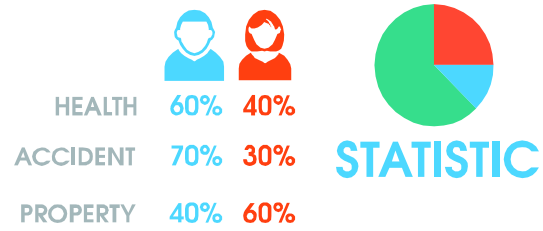
80%
PROPERTY

Lorem ipsum
Lorem ipsum dolor sit amet,
consectetuer adipiscing elit



데이터 마이닝

Caravan 보험 가입 여부에 대하여

김영석 김은태 오지수

SAVE MONEY

Lorem ipsum
Lorem ipsum dolor sit amet,
consectetuer adipiscing elit

ACCIDENT

Lorem ipsum
Lorem ipsum dolor sit amet,
consectetuer adipiscing elit

Index

Ⅰ. 데이터 소개 및 분석목적

The Insurance Company Data

네덜란드 데이터 마이닝 회사
Sentient Machine Research에서 제공한 보험 데이터

관측치(obs) : 9822개 –〉 1172개

독립변수 : 85개

종속변수 : CARAVAN

보험 회사의 고객에 대한
여러 정보가 들어있는 데이터를 분석

어떤 변수가 Caravan 보험 가입에 가장 영향을 많이 끼칠까?

왜 caravan 보험인가?

네덜란드인은 캠핑족!

네덜란드의 캠핑 인구는
전체인구(약 1600만) 중 2/3에 달함

네덜란드 국민 1인당 1년에
한화 약 127만원을 휴가비로 지출

→ Caravan 시장이 매우 큼

II. 데이터 탐색

종속변수

Caravan 보험 가입 여부(0,1) → Logistic Regression

독립변수

- 개인에 대한 변수(1–5)
- Zipcode에 기반한 비율 변수(6–41)
- 수입에 대한 변수(42–43)
- 보험개수 및 보험료에 대한 변수(44–85)

● 개인에 대한 변수

변수	분류	변수 설명	참고
MOSTYPE	Factor	고객 서브타입	표 L0
MOSHOOFD	Factor	고객 메인타입	표 L2

〈표 L0〉

Label	설명
1	고수입자
2	지방 출신 고위급 사람
3	높은 지위의 고위층
⋮	
41	시골 마을의 대가족

〈표 L2〉

Label	설명
1	행복을 중시하는 사람
2	보육을 주도하는 사람
3	일반적인 가족
⋮	
10	농업인

● 개인에 대한 변수

변수	분류	변수 설명	참고
MAANTHUI	Int	집 개수	1–10
MGEMOMV	Int	가족 구성원 수	1–6
MGEMLEEF	Int	가족 구성원 평균 나이	표 L1

〈표 L1〉

Label	설명
1	20 – 30 세
2	30 – 40 세
3	40 – 50 세
⋮	
6	70 – 80 세

- **Zipcode에 기반한 비율 변수**

  : 각 카테고리별 합이 같음 (표 L3 참고)

〈표 L3〉

Label	설명
0	0 %
1	1 – 10 %
2	11 – 23 %
⋮	
9	100 %

〈예시〉

ID	전세/월세	자가	비율 합
A	3	6	9
B	2	7	9

Zipcode에 기반한 비율 변수

〈종교〉

변수	분류	변수 설명	참고
MGODRK	Int	로마 카톨릭 교도	표 L3
MGODPR	Int	신교도	표 L3

⋮

〈결혼여부〉

변수	분류	변수 설명	참고
MRELGE	Int	기혼	표 L3
MRELSA	Int	동거	표 L3
MRELOV	Int	다른 관계	표 L3

Zipcode에 기반한 비율 변수

〈가정형태〉

변수	분류	변수 설명	참고
MFALLEEN	Int	싱글	표 L3
MFGEKIND	Int	아이가 없는 가정	표 L3
MFWEKIND	Int	아이가 있는 가정	표 L3

〈학력수준〉

변수	분류	변수 설명	참고
MOPLHOOG	Int	고학력	표 L3
MOPLMIDD	Int	중간학력	표 L3
MOPLLAAG	Int	저학력	표 L3

Zipcode에 기반한 비율 변수

〈직업〉

변수	분류	변수 설명	참고
MBERHOOG	Int	높은 지위	표 L3
MBERZELF	Int	사업가	표 L3

⋮

〈사회계급〉

변수	분류	변수 설명	참고
MSKA	Int	사회 계급 A	표 L3
MSKB1	Int	사회 계급 B1	표 L3
MSKB2	Int	사회 계급 B2	표 L3

⋮

Zipcode에 기반한 비율 변수

〈자가여부〉

변수	분류	변수 설명	참고
MHHUUR	Int	전세/월세(렌트)	표 L3
MHKOOP	Int	자가	표 L3

⋮

〈차소유〉

변수	분류	변수 설명	참고
MAUT1	Int	차 1대 소유	표 L3
MAUT2	Int	차 2대 소유	표 L3
MAUT0	Int	차 무소유	표 L3

Zipcode에 기반한 비율 변수

〈건강보험〉

변수	분류	변수 설명	참고
MZFONDS	Int	국민건강보험 가입	표 L3
MZPART	Int	사립의료보험 가입	표 L3

〈수입〉

변수	분류	변수 설명(네덜란드 휠던)	참고
MINKM30	Int	수입 〈 30,000	표 L3
MINK3045	Int	30,000 〈= 수입 〈 45,000	표 L3

⋮

● 개별 수입에 대한 변수

변수	분류	변수 설명	참고
MINKGEM	Int	평균 수입	표 L4
MKOOPKLA	Int	구매력 class	1–8

〈표 L4〉

Label	설명(네덜란드 휠던)
0	0
1	1 – 49
2	50 – 99
⋮	
9	20,000 –

• 보험개수 및 보험료에 대한 변수

변수	분류	변수 설명	참고
PWAPART	Factor	제 3자 보험료 (개인)	표 L4
PWABEDR	Factor	제 3자 보험료 (회사)	표 L4
PWALAND	Factor	제3자 보험료 (농업)	표 L4
PPERSAUT	Factor	자동차 보험료	표 L4
⋮			
APLEZIER	Int	보트 보험개수	1–12
AFIETS	Int	자전거 보험개수	1–12
AINBOED	Int	재산 보험개수	1–12
ABYSTAND	Int	사회 보장 보험개수	1–12

〈보험료 변수〉

- 원 데이터의 보험료는 0–9 값
- 휠던의 성격을 어느 정도만 반영

→ Label 별 휠던 중위값으로 변경

〈표 L4〉

Label	설명(네덜란드 휠던)	변경값
0	0	0
1	1 – 49	25
2	50 – 99	75
	⋮	
9	20,000 –	30,000

〈보험료와 보험개수 상관관계표〉

제3자 보험(개인)	제3자 보험(회사)	제3자 보험(농업)	자동차 보험	배달차 보험	오토바이 보험
0.9813	0.9185	0.9861	0.8986	0.9982	0.9626

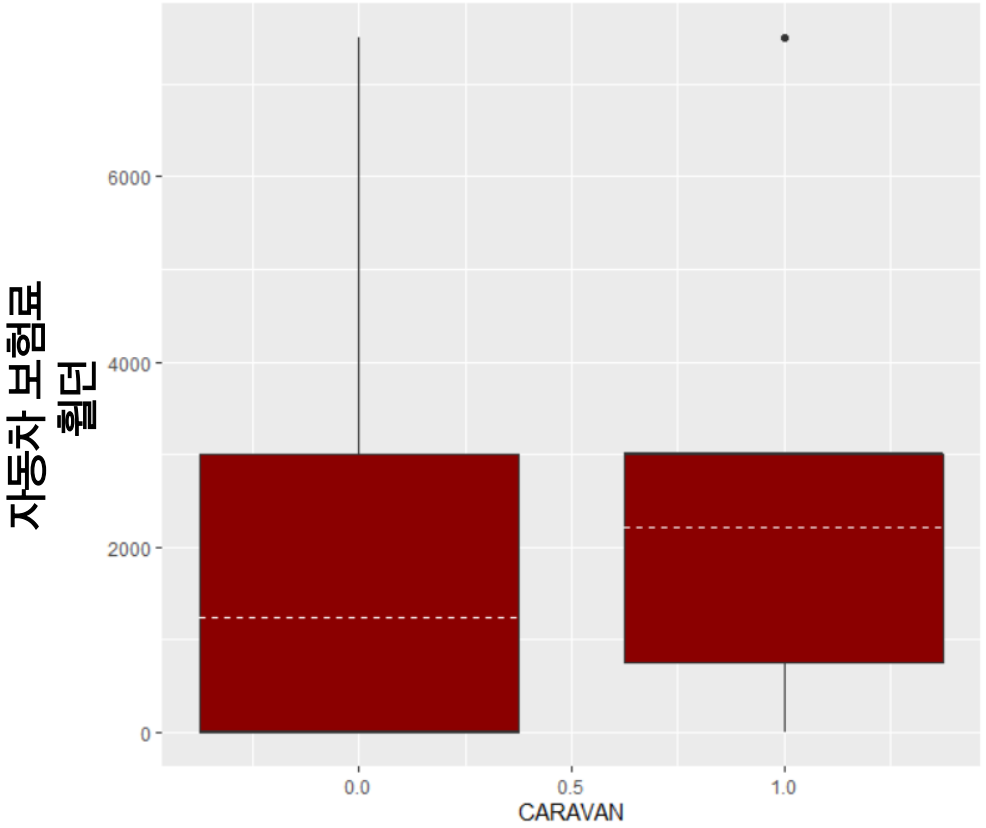
⋮

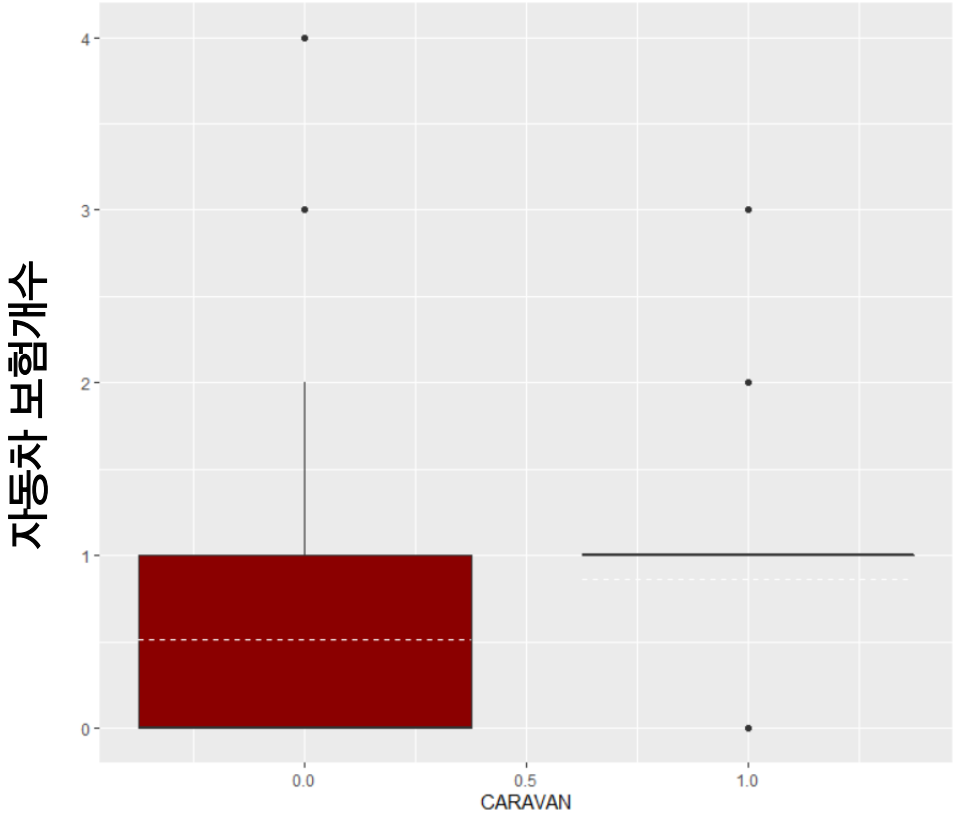
개인상해 보험	가족상해 보험	장애소득 보험	화재 보험	서핑보드 보험	보트 보험
0.8927	0.9808	0.9976	0.8983	0.9621	0.8838

⋮

→ **보험료와 보험개수 중 하나의 변수만 선택**

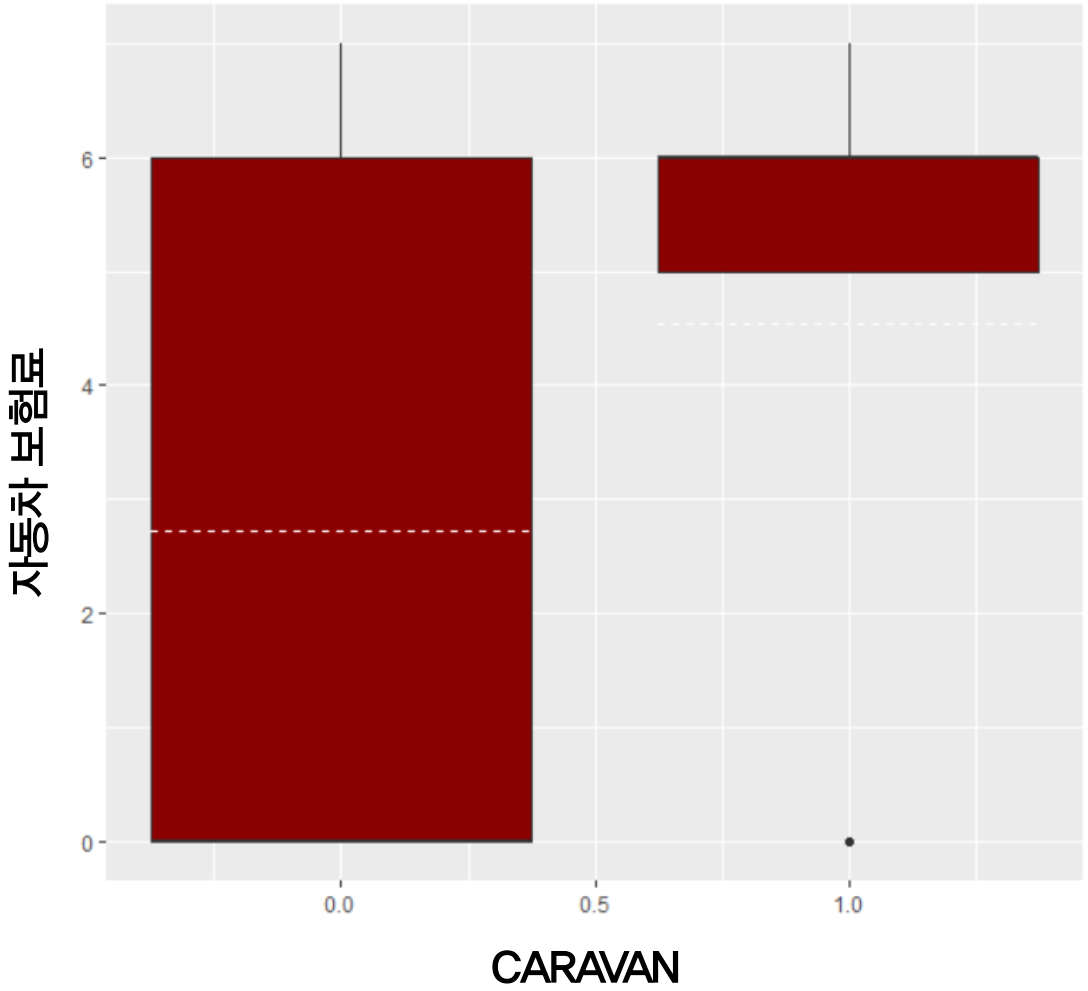
〈보험료와 보험개수 박스플롯 비교〉

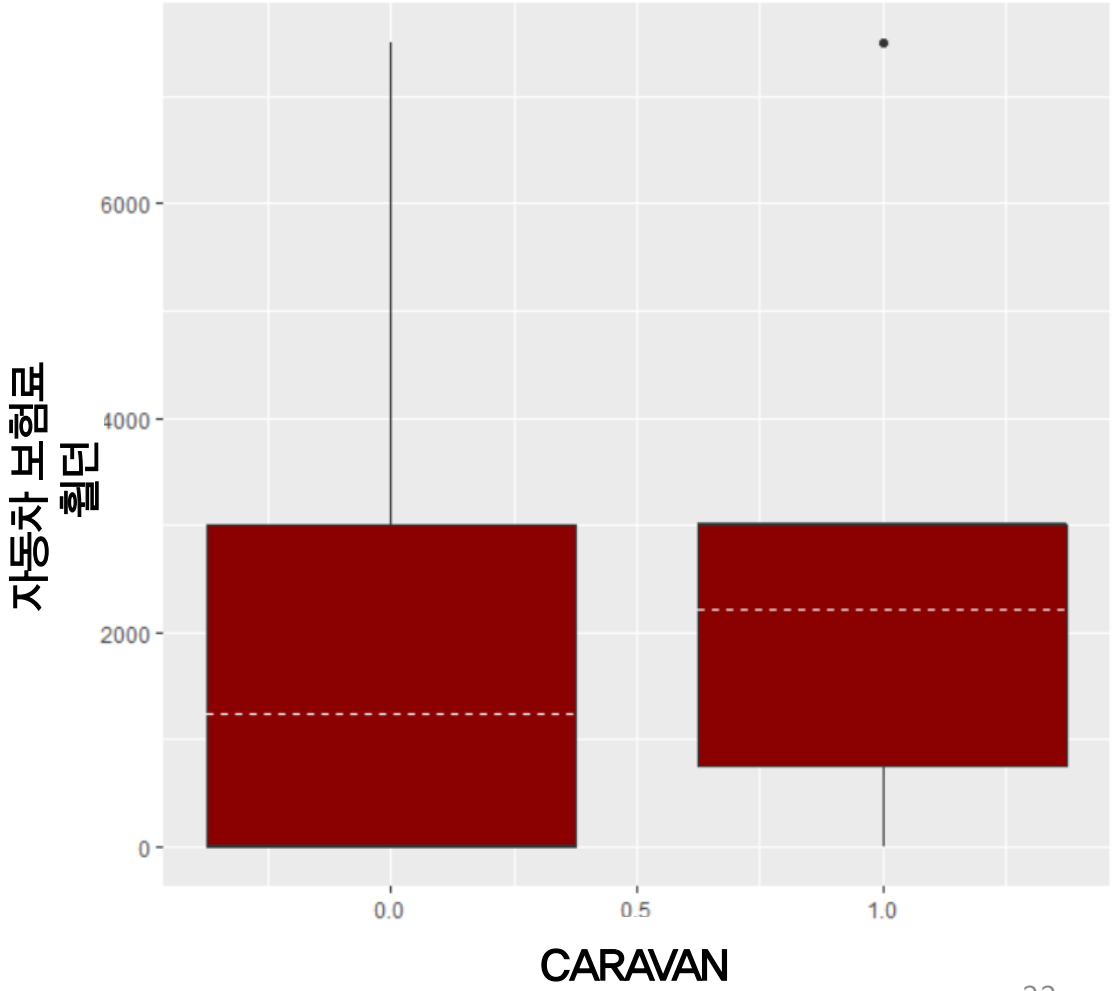




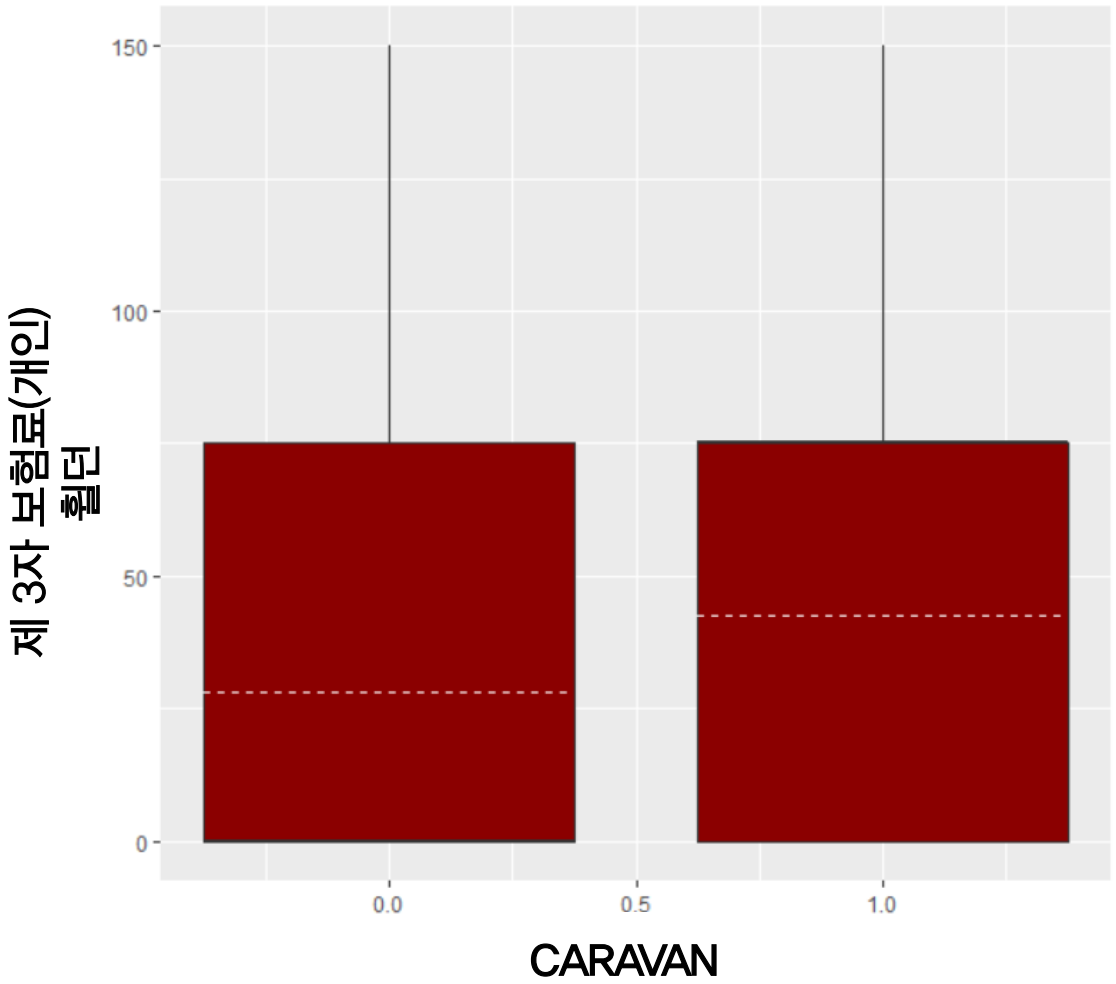
보험개수 제거, 보험료 선택

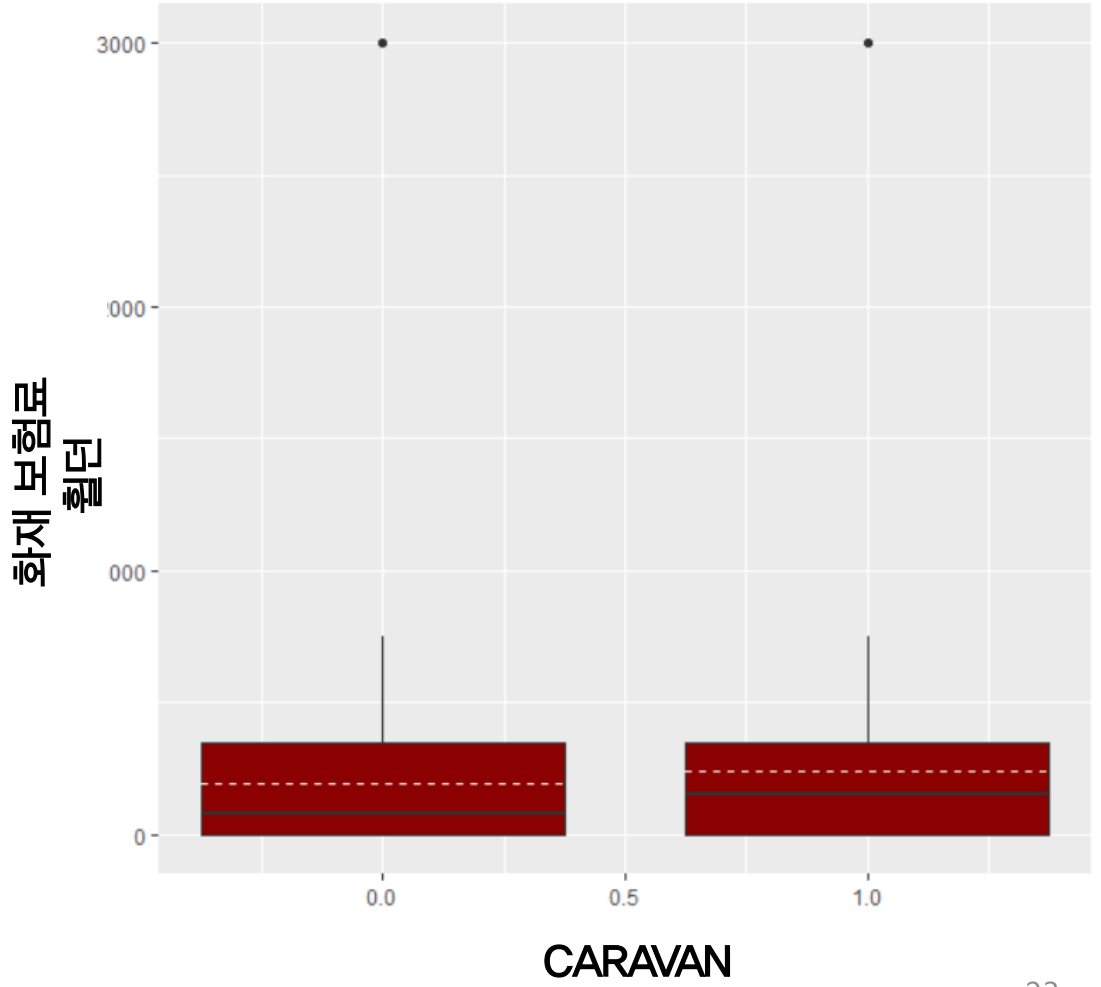
〈보험료 변수〉


자동차 보험료
6
4
2
0
0.0
0.5
1.0
CARAVAN

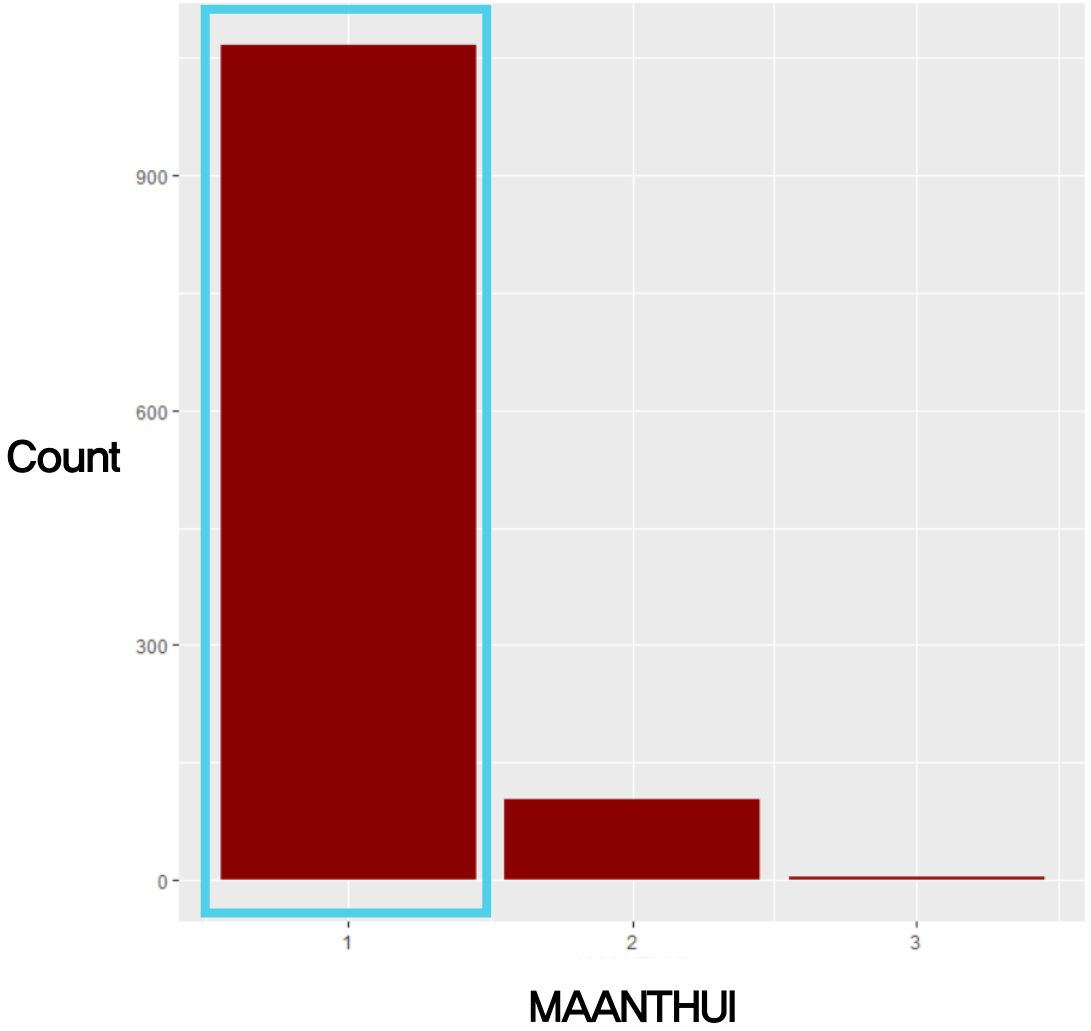

자동차 보험료
훨던
6000
4000
2000
0
0.0
0.5
1.0
CARAVAN

〈보험료 변수〉


150
100
50
0
제 3자 보험료(개인)
훨던
0.0
0.5
1.0
CARAVAN


3000
000
000
0
화재 보험료
훨던
0.0
0.5
1.0
CARAVAN

〈집 개수 변수〉



- 하나의 범주에 대부분의 관측치가 몰려있음

→ **재범주화 진행**

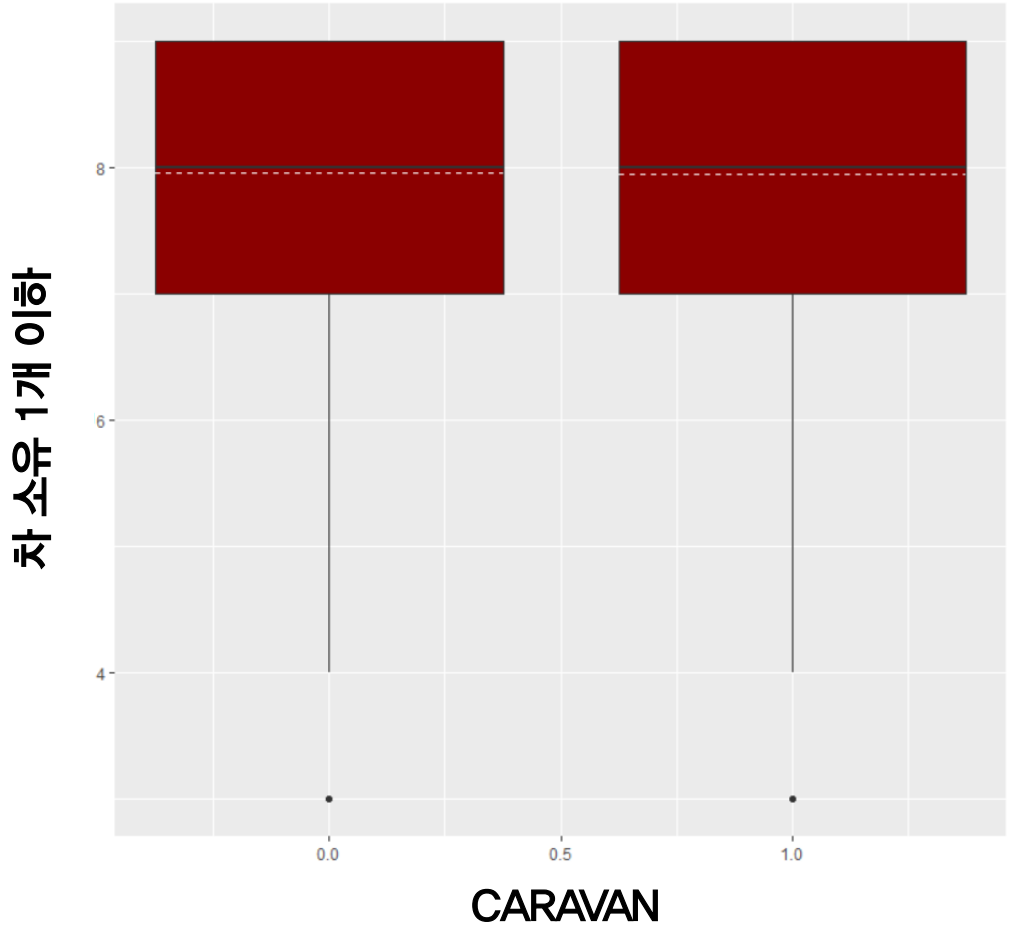
- 범주 2와 3을 합쳐줌

→ **집 1개 소유 / 2개 이상 소유**

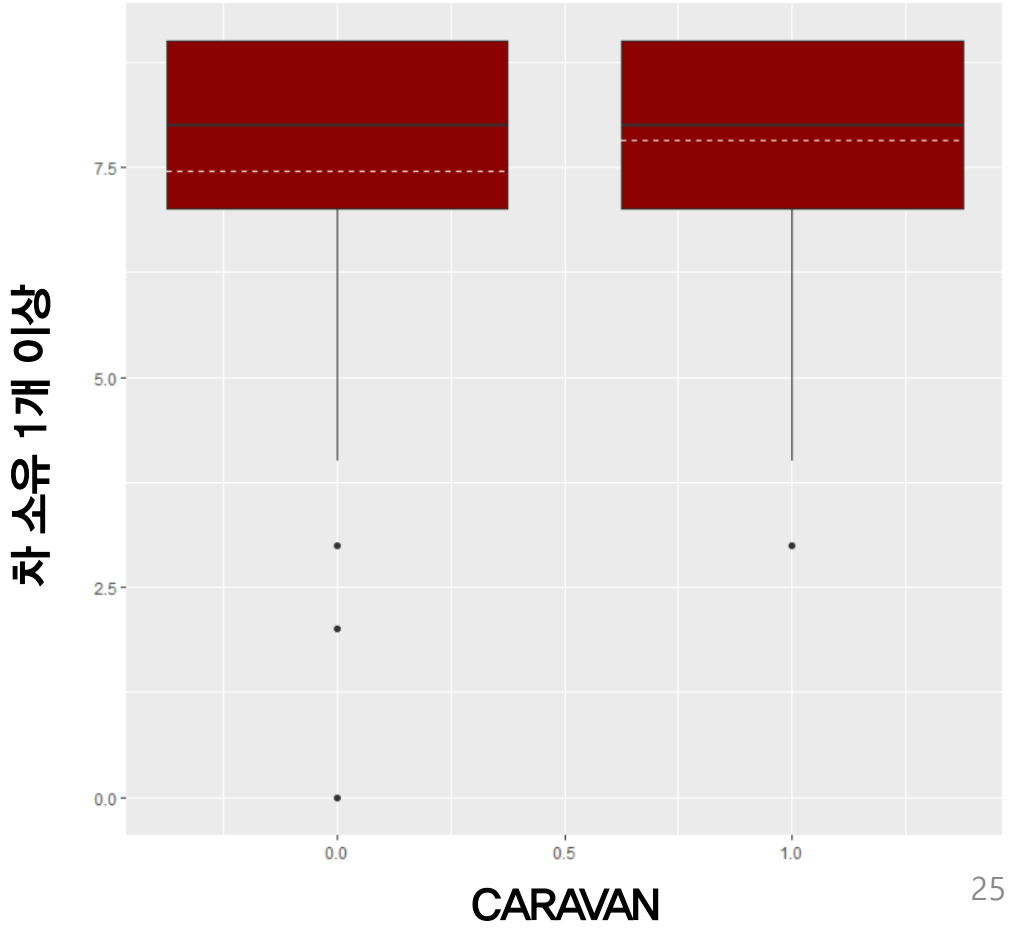
〈차 소유 개수 변수〉 – Zipcode 기반

- MAUT0, MAUT1, MAUT2를 모델에 함께 넣으면 다중공선성 발생 가능

1. MAUT0과 MAUT1를 합하는 방법



2. MAUT1과 MAUT2를 합하는 방법 **(선택)**



〈사회적 위치 관련 변수〉

- MOPL–변수(학력수준)
- MBER–변수(직업)
- MSK–변수(사회계급)

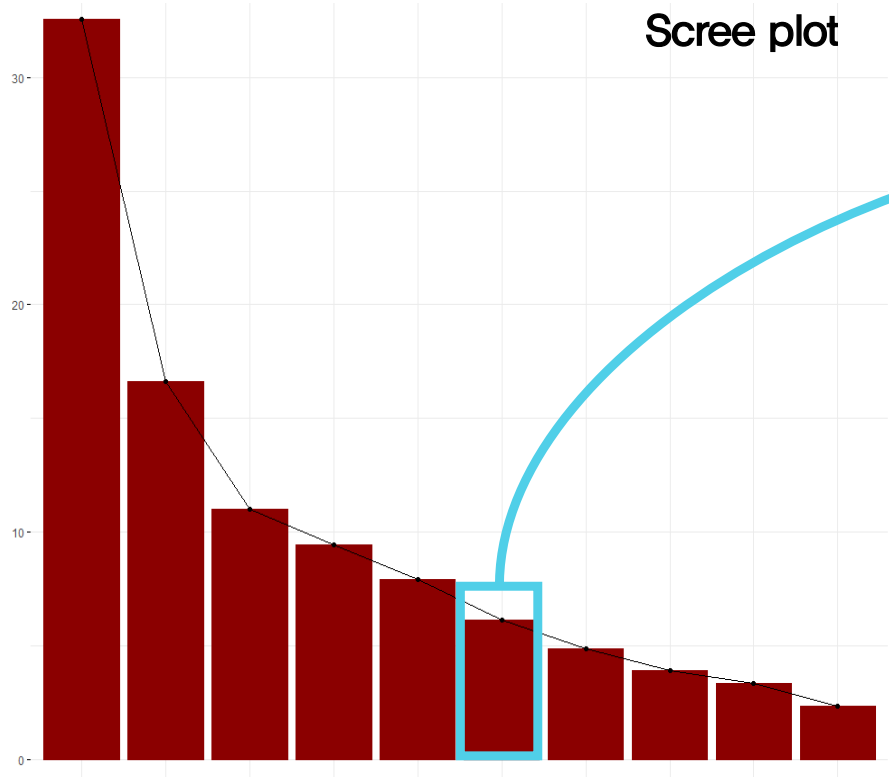
상관계수 살펴봄 → **PCA 진행**

	고학력		높은 지위	사업가	농부	중간 관리직	
고학력	1		0.6038	0.2332	–0.0915	–0.0221	
중간학력	–0.0567		0.0924	0.0683	–0.0245	0.3922	
저학력	–0.669		–0.5072	–0.2146	0.082	–0.2584	
⋮		…					…
사회계급B2	–0.0619		0.064	–0.1595	0.1093	0.1704	
사회계급C	–0.59		–0.5757	–0.1468	–0.1290	–0.1106	
사회계급D	–0.302		–0.2176	–0.1606	0.0544	–0.1994	

〈사회적 위치 관련 변수〉

- 특이값 분해 이용

	PC1	PC2	PC3	…	PC5	PC6	…
분산	0.3258	0.1664	0.1101		0.0789	0.0615	
누적 분산	0.3258	0.4922	0.6023		0.7756	0.8370	



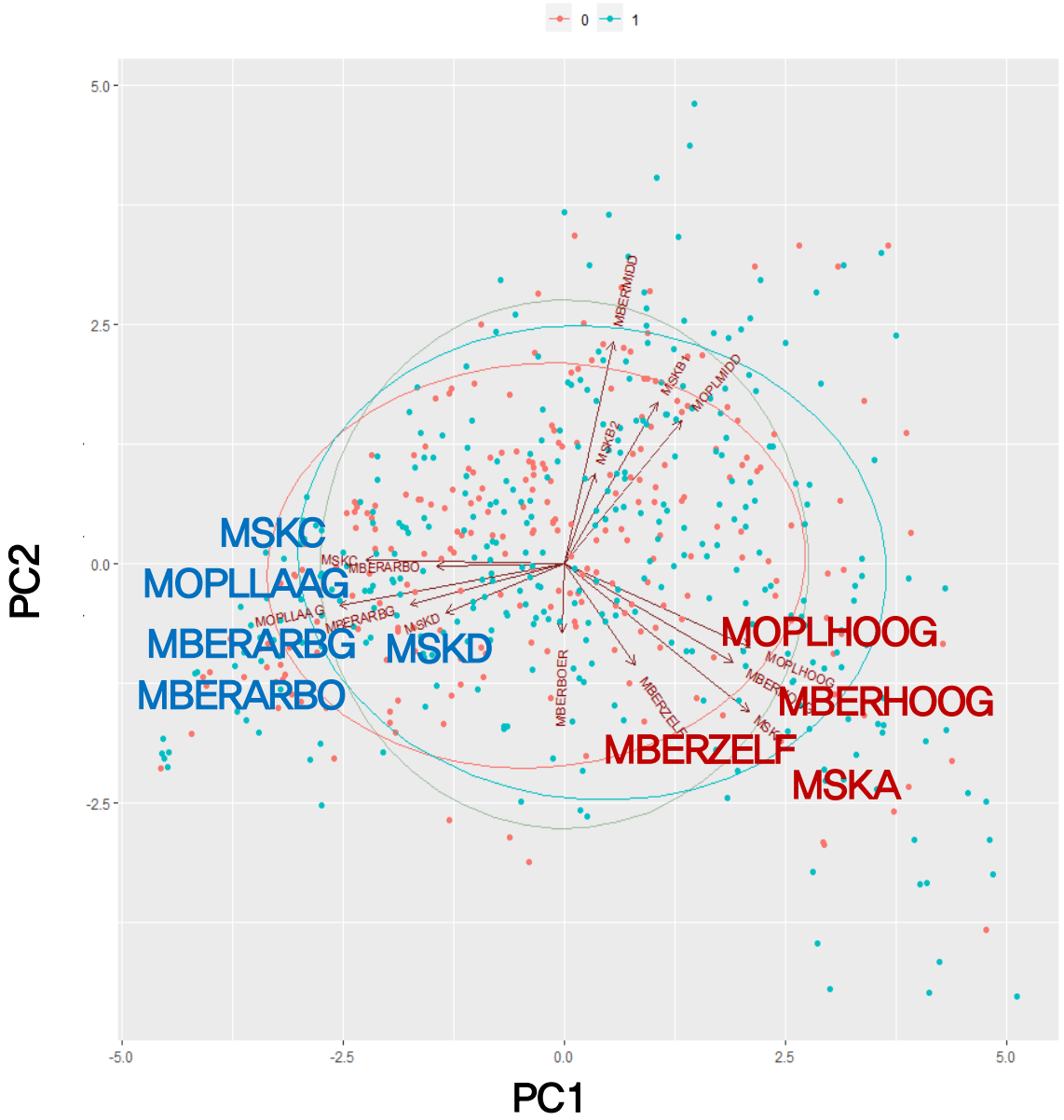
급격히 꺾이는 지점

PC6까지 사용

〈사회적 위치 관련 변수〉

- 주성분 계수

	PC1	PC2	PC3	
고학력	0.3570	–0.2053	0.0609	
중간학력	0.2250	0.3544	0.0069	…
저학력	–0.4309	–0.1040	–0.0444	
높은 지위	0.3232	–0.2426	–0.0908	
⋮				
사회계급C	–0.3803	0.0113	0.3650	
사회계급D	–0.2261	–0.1196	–0.4275	



- 주성분에 대한 해석

주성분	변수 해석
PC1	사회적 위치가 상류층, 하류층과 관련된 변수
PC2	사회적 위치가 중류층과 관련된 변수
PC3	노동자 및 사회계급이 낮은 집단에 대한 변수
PC4	사회계급이 B1인 농부에 대한 변수

⋮

〈가정 형태 관련 변수〉

- MR– (기혼, 동거, 다른 관계)
- MF– (싱글, 아이가 없는 가정, 아이가 있는 가정)

	결혼함	동거중	다른 관계
싱글	–0.6744	0.0473	0.7509
아이가 없는 가정	0.0334	0.1802	–0.1419
아이가 있는 가정	0.4593	–0.1670	–0.4197

→ **MR–변수, MF–변수 함께 처리**

〈가정 형태 관련 변수〉

- MR– (기혼, 동거, 다른 관계)
- MF– (싱글, 아이가 없는 가정, 아이가 있는 가정)

기혼 변수	미혼 변수
결혼함	동거중
아이가 없는 가정	다른 관계
아이가 있는 가정	싱글

기혼 변수 : 기혼 + 아이가 없는 가정 + 아이가 있는 가정

미혼 변수 : 동거 + 다른 관계 + 싱글

최종변수

	변수	라벨	분류	변수 설명	참고
1	MAANTHUI	X1	int	집 개수	1~10
2	MGEMOMV	X2	int	가족 구성원 평균 수	1~6
3	MGEMLEEF	X3	int	가족 구성원 평균 나이	표 L1
4	MOSHOOFD	X4	factor	고객의 메인타입	표 L2
5	MRE	X5	int	기혼비율	
6	PC1	X6	int	사회적 위치가 상류층, 하류층과 관련된 변수	
7	PC2	X7	int	사회적 위치가 중류층과 관련된 변수	
8	PC3	X8	int	노동자 및 사회계급이 낮은 집단에 대한 변수	
9	PC4	X9	int	사회계급이 B1인 농부에 대한 변수	
10	PC5	X10	int	사업가에 대한 변수	
11	PC6	X11	int	사회계급이 B2인 농부에 대한 변수	
12	MHKOOP	X12	int	집 소유 비율	표 L3
13	MAUT	X13	int	차량 소유 비율	
14	MZPART	X14	int	사립의료보험 가입	
15	MINKGEM	X15	int	평균 수입 (0–9)	표 L4
16	MKOOPKLA	X16	int	구매력 class	1~8
17	PWAPART	X17	int	개인 제3자 보험료	표 L4
18	PPERSAUT	X18	int	자동차 보험료	표 L4
19	PBRAND	X19	int	화재보험료	표 L4
20	CARAVAN	Y	factor	caravan 보험 가입 여부	0 or 1

Ⅲ. 모형 구축과정 및 평가

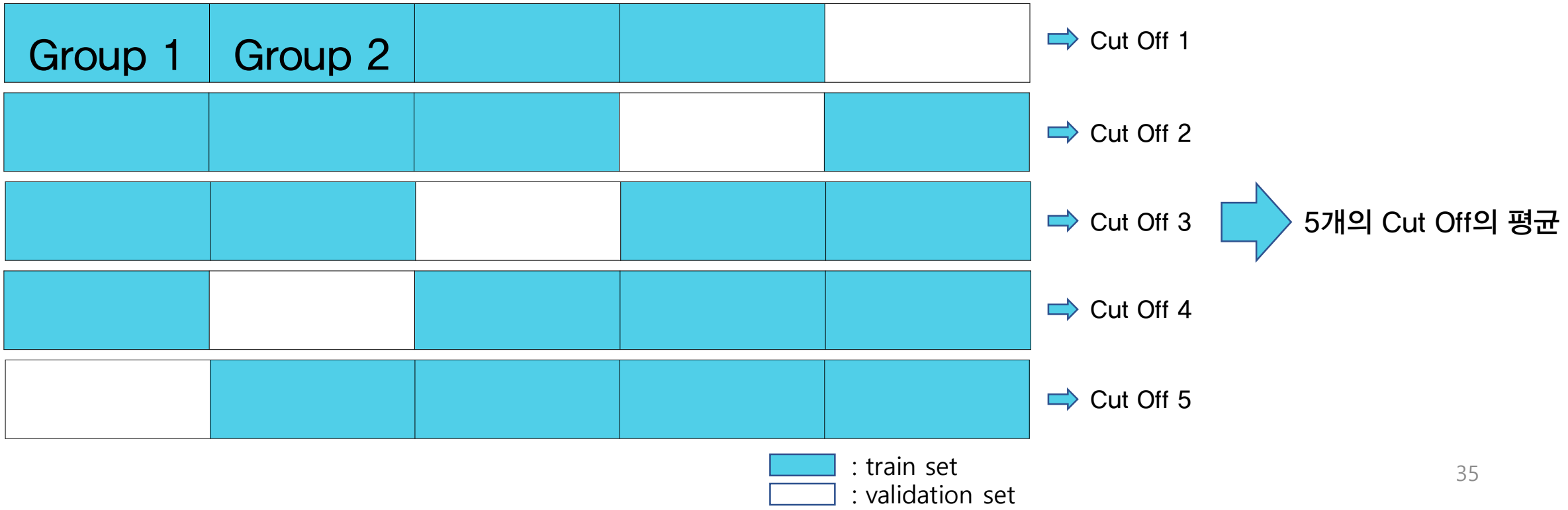
〈1. 모형 구축 알고리즘 선택〉

Logistic Regression	트리 기반 & 신경망
Tree 기반 Bagging, Boosting 모형 혹은 신경망 등에 비해 예측력이 비교적 떨어짐.	예측력은 비교적 좋으나 Black box 기법으로 모형 해석이 불가능함.
모델해석 측면에서 강점	예측력 측면에서 강점

〈2. 데이터 분할〉

- 8:2의 비율로 분할

- 5-Fold Cross Validation



〈3. 모형 구축〉

〈단계적 변수선택법 결과〉

변수명	자유도	AIC
자동차 보험료	1	1217.0
고객 메인 타입	9	1150.5
PC1	1	1141.5
PC6	1	1135.3
PC4	1	1134.9
제3자보험료(개인)	1	1134.4
PC3	1	1133.1

〈3. 모형 구축〉

〈분산팽창인자(VIF) 확인 결과〉

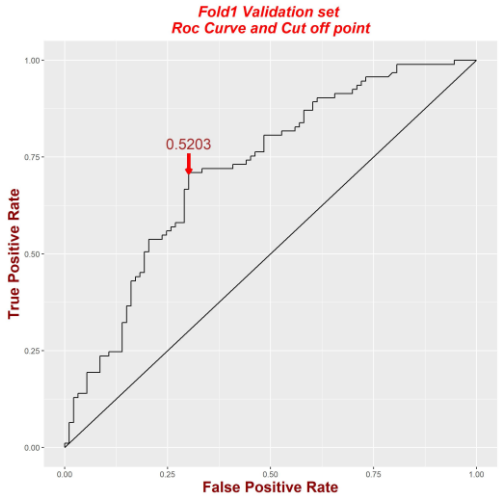
변수명	GVIF	Df	GVIF^(1/(2*Df))
고객 메인 타입	2.3497	9	1.0486
제3자보험료(개인)	1.0733	1	1.0360
자동차 보험료	1.0933	1	1.0456
PC1	1.7258	1	1.3137
PC3	1.0516	1	1.0255
PC4	1.2298	1	1.1090
PC6	1.1039	1	1.0507

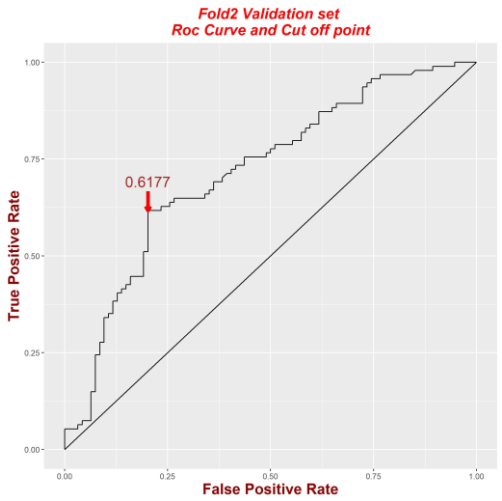
〈3. 모형 구축〉

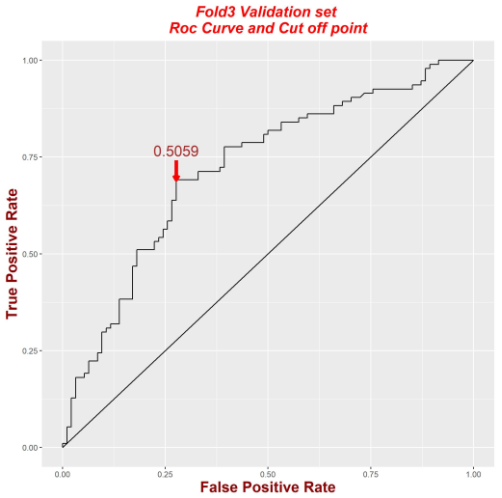
〈Type III Sum Of Square 결과〉

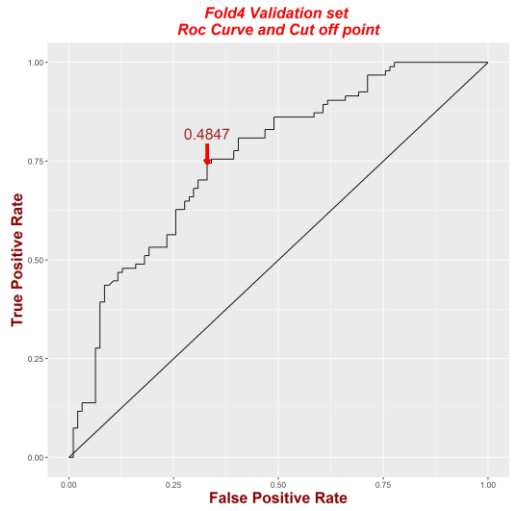
변수명	LR Chisq	Df	GVIF^(1/(2*Df))	
고객 메인 타입	36.746	9	2.921e-05	***
제3자보험료(개인)	4.703	1	0.0301	*
자동차 보험료	87.268	1	<2.2e-16	***
PC1	11.812	1	0.0006	***
PC3	3.406	1	0.0650	.
PC4	5.544	1	0.0185	*
PC6	5.596	1	0.0180	*

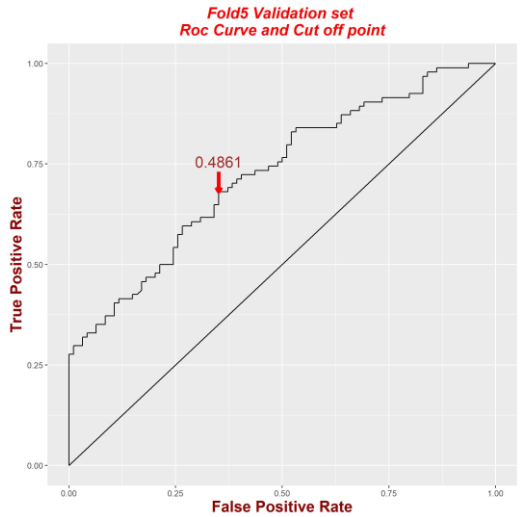
〈4. Cut-Off 도출〉











〈Cut Off 값〉

Fold1	Fold2	Fold3	Fold4	Fold5
0.5203	0.6177	0.5059	0.4847	0.4861.

평균 :0.5229

Confusion Matrix

훈련용 데이터		실제	
		0	1
예측	0	261	116
	1	114	259

정확도 : 0.6933

특이도 : 0.6907

음성 예측도 : 0.6944

평가용 데이터		실제	
		0	1
예측	0	70	39
	1	47	78

정확도 : 0.6325

특이도 : 0.6667

음성 예측도 : 0.6240

이익도표

Decile	Predicted_ prob	Percent_ Active (%)	Cum_ Active (%)	Active_ num	Total_ percent_ active (%)	cum_ active_num	cum_ active_ percent(%)	Lift (%)	Cum_ Lift(%)
0	0.83	66.67	66.67	16	13.68	16	13.68	133	133
1	0.75	56.52	61.70	13	11.11	29	24.79	113	123
2	0.70	82.61	68.57	19	16.24	48	41.03	165	137
3	0.64	50.00	63.83	12	10.26	60	51.29	100	128
4	0.60	56.52	62.39	13	11.11	73	62.40	113	125
5	0.51	56.52	61.43	13	11.11	86	73.51	113	123
6	0.40	33.33	57.32	8	6.84	94	80.35	67	115
7	0.32	43.48	55.61	10	8.55	104	88.90	87	111
8	0.26	43.48	54.29	10	8.55	114	97.45	87	109
9	0.13	12.50	50.00	3	2.56	117	100.00	25	100

모델 평가

- 예측된 확률의 평균이 단조감소
- 십분위0 그룹에서는 평균모형을 이용한 것보다 실제 보험을 가입한 고객을 약 1.33배 더 포함
- 전체 데이터의 50%를 가지고도 약 62%의 가입고객을 찾아낼 수 있다

→ 적절한 모형이다

Ⅳ. 결론

〈최종 모형의 추정값〉

변수명	Estimate	std.Error	z value	Pr(>|z|)	
(Intercept)	−1.362	0.2548	−5.347	8.96e−08	***
MOSHOOFD2	0.8220	0.3033	2.710	0.006721	**
MOSHOOFD3	0.4658	0.2779	1.676	0.093650	.
MOSHOOFD4	−14.12	405.5	−0.035	0.972226	
MOSHOOFD5	−0.7629	0.3992	−1.911	0.055992	.
MOSHOOFD6	−0.3339	0.5405	−0.618	0.536745	
MOSHOOFD7	0.6677	0.3832	1.742	0.081442	.
MOSHOOFD8	0.5355	0.2878	1.861	0.062782	.
MOSHOOFD9	0.7812	0.3258	2.398	0.016491	*
MOSHOOFD10	−0.2083	0.5281	−0.394	0.693256	
PWAPART	4.415e−03	2.0333e−03	2.171	0.029914	*
PPERSAUT	4.720e−04	5.234e−05	9.018	〈2e−16	***
PC1	0.1546	4.536e−02	3.408	0.000655	***
PC3	0.1121	6.093e−02	1.840	0.065764	.
PC4	0.1708	7.319e−02	2.333	0.019645	*
PC6	−0.2040	8.772e−02	−2.325	0.020063	*

Ⅳ. 결론

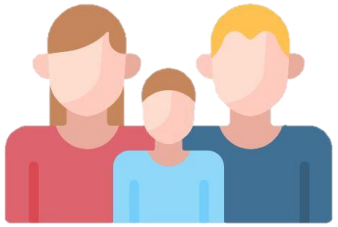
변수명	성공적인 행복을 중시하는 사람에 비교한 오즈 변화
MOSHOOFD2 (보육을 주도하는 사람)	2.27(배)
MOSHOOFD3 (일반적인 가족)	1.6(배)
MOSHOOFD4 (재택근무하는 직업을 가지는 사람)	7.37e−07(배)
MOSHOOFD5 (잘 사는 사람)	0.47(배)
MOSHOOFD6 (여생을 즐기는 노인)	0.72(배)
MOSHOOFD7 (은퇴했으며 신앙심이 깊은 사람)	1.95(배)
MOSHOOFD8 (성인이 된 자식들과 같이 사는 가족)	1.72(배)
MOSHOOFD9 (보수적인 가족)	2.18(배)
MOSHOOFD10 (농업인)	0.81(배)

〈최종 모형 계수의 추정 오즈〉

변수명	한 단위 증가당 오즈 변화
PWAPART (개인 제3자 보험료)	1.004(배)
PPERSAUT (자동차 보험료)	1.0005(배)
PC1 (사회적 위치가 상류층, 하류층과 관련된 변수)	1.16(배)
PC3 (노동자 및 사회계급이 낮은 집단에 대한 변수)	1.12(배)
PC4 (사회계급이 B1인 농부에 대한 변수)	1.19(배)
PC6 (사회계급이 B2인 농부에 대한 변수)	0.82(배)

- 고객 타입에 맞는 결합상품 개발

- 성인이 된 자녀들과 같이 사는 가정
- 보수적인 가족
- 일반적인 가족



가족 단위 고객

ex) 자녀들이 자전거, 서핑보드를 타며 다칠 수 있는 위험 존재

상해 보험 상품을 결합한 caravan 보험 상품을 개발

- 보험 가입의 의사결정자는 부모 세대(40세 – 60세)임을 고려

이메일같은 온라인매체를 통한 마케팅 전략보다는

전화, 방문 판매 등의 오프라인 위주 마케팅 전략 선택

Ⅳ. 결론

- 자동차 회사와 데이터 연계

자동차 보험료를 많이 내는 사람일수록 caravan보험 가입 확률 증가

최근 자동차 구매 고객이 보험 상담을 받을 때, 자동차 보험과 caravan 보험을 함께 추천

한계점과 보완점

Zipcode기반의 지역변수가 아닌, 고객 자체의 특성을 나타내는 변수가 있었다면 정교한 변수 처리가 가능

고객이 보유한 caravan의 종류, 크기, 가격 등의 데이터를 얻을 수 있다면 좀 더 정확한 예측 모형 구축 가능

보험을 가입하는 고객의 성향(위험회피적, 위험중립적, 위험애호적)을 파악할 수 있는 지표를 개발하여 예측변수로 활용한다면 정확도 향상 가능

PCA 과정에서 단순히 설명변수들만의 관계가 아닌 타겟변수와의 관계까지 같이 고려한 PLS를 사용할 경우, 좀 더 유의미한 변수로써의 차원축소가 가능

- 한계점과 보완점

다운샘플링으로 인해 데이터의 손실이 너무 크므로 SMOTE샘플링을 통해 데이터의 손실을 줄여 모형의 성능 향상 기대

잠재 고객에 대한 마케팅 비용을 알 수 있다면 보험에 가입할 것이라고 예측한 사람 중 실제로 보험에 가입한 사람의 비율(Negative predictive value)과 실제 보험에 가입한 사람을 맞춘 비율(Specificity)을 조절해가며 최적의 Cut Off을 찾을 수 있다.

감사합니다