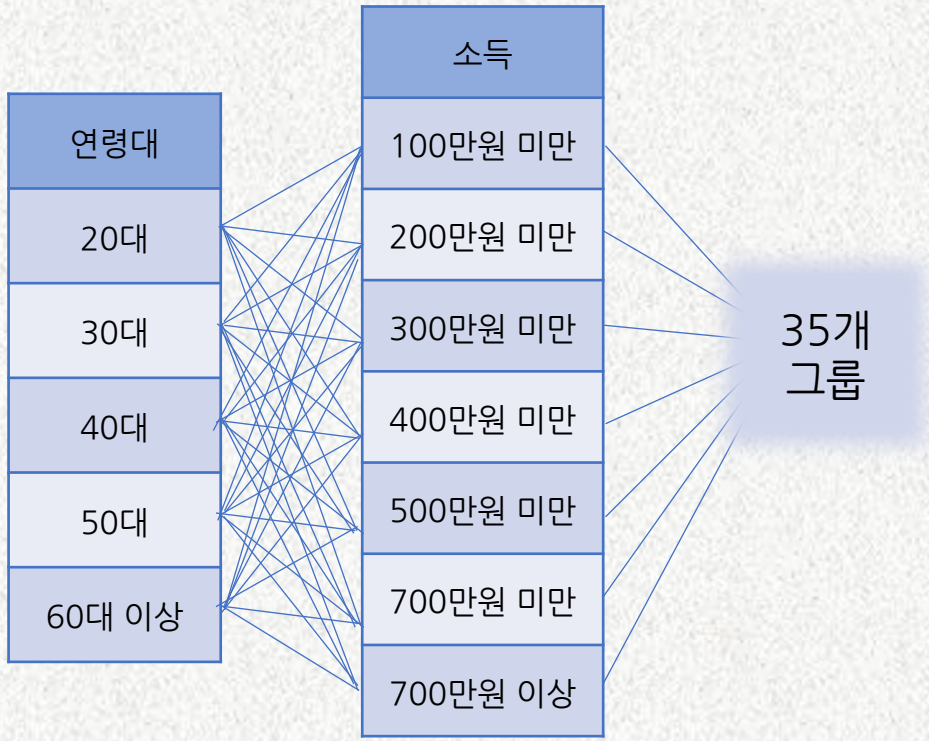
- XGBoost를 통해 총 자산에 대한 기본 변수들의 영향 확인
- 연령대(5개 구간), 소득 구간(7개 구간)에 따른 35개 그룹에서의 이상치 제거

총 자산에 대한 XGBoost 결과



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

두 변수에 따른 그룹 생성



02. 변수 처리

이상치 처리

백분위수 선정

- 소득 분위와 연령대 별 백분위수 확인, '보통 사람'으로 납득 가능한 수치를 고려

연령대 별과 소득 분위 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, 현재까지의 소득을 고려했을 때 타당한 자산 수치가 상위 5%라고 판단
- 저자산층에 분포가 쏠려 있어 하위의 경우 1%가 적정 비율이라 판단

02. 변수 처리

이상치 처리

그룹에 따른 이상치 제거

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

20대에 소득 구간이 4, 5, 6에 해당하는
총자산의 하위 1% 수치

연령대	소득	총 자산 (백분위수 1)
2	4	190
2	5	501.8
2	6	378.8

- 일반적으로 소득 구간이 5인 사람의 총 자산이
  4인 사람보다는 크고, 6인 사람 보다는 작을 것으로 예상
  → 4 구간과 6 구간의 중간 값 수준에서 제거

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

연령대	소득	총 자산 (백분위수 95)
3	1	54165
3	2	22383
3	3	35850

- 마찬가지로 추세를 이탈하는 1 구간의 총 자산 수치
  → 소득 1의 22383 이상을 제거

02. 변수 처리

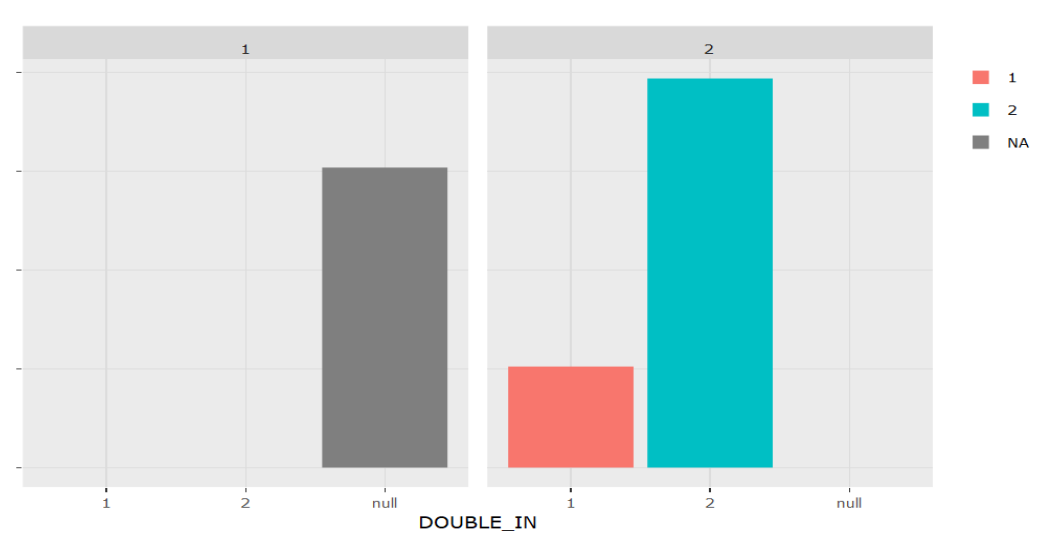
결측치 처리

- 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
- 기본 변수 - 맞벌이 여부 : 미혼으로 인한 결측치

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

맞벌이 여부 변수와 결혼 여부 변수



- 미혼일 때 맞벌이 여부는 항상 NULL, 기혼일 때 값 항상 존재

→ 맞벌이 여부에서의 결측치는 미혼으로 인한 미응답

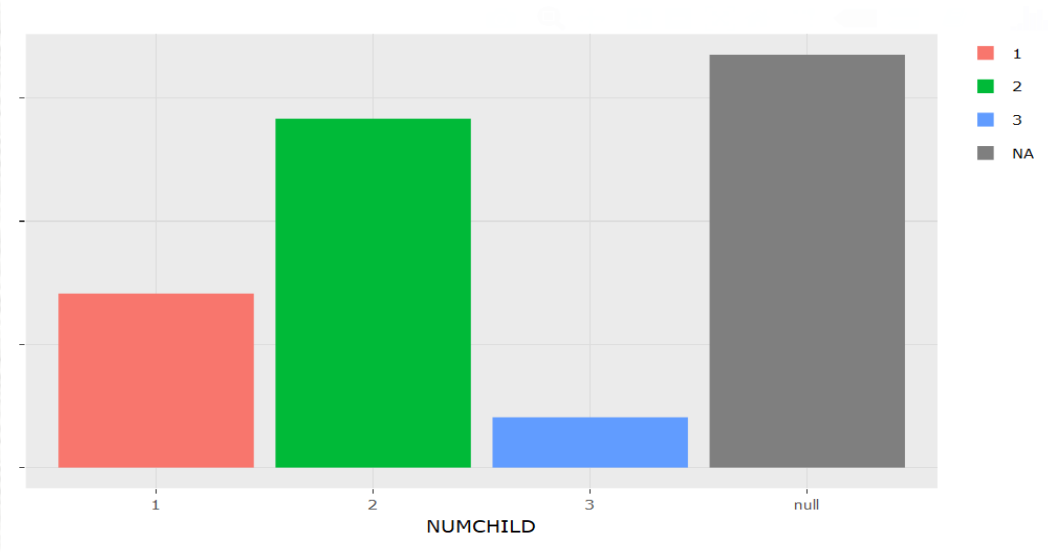
02. 변수 처리

결측치 처리

자녀 수

- 기본 변수 - 자녀 수 : 결혼 여부와 맞벌이 여부를 고려하여 결측치를 0으로 대체

자녀 수 변수



- 0값은 존재하지 않음 →0이 NULL에 포함
- 분석을 위한 NULL값의 적절한 대체 필요

자녀 수가 결측치일 때 결혼 여부와 맞벌이 여부

결혼	맞벌이	자녀 수	빈도 수
미혼	-	-	5946
기혼	외벌이	-	177
기혼	맞벌이	-	1035

- 자녀 수가 결측치일 때 응답을 거부한 경우보다 0이어서 미응답한 경우가 대다수를 차지
  → 결측치를 0으로 처리하는 것이 타당

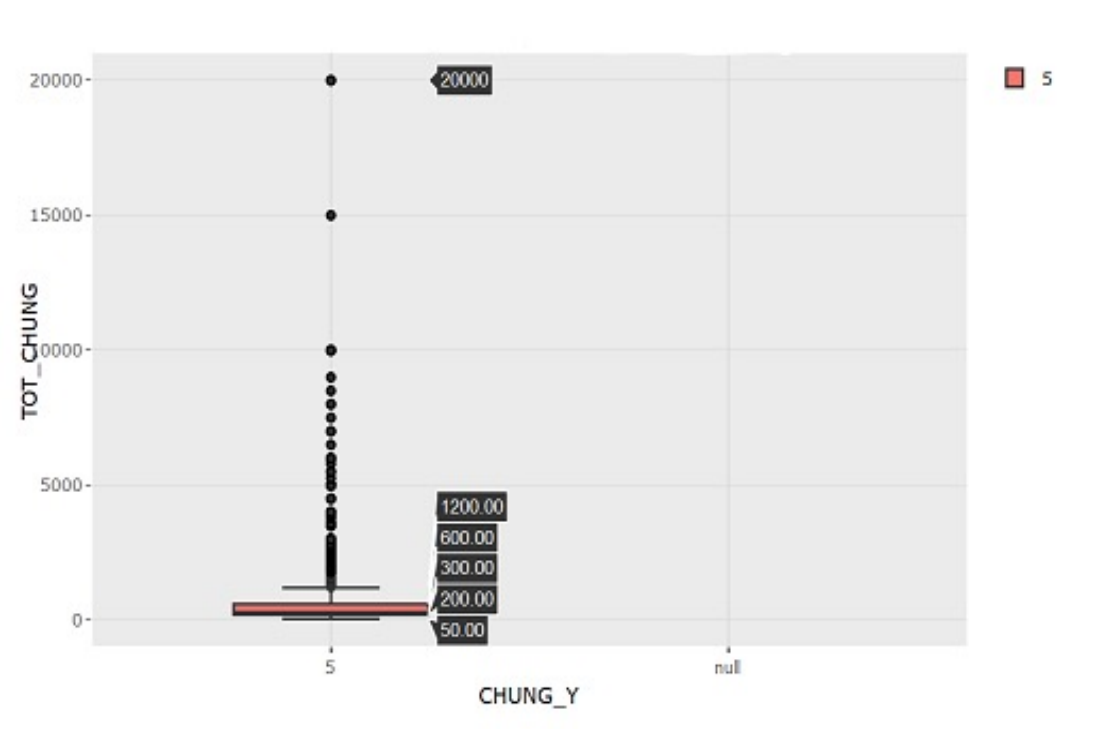
02. 변수 처리

결측치 처리

청약 관련 변수

- 금융 변수 - 청약 보유 여부와 금융상품 잔액_청약 변수 : 결측치를 0으로 대체

청약 보유 여부와 금융상품 잔액_청약 변수



- 청약 보유 여부가 NULL일 때 금융상품 잔액_청약 변수도 NULL
- 금융상품 잔액_청약 변수에서 0값 존재하지 않음

→ 결측치를 0으로 처리하는 것이 타당

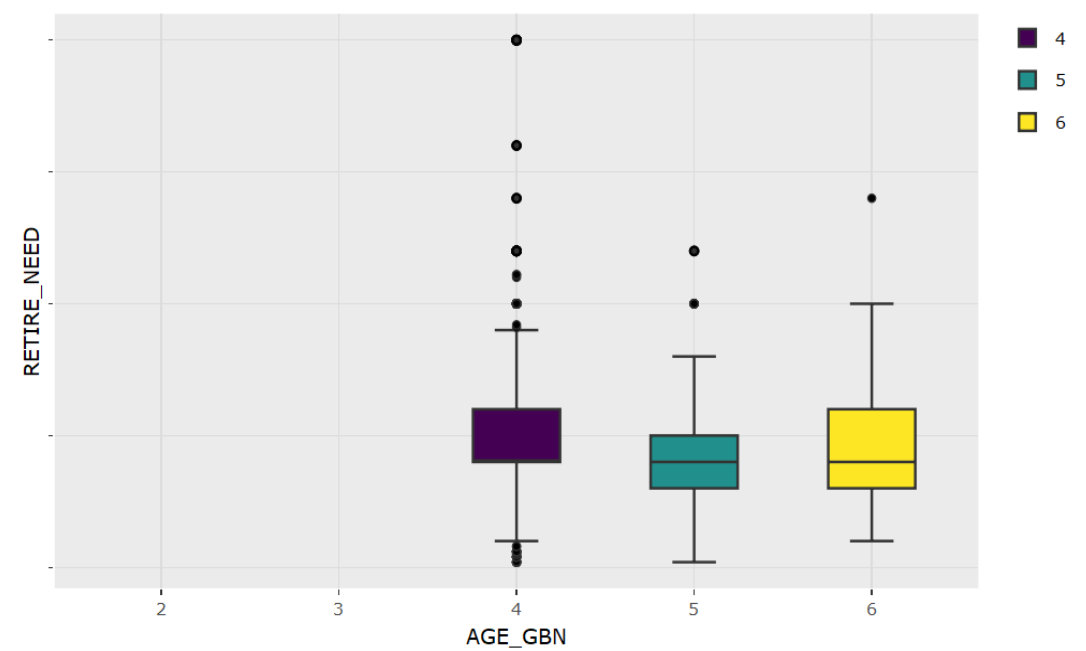
02. 변수 처리

결측치 처리

은퇴 후 필요 자금

- 금융 변수 - 은퇴 후 필요 자금 변수 : 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

	XGboost	Amelia	Mice	Decision tree
RMSE	86	78	94	109

- RMSE : 실제 값과 추정한 값의 차이의 평균
- Amelia의 RMSE가 가장 낮은 것으로 확인
- 이 외 결측치가 있는 타 금융 변수도 위 패키지를 이용해 추정 시도
  → 천 만 단위 이상의 오차 발생으로 적합하지 않다고 판단

02. 변수 처리

결측치 처리

펀드와 적금 잔액

- 금융 변수 - 금융상품 잔액_펀드, 금융상품 잔액_적금 :
  동일한 유형 존재할 때 대푯값으로 처리, 존재하지 않을 때 0으로 처리

금융상품 잔액_펀드와 매월 금액

	TOT_FUND : 결측치	TOT_FUND : 비결측치
M_FUND = 0	13390	1165
M_FUND != 0	0	1447

금융상품 잔액_적금과 매월 금액

	TOT_JEOK : 결측치	TOT_JEOK : 비결측치
M_JEOK = 0	8269	706
M_JEOK != 0	0	7027

- 금융상품 잔액-적금, 펀드에서 0인 값 존재하지 않고,
  결측치일 때 월 저축액_적금, 펀드가 모두 0
  →결측치 0으로 추정 가능

- 하지만 월 저축액_적금, 펀드가 0이어도 금융상품 잔액_적금, 펀드에서 0 이상의 값이 존재하는 경우 약 8% 존재
  →월 저축액이 0일 때의 결측치가 모두 0이라고 단언할 수 없음

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

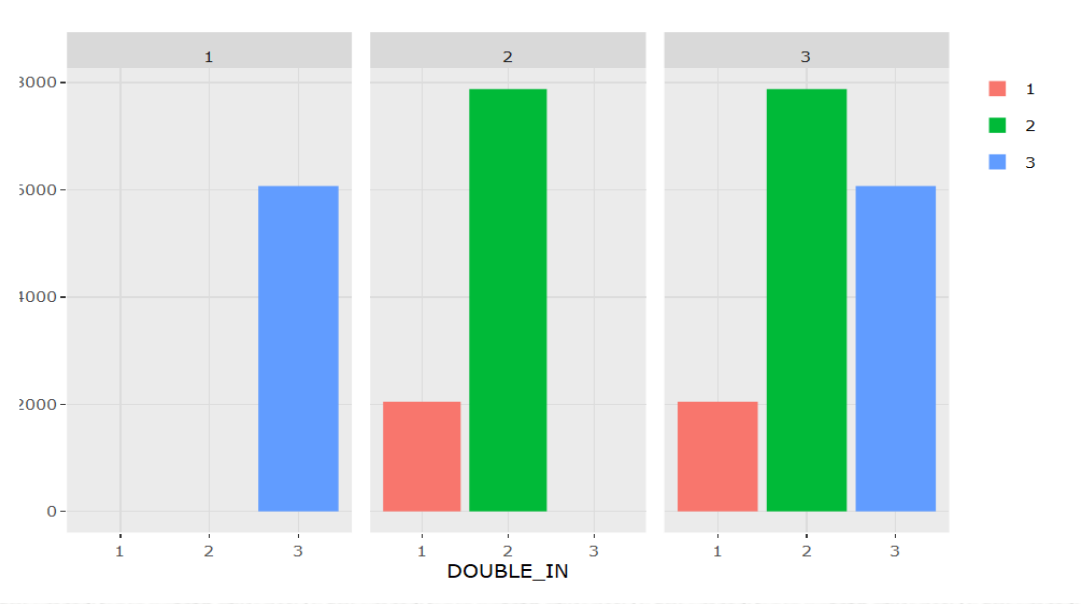
02. 변수 처리

결측치 처리

결혼 여부와 맞벌이 여부

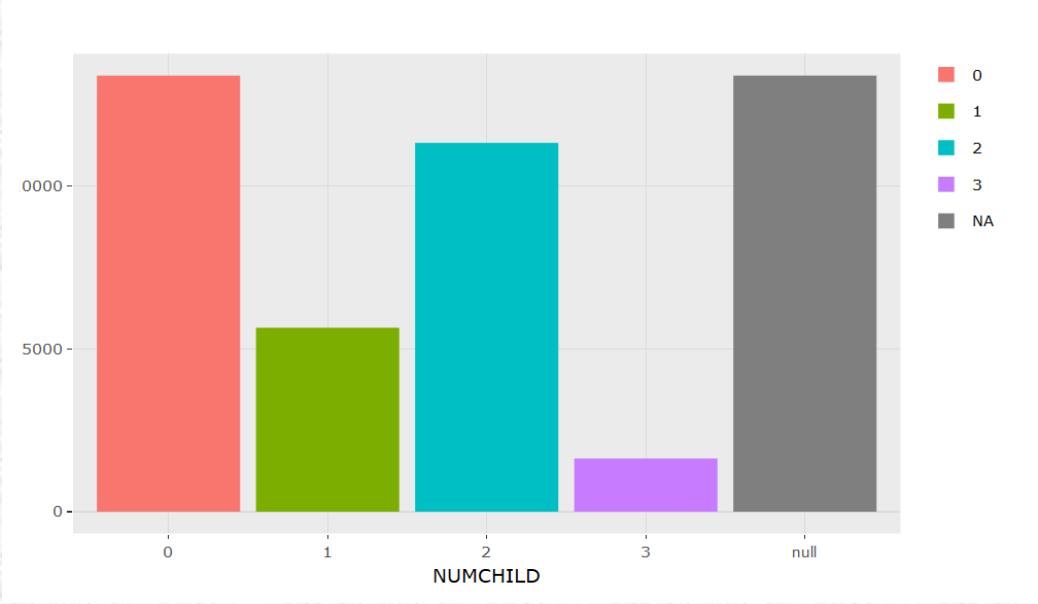
- 기본 변수 - 결혼 여부, 자녀 수 : 데이터에서 존재하지 않은 선택지에 대해 추후 예측을 위해 같은 분포를 가정해 imputation 수행

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 **대푯값**으로 impute함. 그 이유는,
  - 행 별로 예측함으로써 26가지의 금융 정보들의 상관성을 고려
  - 머신러닝의 경우 금융 정보를 변수 별로 하나씩 예측, 상관성을 고려하지 못함

1. 변수 제거 조합 생성

	GROUP1	GROUP2	GROUP3	GROUP4	GROUP5	GROUP6	GROUP7	GROUP8
SEX_GBN	X	O	O	O	O	O	O	O
AGE_GBN	O	X	O	O	O	O	O	O
JOB_GBN	O	O	X	O	O	O	O	O
ADD_GBN	O	O	O	X	O	O	O	O
INCOME_GBN	O	O	O	O	X	O	O	O
MARRY_Y	O	O	O	O	O	X	O	O
DOUBLE_IN	O	O	O	O	O	O	X	O
NUMCHILD	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. RMSE 구하기

1. 제거하는 변수 기준으로 같은 유형들은 같은 대표값 Imputation

- 예를 들어 제공된 데이터에 [표1]에 해당하는 유형이 있지만,
  [표2]에 해당하는 유형이 없을 때, 성별 변수를 제거함으로써
  [표1]에 해당하는 유형의 대푯값으로 [표2]에 해당하는 유형의 값을 채움.

[표1] - 결측치 없는 Instance

성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	…	은퇴자금
1	2	2	2	1	2	N	0	…	100

대표값 Imputation

[표2] - 결측치 있는 Instance

성별	나이	직업	지역	소득	결혼	맞벌이	자녀수	…	은퇴자금
2	2	2	2	1	2	N	0	…	100

2. 모든 조합에 대한 RMSE를 통해 중요하지 않은 변수를 추출

- GROUP별 총 자산과 총 부채의 오차합으로 RMSE 계산
- 특정 변수를 고려하지 않고 생성한 group의 RMSE가 작다면, group이 동질하여 특정 변수는 중요하지 않은 변수라는 가정
  - 아래 표에서는 MARRY_Y - DOUBLE_IN - SEX_GBN -NUMCHILD - AGE_GBN - ADD_GBN - JOB_GBN - INCOME_GBN 순으로 중요하지 않은 변수

	GROUP1 (성별X)	GROUP2 (나이X)	GROUP3 (직업X)	GROUP4 (지역X)	GROUP5 (소득X)	GROUP6 (결혼X)	GROUP7 (맞벌이X)	GROUP8 (자녀수X)
총 자산, 총 부채의 RMSE 합	8917	14344	15722	14532	19290	0	7282	12062

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
총 자산, 총 부채의 RMSE 합	8917	14344	15722	14532	19290	0	7282	12062

	성별+나이	성별+직업	성별+지역	성별+소득	성별+결혼	…	…	맞벌이+자녀수
총 자산, 총 부채의 RMSE 합	19678	20647	19737	24692	10762			16697

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

클러스터링 기준 변수

- 금융 정보 : 총 자산, 총 부채
- 기본 정보 : 가구 소득 구간, 연령, 자녀 수

기준 변수 - 금융 정보 변수

총 자산

금융 자산
부동산 자산
기타 자산

부채 잔액

신용 대출
담보 대출
아파트/주택 담보대출
전세 자금 대출

- 금융 정보 : 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

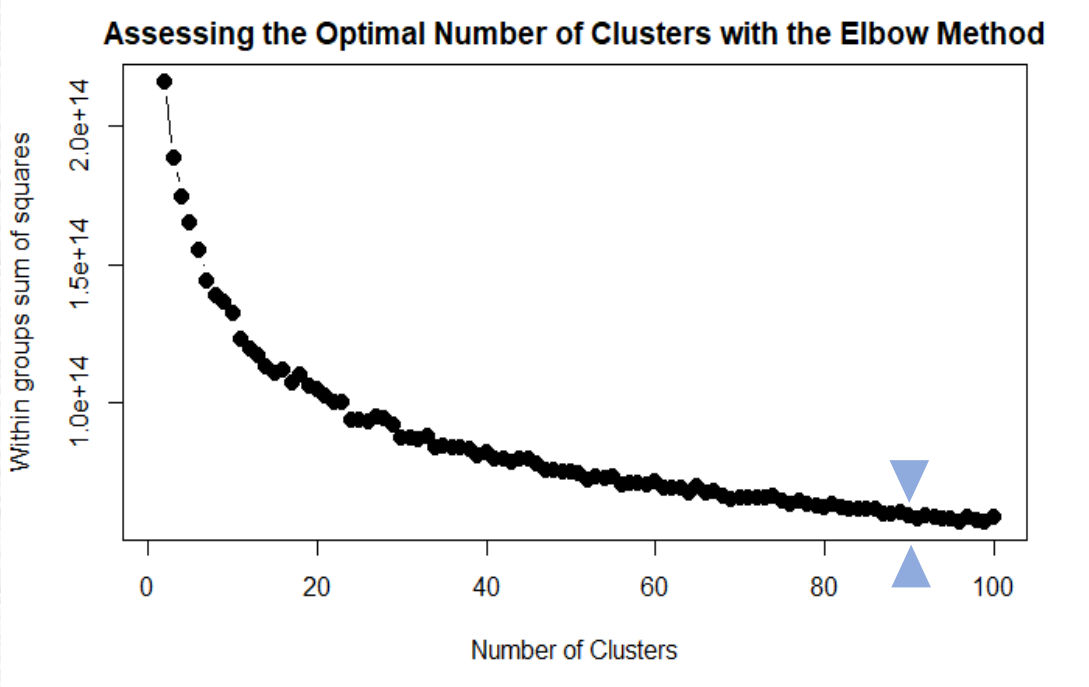
03. 분석 - 문제 2번

금융 거래 정보를 이용한 Peer Group 도출

기준 변수
- 기본정보 변수

- 기본 정보 : 가구 소득 구간, 연령, 자녀 수
- 가구 소득 구간과 연령은 자산에 가장 큰 영향을 미치는 변수
- 자녀 수는 개인의 미래 소비를 예측할 수 있는 변수로 판단, factor로 바꿔 사용

Elbow Method로 최적 군집 개수 도출



- 군집의 개수가 늘어남에 따라 분산 내 거리가 크게 작아지지 않는 지점인 90을 최적 군집 개수로 선택

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.	변수	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

- 후진제거법, GAIN, RMSE 사용해 중요도 점수와 변수 제거 순위 도출
  - 후진제거법, RMSE은 MSE를 구하고 0.5를 곱해 편향을 줄임

후진제거법

- F 통계량을 이용하여 변수의 유의미함 판단

후진제거법을 통한 "총 자산"에 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, 자녀수	0.3735	400590
7	소득, 지역, 나이, 맞벌이, 직업, 자녀수, 성별	0.3737	400585

후진제거법을 통한 "총 부채"에 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, 자녀수	0.091	343879
5	맞벌이, 소득, 지역, 자녀수, 직업 or 나이	0.093	343854
6	맞벌이, 소득, 지역, 자녀수, 직업, 나이	0.095	343830
7	맞벌이, 소득, 지역, 자녀수, 직업, 나이, 성별	0.095	343816

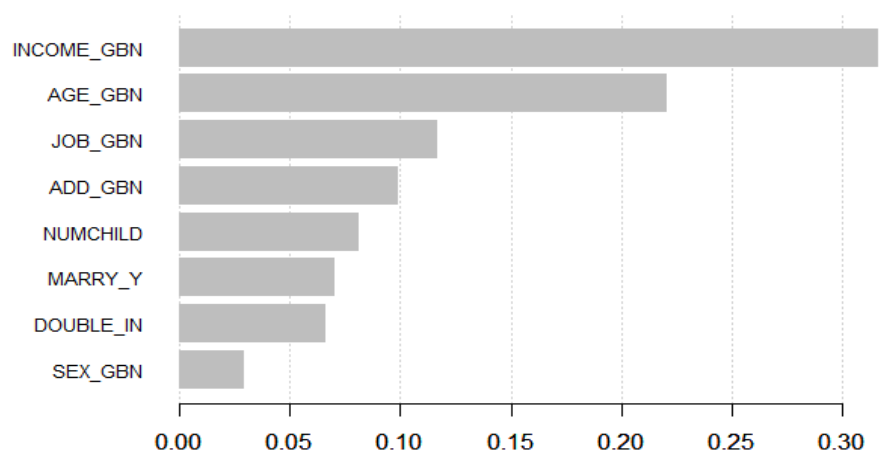
03. 분석 - 문제 4번

고객 기본 정보 수집 최소화 시 필요한 정보

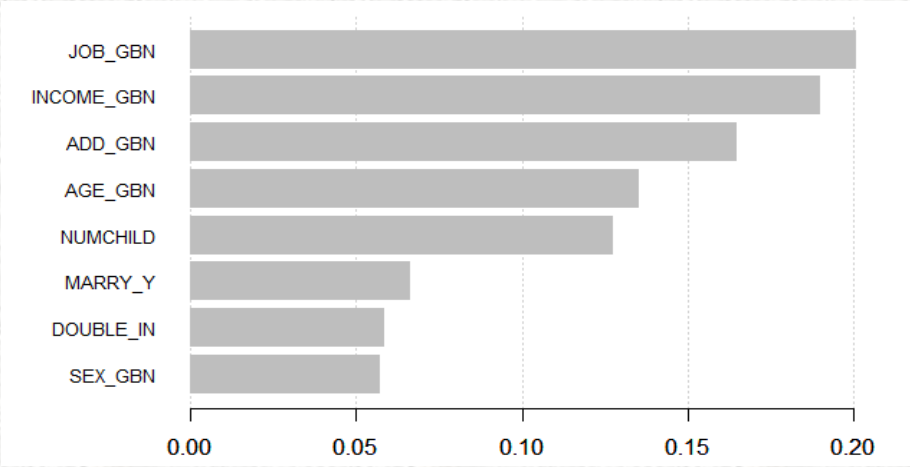
Gain

- 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	자녀수
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 * 후진선택법 점수+ GAIN 점수 + 0.5 * RMSE 점수

전체 중요도 점수와 순위

	후진제거법	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	5
자녀수	8	8	8	12	6

- 순위는 성별 - 결혼 - 자녀 수 - 맞벌이 - 나이 - 지역- 직업 - 소득 순
- 순위를 바탕으로 성별, 결혼 변수를 제외 가능

감사합니다

FRAME