

보험회사의 잠재적 고객 탐색

CARAVAN보험에 가입할 고객특성 알아보기

2019110495 김서영

2019110484 이서현

2019110485 정다혜

2018112625 문수연



CONTENTS LIST

- 1 연구배경 및 목적
- 2 데이터소개
- 3 데이터전처리
- 4 모형구축
- 5 모형평가
- 6 결론



연구배경 및 목적

최근 보험 산업은 경기성장 둔화, 저금리,
시장 포화상태로 인해 성장이 정체



☑ 기존의 사업의 효율을 높이거나 신규 고객을
찾아 사업을 확장하는 전략

☑ 보험 리스크 분석고도화와 서비스 개선, 마케팅과
영업 활동 등 다양한 분야에 빅데이터활용

“

보험회사의 기대 고객층을 분석, 예측하여
동일한 마케팅 비용으로 더 많은 보험 가입자를 확보

”

데이터소개

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	X77	X78	X79	X80	X81	X82	X83	X84	X85	X86
0	33	1	4	2	8	0	6	0	3	5	...	0	0	0	1	0	0	0	0	0	0
1	6	1	3	2	2	0	5	0	4	5	...	0	0	0	1	0	0	0	0	0	1
2	39	1	3	3	9	1	4	2	3	5	...	0	0	0	1	0	0	0	0	0	0
3	9	1	2	3	3	2	3	2	4	5	...	0	0	0	1	0	0	0	0	0	0
4	31	1	2	4	7	0	2	0	7	9	...	0	0	0	1	0	0	0	0	0	0
...
9817	36	1	1	2	8	0	6	1	2	1	...	0	0	0	1	0	0	0	0	0	0
9818	35	1	4	4	8	1	4	1	4	6	...	0	0	0	1	0	0	0	0	0	0
9819	33	1	3	4	8	0	6	0	3	5	...	0	0	0	1	0	0	0	0	0	1
9820	34	1	3	2	8	0	7	0	2	7	...	0	0	0	0	0	0	0	0	0	0
9821	33	1	3	3	8	0	6	1	2	7	...	0	0	0	0	0	0	0	0	0	0

9822 rows × 86 columns

데이터소개

네덜란드 보험회사(TIC)의 데이터
9822개의 관측값과 86개의 변수로 구성

X1 - X43

사회통계학적 데이터

동일한 우편번호 기준으로 사회통계학적 데이터는
동일하다고 가정
X6부터는 백분율로 0부터 9까지로 표현(L3)

X44 - X85

가입된 보험상품과 보험료

보험료 관련 변수들(X44~X64)은
네덜란드 화폐단위(L4)로 표현

X86 - 타겟변수

데이터소개

변수설명

X1	고객하위분류 (L0)	X25-X29	사회계급
X2	같은우편가구수	X30-X31	집소유여부
X3	평균가구크기	X32-X34	차량소유대수
X4	평균나이 (L1)	X35-X36	건강보험 공보험/사보험여부
X5	고객분류 (L2)	X37-X41	수입
X6-X9	종교 (L3)	X42	평균수입
X10-X12	결혼관계	X43	구매력
X13-X15	자녀여부	X44-X64	종류별 보험료 (L4)
X16-X18	교육수준	X65-X85	종류별 가입수
X19-X24	직업	X86	이동주택보험(타겟변수)

L3

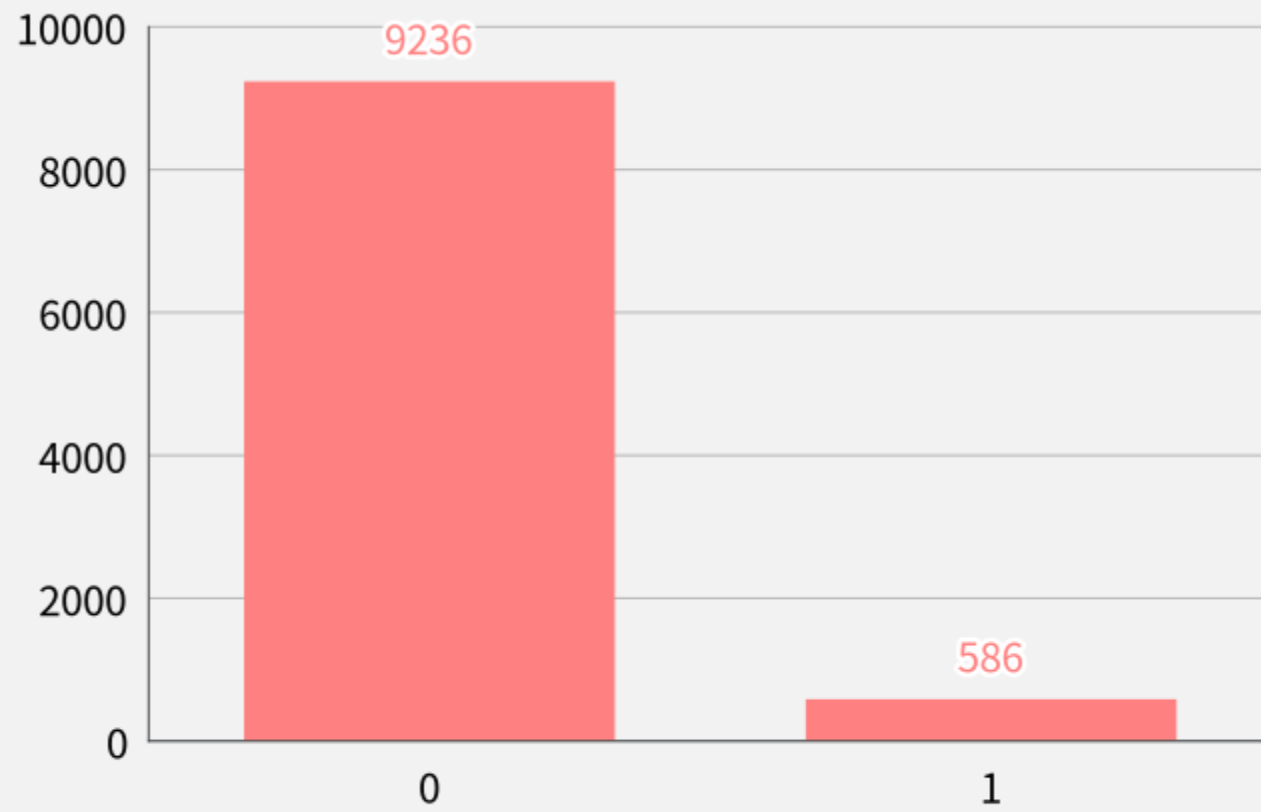
0 0%
1 1 - 10%
2 11 - 23%
3 24 - 36%
4 37 - 49%
5 50 - 62%
6 63 - 75%
7 76 - 88%
8 89 - 99%
9 100%

L4

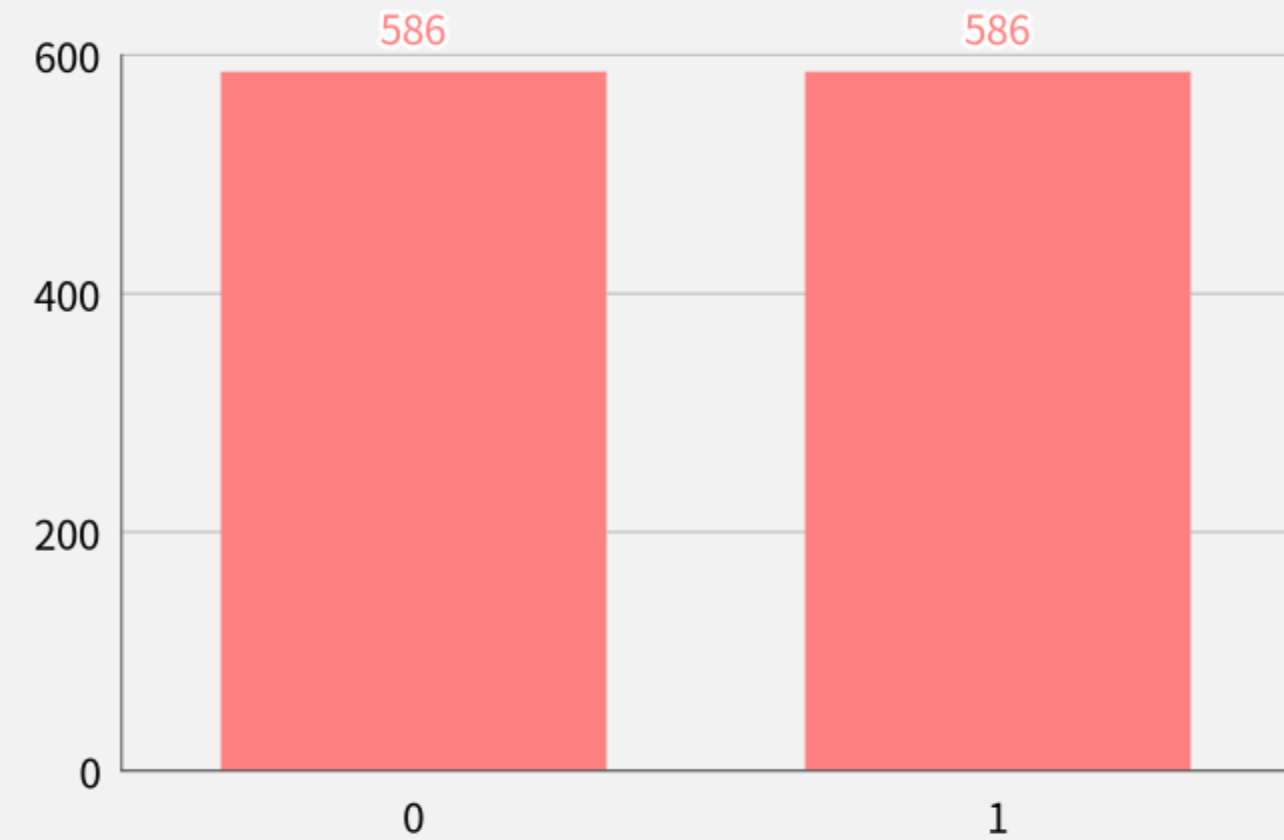
0 f 0
1 f 1 - 49
2 f 50 - 99
3 f 100 - 199
4 f 200 - 499
5 f 500 - 999
6 f 1000 - 4999
7 f 5000 - 9999
8 f 10.000 - 19.999
9 f 20.000 -

데이터소개

불균등한 반응변수 처리



원데이터 반응변수 비율



동일한 비율로 추출한 데이터

반응변수 비율 50:500이 되도록 데이터를 추출함

데이터전처리

새로운변수생성

X4_1	평균 나이	X38_1	최저 평균 수입
X9_1	종교가 있을 확률	X42_1	소득 분위
X10_1	결혼 했을 확률	X47_money~ X64_money	각 보험료 총합
X15_1	아이가 있을 확률	X44_mean	책임보험 평균 보험료
X16_1	교육 수준	X44_mean.money~ X64_mean.money	각 보험료 평균
X19_1~X24_1	직업의 비율	Sum1~Sum3	보험별 총합
X31_1	집을 소유할 확률	Sum_moeny1~ Sum_money3	보험별 보험료 총합
X32_1	차의 평균 개수	Mean_money1~ Mean_money3	보험별 보험료 평균
X35_1	공보험에 가입할 확률	Avg_fee	실제 보험료 평균
X37_1	소득 수준		

데이터전처리

새로운변수생성

- ☑ **X4_1 변수 생성**
자료에 나온 구간의 중앙값을 대입하여 평균 나이 계산
- ☑ **X16_1 변수 생성**
X16~X18변수에 각각 1,0,-1이라는 가중치 부여
가중 평균 계산 후 교육 수준을 -1~1 범주사이에 표현
- ☑ **X9_1, X10_1, X15_1, X31_1, X35_1 변수 생성 (L3변수처리)**
여러 변수를 묶어서 각각 종교유무, 결혼유무, 아이유무 변수 생성

방법1) 기존의 변수들을 결합하여 존재의 확률을 표현
예시) X9_1의 경우 X6~X8=1, X9=0을 부여하여 계산

방법2) 기존의 변수를 활용하여 존재하지 않음의 확률을 표현
예시) X6~X9 중 X9 를 종교를 가지지 않을 확률로 간주
- ☑ **X19_1~X24_1 변수 생성**
X19~X24는 각 직업을 나타내는 변수
전체와의 비율을 계산하여 특정 직업에 해당하는 비율을 계산
예시) X19_1의 경우 X19=1, X20~X24=0을 부여하여 계산

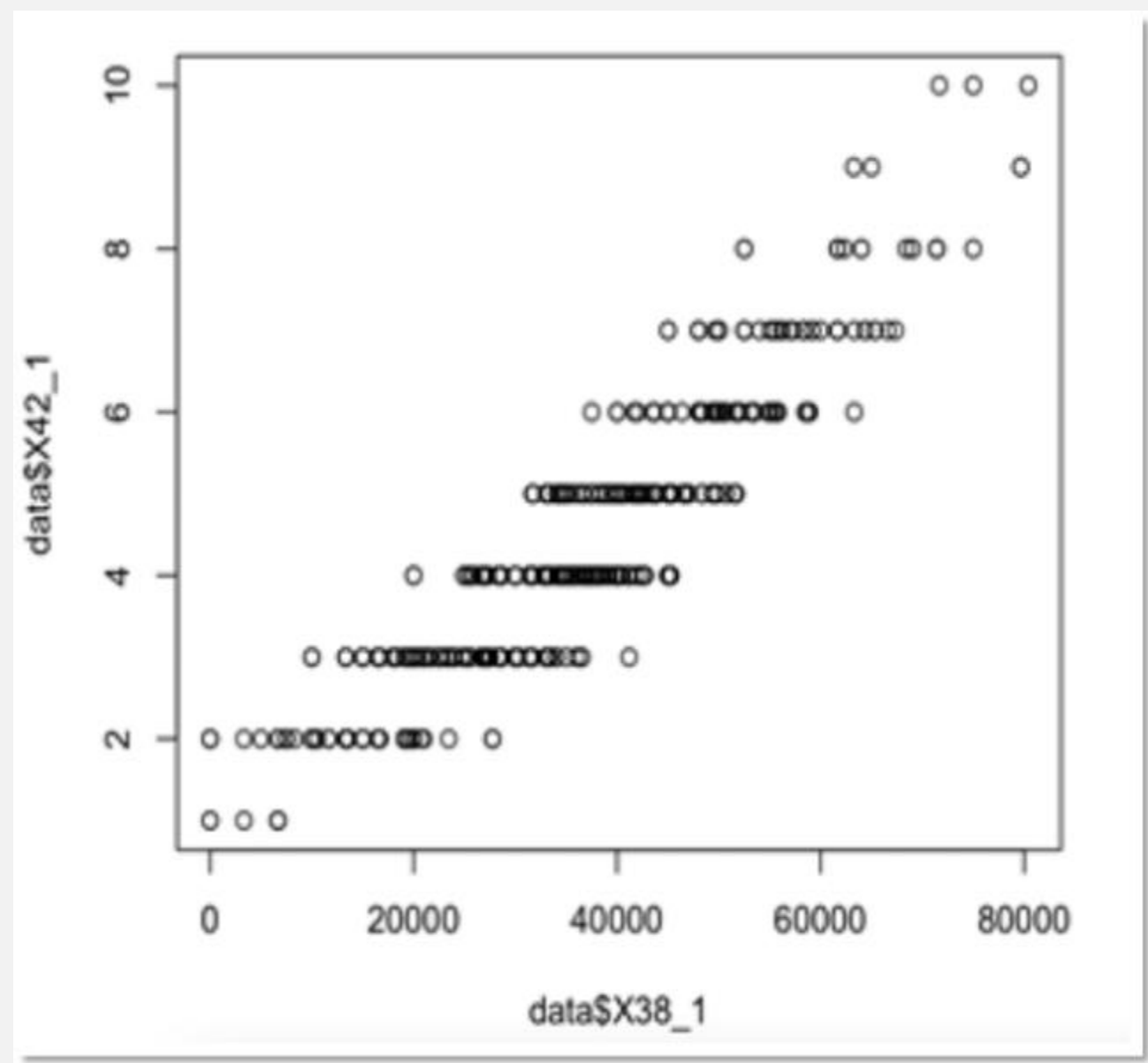
데이터전처리

새로운변수생성

- ☑ X32_1 변수 생성
자료에 나온 소유한 차의 개수를 대입하여 차의 개수 계산
- ☑ X38_1 변수 생성
X37~X41의 최저 수입을 대입하여 최저 평균 수입을 계산
- ☑ X37_1 변수 생성
X37~X41 변수를 1~5단계로 구분하여 연속적인 범주로 표현
- ☑ X42_1 변수 생성
X42 변수의 0을 10으로 변환하여 소득분위로 표현

데이터전처리

새로운변수생성



- ✓ 새로 생성한 X42_1 변수와 X38_1 변수 사이의 산점도
X42_1 변수를 소득분위로 간주하는 것에 대한 타당성 제시

데이터전처리

새로운변수생성

- ☑ X47_money~ X64_money 변수 생성
X47~X64 변수의 최저 보험료를 대입하여 보험료 계산
- ☑ X44_mean 변수 생성
X44와 X65를 이용하여 가구당 평균 책임보험료 계산
- ☑ X47_mean.money~ X64_mean.money 변수 생성
새로 생성한 X47_money~X64_money를 각 보험수로 나누어 평균 보험료 계산
- ☑ Sum1~Sum3 변수 생성
나타난 보험을 생명보험, 손해보험, 제3 보험으로 구분하여 보험의 총합을 계산

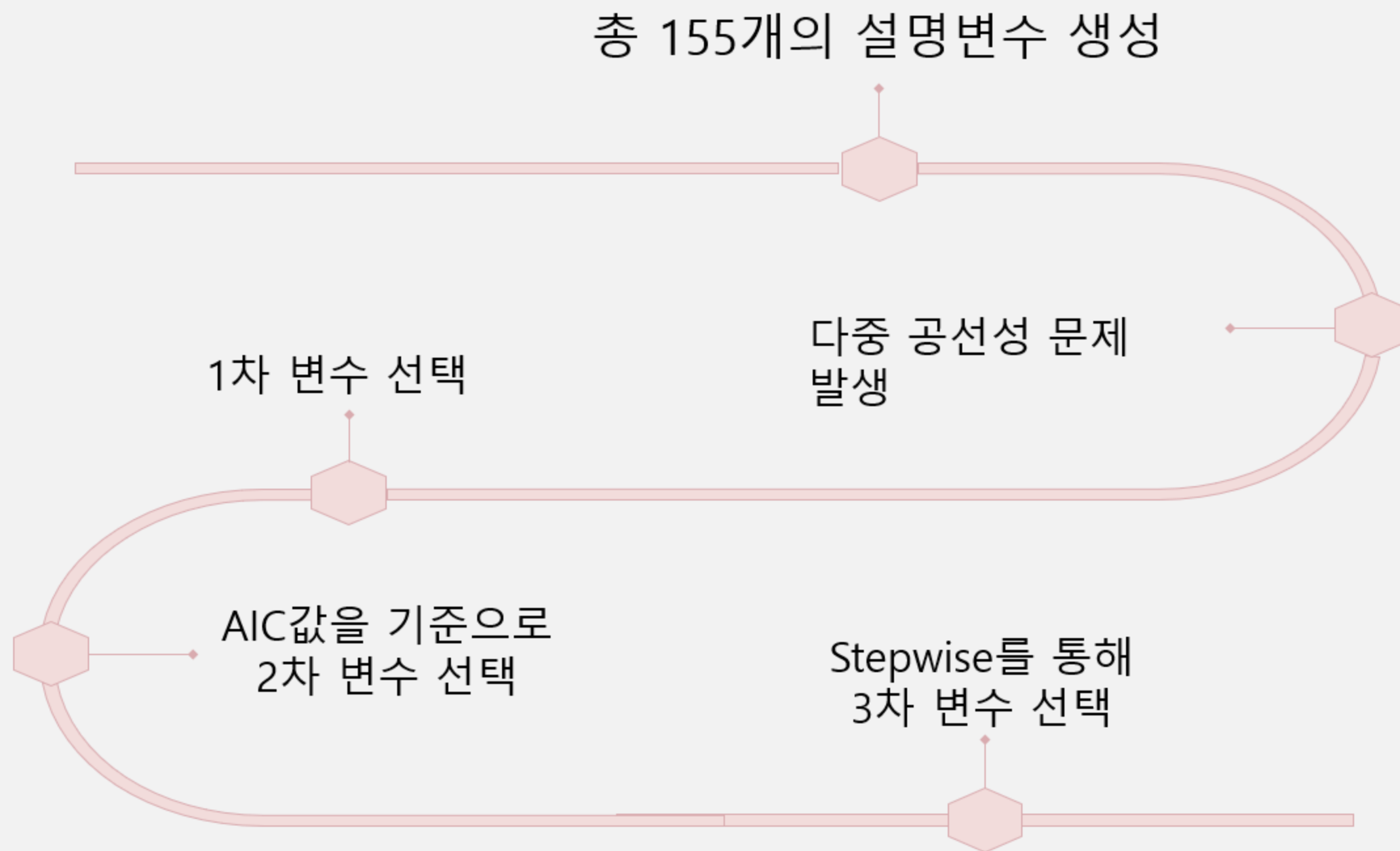
데이터전처리

새로운변수생성

- ☑ Sum_money1~ Sum_money3 변수 생성
각 보험별로 보험료 계산
- ☑ Avg_fee 변수 생성
실제 영향을 주는 요소의 실제 보험료를 대입하여 실제 보험료 계산
- ☑ Mean_money1~ Mean_money3 변수 생성
보험별 보험료를 보험의 수로 나누어 평균 보험료 계산

데이터전처리

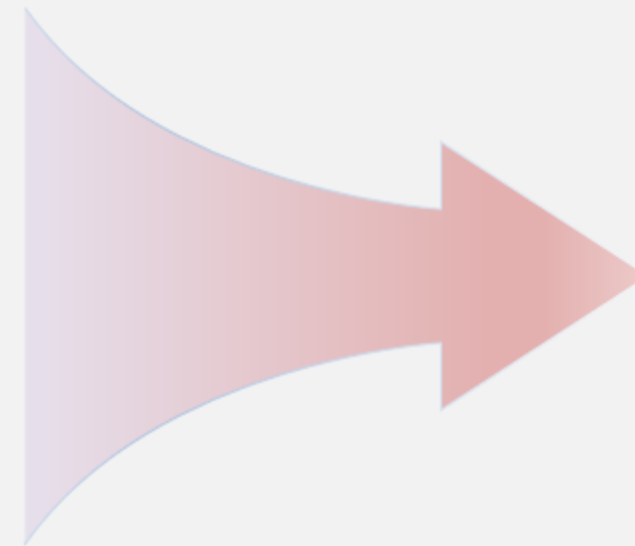
새로운변수생성



데이터전처리

1차변수선택

- ☑ 새로운 변수를 생성할 때 사용한 기존의 변수 삭제
- ☑ X25~X29(계층) 변수 삭제
X25~X29 변수의 기준이 명확하지 않음
- ☑ 종교 변수(X9_1) 삭제
종교의 종류와 유무는 보험가입 여부에 영향 X



1차 변수 선택 결과
108개의 설명변수 선택

데이터전처리

1차변수선택

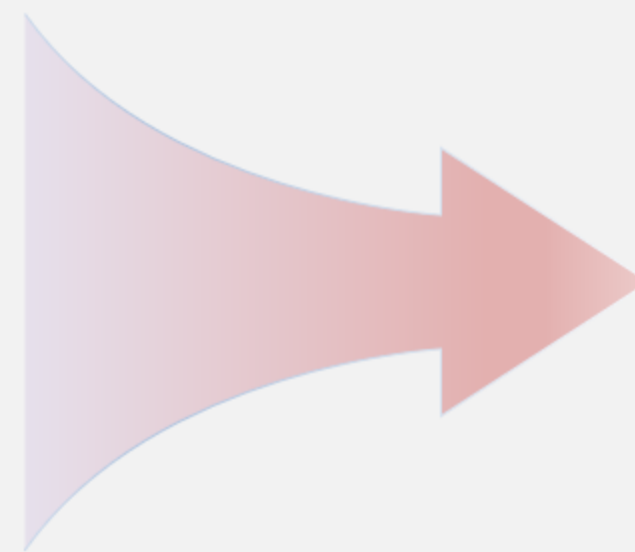
계층변수	가장 높은 빈도수 2가지	가장 낮은 빈도수 2가지
Class A	1 , 8	4 , 6
Class B1	8 , 3	4 , 6
Class B2	8 , 3	4 , 6
Class C	8 , 9	4 , 6
Class D	8 , 7	4 , 10

- ☑ 계층변수(X25~X29)의 계층별 X5 변수 빈도수 계층마다 나타나는 빈도수가 비슷하며 이는 계층에 의미가 없다고 간주

데이터전처리

2차변수선택

- ✓ 1차 변수 선택 후 남은 108개의 변수의 AIC를 계산
- ✓ AIC 가 1608 이하인 변수들만 선택
AIC가 1608을 기준으로 급등
L3처리시 방법1과 방법2를 이용한
변수가 모두 포함된 경우 상대적으로 작은 변수 선택
- ✓ 보험료의 영향을 파악하기 위해 Avg_fee 추가 선택



2차 변수 선택 결과 22개의 설명변수 선택

X68, sum_2, X47_mean.money, X47_money,
sum_money2, X44_mean, X44_mean.money,
X44_money, X65, X5, X1, X16_1,
X21, X43, X80, X37_1, X42_1,
X38_1, X31_1, X10, X24, avg_fee

데이터전처리

2차변수선택

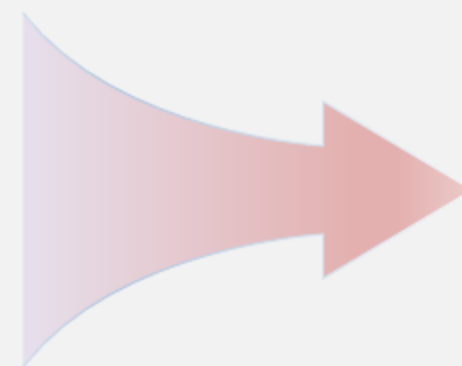
X38_1	1599.421
X31_1	1601.702
X10	1601.718
X31	1602.022
X10_1	1602.275
X24	1605.567
X24_1	1607.321
X85	1614.735
X35	1614.894
X35_1	1614.078

- ☑ AIC 값이 X24_1 과 X85 사이에서 상대적으로 크게 증가
: AIC가 16080이하인 변수들을 선택하는 근거 제시

데이터전처리

3차변수선택

- ✓ 2차 변수 선택 후 남은 22개의 변수를 이용하여 Stepwise 진행
- ✓ 전진 선택법, 후진 선택법, 단계적 선택법 적용



3차 변수 선택 결과 12개의 설명변수 선택

X68, X47_mean.money, X47_money, X44_mean,
X65, X16_1, X21, X80, X10, X24 ,X5, avg_fee

모형구축

Logistic Regression

- 일반화 선형모델의 한 형태로 반응변수가 범주형인 경우 사용
- 모형의 예측력과 해석력을 높일 수 있음

Neural Networks

- 여러 개의 뉴런들이 상호연결하여 입력에 상응하는 최적의 출력 값을 예측
- 매우 안정적이며 예측력이 뛰어남

01

로지스틱회귀

02

의사결정나무

Decision Trees

- 자료들 속에서 나타나는 일정 패턴을 바탕으로 설명모델을 만듦
- 모형을 해석하고 이해하기가 쉬우며 정규성이나 선형성 등의 가정이 불필요

03

인공신경망

04

XGBOOST

XGboost

- 의사결정나무 알고리즘 중 하나로 여러 개의 분류·회귀나무를 묶어 error값을 낮추는 부스팅 기법을 활용한 알고리즘
- 빠른 속도와 확장성

모형구축

70%

TRAIN DATA

30%

TEST DATA

모형평가

		Actual Value	
		True	False
Predicted Value (Classification)	True	True Positive(TP)	False Positive(FP)
	False	False Negative(FN)	True Negative(TN)

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Total}$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1 score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

모형평가

보험가입 여부에 대한 '반응 모형'을 구축하여 예비고객
을 효과적으로 추출하는 것이 목표

실제로 보험을 가입할 고객에게 광고를 하지 않는 것(FN)
실제로 보험에 가입하지 않을 고객에서 광고를 하는 것 (FP)

로지스틱		
Pre/Ref	0	1
0	124	64
1	51	111
Accuracy: 0.6714		

인공신경망		
Pre/Ref	0	1
0	90	42
1	85	133
Accuracy: 0.6371		

의사결정나무		
Pre/Ref	0	1
0	114	47
1	61	128
Accuracy: 0.6914		

XGBOOST		
Pre/Ref	0	1
0	104	56
1	71	119
Accuracy: 0.6371		

Accuracy가 높으면서 혼동행렬 중 FP와 FN중
FP의 비율이 더 큰 모델을 선택

'의사결정나무' 선택

보험료 변수추가

모델의 정확도를 좀 더 높이고자
심리적인 요인을 일부 반영하기 위해 '보험료'라는 새로운 변수를 생성



01
결혼여부



02
교육수준



03
집소유여부




04
연령대별 보험료

보험료 변수추가

변수 유의성확인

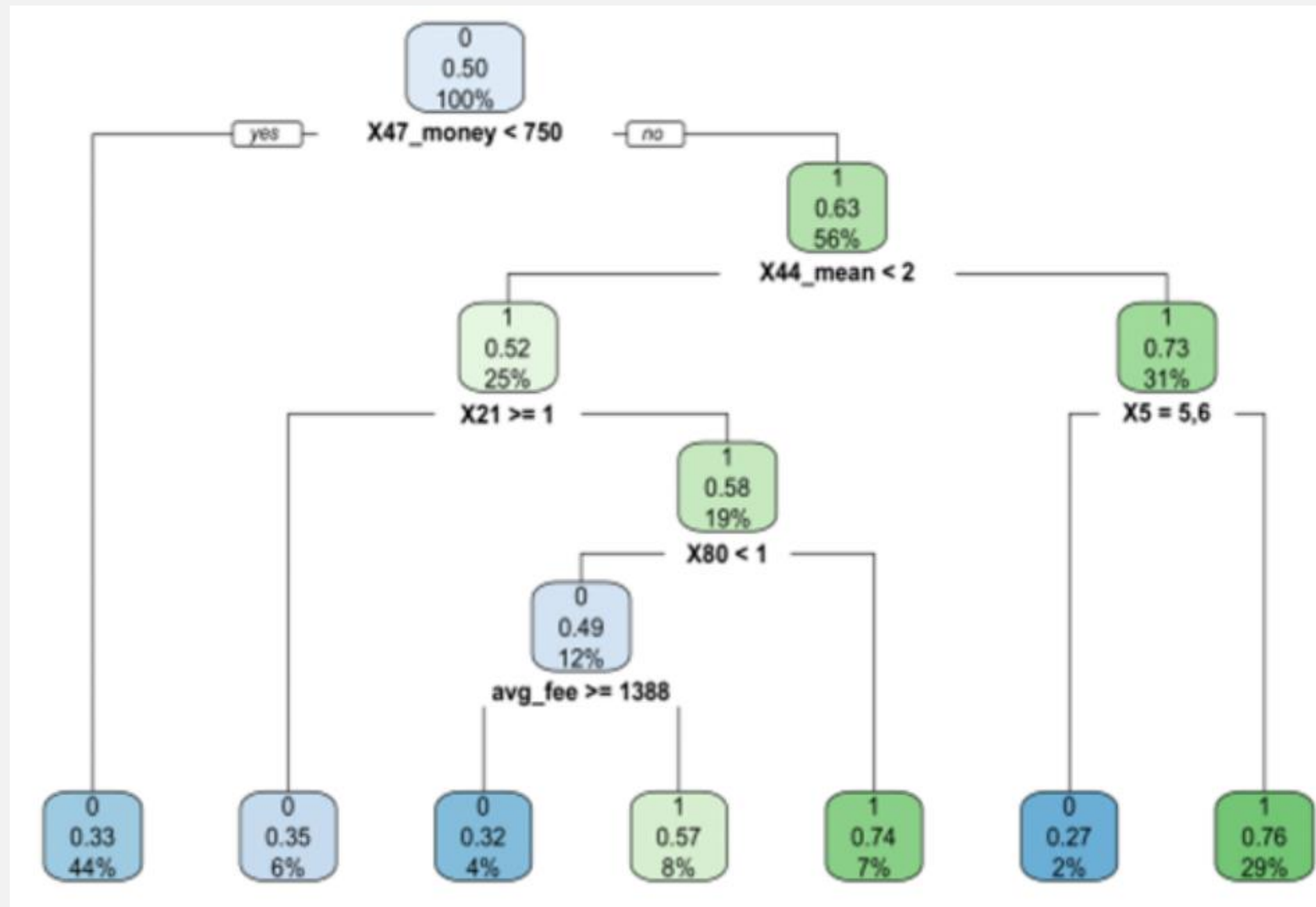
의사결정나무		
Pre/Ref	0	1
0	114	47
1	61	128
Accuracy: 0.6914		

수정된 의사결정나무		
Pre/Ref	0	1
0	138	60
1	37	115
Accuracy: 0.7229		



새로 만든 변수를 정확도를 기준으로 유효성을 판단하면, 정확도가 72%로 높아짐
따라서 '보험료' 변수가 보험가입 여부에 유의미한 영향을 미친다고 판단할 수 있음

보험료 변수추가



최종 채택변수

$X47_money \geq 750$

$x44_mean$

$X21 \leq 1$

$X80 \geq 1$

$Avg_fee \leq 1388$

$X5 \neq 5,6$

결론



채택 변수 : X47_money(차 보험료), X44_mean (가구당 평균 책임 보험료), X21(Farmer),
X5(living well, crusing seniors), X80(화재보험 가입 수), avg_fee(caravan 보험료)

결론



채택 변수 :

X47_money(차 보험료), X44_mean (가구당 평균 책임 보험료), X21(Farmer), X5(living well, crusing seniors), X80(화재보험 가입 수), avg_fee(caravan 보험료)

마케팅 비용과 시간의 효율을 위해 caravan 보험 가입의 가능성이 높은 고객을 타겟으로 설정

차 보험료와 가구당 평균 책임 보험료를 많이 내는, 화재보험의 가입 수가 높고, 책정될 caravan보험료가 낮은 가급적이면 농부의 직업을 가지고 있지 않은 타겟층을 대상으로 우편을 돌리거나 활동반경에서 보험가입을 유도하는 방안을 모색해야함

한계점

01

불균등한 반응변수

반응변수 X86에서 보험에 가입하는 사람의 수(586)가 전체 데이터(9822)에 비해 매우 적음

02

정확하지 않은 데이터

일정한 기준에 따라 분류되어있는 데이터 X
광범위한 범주나 백분율로 표현된 데이터
-> 데이터 해석과 활용에 어려움

03

보험 가입 모형 구축에 필요한 변수 부족

사회적 영향과 잠재고객들의 심리적 변수
의무보험제도 및 요율규제
잔여시장
보험료지원효과
운전자의 특성

THANK YOU!