데이터마이닝 프로젝트

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

2017****** 불교학부 OOO

2018****** 통계학과 OOO

2018****** 통계학과 이윤정

2019****** 통계학과 OOO

목 차

제1장 서 론

네덜란드는 캠핑을 즐기는 인구가 전체 인구의 2/3에 달하고, 일반적인 여행에서도 텐트나 CARAVAN을 이용하는 인구가 75%를 넘는 국가이기 때문에 CARAVAN 시장이 매우 활성화되어있다. 이러한 특성이 반영되어 2000년에 한 네덜란드의 데이터 마이닝 회사가 제공한 'The Insurance Company Data'를 이용하여 CoIL 2000 Challenge 대회가 개최되었다. 대회의 목적은 CARAVAN 보험에 가입한 고객의 특성을 바탕으로 새로운 고객을 획득하는 것이다.

본 연구는 해당 대회 데이터를 이용하여 동일한 목적의 연구를 진행하고자 한다. 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.

데이터의 변수에 대해서 자세히 설명하자면, 86개의 변수는 85개의 설명변수와 1개의 반응변수로 이루어져 있으며 모든 설명변수는 범주형 변수이다. 이때, 인구 사회학적 변수(1~43)는 우편번호를 의미하며 동일한 우편번호를 가진 모든 고객은 동일한 인구사회 학적 속성을 갖는다. 다음으로, 제품 소유권 변수(44~85)는 현재 고객이 소유하고 있는 보험 정보를 알려준다. 예를 들어 P로 시작하는 변수는 고객이 지불하는 특정한 보험료, A로 시작하는 변수는 고객이 가진 특정한 보험 수를 의미한다.

마지막으로 86번 변수는 반응변수로서 CARAVAN 보험을 구매했는지 여부를 구별한다. 보험을 구매한 고객은 1의 값, 구매하지 않은 고객은 0의 값을 가지고 있는 범주형 변수이다. 이는 고차원의 데이터이므로 집단의 특성을 나타내지 않는 변수는 선택하지 않고 분석을 진행하고자 한다.

변수 설명	접두사	Column	비고
인구사회 통계데이터	M	1~43	동일한 우편번호를 지닌 모든 고객은 동일한 인구사회학적 속성을 지님
제품 소유권 데이터	P	44~64	퍼센트
	A	65~85	숫자
CARAVAN	-	86	반응변수

<표 1 - 1>

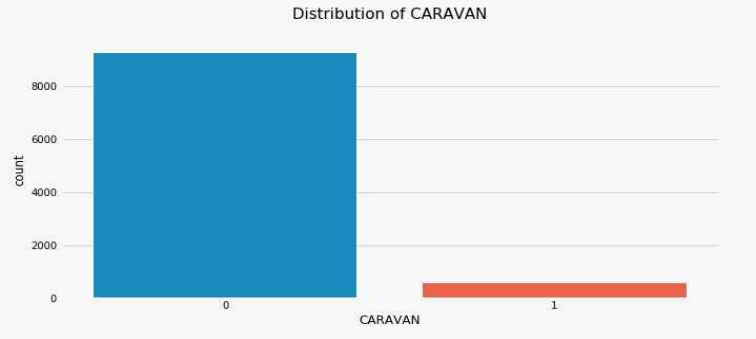
제2장 본 론

제1절 모형 구축을 위한 데이터 준비

가. 탐색적 자료분석 (EDA)

분석에 사용되는 데이터는 고차원의 데이터이므로 데이터를 직관적으로 이해하고 데이터가 표현하는 현상을 이해하고자 탐색적 자료분석을 시행한다. 본 연구의 종속변수인 CARAVAN의 분포를 확인한 후, 각 설명변수별 분포를 확인한다. 이후 각 설명변수의 Level에 따른 반응변수의 분포를 확인한다.

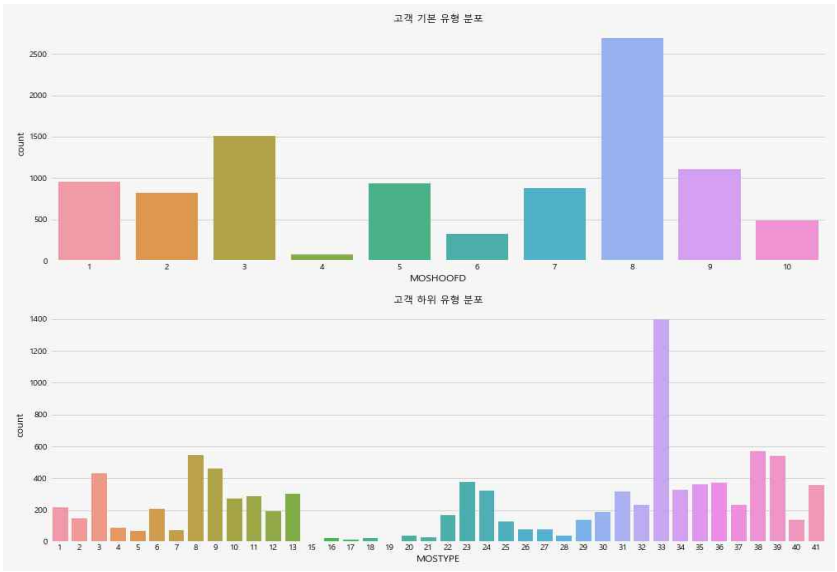
(1) CARAVAN 변수



[그림 2 - 1]

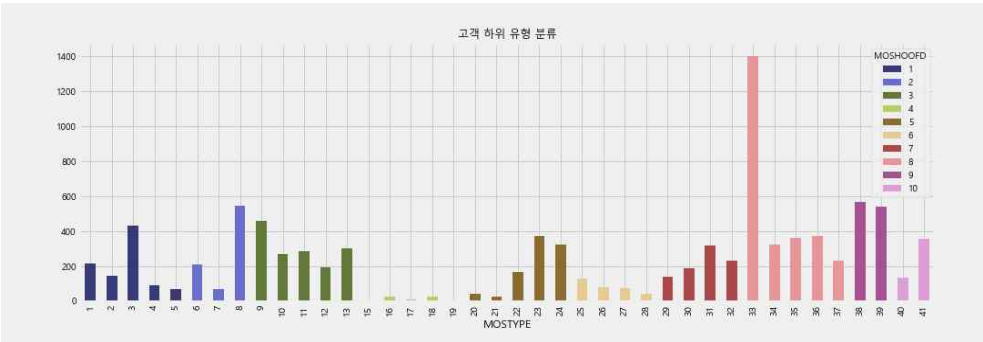
반응변수인 CARAVAN의 경우 0과 1에 대한 비율이 약 564 : 1로 매우 불균형한 분포를 보인다. 데이터가 불균형한 경우 모형 적합과정에서 분포도가 높은 클래스에 대해 가중치를 두기 때문에 예측할 때에도 가중치가 높은 클래스를 더 예측하게 된다. 따라서, 모형 정확도는 높아지지만 가중치가 작은 클래스의 재현율이 낮아지는 문제가 발생한다. 이러한 과적합 문제를 해결하기 위해서는 샘플링 기법을 통해 CARAVAN에 대해 0과 1의 비율을 5 : 5로 맞춰주는 과정이 필요하다. 해당 과정은 나. 모형 구축을 위한 데이터 생성에서 추가 서술한다.

(2) 고객 분류에 대한 변수 (MOSHOOFD, MOSTYPE)



[그림 2 - 2]

고객의 기본 유형 변수(MOSHOOFD)의 경우 유형 8 'Family with grown ups'이 가장 많다. 고객의 하위 유형 변수(MOSTYPE)의 경우 유형 33 'Lower class large families'이 가장 많고, 유형 14 'Junior cosmopolitan'는 존재하지 않는다.

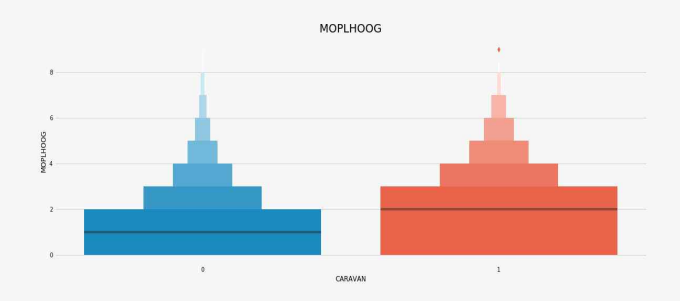


[그림 2 - 3]

고객의 기본 유형을 하위 유형과 맵핑하였을 때, 두 유형은 정확히 대분류(MOSHOOFD)와 소분류(MOSTYPE) 관계를 갖는 것을 알 수 있다. 더 자세한 사항은 부록의 <표 2 - 1>에서 확인할 수 있다.

(3) 교육 수준에 대한 변수

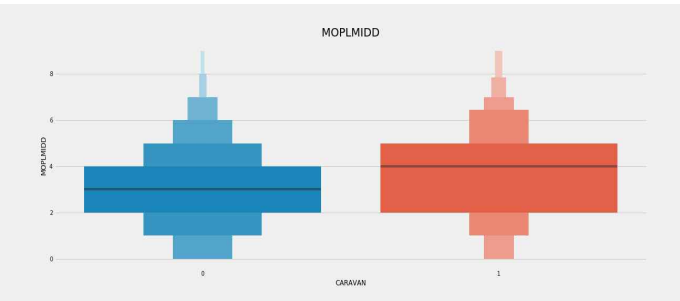
1. MOPLHOOG (High level Education)



[그림 2 - 4]

MOPLHOOG 변수는 High Level의 교육 수준을 의미하는 인구 사회적 변수이다. 부록의 [그림 2 - 5]을 보면, 변수의 분포가 왼쪽으로 치우친 형태를 보인다. 이는 데이터를 구성하는 집단에서 고등교육을 수료한 고객의 비율이 낮은 지역에 거주하는 고객이 많다는 것을 의미한다. CARAVAN 변수에 따른 MOPLHOOG 변수의 분포를 살펴본 결과, 고등교육을 수료한 고객의 비율이 0 ~ 36%인 지역에 거주하는 고객이 CARAVAN 보험을 가장 많이 구매하였다. 또한, 고등교육을 수료한 고객의 비율이 낮은 지역에 거주하는 고객이 높은 지역에 거주하는 고객에 비해 비교적 CARAVAN 보험을 구매한 경우가 많은 것을 알 수 있다.

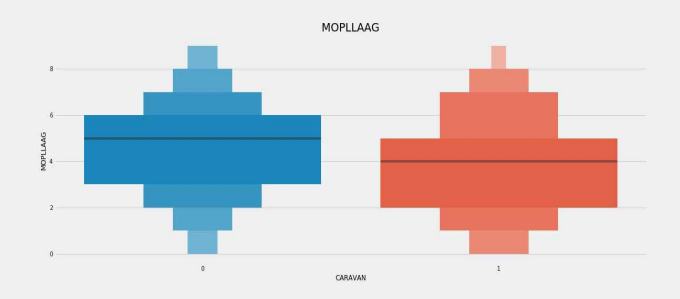
2. MOPLMIDD (Medium Level Education)



[그림 2 - 6]

MOPLMIDD 변수는 Middle Level의 교육 수준을 의미하는 인구 사회적 변수이다. 부록의 [그림 2 - 7]을 보면, 변수의 분포가 비교적 정규분포 형태를 보인다. CARAVAN 변수에 따른 MOPLMIDD 변수의 분포를 살펴본 결과, 중등교육을 수료한 고객의 비율이 11~62%인 지역에 거주하는 고객이 CARAVAN 보험을 가장 많이 구매하였다.

3. MOPLLAAG (Lower Level Education)

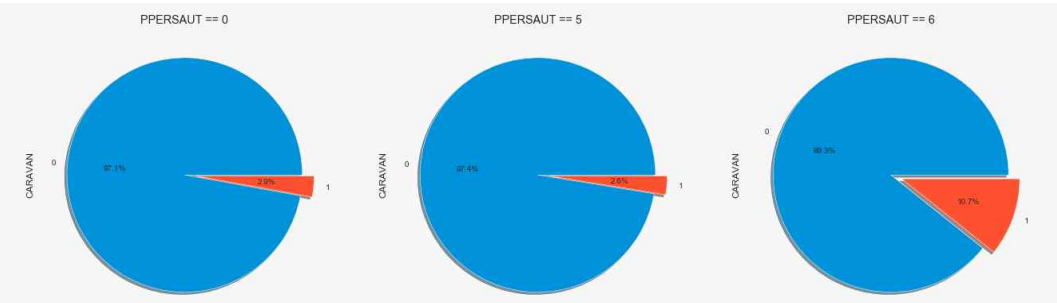


[그림 2 - 8]

MOPLLAAG 변수는 Lower Level의 교육 수준을 의미하는 인구 사회적 변수이다. 부록의 [그림 2 - 9]를 보면, 변수의 분포가 비교적 정규분포 형태를 보인다. CARAVAN 변수에 따른 MOPLLAAG 변수의 분포를 살펴본 결과, 초등교육을 수료한 고객의 비율이 11~62%인 지역에 거주하는 고객이 CARAVAN 보험을 가장 많이 구매하였다.

(4) 자동차 보험에 대한 변수

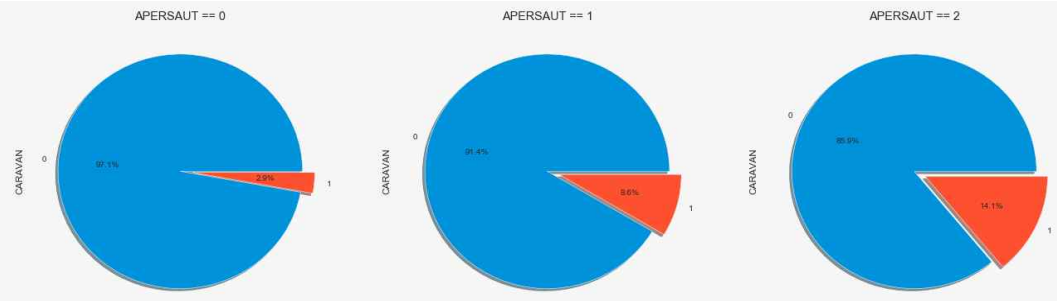
1. PPERSAUT (자동차 보험료)



[그림 2 - 10]

PPERSAUT 변수는 자동차 보험료를 의미하는 제품 소유권변수로, 대체로 보험료를 내지 않는 고객, 보험료를 500~999 내는 고객, 보험료를 1000~4999 내는 고객으로 나뉜다. CARAVAN 변수에 따른 PPERSAUT 변수의 분포를 살펴본 결과, 자동차 보험료가 높을수록 CARAVAN 보험을 많이 구매하였다.

2. APERSAUT (자동차 보험 개수)



[그림 2 - 11]

APERSAUT 변수는 자동차 보험 개수를 의미하는 제품 소유권변수이다. CARAVAN 변수에 따른 APERSAUT 변수의 분포를 살펴본 결과, 자동차 보험을 많이 소유할수록 CARAVAN 보험을 많이 구매하였다.

나. 모형 구축을 위한 데이터 생성

본 연구에서 사용한 데이터는 CoIL 2000 Challenge 데이터로 대회 주최 측에 의한 사전 처리로 인하여 outlier 및 missing value는 존재하지 않는다. 주어진 데이터 파일은 총 3개로 다음과 같으며, 예측 모형을 만들기 위해선 일련의 데이터 변환 과정이 필요하다. 우선, TICEVAL2000.txt와 TICTGTS2000.txt을 옆으로 합쳐 총 86 x 4000 (Col x Row)의 데이터 세트를 생성한다. 이렇게 생성된 데이터 세트를 TICDATA2000.txt와 위아래로 합쳐 총 86 x 9822 (Col x Row)의 최종 데이터 세트를 만든다.

파일명	파일 소개	변수 종류	obs
TICDATA2000.txt	예측 모형을 학습 및 검증하고 설명을 작성하기 위한 데이터 세트	attributes 변수	5822
TICTGTS2000.txt	평가용 데이터 세트 (target 변수)	target 변수	4000
TICEVAL2000.txt	예측용 데이터 세트	attributes, target 변수	4000

<표 2 - 2>

(1) 언더샘플링(Undersampling)

CARAVAN	빈도	백분율	누적빈도	누적 백분율
0	586	50.00	586	50.00
1	586	50.00	1172	100.00

<표 2 - 3>

탐색적 자료분석단계에서 반응 변수(CARAVAN Insurance Purchased)의 분포를 확인한 결과, CARAVAN 보험을 구매하지 않은 고객(0)과 구매한 고객(1)이 매우 불균형하게 분포하였다. 이때, 데이터가 불균형하면 모형은 분포가 높은 클래스에 가중치를 많이 두는 경향이 있다. 즉, 불균형 문제를 해결하지 않으면 모형은 가중치가 높은 클래스를 중심으로 예측하기 때문에 정확도(Accuracy)는 높아질 수 있지만, 분포가 작은 클래스에 대한 정밀도(Precision)은 낮아진다. 또한, 분포가 작은 클래스의 재현율(Recall)이 낮아지는 문제가 발생한다.

이러한 문제를 해결하기 위해 본 연구에서는 표본추출의 방법으로써 언더샘플링(Undersampling)을 사용했다. 언더샘플링이란 불균형한 데이터에서 높은 비율을 차지하는 클래스의 데이터 수를 줄임으로써 데이터 불균형을 해소하는 방법이다.[1] 본 연구의 반응변수는 0의 클래스가 높은 비율을 차지하므로 0의 비율을 1의 클래스의 비율과 같도록 조정하였다. 즉, 낮은 클래스인 1의 개수(586개)에 맞추어 0의 개수를 줄여, 반응변수의 비율을 50 : 50로 맞추는 언더샘플링을 사용한 것이다.

(2) 가중(weight) 방법

언더샘플링을 사용하게 된다면 불균형 데이터의 클래스 비율을 조절할 수 있으나 기존 데이터의 클래스 비율 특성은 사용할 수 없다. 해당 문제를 해결하기 위하여 가중치를 설정하기로 하였다. 이때 가중치는 샘플링된 데이터가 기존 모집단의 클래스 비율 특성을 반영할 수 있도록 한다. 기존 데이터에서 0이 1보다 약 15.76배 많기 때문에 0에 가중치 15.76을 설정하였고, 이를 가중치 변수 (smp_wgt = 15.76)를 생성하였다.

제2절 후보변수 선택 및 변환

가. 변수 선택

(1) 독립성 검정

본 연구의 설명변수는 85개로서 고차원의 데이터이므로 집단의 특성을 나타내지 않는 변수는 선택하지 않고 분석을 진행하고자 한다. 각 설명변수의 예측력을 확인한 후 반응변수와 연관성이 없다고 판단되는 변수를 선택하지 않기 위해 카이스퀘어 검정을 시행했다. 즉, 카이스퀘어 검정 통계량의 P-value가 0.5 이상이면 각 변수가 통계적으로 독립이므로 유의하지 않다고 판단하여 해당 변수를 제거한 것이다.

---

1) 파이썬 머신러닝 완벽 가이드(권철민, 2019)

귀무가설(H_0): 두 변수 사이에는 연관이 없다. (독립이다)

대립가설(H_1): 두 변수 사이에는 연관이 있다. (종속이다)

이때, 교차표에서 기대 빈도가 5보다 작은 셀이 20% 이상일 때는 카이스퀘어 검정 결과를 신뢰할 수 없기 때문에 Fisher의 정확 검정 방법을 사용하여 독립 여부를 판단하였다.

변수명	카이스퀘어 검정통계량 (P-value)	교차표에서 기대 빈도가 5보다 작은 셀 비율(%)	Fisher의 정확 검정 통계량 (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	보류
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	0.7684	독립
AZEILPL	0.5257	25	0.4667	보류

<표 2 - 4>

예를 들어, 고등 교육을 수료한 고객을 의미하는 MOPLHOOG 변수의 카이스퀘어 검정 통계량의 P-value는 0.0001보다 작은 매우 유의한 값이 나왔다. 즉, 유의수준 0.5 이하에서 MOPLHOOG 변수와 CARAVAN 변수는 연관이 있다고 할 수 있다.

고객이 지불하는 화물차 보험료를 의미하는 PVRAAUT 변수의 카이스퀘어 검정 통계량의 P-value은 0.3108로서 유의수준 0.5 이하에서 유의한 결과가 나왔다고 할 수 있다. 그러나 해당 변수의 교차표의 기대 빈도가 5보다 작은 셀이 25%이므로 Fisher의 정확 검정을 시행하였다. 그 결과 검정 통계량의 P-value는 0.6201로서 유의수준 0.5 이하에서 PVRAAUT 변수는 CARAVAN 변수와 연관이 없으므로 해당 변수는 선택하지 않는다.

고객이 지불하는 서핑보드 보험료를 의미하는 PZEILPL 변수의 Fisher의 정확 검정 결과 검정 통계량의 P-value는 0.4887로서 유의수준 0.5 경계선 근처이다. 이러한 결과가 나오는 변수는 자세한 분석을 위해 후보 변수로 선택하기로 판단했다.

모든 설명변수에 대해서 위와 같은 검정을 반복 실행한 결과 총 9개의 설명변수가 통계적으로 유의하지 않았다. 이러한 결과를 바탕으로 본 연구는 85개의 설명변수 중 9개의 변수를 후보 변수로 선택하지 않았다.

나. 변수 변환

(1) 지시변수 생성

빈도
백분율
행백분율
칼럼 백분율

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

<표 2 - 5>

본 연구에서 사용할 로지스틱 회귀모형에서는 수치형 변수만 다룬다. 따라서 범주형 변수의 경우에는 변수를 다시 조정해야 한다. 일반적으로 가장 많이 사용하는 방법인 지시변수를 생성하는 방법을 통해 범주형 변수를 조정하였다. 먼저, 고객의 주요 유형과 하위 유형을 설명하는 MOSHOOFD, MOSTYPE 변수를 제외한 설명변수에 대해 각 분할표를 생성했다. 이때 생성한 분할표를 통해 각 변수들의 클래스별 고객의 CARAVAN 보험을 구매한 비율이 서로 유의하게 다른지에 따라 새로운 지시변수를 생성하였다. 만약 CARAVAN 보험을 구매한 고객의 비율이 같거나 비슷한 경우 (차이가 약 0.3 내외), 해당 클래스를 하나의 클래스로 합친 뒤 그 클래스에 대해 0 또는 1의 값을 갖는 새로운 지시변수를 생성하였다.

예를 들어, 연령대를 설명해주는 MGEMLEEF 변수의 분할표를 생성한 결과 level 3과 level 4에 해당되는 고객이 CARAVAN 보험을 구매한 비율은 각각 6.08, 6.04로 그 차이가 0.3 이내이므로 비슷하다. 그러므로 두 level를 하나의 level로 조정한 뒤, 해당 변수의 나머지 level에 대한 지시변수를 생성하였다.

인구통계학적 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			

<표 2 - 6>

위와 같은 과정을 독립성 검정 결과로 선택한 76개의 설명변수에 반복 실행하여 총 33개의 변수를 다시 조정했다. 이때, 28개의 변수는 인구 사회적 변수이며 5개의 변수는 제품 소유

권과 관련된 변수이다. 다시 조정한 변수에 지시변수를 생성한 결과, 최종모형을 위한 후보 변수 군에 속해있는 지시변수들을 포함한 총 278개의 설명변수가 생성되었다.

제3절 모형 구축 과정

가. 데이터 분할

모든 데이터를 학습에 사용하는 경우 과적합이 발생하여 학습하지 않은 데이터에 대한 모형의 성능이 좋지 않을 수 있다. 이러한 과적합을 방지하기 위하여 전처리가 완료된 데이터 세트를 Train set, Test set으로 분할하여 학습과 모형 평가를 하고자 하였다. 이때 분할 비율은 통계 분석에서 일반적으로 많이 사용되는 7 : 3 비율로 설정하였다.

나. 로지스틱 회귀모형

(1) 최종 변수 선택

1. 후진 제거법

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

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

여기서, $MFWEKIND8_i$ = $MFWEKIND$ 변수 중 level 8에 해당하는 지시 변수,
$MOPLHOOG2_i$ = $MOPLHOOG$ 변수 중 level 2에 해당하는 지시 변수,
$MSKB13_i$ = $MSKB1$ 변수 중 level 3에 해당하는 지시 변수,
$MSKC1_i$ = $MSKC$ 변수 중 level 1에 해당하는 지시 변수,
$MINKM308_i$ = $MINKM30$ 변수 중 level 8에 해당하는 지시 변수, ϵ_i = 오차항

2. 전진 선택법

전진 선택법은 고려된 설명변수 중 설명력이 가장 높고 설명력이 유의하면 변수를 선택하는 방법이다. 이미 선택된 설명변수를 제외하고 남은 변수들을 비교하여 설명력이 유의한 경우를 선택한다. 후진 제거법과 마찬가지로 남은 변수 중 유의한 설명변수가 존재하지 않을 때까지 반복한다. 그 결과 반응변수과 유의한 13개의 변수가 선택되었다.

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

여기서, $MOPLHOOG2_i$ = MOPLHOOG 변수 중 level 2에 해당하는 지시 변수,
$MOPLHOOG8_i$ = MOPLHOOG 변수 중 level 8에 해당하는 지시 변수,
$MSKB13_i$ = MSKB1 변수 중 level 3에 해당하는 지시 변수,
$MSKB17_i$ = MSKB1 변수 중 level 7에 해당하는 지시 변수,
$MSKB19_i$ = MSKB1 변수 중 level 9에 해당하는 지시 변수,
$PPLEZIER4_i$ = PPLEZIER 변수 중 level 4에 해당하는 지시 변수,
$MINKM302_i$ = MINKM30 변수 중 level 2에 해당하는 지시 변수, ϵ_i = 오차항

3. Score

사용 변수	Score 점수
17	723.3377
17	717.3686
18	724.7781

<표 2 - 7>

최종 변수 선택의 효율성을 위하여 score 점수가 가장 높은 최적의 설명변수 조합으로 모형을 적합시키기 위해 Score 함수를 이용해서 각 설명변수에 대한 최적의 조합을 2개씩 확인하였다. 이때, 모형 적합 후 더 자세한 분석을 위하여 후진 제거법과 전진 선택법에 의한 변수 선택 결과에서 한 번이라도 나온 변수를 모두 사용하기로 하였다. 변수에 대해 자세히 설명하자면, 후진 제거법에서 남은 11개의 변수와 전진선택법에서 선택된 13개의 변수 중 6개의 변수가 일치하며 총 18개의 변수에 대한 최적의 조합을 확인하였다.

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

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

(2) 결과물 및 해석

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

<표 2 - 8>

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

<표 2 - 9>

<표 2 - 8>과 <표 2 - 9>는 최종 적합한 로지스틱 모형의 오즈비 추정 결과로써 반응변수에 대한 양의 상관관계와 음의 상관관계를 나누어 유의한 크기의 순서대로 나열한 것이다. 총 17개의 최종 적합변수 중 13개의 변수가 CARAVAN 보험 상품 구매를 승법적으로 증가시키는 영향력을 가지고 있고 4개의 변수가 CARAVAN 보험 상품 구매를 승법적으로 감소시키는 영향력을 가지고 있다.

먼저, <표 2 - 8>에서 CARAVAN 보험 상품 구매를 승법적으로 증가시키는 오즈비 추정값이 매우 크게 나온 변수 MOPLHOOG8, PPLEZIER4, MSKB19를 확인할 수 있다. 이는 모두 level을 다시 조정한 설명변수로써 각각 고등교육을 수료한 고객의 비율이 88~99%인 지역에 거주하는 고객, 보트 보험료를 200~499 내는 고객, 사회적 계급이 B1에 속하는 고객의 비율이 100%인 지역에 거주하는 고객을 의미한다. 이러한 특성을 가지는 고객이 CARAVAN 보험을 구매한 오즈 추정값은 그렇지 않은 고객의 오즈 추정값의 999배보다 더 높게 추정된다고 해석할 수 있다.

다음으로, <표 2 - 9>에서 CARAVAN 보험 상품 구매를 승법적으로 증가시키는 오즈비 추정값이 다른 변수의 오즈비 추정값과 비교해서 유의미하게 크게 나온 2개의 변수인 PBRAND, AWAOREG를 확인했다. 이는 모두 level을 조정하기 전의 설명변수로써 각각 고객이 지불하는 화재 보험료, 고객이 가진 장애 보험 수를 의미한다. 이는 주어진 최종 변수에서 고객이 지불하는 화재 보험료가 한 단위씩 증가할 때마다 CARAVAN 보험을 소유한 고객의 오즈는 $e^{3.6498}$ = 38.467배만큼 승법적으로 증가한다고 할 수 있다. 또한, 주어진 최종 변수에서 고객이 가진 장애 보험 수가 한 단위씩 증가할 때마다 CARAVAN 보험을 소유한 고객의 오즈는 $e^{3.6498}$ = 23.015배만큼 승법적으로 증가한다고 할 수 있다.

마지막으로, <표 2 - 9>에서 CARAVAN 보험 상품 구매를 승법적으로 감소시키는 오즈비 추정값이 매우 크게 나온 2개의 변수인 MOPLLAAG, MOPLHOOG2를 확인할 수 있다. 우선, level을 조정하기 전의 설명변수인 초등교육을 수료한 고객의 비율을 의미하는 MOPLLAAG 변수의 오즈비 추정값은 CARAVAN 보험 상품 구매를 가장 크게 승법적으로 감소시킨다. 즉, 이는 주어진 최종 변수에서 고객이 초등교육을 수료한 비율이 한 단위 증가하는 지역에 거주할 때마다 CARAVAN 보험을 소유한 고객의 오즈 추정값이 $e^{-0.125}$=0.883배만큼 승법적으로 감소한다고 할 수 있다. 말하자면, 초등교육을 수료한 고객이 많은 지역일수록 CARAVAN 보험을 구매하지 않는 경향을 보이는 것이다. 이는 고등교육을 수료한 고객의 비율이 88~99%인 지역의 오즈 추정값이 그렇지 않은 지역보다 999배보다 더 높게 추정되는 것과 비교됨으로써 매우 유의미한 결과를 보여준다. 다음으로, MOPLHOOG2는 다시 조정한 설명변수로써 고등교육을 수료한 고객의 비율이 11~23%인 지역에 거주하는 고객을 나타낸다. 이는 주어진 최종 변수에서 고등교육을 수료한 고객의 비율이 11~23%인 지역에 거주하는 고객이 CARAVAN 보험을 구매한 오즈 추정값은 그렇지 않은 고객의 오즈 추정값의 $e^{-0.6513}$=0.521배이다. 즉, 고등교육을 수료한 고객의 비율이 적은 지역은 그렇지 않은 지역보다 CARAVAN 보험을 구매하지 않은 경향을 보이는 것이다. 이는 고등교육을 수료한 고객의 비율이 88~99%인 지역에 거주하는 고객의 오즈 추정값이 그 외 지역에 거주하는 고객의 오즈 추정값의 999배보다 더 높은 것과 비교되어 매우 유의미한 결과를 보여준다.

제4절 모형 평가 과정

가. 기본 개념

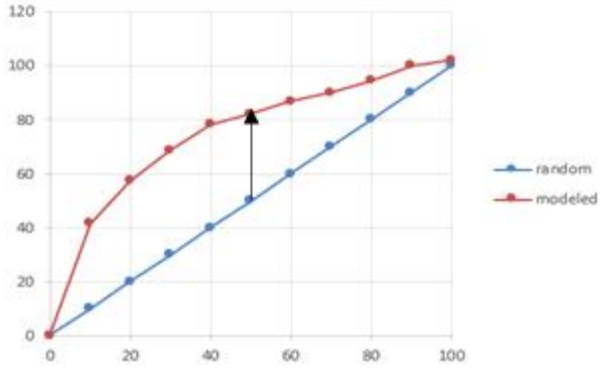
모형 평가란 모형을 구축하지 않았을 때(random approach) 또는 평균적인 모형보다 구축된 모형이 얼마나 더 적합하고 잘 적용되는지를 각 십분위에서 알아보는 것을 말한다. 이때, 모형 평가의 결과가 좋지 않다면 잘못된 변수 선택 혹은 전처리 과정에서의 문제 등을 고려해야하기에 모형 평가를 올바르게 해석하는 것이 중요하다.

나. 모형 평가의 기준

(1) 이익도표

일반적으로 모형을 평가하는 가장 기본적인 방법은 이익도표를 사용하는 것이다. 실제로 마케팅이나 경영에서 전문가들은 모형 결과를 쉽고 전체적으로 확인할 수 있다는 점 때문에 이익도표를 많이 활용한다. 이익도표를 사용해 십분위 분석을 진행할 경우, 구축된 모형이 얼마나 효율적인 결과를 낼 수 있는지 판단할 수 있어 마케팅 비용을 절감할 수 있다.

1. Train 데이터 십분위



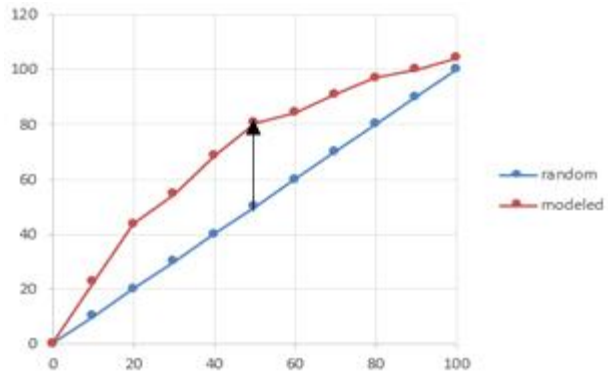
[그림 2 - 12]

예측확률을 내림차순으로 정렬해 CARAVAN 보험 구매 확률이 높은 순서로 정렬한 뒤, 가중치를 고려해 전체 7017개를 10개의 그룹으로 나누는 십분위 분석을 진행하였다. 이때, 십분위(0)은 상위 10프로 (약 688명)을 의미한다. 십분위 분석 결과, 본 연구의 train 데이터에서 가장 활성화 비율이 높은 십분위(0)의 활성화 비율(0.24416)은 가장 낮은 십분위 (9)의 활성화 비율(0.01281)보다 약 19배 크며, 모형 구축을 하지 않았을 때 (혹은 평균적인 모형을 사용한 경우)보다 실제로 CARAVAN 보험을 가입한 고객을 약 4.14배 더 포함하고 있다.

십분위 분석 결과를 누적 리프트(Cumulative Lift)의 관점에서 분석한 결과 전체 데이터의 50%만으로도 CARAVAN 보험에 가입한 고객 중 약 82%를 찾을 수 있다. 이는 모형을 구축하지 않았을 때보다 61% 증가한 것이다.

CARAVAN 보험에 가입한 고객의 비율이 감소하다가 증가하는 즉, 십분위로써 역전되는 부분 없이 활성화 비율이 단조감소하고 있다. 이를 토대로 현재 모형이 적합하다고 생각하고 test 데이터에 적용하였다.

2. Test Data 십분위



[그림 2 - 13]

본 연구의 train 데이터에서 가장 활성화 비율이 높은 십분위(0)의 활성화 비율(0.52577)은 가장 낮은 십분위(9)의 활성화 비율(0.00747)보다 약 70배 크며, 모형 구축을 하지 않았을 때 혹은 평균적인 모형을 사용한 경우보다 실제로 CARAVAN 보험을 가입한 고객을 약 2.21배 더 포함하고 있다.

십분위 분석 결과를 누적 리프트(Cumulative Lift)의 관점에서 분석한 결과, 전체 데이터의 50%만으로도 CARAVAN 보험에 가입한 고객 중 약 81%를 찾을 수 있다. 이는 모형을 구축하지 않았을 때 보다 56% 증가한 것이다. CARAVAN 보험에 가입한 고객의 비율이 감소하다가 증가하는 즉, 십분위로써 역전되는 부분 없이 활성화 비율이 단조감소하고 있다. 이를 토대로 현재 모형은 효율적인 결과를 가져오며 적합한 모형이라고 판단된다.

(2) 재표본

구축된 모형들의 안정성이나 로버스트 정도를 검정하기 위해 재표본 방법을 진행하였다. 이는 반복해서 샘플을 추출하고 각 추출에 대해 모형을 적합시키는 방식으로 과대적합을 파악할 수 있다. 본 연구에서는 재표본을 통해 추정값의 신뢰구간과 추정값을 구축된 모형과 비교해 모형의 안정성을 판단하였다. 이때, 재표본은 잭나이프 방법(jackknifing)과 부스트랩 방법(bootstrapping)으로 진행하였다.

1. 잭나이프

Decile	JK Est Prob	JK Lower CI Prob	JK Upper CI Prob	JK Est % Active	JK Lower CI % Active	JK Upper CI % Active	JK Est Lift	JK Lower CI Lift	JK Upper CI Lift
0	339.416%	322.828%	356.005%	206.211%	195.476%	216.946%	315	299	332
1	121.584%	121.086%	122.082%	123.315%	11.654%	134.976%	188	173	204
2	83.984%	83.575%	84.392%	72.504%	67.850%	77.158%	111	104	118
3	63.955%	63.709%	64.201%	65.344%	60.604%	70.083%	100	93	107
4	50.563%	50.377%	50.748%	68.522%	64.303%	72.740%	105	99	110
5	37.685%	37.559%	37.811%	32.041%	29.473%	34.609%	49	45	53
6	33.233%	33.160%	33.305%	32.046%	29.494%	34.598%	49	45	53
7	25.860%	25.726%	25.994%	35.480%	32.728%	38.232%	54	50	58
8	18.282%	18.199%	18.365%	3.460%	2.756%	4.164%	5	4	6
9	10.786%	10.544%	11.028%	28.771%	25.960%	31.581%	44	40	48
Total	78.535%	76.676%	80.393%	66.769%	62.030%	71.509%	102	95	109

<표 2 - 10>

잭나이프 방법이란 하나의 관측값만 남겨두는 'leave-one-out' 방식에 근거를 둔다. 본 연구에서는 고차원의 데이터를 사용하므로, 하나의 관측값만 제외하는 잭나이프의 일반적인 방식이 아닌 다수의 관측값을 제외하는 변형된 형태의 잭나이프를 실시하였다. 이때, 활성화 확률, 실제 활성화 비율, 각 십분위 리프트값들을 추정하기 위하여 표본의 크기가 전체의 99% 인 표본을 100개 사용하여 재표본을 진행하였다.

[그림 2 - 14]

일반적으로 부스트랩은 잭나이프보다 신뢰구간의 범위가 넓게 추정되기 때문에 잭나이프 방법을 통해 추정치의 신뢰범위를 통한 모형의 안정성를 파악하였다. 잭나이프 방법을 통한 재표본 결과, 잭나이프 추정값 (JK Est Lift)이 구축된 모형에서 증가하는 경우 즉, 역전되는 경우가 두 번 발생하지만 전반적으로 단조 감소하는 형태이다. 또한, 신뢰구간 하한선(JK Lower CI Lift)와 상한선(JK Upper CI Lift)으로 구한 신뢰구간의 범위가 40이내로 비교적 작은 범위를 유지한다는 점을 통해 본 모형이 로버스트하다는 것을 알 수 있다.

2. 부스트랩

Decile	Prob	BS Est Prob	BS Lower CI Prob	BS Upper CI Prob	% Y=1	BS est % Y=1	BS Lower CI % Y=1	BS Upper CI % Y=1	Lift	BS Est Lift	BS Lower CI Lift	BS Upper CI Lift
0	0.53	0.51	0.35	0.68	0.15	0.13	0.03	0.24	251	225	64	386
1	0.15	0.15	0.14	0.15	0.12	0.11	0.04	0.18	194	186	81	291
2	0.09	0.09	0.08	0.09	0.09	0.08	0.02	0.15	152	142	54	230
3	0.06	0.06	0.08	0.06	0.05	0.05	0.02	0.08	81	82	37	128
4	0.04	0.04	0.04	0.05	0.05	0.04	0.01	0.08	75	73	15	132
5	0.03	0.03	0.03	0.04	0.04	0.04	0.01	0.07	70	70	21	120
6	0.03	0.03	0.03	0.03	0.04	0.04	-0.00	0.08	64	64	5	121
7	0.02	0.02	0.02	0.02	0.05	0.04	-0.01	0.08	75	75	-2	125
8	0.01	0.01	0.01	0.01	0.00	0.00	-0.00	0.01	6	6	-5	19
9	0.01	0.01	0.01	0.01	0.02	0.02	0.00	0.04	39	39	5	64
Total	0.10	0.10	0.08	0.11	0.06	0.06	0.01	0.10	101	101	27	161

<표 2 - 11>

전체 표본 데이터 세트에서 표본을 복원추출하는 부스트랩 방식을 통해 분석을 진행하였다. 잭나이프 방식과 마찬가지로 고차원의 데이터이므로 변형된 형식의 부스트랩을 통해 총 25개의 부스트랩 표본을 생성하였다. 부스트랩 추정치를 통해 모형의 안정성을 분석해본 결과, 부스트랩 이익도표에서 모든 부스트랩 추정값(BS est Lift)이 구축된 모형에서 구한 리

프트값(Lift)과의 차이가 작은 편이기 때문에 구축된 모형이 안정적이고 로버스트하다고 볼 수 있다.

3. 결과물 및 해석

추정치와 추정치의 신뢰구간을 통해 현재 모형을 비교해본 결과, 추정값과 구축된 모형에서의 리프트 값의 차이가 작고 이러한 추정값이 단조감소한다는 점과 신뢰구간의 범위가 짧다는 점을 바탕으로 구축된 모형이 안정적인 형태를 보인다고 판단 된다.

제3장 결 론

제1절 추가 연구

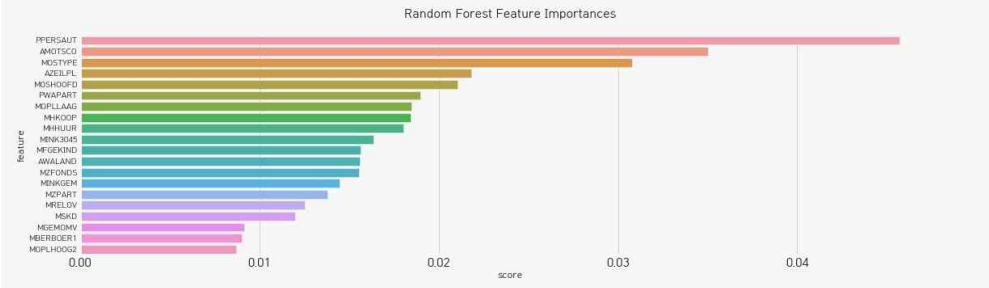
Logistic Regression 모형의 경우 최초 모형 식은 비선형적인 성질을 보이지만 로그 변환을 통해 계수가 선형적인 성질을 갖도록 한 일반화 선형 모형이다. 따라서, 설명변수와 반응변수 간의 관계가 비선형의 경우를 고려하기 위하여 본 연구에서 사용한 최종 Logistic Regression 모형과 다른 모형 간의 비교평가를 진행하였다.

Model	Accuracy	F1-score	Precision	Recall	Specificity
Random Forest	0.6420	0.6480	0.6373	0.6590	0.6250
Support Vector	0.6335	0.6465	0.6243	0.6704	0.5965
Ridge Regression	0.6250	0.6353	0.6182	0.6534	0.5965
Gradient Boosting	0.6335	0.6342	0.6379	0.6306	0.6420
LGBM	0.6335	0.6406	0.6284	0.6534	0.6136

<표 3 - 1>

비교평가를 위해 사용된 모형은 Random Forest,Support Vector Machine, Ridge Regression, Gradient Boosting, LGBM이며, 프로젝트 수행 목표가 반응변수 CARAVAN을 이진 분류하는 것이므로 혼동행렬로부터 계산된 정확도(Accuracy)[2], 특이도(Specificity)[3], 정밀도(Precision)[4], 재현율(Recall)[5], 정밀도와 재현율의 조화평균인 F1-Score를 모형 평가 지표로 사용하였다. 다른 모형을 사용했을 때 결과는 <표 3 - 1>를 통해 알 수 있다.

2) 정확도(accuracy)는 전체 샘플 중 맞게 예측한 샘플 수의 비율을 뜻한다.
3) 특이도(specificity)는 음성 클래스에 속한다고 출력한 샘플 중 실제로 음성 클래스에 속하는 샘플 수의 비율을 말한다.
4) 정밀도(precision)은 양성 클래스에 속한다고 출력한 샘플 중 실제로 양성 클래스에 속하는 샘플 수의 비율을 말한다.
5) 재현율(recall)은 실제 양성 클래스에 속한 표본 중에 양성 클래스에 속한다고 출력한 표본의 수의 비율을 뜻한다.



[그림 3 - 1]

이때, Random Forest 모형의 상위 20개의 변수 기여도는 [그림 3 - 1]과 같다. 따라서, Random Forest 모형에서 고객의 CARAVAN 보험 구매 여부를 예측하는 데 기여하는 변수는 자동차 보험료(PPERSAUT), 오토바이 및 스쿠터 보험 개수(AMOTSCO), 고객 하위유형(MOSTYPE), 서핑 보드 보험 개수(AZEILPL) 등이 있다. Random Forest을 제외한 이외 모형의 상위 20개 변수 중요도는 부록에서 확인할 수 있다.

Variable	Random Forest	Support Vector	Ridge Regression	Gradient Boosting	LGBM	SUM
AZEILPL	O	X	O	O	O	4
PPERSAUT	O	X	X	O	O	3
MOSTYPE	O	X	X	O	O	3
PWAPART	O	X	X	O	O	3
MINK3045	O	X	X	O	O	3
MFGEKIND	O	X	X	O	O	3
MINKGEM	O	X	X	O	O	3
MRELOV	O	X	X	O	O	3
MBERBOER1	O	X	O	O	X	3
MRELGE9	X	O	O	O	X	3
MINKM305	X	O	O	O	X	3
MBERMIDD4	X	X	O	O	O	3

<표 3 - 2>

다음은 각 모형의 상위 20개의 중요 변수 중 3번 이상 중복으로 선정된 변수들에 대한 표이다. 서핑 보드 보험 개수에 대한 변수(AZEILPL)의 경우 Support Vector Machine을 제외한 모든 모형에서 중요 변수로 선정되었으며, 이외에도 자동차 보험료(PPERSAUT), 고객 하위유형(MOSTYPE), 제 3자 보험료(PWAPART), 평균 수입(MINKGEM) 등이 다수의 모형에서 중요 변수로 선정되었다. 이 중 자동차 보험료(PPERSAUT), 서핑 보드 보험 개수에

대한 변수(AZEILPL), 아이가 없는 가구(MFGEKIND), 수입과 관련된 변수(MINKM30)의 경우 최종 Logistic Regression 모형의 유의한 변수와도 중복되는 것을 보아 고객의 CARAVAN 보험 구매 여부를 예측하는 데 기여한 변수들 중 더 중요하다는 것을 알 수 있다.

제2절 결론 및 향후과제

가. 결론

여러 모형의 결과를 비교하여, CARAVAN보험에 유의미한 결과를 나타내는 변수들을 중점으로 CARAVAN보험을 구입한 고객의 특성을 분석해보았다.

먼저, 고객이 지불하고 있는 화재 보험료와 자동차 보험료는 같은 맥락에서 분석할 수 있다. 왜냐하면 고객이 화재 보험료와 자동차 보험료에 돈을 많이 지불한다는 것은 그만큼 고객이 소유하고 있는 재산의 가치가 크다는 것을 의미하기 때문이다. 이는 캠핑과 같은 여가생활에 소비하는 돈이 많다는 해석으로 이어질 수 있다. 또한, 이러한 특성은 고객이 소유하고 있는 서핑 보드 보험의 개수에서도 이어진다. 왜냐하면 보드를 타는 것 또한 캠핑과 같은 하나의 여가생활이기 때문이다. 그러므로 고객이 서핑 보드 보험에 많이 가입했다는 것은 소유하고 있는 재산이 비교적 수준이 많다는 것으로 해석할 수 있다. 그렇기에 비교적 낮지 않은 가격대의 CARAVAN을 소유하고 이와 관련된 보험에 관심을 갖을 확률이 높다고 판단된다.

앞선 로지스틱 분석에서 고등교육을 받은 고객이 많을수록 CARAVAN보험 구매가 많아지고, 초등교육만 받은 고객이 많을수록 CARAVAN보험 구매는 줄어들었다. 이는 교육수준이 높을 경우, 일반적으로 평균 월급이 높은 직업을 갖게 되고 자연스럽게 사회적 계급도 높아지기 때문으로 해석할 수 있다.

로지스틱 분석에서 오즈비 추정값이 상대적으로 크게 나온 장애보험 변수와 CARAVAN보험의 관계는 다음과 같이 해석할 수 있다 . 장애보험은 장애를 갖고 있는 고객이 구매하는 경우도 있으나 장애를 갖게 될 상황을 대비해 구입하는 경우도 있다. 일반적으로 소득이 평범한 가정의 경우, 보험에 큰 비용을 사용하기 힘들기에 장애 보험을 필수적으로 가입하지 않는 반면 소득수준이 높은 가정의 경우, 혹시 모를 위험에 대비해 장애 보험을 가입한다. 마지막으로 소득과 관련된 변수는 소득이 높은 그룹이 CARAVAN보험을 많이 구입하는 것을 보여주며 직접적으로 소득과 CARAVAN보험이 연관되어 있다는 것을 알려준다.

이러한 변수들과 CARAVAN보험의 관계를 분석해본 결과, 대다수의 특성이 고객이 소유한 재산 수준과 관련되어 있다는 것을 알 수 있다. 또한, 재산 수준이 높을 뿐만 아니라 다양한 종류의 보험을 가입하면서 혹시 모를 위험에 대비하는 모습을 발견할 수 있었다.

나. 마케팅 전략

앞서 분석한 결과를 토대로 보험회사에서 고려할 만한 마케팅 전략은 다음과 같다.

첫 번째는 결합상품이다. 앞서 분석한 것에 따르면 높은 수준의 보트 보험료와 자동차 보험료를 지불할 수 있는 고객이 CARAVAN보험에 가입한 경우가 많았다. 따라서 CARAVAN과 보트를 결합한 보험 상품, 자동차와 CARAVAN을 결합한 보험 상품을 통한 CARAVAN보험 구매율을 높이고, CARAVAN을 소유한 기존의 보트 보험, 자동차 보험 가입자들을 대상으로 보험을 갱신할 때, 새로운 결합 형태의 보험으로 변경할 수 있는 기회를 제공하여 CARAVAN보험 구매를 유도할 수 있다.

두 번째는 기존의 상품을 보완하는 방식이다. 아이가 있는 가족과 CARAVAN보험 간의 연관을 고려하여 CARAVAN보험에 아이 관련 특약을 추가하고 이를 통해 아이가 있는 가정의 보험 가입을 유도한다.

마지막으로 소득수준, 교육수준이 높은 지역에서 CARAVAN구매율이 높다는 점을 통해 해당 지역을 대상으로 CARAVAN보험 관련 프로모션을 진행한다면 평소의 CARAVAN보험에 관심 있었던 고객뿐만 아니라 잘 알지 못했던 고객 또한, 유입시킬 수 있을 것이다.

다. 한계점

본 연구의 한계점은 다음과 같다.

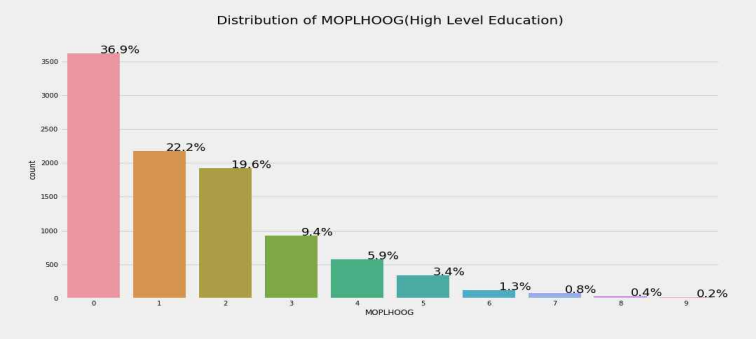
먼저, 샘플링 방식이다. 분석을 진행하는 과정에서 종속변수의 불균형을 해결하기 표본추출 방법으로써 언더샘플링을 진행하였다. 오버샘플링이나 SMOTE와 같은 다른 샘플링 방식으로도 샘플링을 진행해보고 결과를 비교하였다면, 보다 예측력이 높은 모형을 구축할 수도 있었을 것이다.

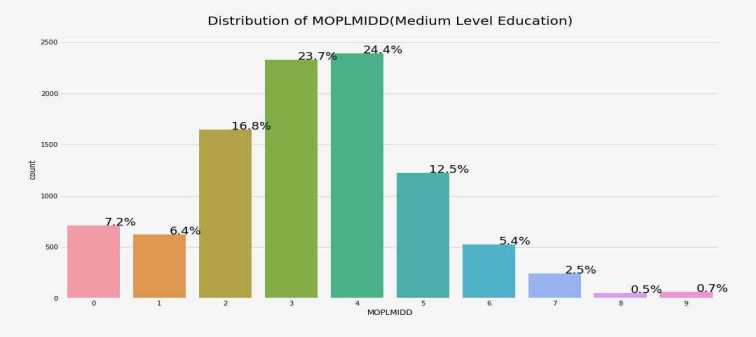
다음으로 ,리스크 모형에 관한 것이다. 보험료 미지급이나 연체와 같은 지불과 관련한 데이터가 추가된다면, 위험이 큰 고객을 피할 수 있는 리스크모형을 만들어서 손실을 줄일 수 있을 것으로 기대할 수 있을 것이다.

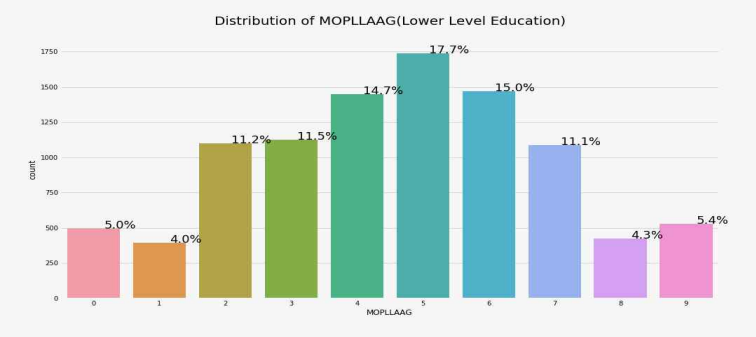
마지막으로, 고객 개인별 특성에 관한 문제다. 본 연구에서 사용한 인구사회학적 변수는 우편 번호를 기반으로 한 지역을 단위로 했기 때문에 개인의 특성에 대해서는 파악하기 어려웠다. 즉, 집단의 특성을 통하여 고객 개개인의 특성을 파악했기 때문에 고객 개인별 특성을 통 구축한 모형보다 설득력이 떨어진다고 할 수 있다. 이때, 최신 기술인 개인정보의 비식별화를 통하여 개인별 특성을 가지는 변수를 제공받을 수 있다면 예측력과 설득력이 향상 될 수 있을 것이다.

ABSTRACT

[그림목차]







<표목차>

고객 기본 유형	고객 하위 유형
유형 1 Successful hedonists	* 유형 1 High Income, expensive child * 유형 2 Very Important Provincials * 유형 3 High status seniors * 유형 4 Affluent senior apartments * 유형 5 Mixed seniors
유형 2 Driven Growers	* 유형 6 Career and childcare * 유형 7 Dinki's * 유형 8 Middle class families
유형 3 Average Family	* 유형 9 Modern, complete families * 유형 10 Stable family * 유형 11 Family starters * 유형 12 Affluent young families * 유형 13 Young all american family
유형 4 Career Loners	* 유형 15 Senior cosmopolitans * 유형 16 Students in apartments * 유형 17 Fresh masters in the city * 유형 18 Single youth * 유형 19 Suburban youth
유형 5 Living well	* 유형 20 Etnically diverse * 유형 21 Young urban have-nots * 유형 22 Mixed apartment dwellers * 유형 23 Young and rising * 유형 24 Young, low educated
유형 6 Cruising Seniors	* 유형 25 Young seniors in the city * 유형 26 Own home elderly * 유형 27 Seniors in apartments * 유형 28 Residential elderly
유형 7 Retired and Religeous	* 유형 29 Porchless seniors * 유형 30 Religious elderly singles * 유형 31 Low income catholics * 유형 32 Mixed seniors
유형 8 Family with grown ups	* 유형 33 Lower class large families * 유형 34 Large family, employed child * 유형 35 Village families * 유형 36 Couples with teens 'Married with children' * 유형 37 Mixed small town dwellers
유형 9 Conservative families	* 유형 38 Traditional families * 유형 39 Large religous families
유형 10 Farmers	* 유형 40 Large family farms * 유형 41 Mixed rurals