데이터 마이닝

The Insurance Company Data 기반,
고객의 CARAVAN 보험 구매 여부 예측에 관한 연구

3조

통계학과 2018****** 이윤정

불교학부 2017****** OOO

통계학과 2018****** OOO

통계학과 2019****** OOO

분석 목적

분석 목적

네덜란드는 캠핑을 즐기는 인구가 전체 인구의 2/3에 달하고, 일반적인 여행에서도 텐트나 CARAVAN을 이용하는 인구가 75%를 넘는 국가이기 때문에 CARAVAN 시장이 매우 활성화 되어있다.

고객의 사회학적 특징과 제품 소유권에 대한 속성을 분석하고, 어떤 특성이 CARAVAN 보험 구매에 영향을 주는지 파악하는 것을 목적으로 한다. 최종적으로 해당 특성을 고려하여 고객에 따른 보험 구매 여부를 예측한다.

데이터 소개

모든 설명변수는 범주형 변수이며,
86개의 변수 = 85개의 설명변수 + 1개의 반응변수로 이루어져 있다.

설명 변수1

인류사회 통계데이터

동일한 우편번호를 가진
모든 고객은 동일한
인구사회학적 속성을
지닌다.

반응변수

CARAVAN

CARAVAN 보험을
구매했는지 여부를
판단한다.

설명 변수2

제품 소유권 데이터

현재 고객이
소유하고 있는
보험 정보를 알려준다.

변수명	변수 설명	비고
MOSTYPE	고객 하위 유형	L0
MOSHOOFD	고객 기본 유형	L2
MAANTUHI	집 개수	1~10
MGEMOMV	가족 구성원 수	1~6
MGEMLEEF	가족 구성원 평균 나이	L1
…	…	…
MOPLHOOG	고등교육수준	L3
…	…	…

접두사 M

설명변수 1

인구사회 통계데이터

주어진 데이터셋에서 행 1부터 43까지의 변수는 인구사회 통계데이터를 의미한다.

이는 동일한 우편변호를 지닌 모든 고객은 동일한 인구사회학적 속성을 지닌다는 특성을 지닌다.

고객 기본 및 하위 유형부터 수료한 교육수준, 사회계급, 자동차 대수, 수입 등 다양한 인구사회학적 속성을 보인다.

변수명	변수 설명	비고
PPERSAUT	자동차 보험료	L4
PBRAND	화재 보험료	L4
PZEILPL	서핑보드 보험료	L4
…	…	…
APERSAUT	자동차 보험 개수	1~12
ABRAND	화재 보험 개수	1~12
AZEILPL	서핑보드 보험 개수	1~12
…	…	…

접두사 P # 접두사 A

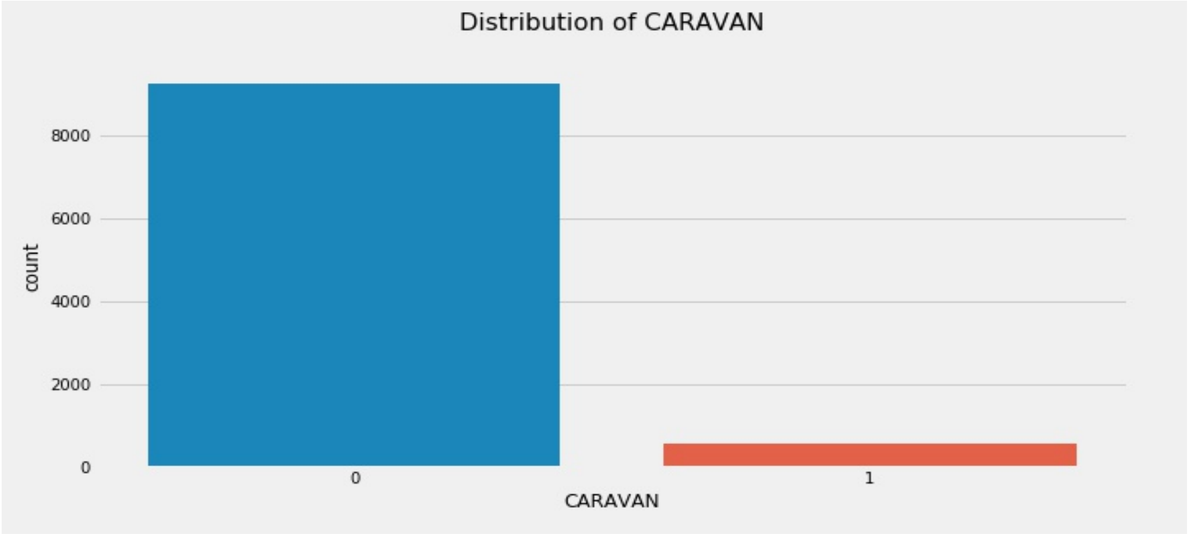
설명변수 2

제품 소유권데이터

주어진 데이터셋에서 행 44부터 85까지의 변수는 제품 소유권데이터를 의미한다.
이는 고객이 소유하고 있는 보험 정보를 알려준다.

P로 시작하는 변수는 고객이 지불하는 특정한 보험료, A로 시작하는 변수는 고객이 가진 특정한 보험 수를 의미한다.

자동차, 오토바이, 보트 등의 다양한 운송수단 보험부터 서드파티, 건강, 화재, 장애 보험 등 다양한 종류의 보험과 관련된 변수이다.



CARAVAN

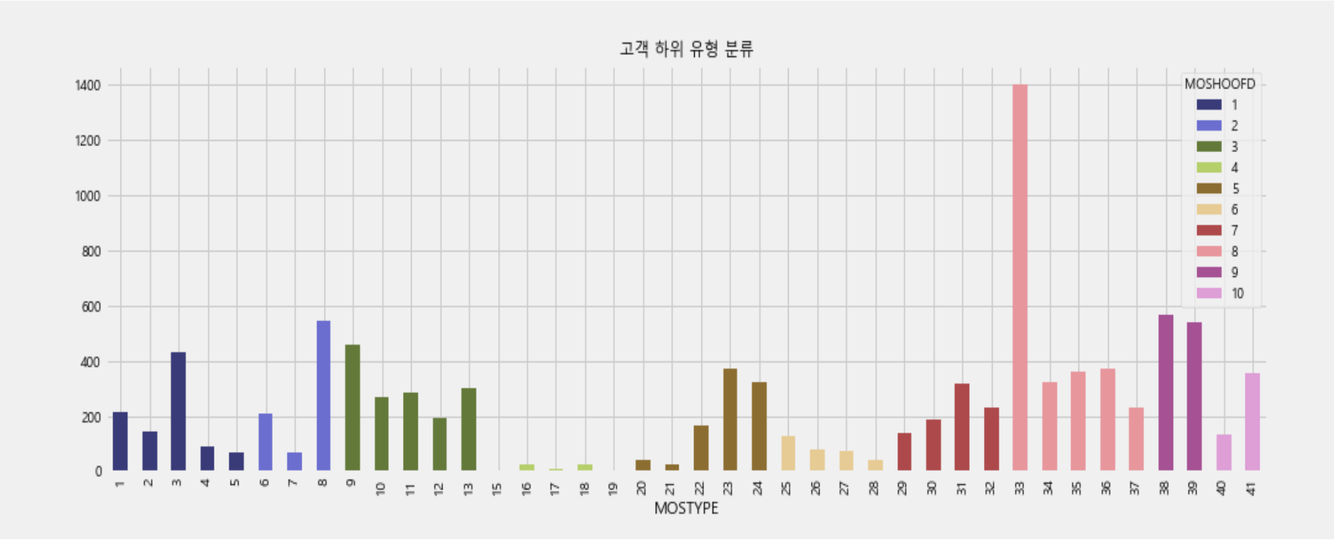
탐색적 자료분석

CARAVAN 변수

01 CARAVAN의 경우 0과 1에 대한 비율이 약 15.76 : 1로 매우 불균형한 분포를 보인다.

02 데이터가 불균형한 경우 과적합 문제가 발생할 수 있으므로,

가중치 혹은 샘플링 기법을 사용해야한다.

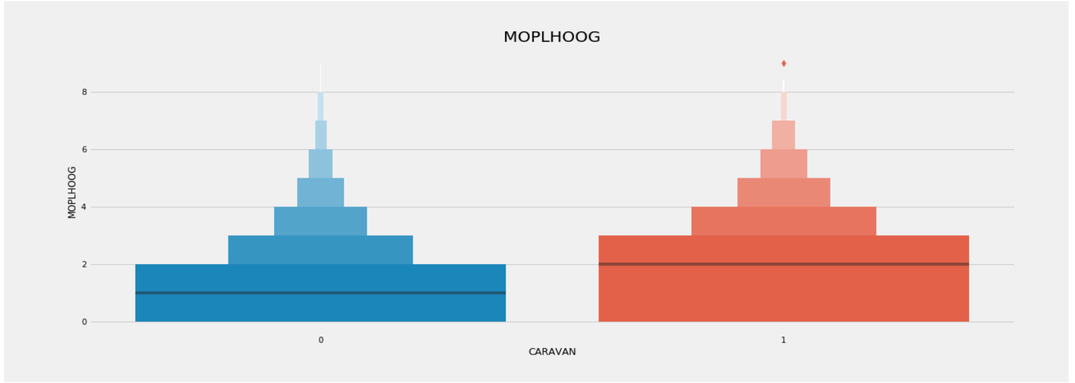


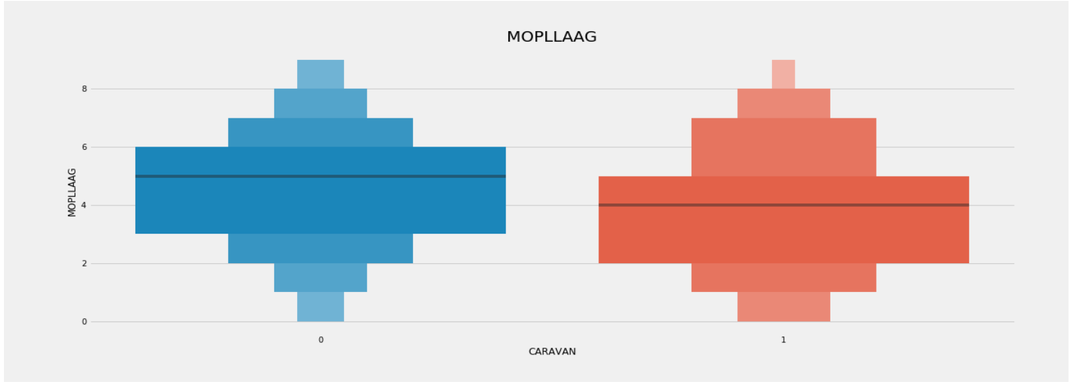
MOSTYPE #MOSHOOFD

탐색적 자료분석

고객 유형에 대한 변수

01 고객의 기본 유형을 하위 유형에 맵핑 하였을 때, 각 기본 유형은 정확히 하위 유형에 맵핑 되었다. 이를 통해, 두 유형은 정확히 대분류(MOSHOOFD)와 소분류(MOSTYPE) 관계를 갖는 것을 알 수 있다.





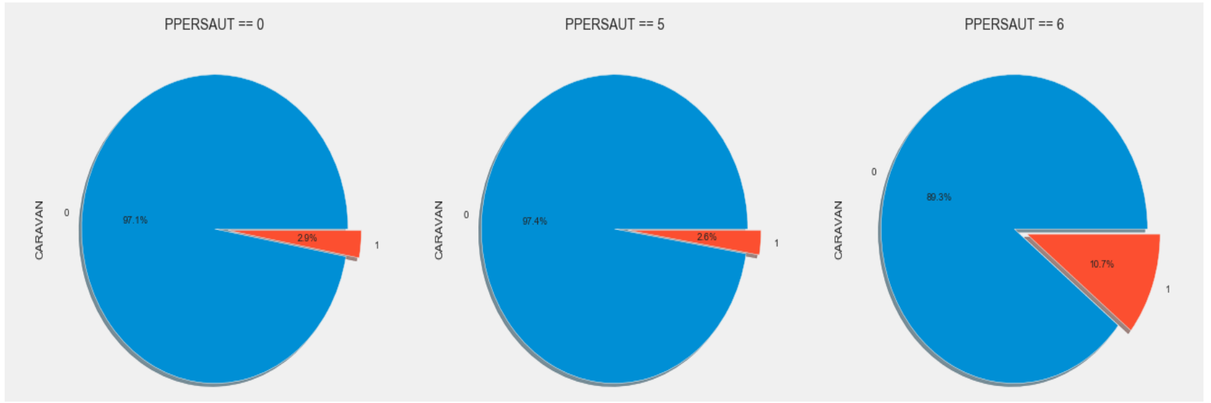
MOPLHOOG # MOPLLAAG

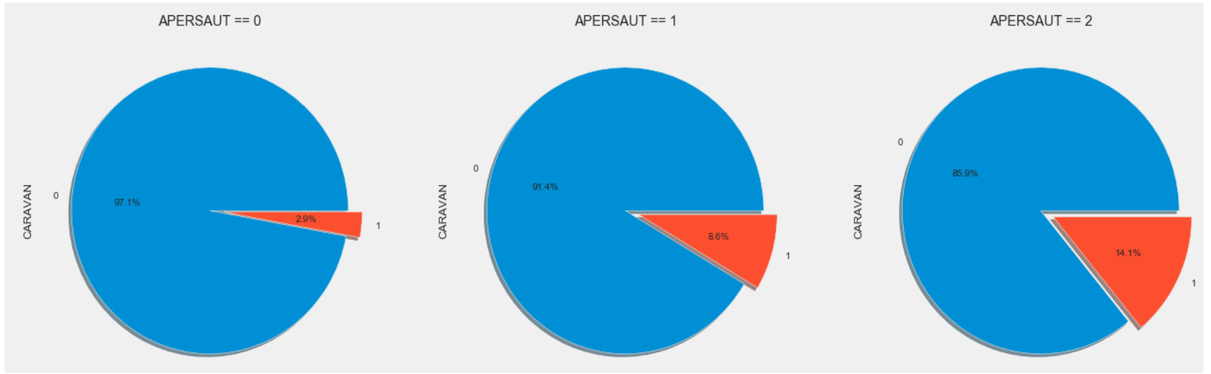
탐색적 자료분석

교육 수준에 대한 변수

01 고등교육을 수료한 고객의 비율이 0~36%인 지역에 거주하는 고객이 CARAVAN 보험을 가장 많이 구매하였다.

02 초등교육을 수료한 고객의 비율이 11~62%인 지역에 거주하는 고객이 CARAVAN 보험을 가장 많이 구매하였다.





PPERSAUT # APERSAUT

탐색적 자료분석

자동차 보험에 대한 변수

01 자동차 보험료가 높을수록 비교적 CARAVAN 보험을 많이 구매하였다.

02 자동차 보험을 많이 소유할수록 비교적 CARAVAN 보험을 많이 구매하였다.

데이터 생성

데이터셋 구축 # 언더샘플링 # 가중방법

01

CoIL 2000 Challenge 데이터로 대회 주최 측에 의한 사전 처리로 인하여 outlier 및 missing value는 존재하지 않는다. 주어진 파일 병합 과정을 통해 총 86 x 9822 (Col x Row)의 최종 데이터 세트를 만든다.

02

언더샘플링이란 불균형한 데이터에서 높은 비율을 차지하는 클래스의 데이터 수를 줄임으로써 데이터 불균형을 해소하는 방법이다. 언더샘플링을 통해 0의 개수를 비율이 낮은 1의 개수(586개)에 맞추어 반응변수의 비율을 50 : 50로 맞추어 과적합 문제를 방지하였다.

03

언더샘플링을 사용하게 된다면 기존 데이터의 클래스 비율 특성은 사용할 수 없다. 해당 문제를 해결하기 위하여 가중치 변수를 설정하였다. 기존 데이터에서 0이 1보다 약 15.76배 많기 때문에 0에 가중치 변수는 15.76의 값을 갖는다.

변수 선택

독립성 검정

변수명	카이스퀘어 검정통계량 (P-value)	교차표에서 기대빈도가 5보다 작은 셀 비율(%)	Fisher 의 정확검정 통계향 (P-value)	유의성
MOPLHOOG	<0.0001	<20	-	연관
PBESAUT	0.5935	25	0.7979	독립
PVRAAUT	0.3108	25	0.6201	독립
PWERKT	0.3108	25	0.6201	독립
PPERSONG	0.7692	33	0.8939	독립
PZEILPL	0.6620	33	0.4887	연관
AWABEDR	0.7406	25	0.6763	독립
ABESAUT	0.5935	25	0.7979	독립
AVRAAUT	0.3108	25	0.6201	독립
AWERKT	0.3108	25	0.6201	독립
APERSONG	0.5465	25	0.7664	독립
AZEILPL	0.5257	25	0.4667	연관

01

카이스퀘어 검정 통계량의 P-value가 0.5 이상이면 각 변수가 통계적으로 독립이므로 유의하지 않다고 판단하여 해당 변수 제거

귀무가설(H0) : 두 변수 사이에는 연관이 없다. (독립이다)

대립가설 (H1) : 두 변수 사이에는 연관이 있다. (종속이다)

02

교차표에서 기대 빈도가 5보다 작은 셀이 20% 이상일 때는 카이스퀘어 검정 결과를 신뢰할 수 없다.

→ Fisher의 정확 검정 방법을 사용하여 독립 여부를 판단

변수 선택

독립성 검정

변수명	카이스퀘어 검정통계량 (P-value)	교차표에서 기대빈도가 5보다 작은 셀 비율(%)	Fisher 의 정확검정 통계향 (P-value)	유의성
MOPLHOOG	<0.0001	<20	-	연관
PBESAUT	0.5935	25	0.7979	독립
PVRAAUT	0.3108	25	0.6201	독립
PWERKT	0.3108	25	0.6201	독립
PPERSONG	0.7692	33	0.8939	독립
PZEILPL	0.6620	33	0.4887	연관
AWABEDR	0.7406	25	0.6763	독립
ABESAUT	0.5935	25	0.7979	독립
AVRAAUT	0.3108	25	0.6201	독립
AWERKT	0.3108	25	0.6201	독립
APERSONG	0.5465	25	0.7664	독립
AZEILPL	0.5257	25	0.4667	연관

01

카이스퀘어 검정 통계량의 P-value가 0.5 이상이면 각 변수가 통계적으로 독립이므로 유의하지 않다고 판단하여 해당 변수 제거

귀무가설(H0) : 두 변수 사이에는 연관이 없다. (독립이다)

대립가설 (H1) : 두 변수 사이에는 연관이 있다. (종속이다)

02

교차표에서 기대 빈도가 5보다 작은 셀이 20% 이상일 때는 카이스퀘어 검정 결과를 신뢰할 수 없다.

→ Fisher의 정확 검정 방법을 사용하여 독립 여부를 판단

빈도
백분율
행 백분율
칼럼 백분율

변수 변환

지시변수 생성 # 분할표

테이블 CARAVAN * MGEMLEEF							
CARAVAN	MGEMLEEF						
	1	2	3	4	5	6	합계
0	112	2272	4912	1632	368	80	
	1.12	22.77	49.23	16.36	3.69	0.80	9376
	1.19	24.23	52.39	17.41	3.92	0.85	93.98
	99.12	93.57	93.92	93.96	94.85	98.77	
1	1	156	318	105	20	1	
	0.01	1.56	3.19	1.05	0.20	0.01	601
	0.17	25.96	52.91	17.47	3.33	0.17	6.02
	0.88	6.43	6.08	6.04	5.15	1.23	
합계	113	2428	5230	1737	388	81	9977
	1.13	24.34	52.42	17.41	3.89	0.81	100.00

01

로지스틱 회귀 모형은 수치형 변수만 다룬다.

→ 범주형의 경우 변수를 다시 조정해야 한다.

02

설명변수에 대해 분할표를 생성하고 구매한 비율이 서로 유의하게 다른 지에 따라 새로운 지시변수 생성한다.

→ 만약 CARAVAN 보험을 구매한 고객의 비율의 차이가 (0.3)내외이면 해당 클래스를 하나의 클래스로 합친 뒤 그 클래스에 대해 0 또는 1값을 갖는 새로운 지시변수 생성

빈도
백분율
행 백분율
칼럼 백분율

변수 변환

지시변수 생성 # 분할표

테이블 CARAVAN * MGEMLEEF							
CARAVAN	MGEMLEEF						
	1	2	3	4	5	6	합계
0	112	2272	4912	1632	368	80	
	1.12	22.77	49.23	16.36	3.69	0.80	9376
	1.19	24.23	52.39	17.41	3.92	0.85	93.98
	99.12	93.57	93.92	93.96	94.85	98.77	
1	1	156	318	105	20	1	
	0.01	1.56	3.19	1.05	0.20	0.01	601
	0.17	25.96	52.91	17.47	3.33	0.17	6.02
	0.88	6.43	6.08	6.04	5.15	1.23	
합계	113	2428	5230	1737	388	81	9977
	1.13	24.34	52.42	17.41	3.89	0.81	100.00

01

로지스틱 회귀 모형은 수치형 변수만 다룬다.

→ 범주형의 경우 변수를 다시 조정해야 한다.

02

설명변수에 대해 분할표를 생성하고 구매한 비율이 서로 유의하게 다른 지에 따라 새로운 지시변수 생성한다.

→ 만약 CARAVAN 보험을 구매한 고객의 비율의 차이가 (0.3)내외이면 해당 클래스를 하나의 클래스로 합친 뒤 그 클래스에 대해 0 또는 1값을 갖는 새로운 지시변수 생성

최종 후보 변수

독립성 검정 결과
지시변수 생성 결과

인구사회학적 변수	MGEMLEEF	MFALLEEN	MFWEKIND	MOPLHOOG
	MGODRK	MOPLMIDD	MBERHOOG	MBERZELF
	MGODPR	MBERBOER	MBERMIDD	MBERARBG
	MGODOV	MBERARBO	MSKA	MSKB1
	MGODGE	MSKB2	MSKC	MAUT1
	MRELGE	MAUT2	MAUT0	MINKM30
	MRELSA	MINK4575	MINK7512	MKOOPKLA
제품 소유권변수	PWABEDR, PTRACTOR, PZEILPL, PPLEZIER, ALEVEN			

01

독립성 검정 결과로 선택한 76개의 설명변수에 변수 조정을 반복 실행하여 총 33개의 변수를 다시 조정하였다.

→ 28개의 인구 사회학적 변수와 5개의 제품소유권 관련 변수

02

33개의 변수에 지시변수를 생성한 결과 최종 모형을 위한 후보 변수군에 속해 있는 지시 변수들을 포함한 총 278개의 설명변수가 생성되었다.

로지스틱 회귀 모형 # 최종 변수 선택

01 후진제거법

모든 설명변수를 포함한 모형을 생성한 후, F값을 기준으로 유의하지 않은 변수들을 하나씩 제거하며 유의한 변수만 채택하는 방법이다.
변수가 278개의 고차원이므로 유의수준 알파를 0.0001로 설정하였다.
그 결과 반응변수와 매우 유의한 11개의 설명변수가 채택되었다.

$$model\ 1:$$
$$logit[P(Y=1)] = \alpha_0 + \alpha_1 * MOPLLAAG_i + \alpha_2 * MHKOOP_i + \alpha_3 * PPERASUT_i + \alpha_4 * AWAOREG_i + \alpha_5 * AZEILPL_i + \alpha_6 * AFIETS_i + \alpha_7 * MFWEKIND8_i + \alpha_8 * MOPLHOOG2_i + \alpha_9 * MSKB13_I + \alpha_{10} * MSKC1_i + \alpha_{11} * MINKM308_i + \epsilon_i$$

여기서, MFWEKIND 8i= MFWEKIND 변수 중 level 8에 해당하는 지시 변수,
MOPLHOOG 2i = MOPLHOOG 변수 중 level 2에 해당하는 지시 변수,
MSKB1 3i = MSKB1 변수 중 level 3에 해당하는 지시 변수,
MSKC 1i= MSKC 변수 중 level 1에 해당하는 지시 변수,
MINKM 30i= MINKM30 변수 중 level 8에 해당하는 지시 변수, 오차항

02 전진 선택법

전진 선택법은 고려된 설명변수 중 설명력이 가장 높고 설명력이 유의하면 변수를 선택하는 방법이다.
그 결과 반응변수과 유의한 13개의 변수가 선택되었다.

$$model\ 2:$$
$$logit[P(Y=1)] = \alpha_0 + \alpha_1 * PPERSAUT_i + \alpha_2 * PBRAND_i + \alpha_3 * PFIETS_i + \alpha_4 * AWAOREG_i + \alpha_5 * AZEILPL_i + \alpha_6 * MFWEKIND8_i + \alpha_7 * MOPLHOOG2_i + \alpha_8 * MOPLHOOG8_i + \alpha_9 * MSKB13_i + \alpha_{10} * MSKB17_i + \alpha_{11} * MSKB19_i + \alpha_{12} * PPLEZIER4_i + \alpha_{13} * MINKM302_i + \epsilon_i$$

여기서, MOPLHOOG 2i = MOPLHOOG 변수 중 level 2에 해당하는 지시 변수,
MOPLHOOG 8i= MOPLHOOG 변수 중 level 8에 해당하는 지시 변수,
MSKB1 3i= MSKB1 변수 중 level 3에 해당하는 지시 변수,
MSKB1 7i= MSKB1 변수 중 level 7에 해당하는 지시 변수,
MSKB1 9i= MSKB1 변수 중 level 9에 해당하는 지시 변수,
PPLEZIER 4i= PPLEZIER 변수 중 level 4에 해당하는 지시 변수,
MINKM30 2i= MINKM30 변수 중 level 2에 해당하는 지시 변수, 오차항

로지스틱 회귀 모형 # 최종 변수 선택

03 Score

Score 점수가 가장 높은 최적의 설명변수 조합으로 모형을 적합 시키기 위하여 Score함수를 이용하여 각 설명변수에 대한 최적의 조합을 2개씩 확인하였다.

후진제거법과 전진선택법 중 한번이라도 나온 변수 모두 사용

⟶ 후진제거법에서 남은 11개의 변수와 전진선택법에서 선택된 13개의 변수에서 중복 제외 총 18개 사용

사용 변수	Score 점수
17	723.3377
17	717.3686
18	724.7781

$$
\begin{aligned}
&model\ 3: \\
&logit[P(Y=1)] = \alpha_0 + \alpha_1 MOPLHOOG8_i + \alpha_2 PPERSAUT_i + \alpha_3 PPLEZIER4_i \\
&\quad + \alpha_4 MOPLHOOG2_i + \alpha_5 AWAOREG_i + \alpha_6 AZEILPL_i + \alpha_7 MSKB17_i \\
&\quad + \alpha_8 MFWEKIND8_i + \alpha_9 MSKB19_i + \alpha_{10} PBRAND_i + \alpha_{11} MSKB13_i \\
&\quad + \alpha_{12} MINKM302_i + \alpha_{13} MOPLLAAG_i + \alpha_{14} MHKOOP_i \\
&\quad + \alpha_{15} AFIETS_i + \alpha_{16} MSKC1_i + \alpha_{17} MINKM308_i + \epsilon_i
\end{aligned}
$$

$$
\begin{aligned}
&model\ 4: \\
&logit[P(Y=1)] = \alpha_0 + \alpha_1 MOPLHOOG8_i + \alpha_2 PPERSAUT_i + \alpha_3 PPLEZIER4_i \\
&\quad + \alpha_4 MOPLHOOG2_i + \alpha_5 AWAOREG_i + \alpha_6 AZEILPL_i + \alpha_7 MSKB17_i \\
&\quad + \alpha_8 MFWEKIND8_i + \alpha_9 MSKB19_i + \alpha_{10} PBRAND_i + \alpha_{11} MSKB13_i \\
&\quad + \alpha_{12} MINKM302_i + \alpha_{13} MOPLLAAG_i + \alpha_{14} MHKOOP_i \\
&\quad + \alpha_{15} AFIETS_i + \alpha_{16} MSKC1_i + \alpha_{17} MINKM308_i + \alpha_{18} PFIETS + \epsilon_i
\end{aligned}
$$

다중 공선성 # 다중 공선성 고려

다중 공선성 이란?

독립 변수들 간에 높은 선형관계가 존재하여 강한 상관관계를 지니는 문제

본 연구에서 변수선택법을 통해 유의미하게 채택된 설명변수들은 약 20여개로 많은 변수들이 모델에 사용되었다.

다중 공선성은 회귀계수 추정오차에 영향을 주지만 모형의 정확, 예측에는 영향을 주지 않는다. 오히려 예측력을 높이는 효과가 존재한다. 따라서, 본 연구의 목적은 CARAVAN 보험 구매여부 예측이므로 다중 공선성을 고려하지 않았다.

로지스틱 회귀 모형 # 결과물 및 해석

01 최종 모형

$$
\begin{aligned}
logit[\hat{P}(Y=1)] = & -3.4395 + 20.1194 * MOPLHOOG8_i + 0.221 * PPERSAUT_i \\
& + 20.0265 * PPLEZIER4_i - 0.6513 * MOPLHOOG2_i + 3.1362 * AWAOREG_i \\
& + 0.7522 * AZEILPL_i + 2.396 * MSKB17_i + 1.0728 * MFWEKIND8_i \\
& + 23.4767 * MSKB19_i + 3.6498 * PBRAND_i - 0.8652 * MSKB13_i \\
& + 0.5306 * MINKM302_i - 0.125 * MOPLLAAG_i + 0.0623 * MHKOOP_i \\
& + 2.1453 * AFIETS_i - 0.956 * MSKC1_i + 1.3214 * MINKM308_i
\end{aligned}
$$

여기서, MOPLHOOG 8i = MOPLHOOG 변수 중 level 8에 해당하는 지시 변수,
PPLEZIER 4i= PPLEZIER 변수 중 level 4에 해당하는 지시 변수,
MOPLHOOG 2i = MOPLHOOG 변수 중 level 2에 해당하는 지시 변수,
MSKB1 7i = MSKB1 변수 중 level 7에 해당하는 지시 변수,
MFWEKIND 8i = MFWEKIND 변수 중 level 8에 해당하는 지시 변수,
MSKB1 9i = MSKB1 변수 중 level 9에 해당하는 지시 변수,
MSKB1 3i = MSKB1 변수 중 level 3에 해당하는 지시 변수,
MINKM30 2i = MINKM30 변수 중 level 2에 해당하는 지시 변수,
MSKC 1i = MSKC 변수 중 level 1에 해당하는 지시 변수,
MINK30 8i= MINKM30 변수 중 level 8에 해당하는 지시 변수

로지스틱 회귀 모형 # 결과물 및 해석

02 해석

증가

고등 교육 수료 비율 고객이 89~99%인 지역 ⟷ 초등 교육수료 고객 비율이 많은 지역

사회 계급 B1 고객 비율이 100%인 지역 ⟷ 사회 계급 B1 고객 비율이 24~36% 인 지역

감소

고등 교육 수료 고객 비율이 적은 지역

사회 계급 C 고객 비율이 1~10% 인 지역

Dictionary	Effect	Point Estimate
*고등교육수료 고객 비율 89 - 99%	MOPLHOOG8	>999.999
*보트 보험료 200-499 내는 고객	PPLEZIER4	>999.999
*사회계급 B1 고객 비율 100%	MSKB19	>999.999
고객이 지불하는 화재 보험료	PBRAND	38.467
고객이 가진 장애 보험 수	AWAOREG	23.015
*사회계급 B1 고객 비율 76-88%	MSKB17	10.98
고객이 가진 자전거 보험 수	AFIETS	8.545
*수익 30,000 미만 고객 비율 89-99%	MINKM308	3.749
*아이가 있는 가정의 고객 비율 89-99%	MFWEKIND8	2.923
고객이 가진 서핑 보험 수	AZEILPL	2.122
*수익 30,000 미만 고객 비율 11-23%	MINKM302	1.7
고객이 지불하는 자동차 보험료	PPERSAUT	1.247
집을 소유하고 있는 고객의 비율	MHKOOP	1.064

Dictionary	Effect	Point Estimate
초등교육수료 고객 비율	MOPLLAAG	0.883
*고등교육수료 고객 비율 11-23%	MOPLHOOG2	0.521
*사회계급 B1 고객 비율 24-36%	MSKB13	0.421
*사회계급 C 고객 비율 1-10%	MSKC1	0.384

로지스틱 회귀 모형 # 결과물 및 해석

02 해석

증가

고등 교육 수료 비율 고객이 89~99%인 지역 ⟷ 초등 교육수료 고객 비율이 많은 지역

사회 계급 B1 고객 비율이 100%인 지역 ⟷ 사회 계급 B1 고객 비율이 24~36% 인 지역

감소

초등 교육수료 고객 비율이 많은 지역

고등 교육 수료 고객 비율이 적은 지역

사회 계급 B1 고객 비율이 24~36% 인 지역

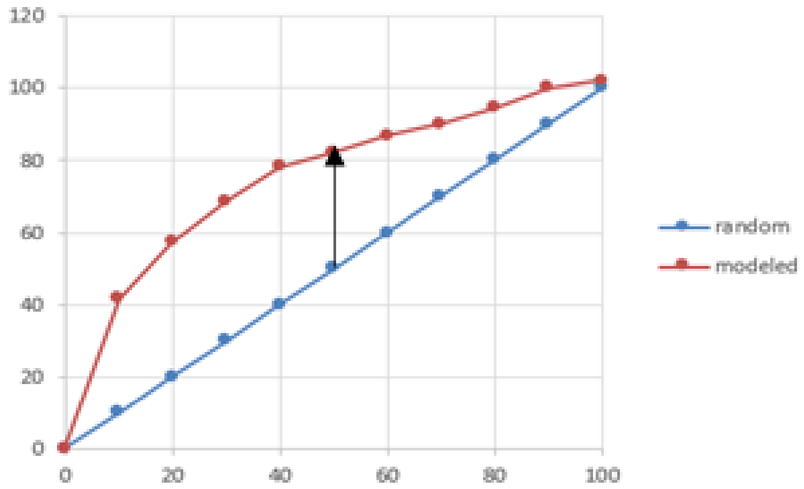
사회 계급 C 고객 비율이 1~10% 인 지역

Dictionary	Effect	Point Estimate
*고등교육수료 고객 비율 89 - 99%	MOPLHOOG8	>999.999
*보트 보험료 200-499 내는 고객	PPLEZIER4	>999.999
*사회계급 B1 고객 비율 100%	MSKB19	>999.999
고객이 지불하는 화재 보험료	PBRAND	38.467
고객이 가진 장애 보험 수	AWAOREG	23.015
*사회계급 B1 고객 비율 76-88%	MSKB17	10.98
고객이 가진 자전거 보험 수	AFIETS	8.545
*수익 30,000 미만 고객 비율 89-99%	MINKM308	3.749
*아이가 있는 가정의 고객 비율 89-99%	MFWEKIND8	2.923
고객이 가진 서핑 보험 수	AZEILPL	2.122
*수익 30,000 미만 고객 비율 11-23%	MINKM302	1.7
고객이 지불하는 자동차 보험료	PPERSAUT	1.247
집을 소유하고 있는 고객의 비율	MHKOOP	1.064

Dictionary	Effect	Point Estimate
초등교육수료 고객 비율	MOPLLAAG	0.883
*고등교육수료 고객 비율 11-23%	MOPLHOOG2	0.521
*사회계급 B1 고객 비율 24-36%	MSKB13	0.421
*사회계급 C 고객 비율 1-10%	MSKC1	0.384

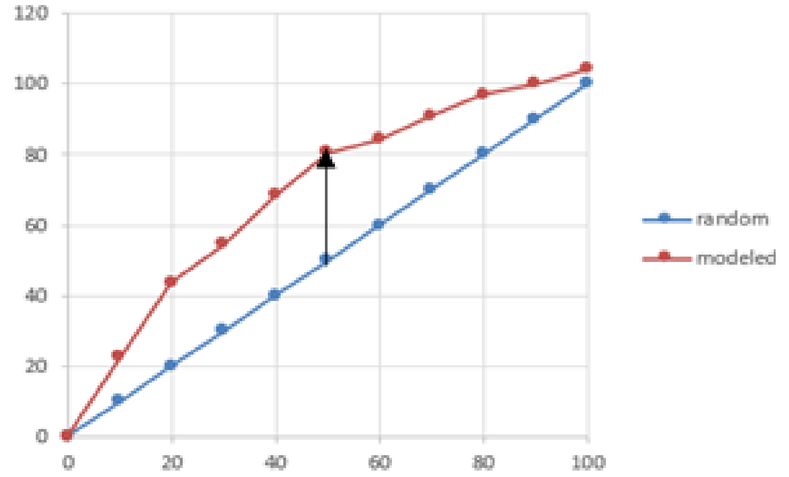
모형 평가 # 이익 도표

01 Train 데이터 십분위



가장 활성화 비율이 높은 십분위(0)의 활성화 비율(0.24416)은 가장 낮은 십분위 (9)의 활성화 비율(0.01281)보다 약 19배 크다,
모형 구축을 하지 않았을 때 보다 실제로 CARAVAN 보험을 가입한 고객을 약 4.14배 더 포함하고 있다.

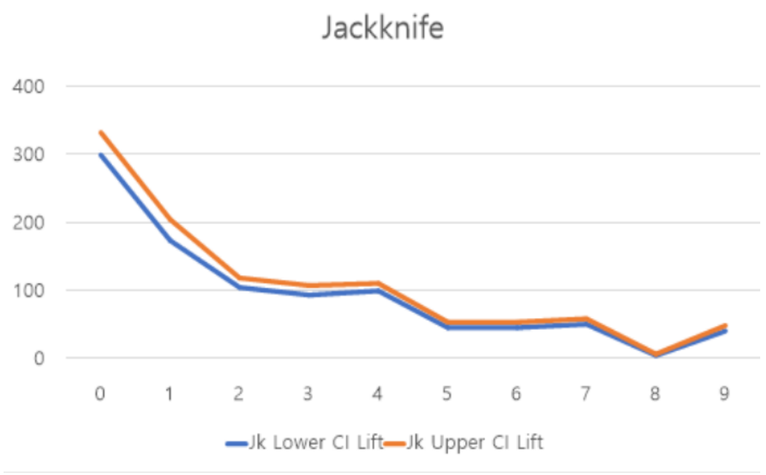
02 Test 데이터 십분위



가장 활성화 비율이 높은 십분위(0)의 활성화 비율(0.52577)은 가장 낮은 십분위(9)의 활성화 비율(0.00747)보다 약 70배 크다,
모형 구축을 하지 않았을 때 보다 실제로 CARAVAN 보험을 가입한 고객을 약 2.21배 더 포함하고 있다.

모형 평가 # 재표본

01 잭나이프



Decile	JK Est Lift	JK Lower CI Lift	JK Upper CI Lift
0	315	299	332
1	188	173	204
2	111	104	118
3	100	93	107
4	105	99	110
5	49	45	53
6	49	45	53
7	54	50	58
8	5	4	6
9	44	40	48
Total	102	95	109

잭나이프 추정값이 역전되는 경우가 두 번 발생하지만 전반적으로 단조 감소하는 형태이다.

신뢰구간 하한선와 상한선으로 구한 신뢰구간의 범위가 40이내로 비교적 작은 범위를 유지한다는 점을 통해 본 모형이 로버스트하다는 것을 알 수 있다.

02 부스트랩

Decile	Lift	BS Est Lift	BS Lower CI Lift	BS Upper CI Lift
0	251	225	64	386
1	194	186	81	291
2	152	142	54	230
3	81	82	37	128
4	75	73	15	132
5	70	70	21	120
6	64	64	5	121
7	75	75	-2	125
8	6	6	-5	19
9	39	39	5	64
Total	101	101	27	161

부스트랩 이익도표에서 모든 부스트랩 추정값(BS est Lift)이 구축된 모형에서 구한 리프트값(Lift)과의 차이가 작은 편이기 때문에 구축된 모형이 안정적이고 로버스트하다고 볼 수 있다.

추가 연구 # 변수 중요도

Variable	Random Forest	Support Vector	Ridge Regression	Gradient Boosting	LGBM
AZEILPL	O	X	O	O	O
PPERSAUT	O	X	X	O	O
MINK30	O	X	X	O	O
MINKGEM	O	X	X	O	O
MBERBOER1	O	X	O	O	X
MRELGE9	X	O	O	O	X

Logistic Regression 모형의 경우 로그 변환을 통해 계수가 선형적인 성질을 갖도록 한 일반화 선형 모형이다.
따라서, 설명변수와 반응변수 간의 관계가 비선형의 경우를 고려하기 위하여 본 연구에서 사용한
최종 Logistic Regression 모형과 다른 모형 간의 비교평가 및 유의한 변수를 도출하였다.

CARAVAN 보험을 구매한 고객의 특성

대다수의 특성이 고객이 소유한 재산 수준과 관련되어 있다는 것을 알 수 있다.

여가 생활을 즐기며, 다양한 종류의 보험을 가입하면서 혹시 모를 위험에 대비하는 모습을 발견할 수 있다.

마케팅 전략

결합상품 # 기존 상품 보완 # 프로모션

01

높은 수준의 보트 보험료와 자동차 보험료를 지불할 수 있는 고객이 CARAVAN보험에 가입한 경우가 많았다. 따라서, CARAVAN 보험과 보트 및 자동차 보험을 결합하여 새로운 결합 형태의 보험 제공한다.

02

아이가 있는 가족과 CARAVAN 보험 간의 연관성을 고려하여 아이 관련 특약을 추가하여 아이가 있는 가정의 보험 가입을 유도한다.

03

소득 수준 및 교육 수준이 높은 지역에서 CARAVAN 구매율이 높다는 점을 고려하여 해당 지역 대상 프로모션을 진행하여 CARAVAN 보험에 관심있던 고객 뿐만 아니라 그렇지 않은 고객을 유입한다.

한계점

샘플링 # 리스크 모형 # 고객 개인별 특성

01

본 연구에서 사용한 언더샘플링 기법 이외의 오버샘플링 및 SMOTE 샘플링과 같은 다양한 샘플링 기법을 사용하고 결과를 비교한다면 예측력이 높은 모형을 구축할 수 있다.

02

보험료 미지급이나 연체와 같은 지불과 관련한 데이터가 추가된다면, 위험이 큰 고객을 피할 수 있는 리스크 모형을 만들어서 손실을 줄일 수 있을 것으로 기대할 수 있을 것이다.

03

주요 변수인 인구사회적 변수는 우편 번호를 기반으로 한 지역 단위 변수이므로 개인의 특성에 대해서는 파악하기 어렵다. 집단의 특성을 통하여 고객 개개인의 특성을 파악했기 때문에 고객 개인별 특성을 통 구축한 모형보다 설득력이 떨어진다.
개인정보의 비식별화를 통하여 개인별 특성을 가지는 변수를 제공받을 수 있다면 예측력과 설득력이 향상 될 수 있다.

감사합니다.