dongguk UNIVERSITY

CARAVAN INSURANCE

과 목 명 : 데이터마이닝

교 수 명 : 김희경

제 출 일 : 2019.05.31.

학번_이름 : 2013110508 한태규
2014110482 장현석
2017111748 임정민

[목차]

Ⅰ. 개요

Ⅱ. 서론

ⅰ. 데이터 설명
ⅱ. 분석 목적
ⅲ. 분석 방법 및 절차

Ⅲ. 본론1

ⅰ. EDA
ⅱ. 데이터 변환 및 전처리
ⅲ. Under Sampling
ⅳ. 중요변수 선택

Ⅳ. 본론2

ⅰ. 모형적합 준비단계
ⅱ. Logistic Regression
ⅲ. 나이브베이즈
ⅳ. 신경망
ⅴ. 랜덤포레스트
ⅵ. 아다부스트
ⅶ. 그래디언트 부스트
ⅷ. SVM

Ⅳ. 결론

Ⅴ. 부록

Ⅰ. 개요.

우리팀이 분석할 데이터는 TIC데이터로서 한 보험회사가 데이터 분석 기법을 통해 기증한 자료이다. 이 데이터는 인구통계학적 변수들과 회사가 가지고 있는 고객들의 구매정보를 포함하고 있다. 우리 분석의 주요목적은 이동용 캠핑카 보험을 구매한 고객들의 주요한 특징을 파악하고 다른 고객 중에 잠재적인 구매자를 효율적으로 구분하고자 한다.

Ⅱ. 서론

ⅰ. 데이터 설명

우리는 9822명의 고객에 대한 데이터를 가지고 있다. 이 데이터는 85개의 설명변수와 1개의 반응변수로 총 86개 변수로 이루어져 있다.

	MOSTYPE	MAANTHUI	MGEMOMV	MGEMLEEF	MOSHOOFD	...	AFIETS	AINBOED	ABYSTAND	CARAVAN
0	33	1	4	2	8	...	0	0	0	0
1	6	1	3	2	2	...	0	0	0	1
2	39	1	3	3	9	...	0	0	0	0
3	9	1	2	3	3	...	0	0	0	0
4	31	1	2	4	7	...	0	0	0	0

	MOSTYPE	MAANTHUI	MGEMOMV	MGEMLEEF	MOSHOOFD	...	AFIETS	AINBOED	ABYSTAND	CARAVAN
9817	36	1	1	2	8	...	0	0	0	0
9818	35	1	4	4	8	...	0	0	0	0
9819	33	1	3	4	8	...	0	0	0	1
9820	34	1	3	2	8	...	0	0	0	0
9821	33	1	3	3	8	...	0	0	0	0

그림 1

위의 [그림1]은 데이터의 첫 5개의 값과 마지막 5개의 값을 보여준다.

설명변수는 크게 2가지 범주, ‘M’으로 시작하는 변수와 ‘P’ 또는 ‘A’로 시작하는 변수로 나눌 수 있다. ‘M’으로 시작하는 변수는 고객 정보와 고객의 우편번호로부터 추출된 사회-인구 데이터이다. 즉, 같은 우편번호에 사는 모든 고객들은 같은 사회 인구통계 데이터를 가진다. ‘P’ 또는 ‘A’로 시작하는 변수는 고객이 소유한 보험에 관한 데이터이다. ‘P’로 시작하는 변수는 21개의 보험금 데이터이고 ‘A’로 시작하는 변수는 21개의 보험 수 데이터이다.

86개 변수는 23개의 수치형 변수와 63개의 범주형 변수로 구성된다. 63개 범주형 변수는 우리의 관심변수를 포함한 3개의 명목형 변수와 60개의 순위형 변수로 구성된다.

더 상세한 데이터 설명은 [부록1]에 제시되어 있다.

ⅱ. 분석 목적 (check / 보완수정하기)

한 보험 회사의 기존 고객을 기반으로 카라반 보험을 추가로 구입할 수 있도록 유도하는, ‘Cross-Selling’ 전략을 위한 모형을 만들고자 한다. 이를 위해 기존 고객 중 어떤 고객이 카라반 보험에 잠재적으로 관심이 있는 고객인지 예측하고 이러한 고객들이 왜 카라반 보험을 구매하는지 이유를 설명하고자 한다.

ⅲ. 분석 방법 및 절차

데이터 탐색:

우리는 데이터 탐색에서 이상치의 존재유무를 확인하고 변수들마다의 평균, 표준편차, 사분위수, 그리고 관측값의 분포를 통해 변수들이 어떤 특징을 가지고 어떤 변수들이 서로 묶일 수 있고 중복된 정보를 가질 수 있는지 간단하게 확인하고 시각적으로 표현해 보았다. 그리고 일차적으로 반응변수와 나머지 독립변수들과의 관계를 상관계수를 통해 순위를 매기고 간단하게 시각적으로 반응변수 카라반 보험이 그 변수에 의해 어떻게 나뉘는지를 시각화 해서 살펴 보았다.

데이터 전처리:

TIC 데이터는 1개의 반응변수와 85개의 독립변수로 이루어져 있는데 2개는 범주형 변수다. 우리는 이 변수들을 지역 PCA, 그리고 중간값 넣기, 그리고 범주형 변수들의 분포의 비교를 통해 하나를 제거 하고 하나만을 더미 변수로 만들어 놓는 판단을 할 것이다.

변수 선택:
랜덤포레스트의 변수 중요도 평가, 공분산 평가, AIC,BIC 변수 규제법을 활용해서 모델을 적합시키기전

모형 적합:
십분위수 분석을 하기에 용이한 머신러닝 기법들 7가지를 적용해 모델들의 점수를 시각화하고 비교해본다

모델 선택 :
최종 모델을 선택해 이익도표를 통한 10분위수 분석을 진행한다.

최종 변수평가:
마지막 모델에서 쓰인 Logistic 회귀 계수와 Random Forest를 활용해 변수들이 가지는 의미 . 실제로 어떤 사람들이 이동식주택보험인 CARAVAN을 구입하는지 분석한다.

Ⅲ. 본론1

ⅰ. EDA

(1) 결측값 탐색 및 처리

```
1 print('Missing values: %i' % ticdata.isnull().sum().sum())
Missing values: 0
```

우리의 데이터는 결측값이 없음을 확인했다.

(2) 이상치 탐색 및 처리

우리의 데이터는 다음과 같이 각 설명변수에 대한 범위가 주어져있다.

- 'MOSTYPE: Customer Subtype' : 1 - 41
- 'MAANTHUI: Number of house' : 1 - 10
- 'MGEMOMV: Average size household' : 1 - 6
- 'MGEMLEEF: Avg age' : 1 - 6
- 'MOSHOOFD: Customer main type' : 1- 10
- 'MGODRK: Roman catholic'
  ~ 'MKOOPKLA: Purchasing power class' : 0 - 9
- 'PWAPART: Contribution private third party insurance polices'
  ~ 'PBYSTAND: Contribution social security insurance polices' : 0 - 9
- 'AWAPART: Number of private third party insurance polices'
  ~ 'ABYSTAND: Number of social security insurance polices' : 1 - 12

우리의 데이터 값이 모두 위의 범위 안에 포함되는지 확인하기 위해 min-max 그래프를 그려보았다.

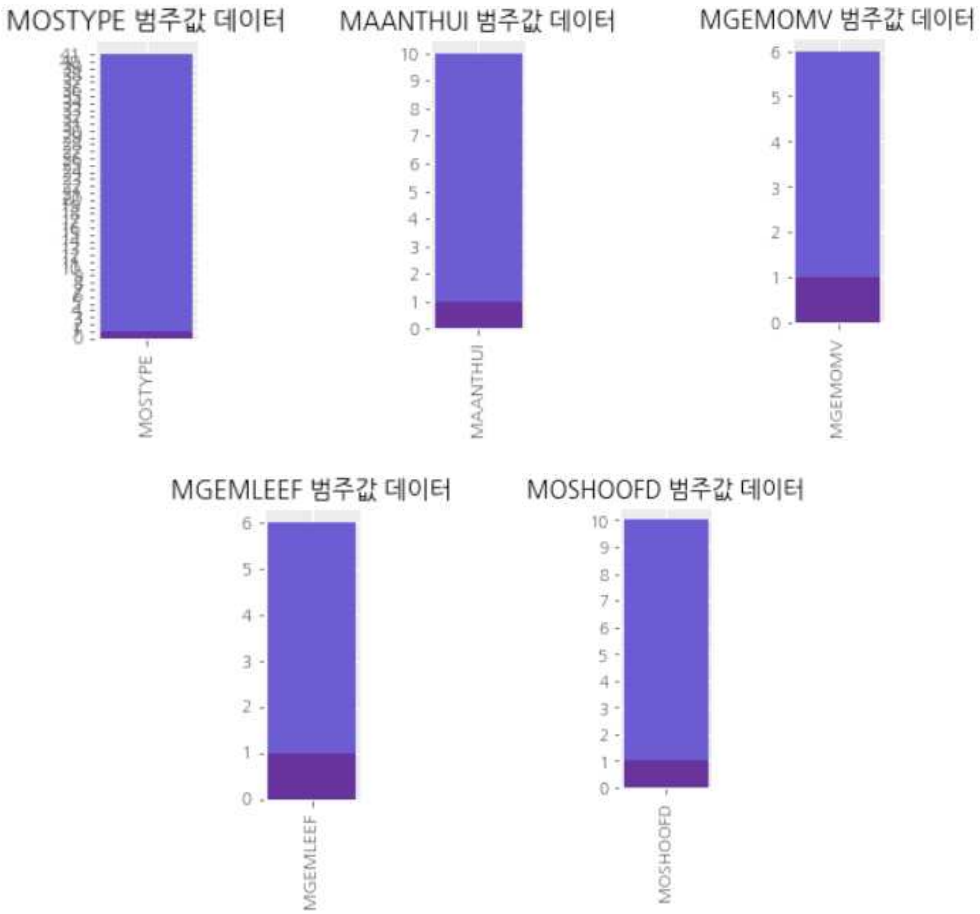

MOSTYPE 범주값 데이터
MAANTHUI 범주값 데이터
MGEMOMV 범주값 데이터
MOSTYPE
MAANTHUI
MGEMOMV
MGEMLEEF 범주값 데이터
MOSHOOFD 범주값 데이터
MGEMLEEF
MOSHOOFD

그림 2

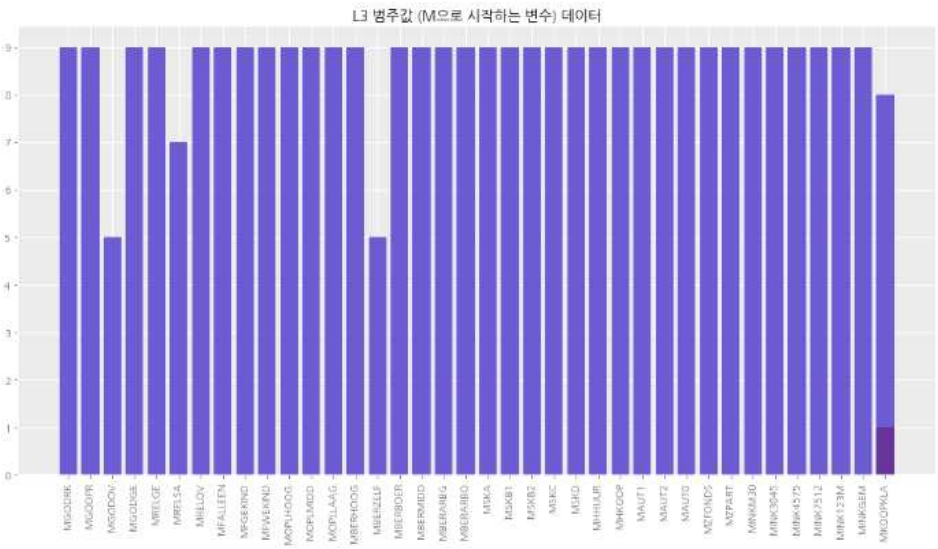
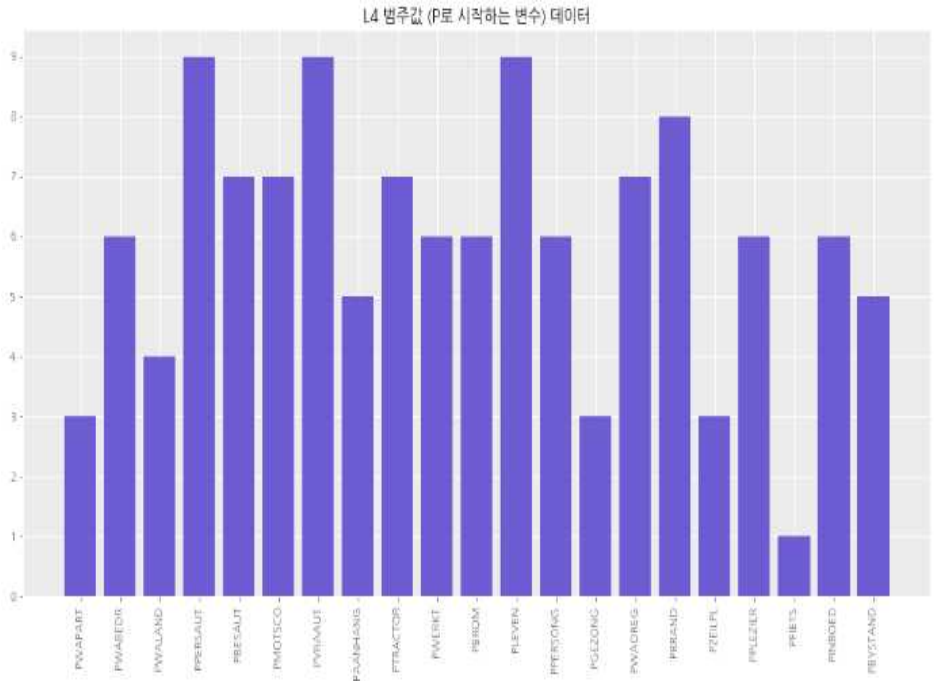

L3 범주값 (M으로 시작하는 변수) 데이터

그림 3



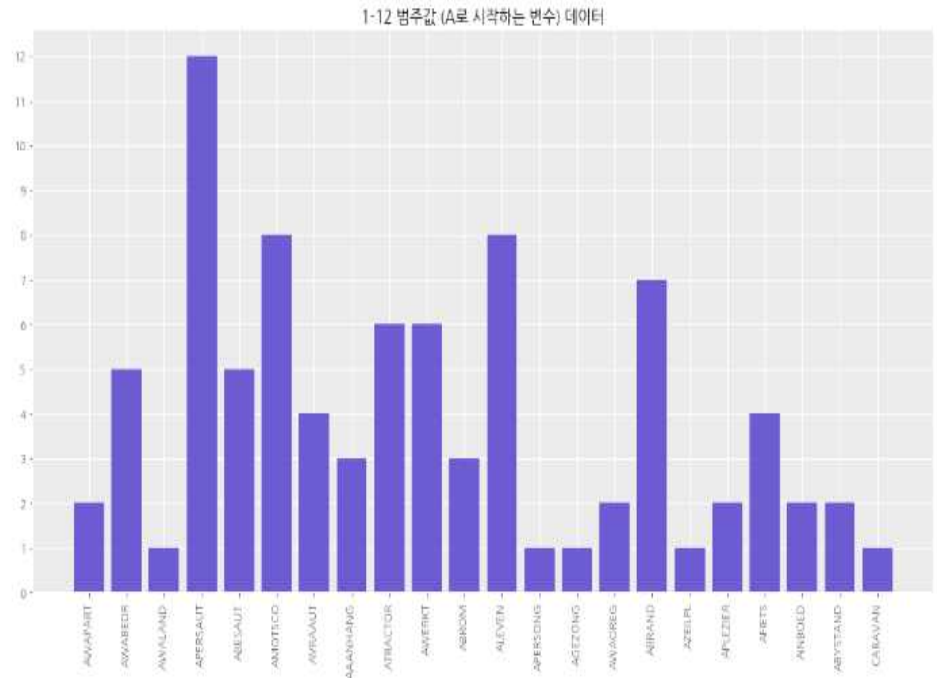


그림 4

위의 [그림2], [그림3], [그림4]을 보면 모든 데이터의 최소-최대값 그래프가 주어진 범위에 포함되는 것을 확인할 수 있다. 또한 정수가 아닌 값이 있는지 확인하기위해 [부록2]에서 각각의 값을 모두 제시했다. 따라서 우리 데이터에서는 이상치가 없음을 확인할 수 있다.

(3) 관심변수 비율 확인

이번에는 반응변수인 CARAVAN 보험 유무 차이를 보기위해 그래프를 그려보았다.

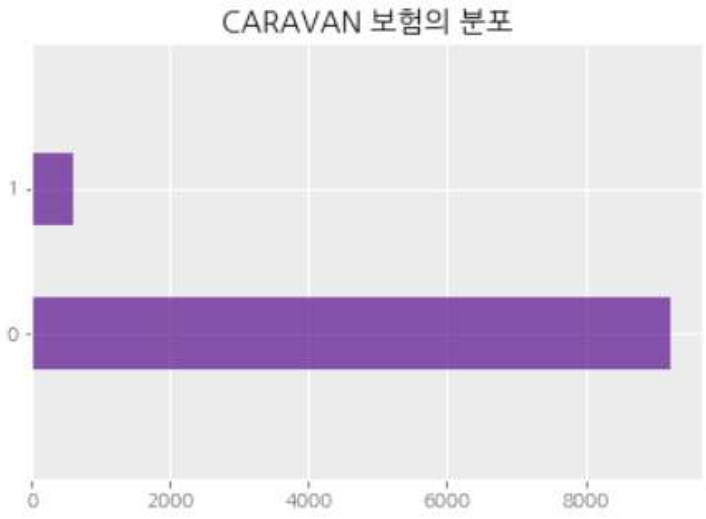


그림 5

위의 [그림5]에서도 볼 수 있듯이, 우리의 데이터는 심한 불균형 데이터이다. 9822명 고객 중 반응변수 값이 1인 고객은 즉, Caravan 보험을 가지고 있는 고객은 586명, 전체 고객 중 6%에 불과하다.

(4) 설명변수 분포

이제 각 변수별로 평균, 표준편차 그리고 분위수를 시각적으로 확인해보자

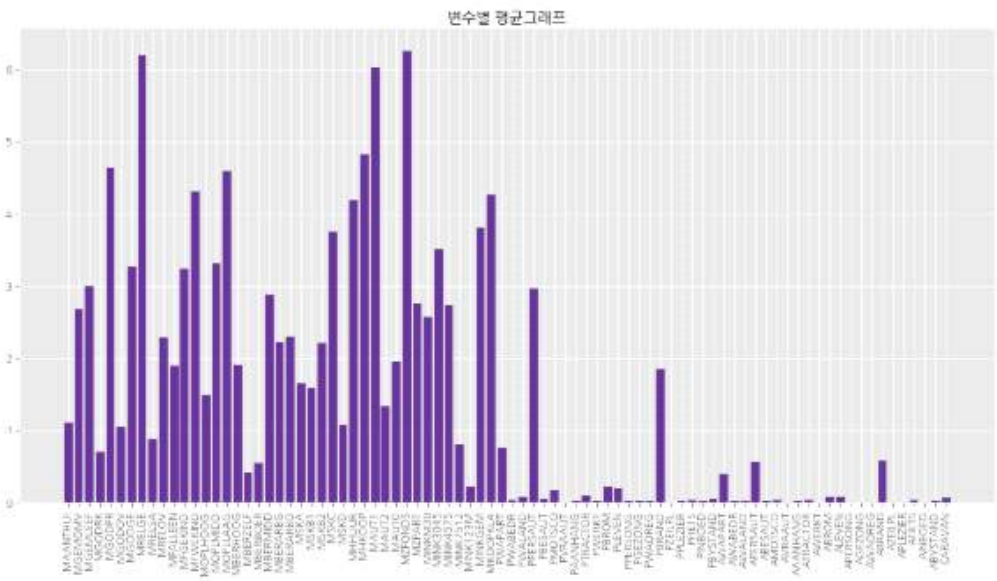


그림 6

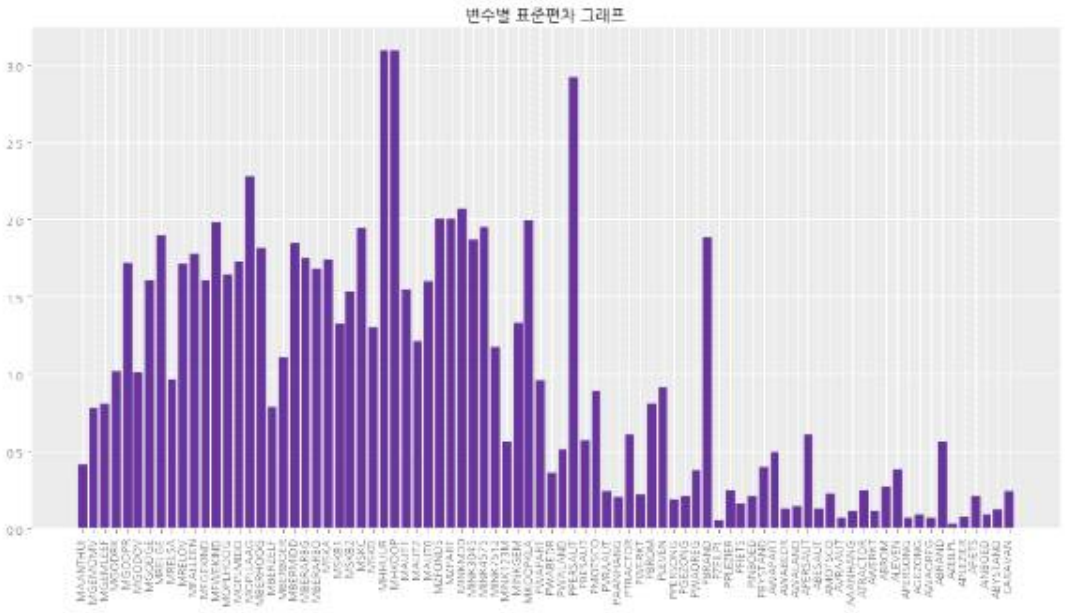


그림 7

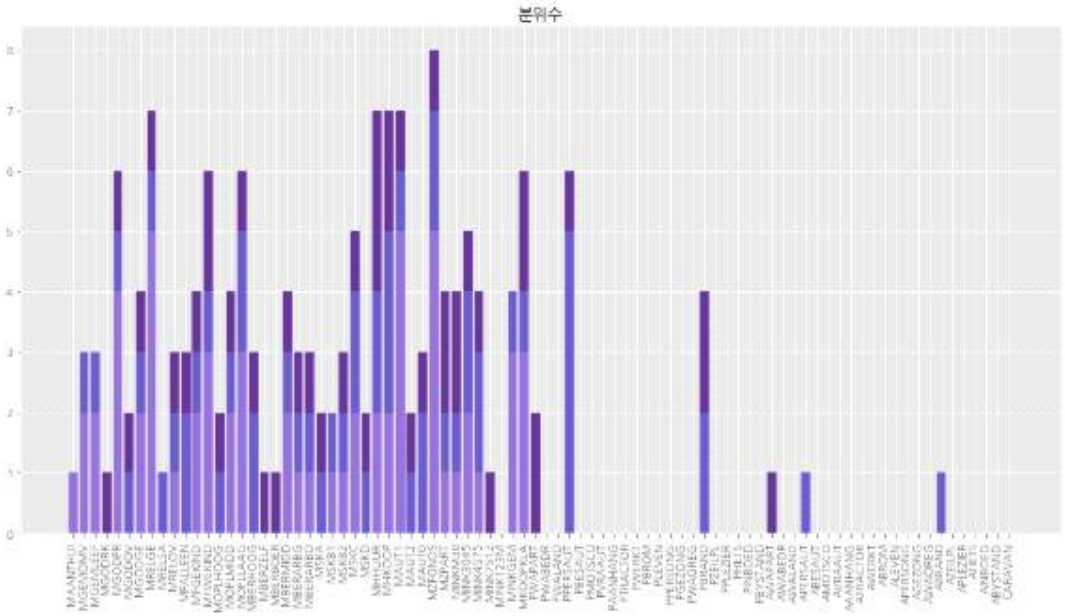


그림 8

위의 [그림6], [그림7], [그림8]은 순서대로 변수별로 평균, 표준편차 그리고 분위수를 보여준다. 분위수 그래프 [그림8]를 보면 75분위수마저 0인 변수가 많다. 따라서 변수별로 0의 개수를 파악해보았다.

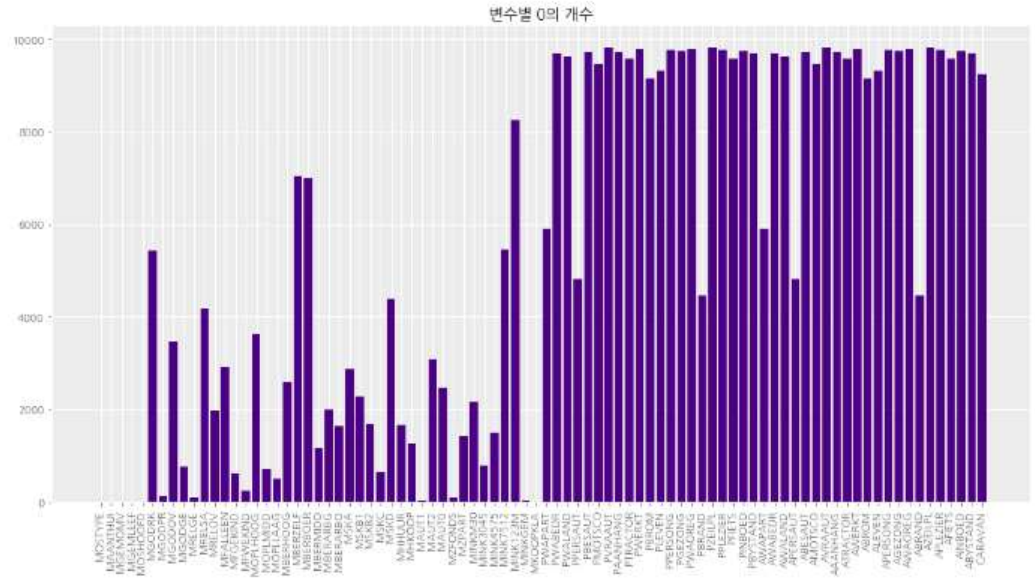


그림 9

위의 변수별 0의 개수를 보여주는 [그림9]을 보면 P와 A로 시작하는 보험금과 보험수 변수들은 0의개수가 매우 많은걸 볼 수 있다.

다음으로 반응변수를 제외한 'MOSTYPE: Customer Subtype' 과 'MOSHFOOD: Customer main type', 두 명목형 변수의 분포를 파악하기 위해 히스토그램을 그렸다.

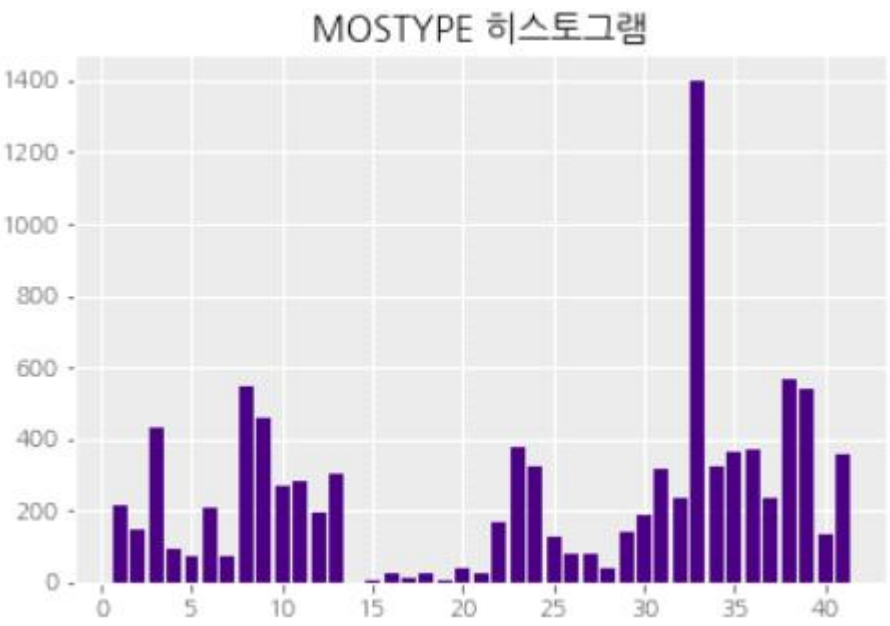


그림 10

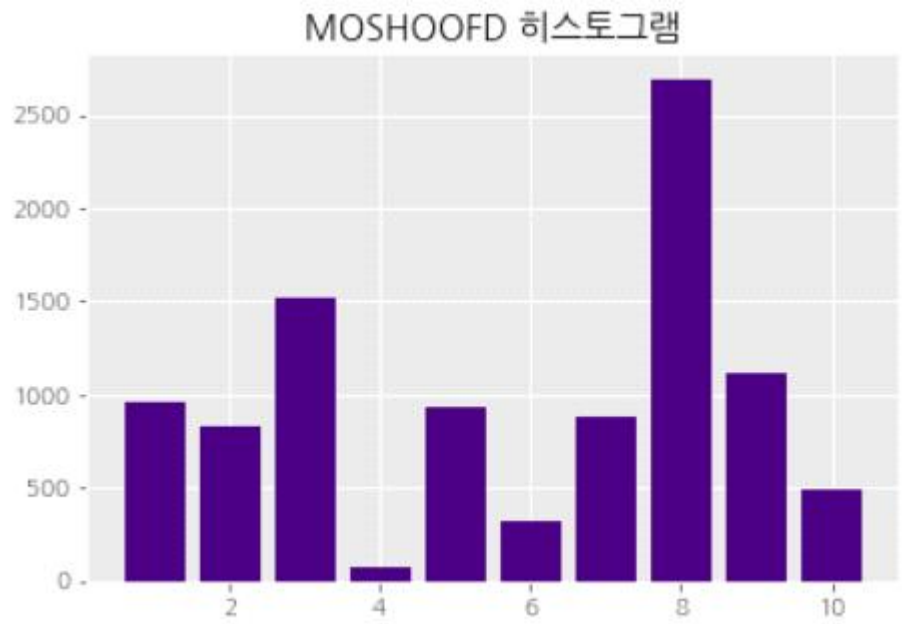


그림 11

[그림10], [그림11]을 보면 고객 유형별 분포가 일정하지 않고 관측값이 적은 범주도 있음을 확인했다.

(5) 주요한 변수 후보 탐색

우선 우리의 반응변수 'CARAVAN: Number of mobile home polices'와 모든 설명변수와의 상관관계를 보았다. 다음은 관심변수와 상관관계에 있어 상위 20개의 설명변수이다.

PPERSAUT	0.38	MOPLLAAG	0.15
APERSAUT	0.32	MHKOOP	0.14
PWAPART	0.22	MBERBOER	0.13
AWAPART	0.21	ABRAND	0.13
MKOOPKLA	0.18	MRELOV	0.13
MAUT1	0.17	MZFONDS	0.13
MRELGE	0.17	MGEMOMV	0.12
MINKGEM	0.17	MINKM30	0.12
MAUT0	0.17	MFALLEEN	0.11
MHHUUR	0.16	MGODOV	0.10

상위 8개의 설명변수에 대해서, CARAVAN 보험 유무에 따른 분포를 확인하고 각 value별 odds를 확인해보겠다.

① ‘PPERSAUT: Contribution car polices’

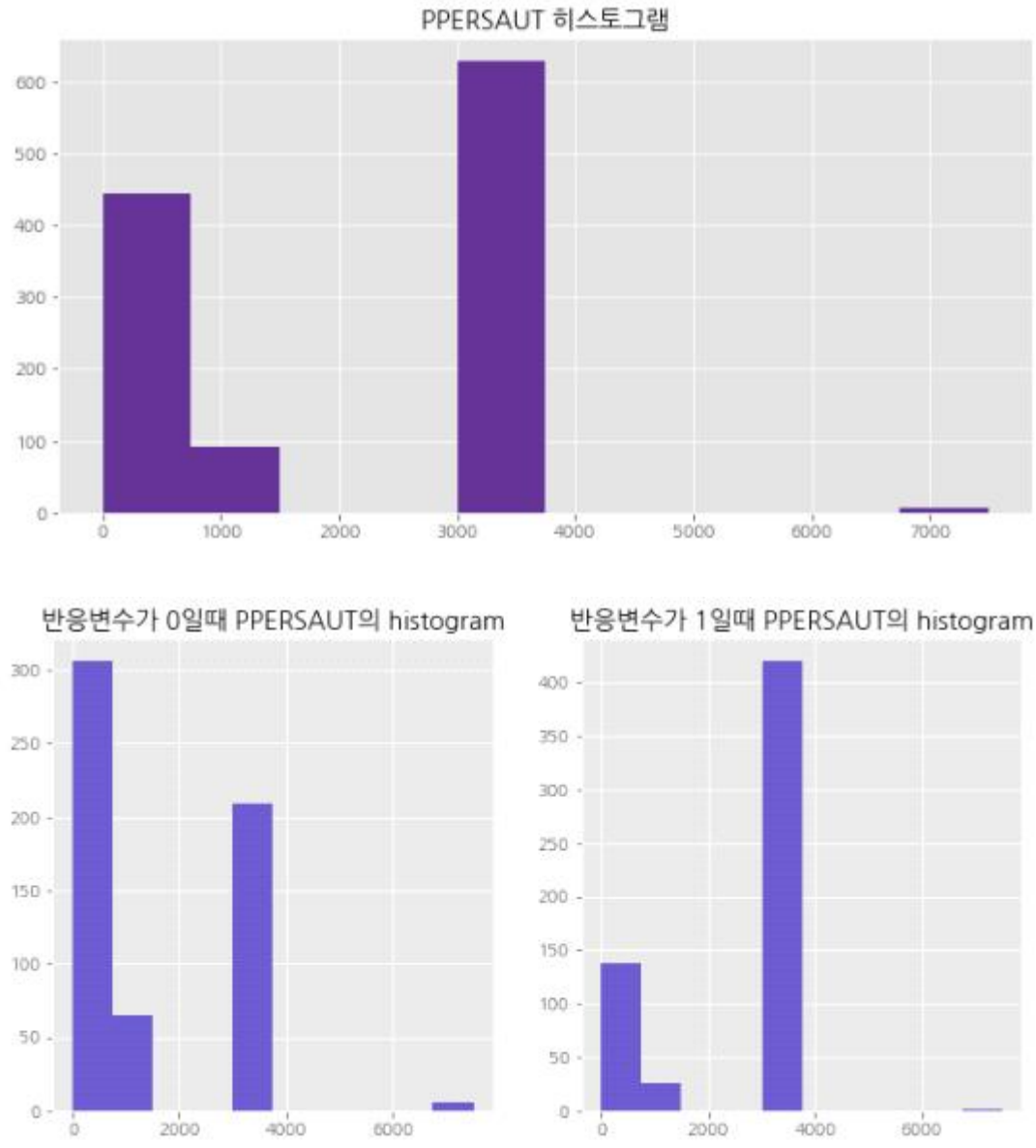


그림 12

'PPERSUAT: Contribution delivery van polices' 변수는 고객이 지불한 차 보험금을 나타낸다. 반응변수가 0인 경우와 반응변수가 1인 경우의 분포 차이를 확인할 수 있다. 반응변수가 1일때, 히스토그램을 보면 지불한 차 보험금이 3000인 경우가 제일 많다. odds를 볼 때, 지불한 차 보험금이 3000일 때, odds가 0.1203으로 0.0638(=6/94) (반응변수 비율고려한 odds)보다 큰 것으로 보아 CARAVAN보험을 예측할 때, 지불한 차 보험금이 3000인 경우는 중요한 요인이 될 수 있다고 본다.

② ‘APERSAUT: Number of car’

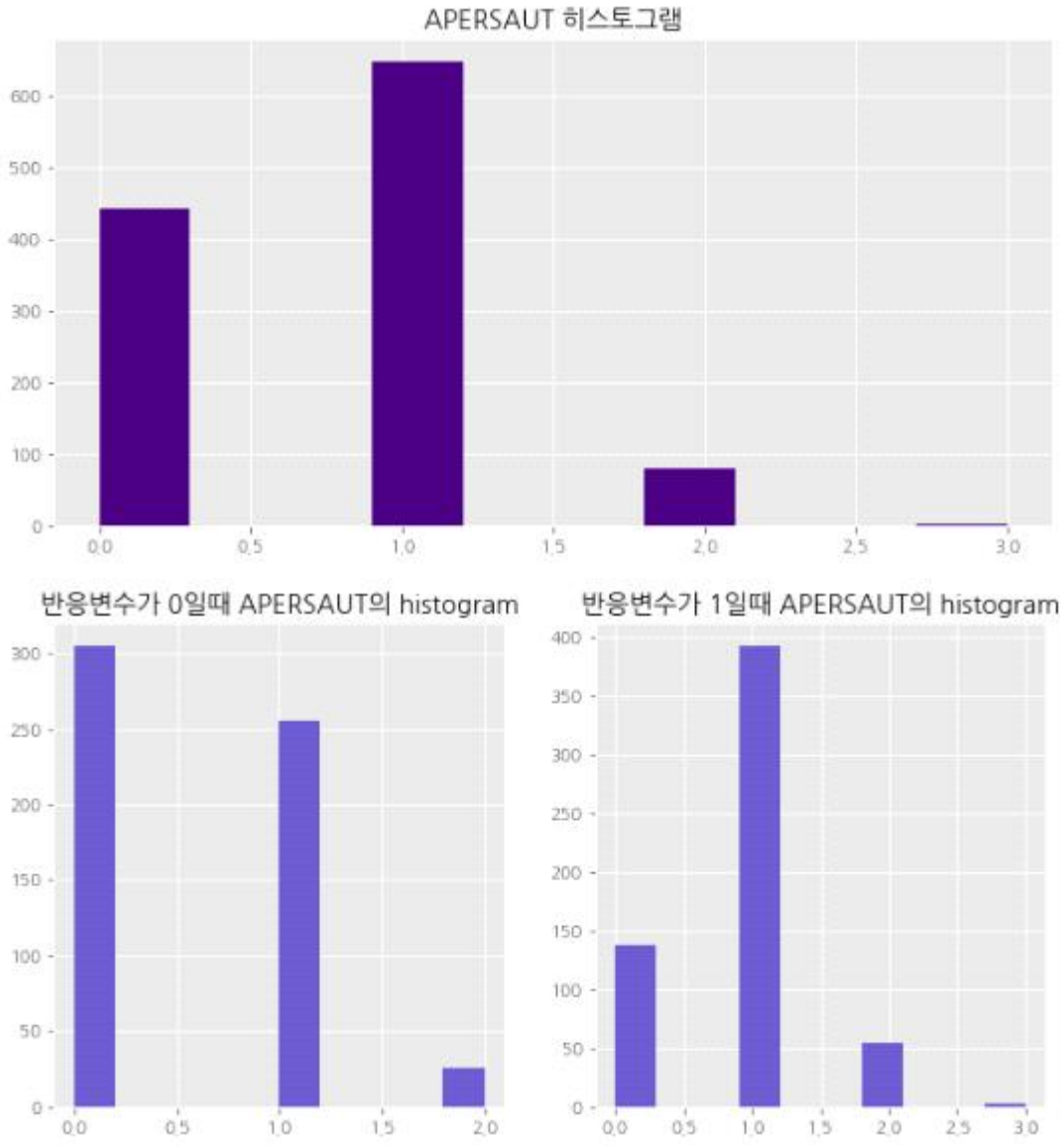


그림 13

'APERSAUT: Number of Car polices' 변수는 고객이 소유한 차 보험의 수를 나타낸다. 반응변수가 0인 경우와 반응변수가 1인 경우 분포에 있어 차이를 확인할 수 있다. 반응변수가 1일 때, 히스토그램을 보면 소유한 차 보험의 수가 1인 경우가 제일 많다. odds를 볼 때, 소유한 차 보험수가 1, 2일 때, odds가 0.0936, 0.1636으로 0.0638(=6/94)보다 큰 것으로 보아 CARAVAN보험을 예측할 때, 소유한 차 보험수가 1, 2인 경우는 중요한 요인이 될 수 있다고 본다.

③ ‘PWAPART: Contribution private third party insurance’

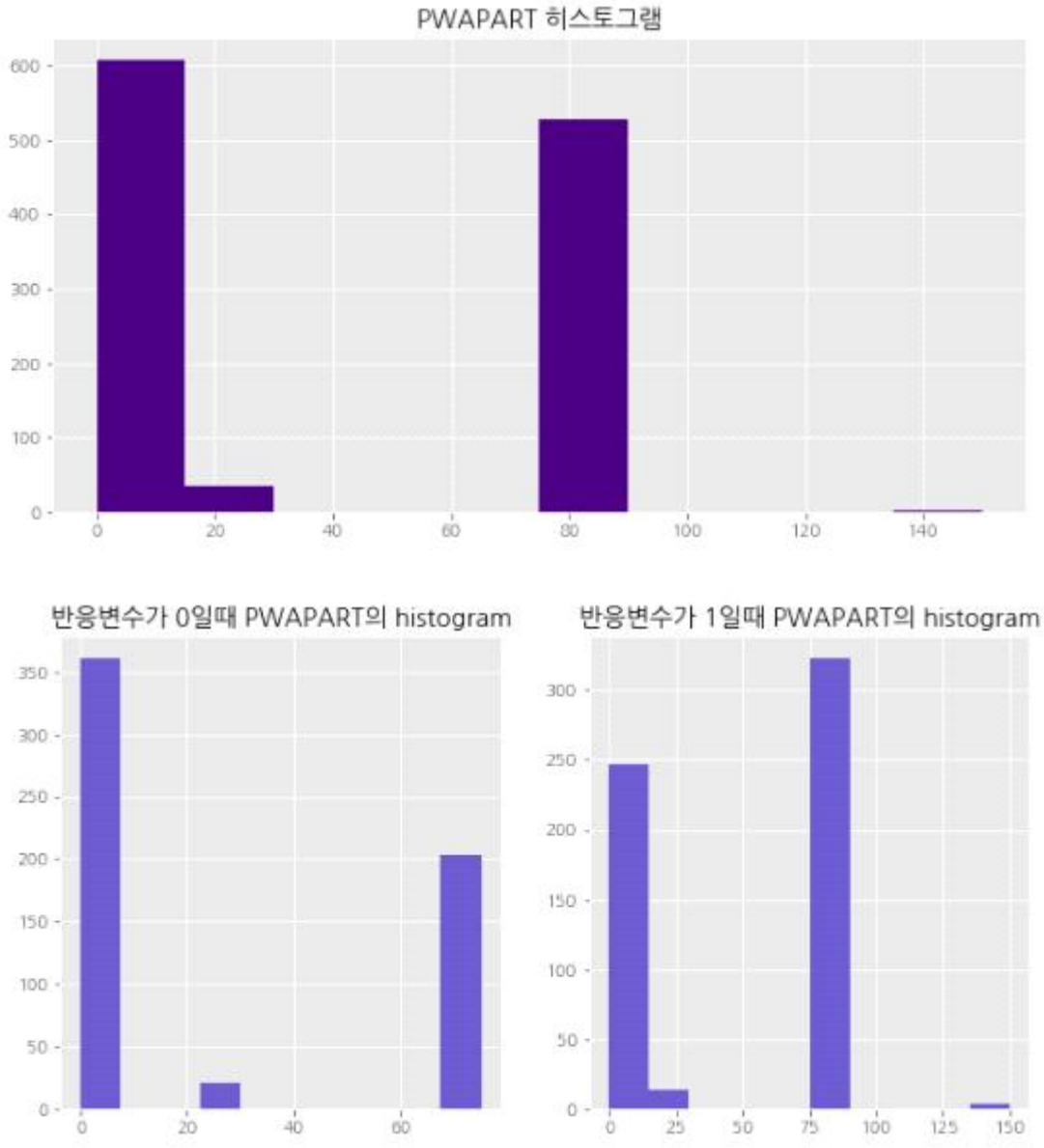


그림 14

'PWAPART: Contribution private third party insurance' 변수는 고객이 지불한 제 3자 보험금을 나타낸다. 반응변수가 0인 경우와 반응변수가 1인 경우 분포에 있어 차이를 확인할 수 있다. 반응변수가 1일때, 히스토그램을 보면 지불한 제 3자 보험금이 75인 경우가 제일 많다. odds를 볼 때, 지불한 제 3자 보험금이 75일때, odds가 0.0997로 0.0638(=6/94)보다 큰 것으로 보아 CARAVAN보험을 예측할 때, 지불한 제 3자 보험금이 75인 경우는 중요한 요인이 될 수 있다고 본다.

④ ‘AWAPART: Number of private third party insurance’

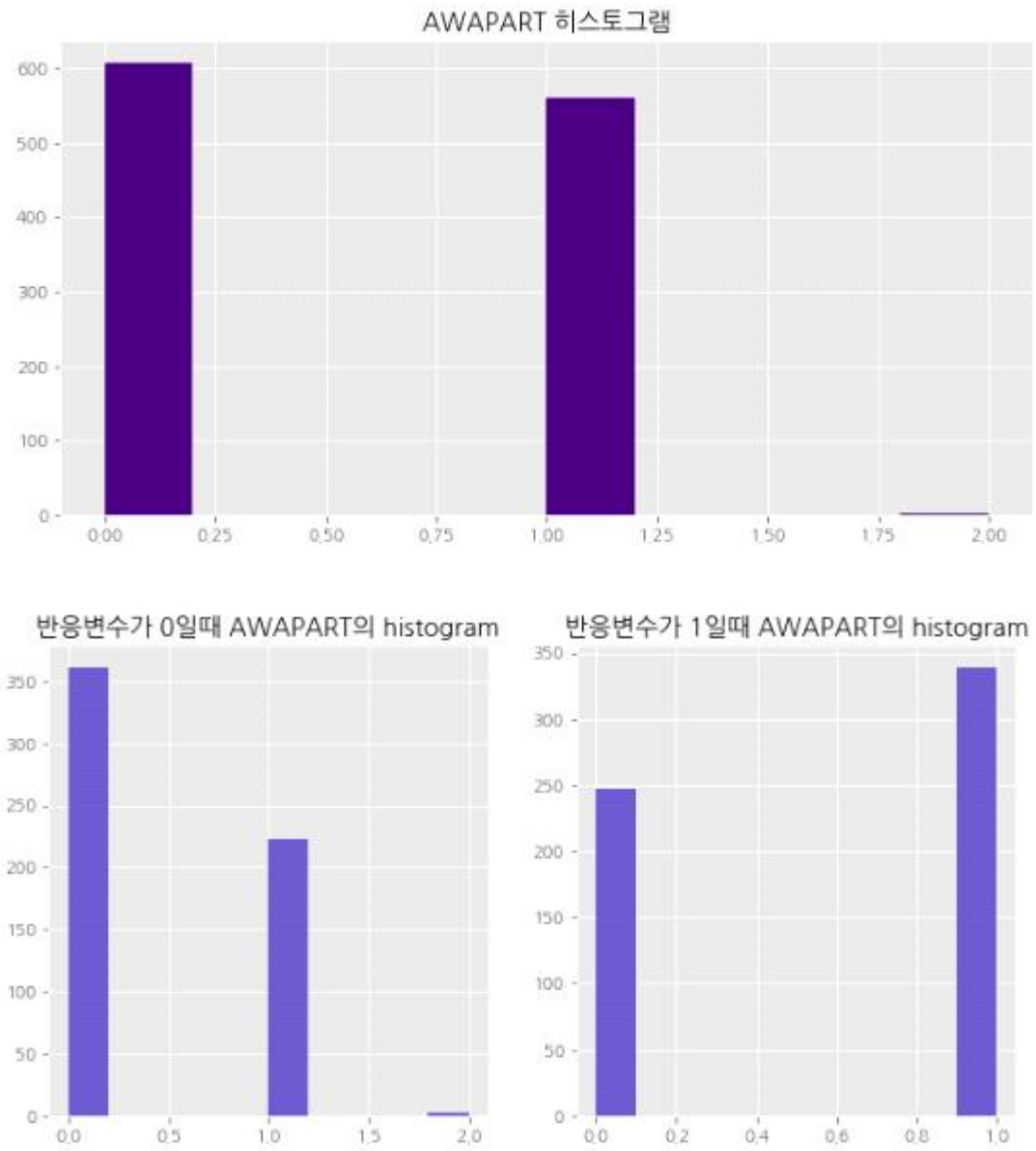


그림 15

'AWAPART: Number of private third party insurance' 변수는 고객이 소유한 제 3자 보험의 수를 나타낸다. 반응변수가 0인 경우와 반응변수가 1인 경우 분포에 있어 차이를 확인할 수 있다. 반응변수가 1일때, 히스토그램을 보면 소유한 제 3자 보험수가 1인 경우가 제일 많다. odds를 볼 때, 소유한 제 3자 보험의 수가 1일때, odds가 0.0949로 0.0638(=6/94)보다 큰 것으로 보아 CARAVAN보험을 예측할 때, 소유한 제 3자 보험의 수가 1인 경우는 중요한 요인이 될 수 있다고 본다.

⑤ ‘MKOOPKLA: Purchasing power class’

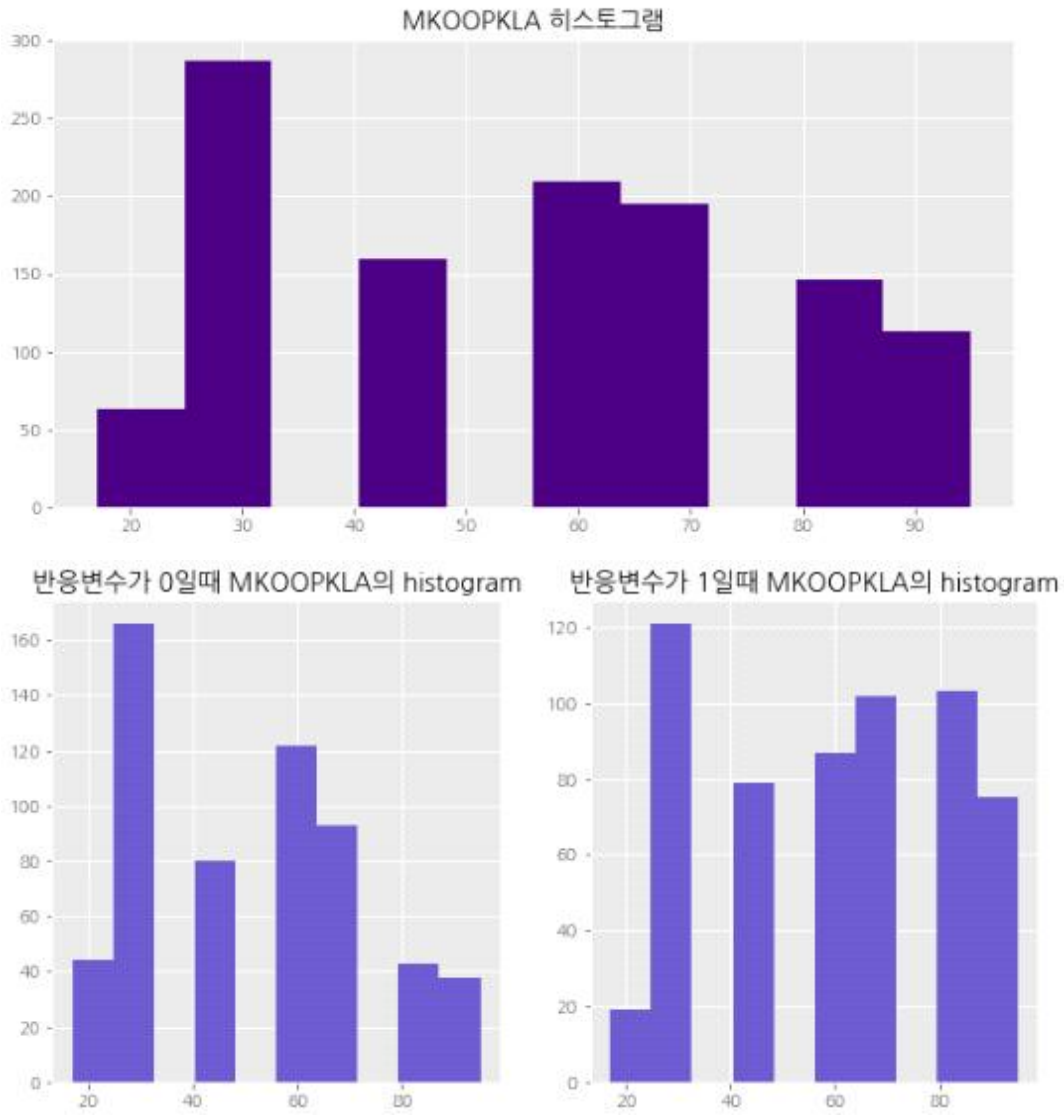


그림 16

'MKOOPKLA: Purchasing power class' 변수는 고객이 속한 지역의 구매력을 나타내고 있다. 반응변수가 0인 경우와 반응변수가 1인 경우 분포에 있어 차이를 확인할 수 있다. 반응변수가 1일때, 히스토그램을 보면 고객이 속한 지역의 구매력이 클수록 즉, 82, 95인 경우가 제일 많다. odds를 볼 때, 구매력의 값이 82, 95일 때, odds가 각 0.1528, 0.1145로 0.0638(=6/94)보다 큰 것으로 보아 CARAVAN보험을 예측할 때, 구매력이 높은 지역은 중요한 요인이 될 수 있다고 본다.

⑥ ‘MAUT1: 1 car’

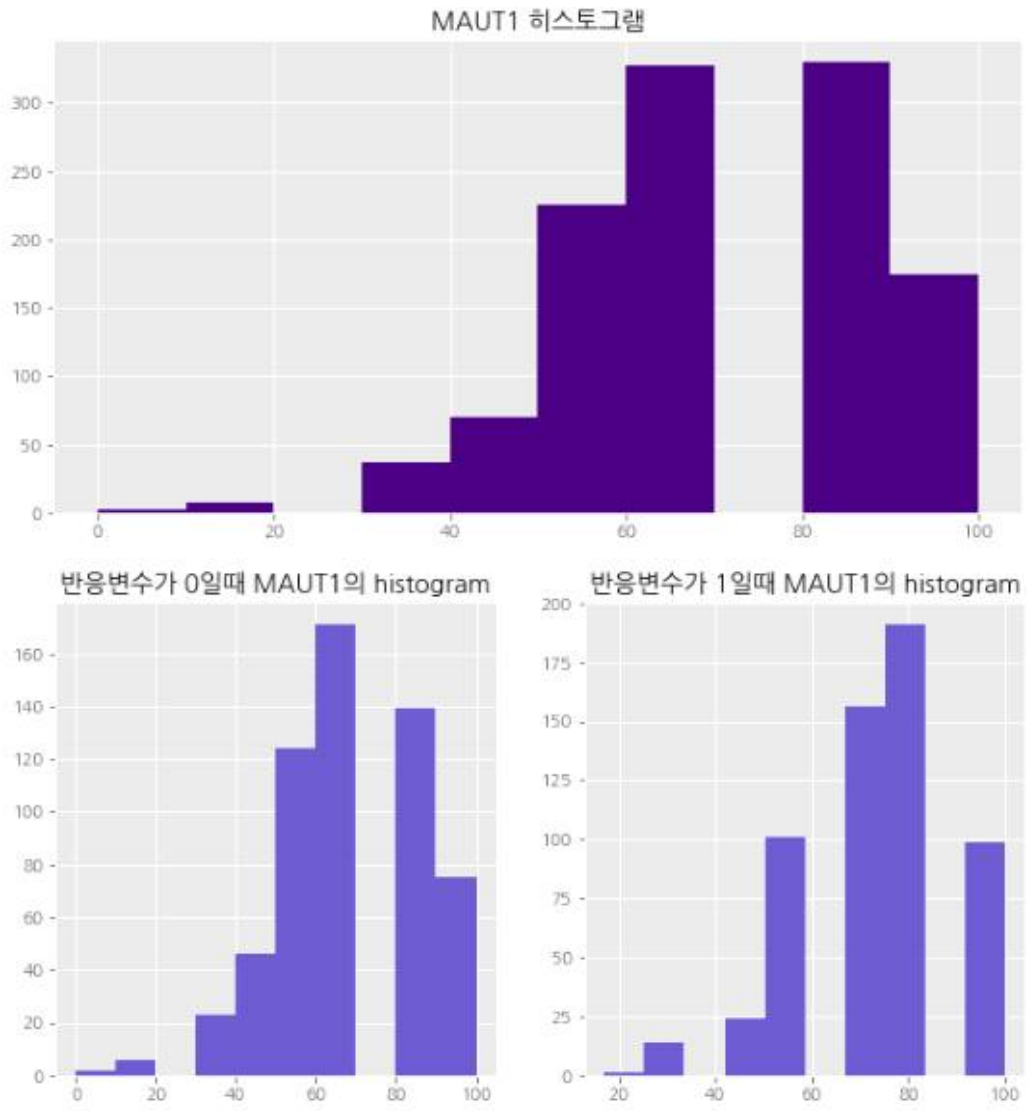


그림 17

'MAUT1: 1 car' 변수는 고객이 속한 지역의 차 1개를 가진 비율을 나타낸다. 반응변수가 0인 경우와 반응변수가 1인 경우 분포에 있어 차이를 확인할 수 있다. 반응변수가 1일때, 히스토그램을 보면 고객이 속한 지역의 차 1개를 가진 비율이 82인 경우가 제일 많다. odds를 볼 때, 차 1개를 가진 비율이 82, 100일 때, odds가 각 0.0896, 0.0965로 0.0638(=6/94)보다 큰 것으로 보아 CARAVAN보험을 예측할 때, 차 1개를 보유한 비율이 높은 지역은 중요한 요인이 될 수 있다고 본다.

⑦ ‘MRELGE: Married’

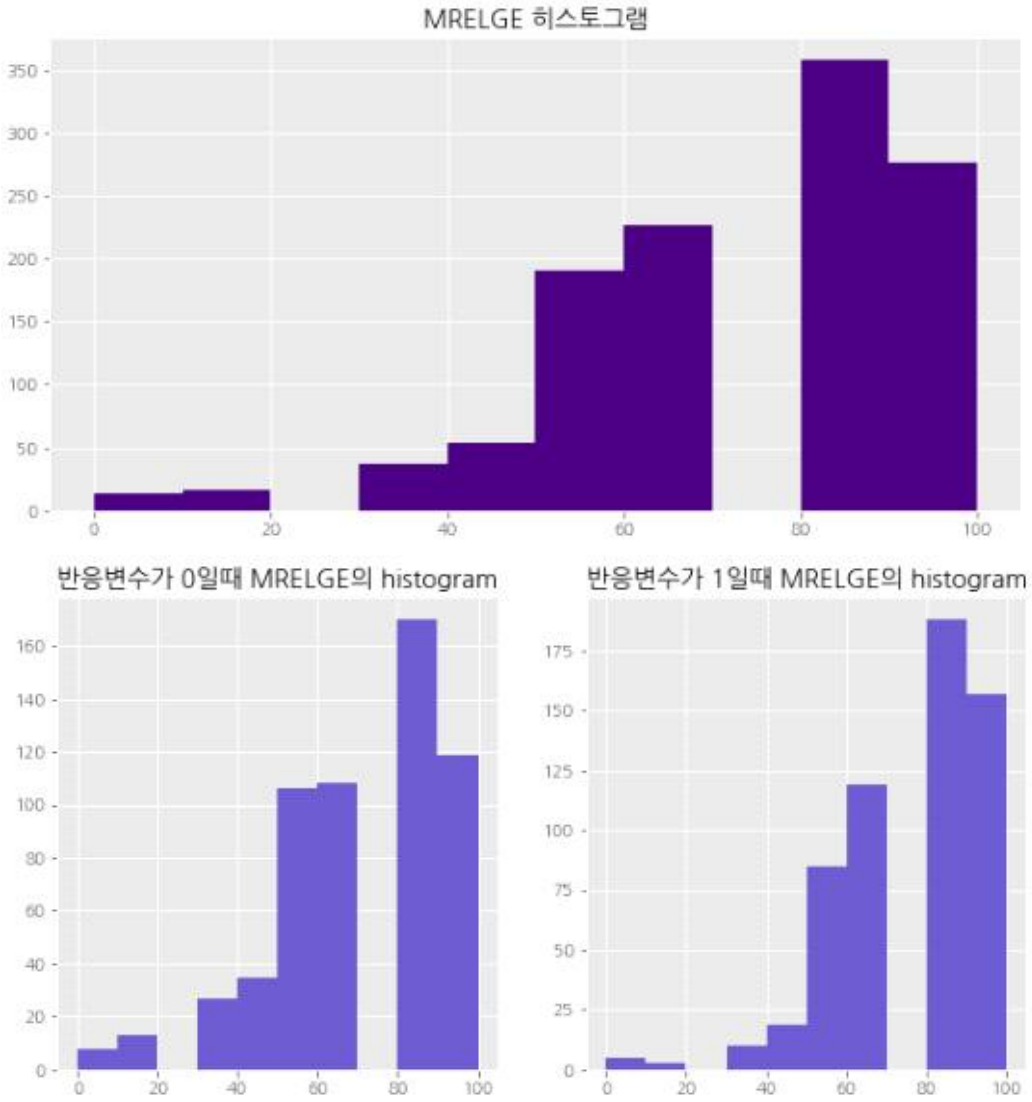


그림 18

'MRELGEL: Married' 변수는 고객이 속한 지역의 결혼 사람들의 비율을 나타낸다. 반응변수가 1일때, 히스토그램을 보면 고객이 속한 지역의 결혼 사람들의 비율이 95인 경우가 제일 많다. odds를 볼 때, 결혼 사람의 비율이 95일 때, odds가 0.0806으로 0.0638(=6/94)보다 큰 것으로 보아 CARAVAN보험을 예측할 때, 결혼 사람들의 비율이 높은 지역은 중요한 요인이 될 수 있다고 본다.

⑧ ‘MINKGEM: Average income’

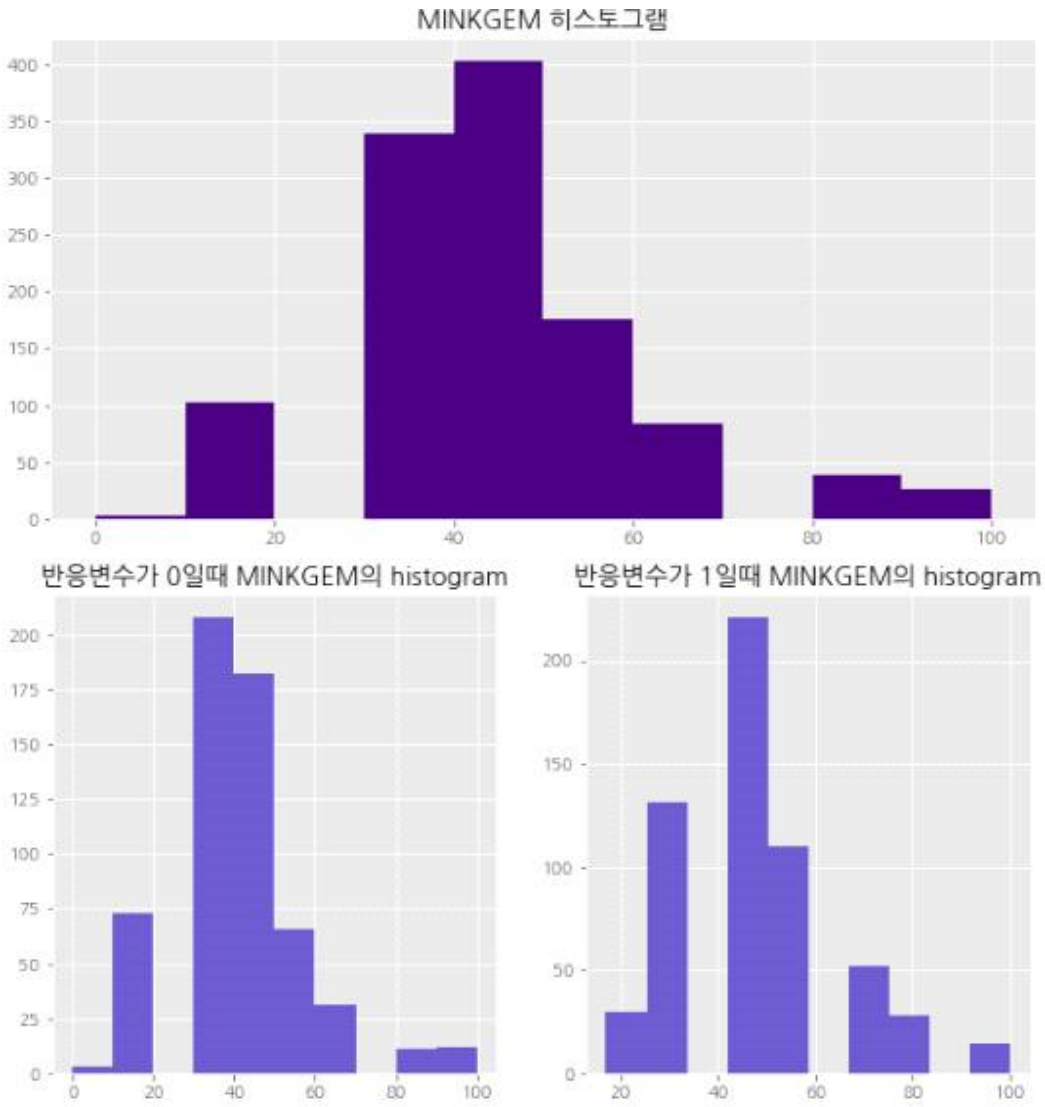


그림 19

'MINKGEM: Average income' 변수는 고객이 속한 지역의 평균 수입 수준을 나타낸다. 반응변수가 1일때, 히스토그램을 보면 고객이 속한 지역의 평균 수입 수준이 43인 경우가 제일 많다. odds를 볼 때, 평균 수입 수준이 43인 경우가 제일 많다. odds를 볼 때, 평균 수입 수준이 43, 56, 69, 82, 95일 , odds값이 모두 0.0638(=6/94)보다 큰 것으로 보아 CARAVAN보험을 예측할 때, 그 지역의 평균 소득 수준이 중, 상인 지역은 중요한 요인이 될 수 있다고 본다.

반응변수와의 직접적인 상관관계를 보았을때 자동차보험 비용, 자동차 보험개수, 제 3자 보험 비용, 제 3자 보험 개수, 구매력, 차 소유 여부, 화재 보험 가입여부 등이 유의한 상관관계를 가졌으리라 유추할 수 있다. 하지만 동시에 유의한 변수들 사이에 공통요인이 보이는 만큼 (차 소유여부, 구매력등) 상관관계가 많으리라 유추해 볼 수 있으므로 최대한 중요한 변수들의 정보를 잃지 않으면서 변수간의 공분산은 최소화 하는 변수선택을 할 필요가 있다.

ⅱ. 데이터 변환 및 전처리

(1) 순위형 변수 변환

앞서 우리의 데이터에 60개의 순위형 변수가 있다고 언급했다. 60개 순위형 변수 중 'MINKGEM: Averae income' 변수와 'MKOOPKLA: Purchasing power class' 변수를 제외한 58개의 변수에 대하여 다음과 같은 중간값 변환을 진행하고 연속형 변수로 간주한다.

- 'MGEMLEEF: Avg age'

원자료		변환
순위형	label	
1	20-30 years	25
2	30-40 years	35
3	40-50 years	45
4	50-60 years	55
5	60-70 years	65
6	70-80 years	75

- 'MGODRK: Roman catholic' ~ 'MKOOPKLA: Purchasing power class' (L3)

원자료		변환
순위형	label	
0	0%	0
1	1-10%	5
2	11-23%	17
3	24-36%	30
4	37-49%	43
5	50-62%	56
6	63-75%	69
7	76-88%	82
8	89-99%	95
9	100%	100

- 'PWAPART: Contribution private third party insurance'
  ~ 'PBYSTAND: Contribution social security insurance polices' (L4)

원자료		변환
순위형	label	
0	0	0
1	1-49	25
2	50-99	75
3	100-199	150
4	200-499	350
5	500-999	750
6	1000-4999	3000
7	5000-9999	7500
8	10000-19999	15000
9	20000-	20000

(2) 명목형 변수 처리

우리는 앞서 관심변수 'CARAVAN: Number of mobile home policies' 를 제외한 2개의 명목형 변수가 있음을 설명하고 두 변수의 분포를 히스토그램으로 확인했다. 'MOSTYPE: Customer Subtype' 과 'MOSHOOFD: Customer main type' 은 각각 41개, 10개의 범주를 가지고 있다. 그리고 'MOSHOOFD'의 10개의 범주는 다음과 같이 'MOSTYPE'의 41개의 범주로 세분화된다.

우리는 각 범주에 대한 더미변수를 만들어 준 후 분석하고자 한다. 단, 'MOSTYPE'의 41개의 범주 중 몇 개의 범주는 관측값이 매우 작고 ([부록3]에 제시되어 있다.) 그에 반해 'MOSHOOFD'의 범주 중 작은 관측값은 적어도 70개는 넘는다. 따라서 'MOSHOOFD' 변수의 각 범주에 대한 10개의 더미변수만을 사용하고 MOSTYPE 변수는 제거 해 주었다.

MOSHOOFD	MOSTYPE
(1) Successful hedonists	(1) High Income, expensive child (2) Very Important Provincials (3) High status seniors (4) Affluent senior apartments (5) Mixed seniors
(2) Driven Growers	(6) Career and childcare (7) Dinki's (double income no kids) (8) Middle class families
(3) Average Family	(9) Modern, complete families (10) Stable family (11) Family starters (12) Affluent young families (13) Young all american family
(4) Career Loners	(14) Junior cosmopolitan (15) Senior cosmopolitans (16) Students in apartments (17)Fresh masters in the city (18) Single youth (19) Suburban youth
(5) Living well	(20) Etnically diverse (21) Young urban have-nots (22) Mixed apartment dwellers (23) Young and rising (24) Young, low educated
(6) Cruising Seniors	(25) Young seniors in the city (26) Own home elderly (27) Seniors in apartments (28) Residential elderly

(7) Retired and Religeous	(29) Porchless seniors: no front yard (30) Religious elderly singles (31) Low income catholics (32) Mixed seniors
(8) Family with grown ups	(33) Lower class large families (34) Large family, employed child (35) Village families (36) Couples with teens 'Married with children' (37) Mixed small town dwellers
(9) Conservative families	(38) Traditional families (39) Large religous families
(10) Farmers	(40) Large family farms (41) Mixed rurals

다음 그림을 보면 MOSHOOFD가 MOSTYPE에 속한다는 것을 관측값 개수 표를 통해 알 수 있다.

MOSTYPE	1	2	3	4	5	6	7	8	9	10	...	32	33	34	35	36	37	38	39	40	41
MOSHOOFD																					
1	218	148	433	90	70	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	209	72	546	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	460	271	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	...	234	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	...	0	1401	325	362	373	233	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	569	542	0	0
10	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	137	355

그림 20

(3) 데이터 차원축소 - PCA

데이터의 차원이 클 경우, 학습 속도가 느릴 뿐만 아니라 성능 또한 좋지 않을 가능성이 크다. 또한 모형 개발 시, 설명변수들 간의 상관관계가 높은 다중공선성이 존재할 경우 모형이 잘못 만들어지고 해석에도 문제가 생기게 된다. 따라서 설명변수들 간에 내재하는 상관관계, 연관성을 이용해 소수의 주성분으로 차원을 축소하고자 한다.

PCA는 큰 분산을 갖는 방향이 중요한 정보를 담고 있다고 가정한다. 하지만 큰 분산을 갖는 방향이 정말로 우리가 찾고자하는 중요한 방향인지 생각해봐야한다. 즉, 전체 변동을 거의 설명하지 못하고 있는 주성분들이 예측하는데 있어 중요한 역할을 할 수도 있다. 그리고 85개 변수

들에 대해 주성분 분석을 했을 때, 각 주성분에 대한 해석이 어려울 수 있다. 따라서 우리는 85개의 설명변수에 대한 PCA보다 설명변수들을 세분화하여 부분적으로 PCA를 진행하고자 한다.

예를 들어, 'L3' 범주를 따르는 설명변수 중 'MOPLHOOG: High level education', 'MOPLMIDD: Medium level education', 'MOPLLAG: Lower level education'는 고객이 속한 지역에 대하여, 3가지 교육수준에 따른 비율을 나타내고 있고 이 비율의 합은 100이 된다. 이렇게 합이 100이 되는 변수들에 대해 PCA를 진행한다.

	High level education	Medium level education	Lower level education	합계
'L3'범주	2	2	5	
label	11 - 23 %	11 - 23 %	50 - 62 %	100 %

아래에서 보여주는 설명은 'MINKGEM: Averae income' , 'MKOOPKLA: Purchasing power class' 를 제외한 'L3'범주를 따르는 36개의 설명변수를 다음과 같이 세분화 한 후 부분적으로 진행한 PCA결과에 대한 요약이다. 자세한 사항은 부록에 제시해두었으며 36개의 설명변수가 24개의 설명변수로 차원축소 되었다.

- 'MGODRK: Roman catholic' ~ 'MGODGE: No religion'

위 4개의 설명변수는 2개의 주성분으로 차원축소 되었다. 주성분 1,2는 전체 변동의 (71.64)%를 설명하고 있으며 각 주성분의 변수명은 religion1, religion2이다.

- 'MRELGE: Married' ~ 'MRELOV: Other relatioin'

위 3개의 설명변수는 2개의 주성분으로 차원축소 되었다. 주성분 1,2는 전체 변동의 ()%를 설명하고 있으며 각 주성분의 변수명은 married1, married2이다.

- 'MFALLEN: Singles' ~ 'MFWEKIND: Household with children'

위 3개의 설명변수는 2개의 주성분으로 차원축소 되었다. 주성분 1,2는 전체 변동의 ()%를 설명하고 있으며 각 주성분의 변수명은 single1, single2이다.

- 'MOPLHOOG: High level education' ~ 'MOPLLAAG: Lower level education'

위 3개의 설명변수는 2개의 주성분으로 차원축소 되었다. 주성분 1,2는 전체 변동의 ()%를 설명하고 있으며 각 주성분의 변수명은 Edu1, Edu2이다.

- ‘MBERHOOG: High status’ ~ ‘MBERARBO: Unskilled labourers’

위 6개의 설명변수는 4개의 주성분으로 차원축소 되었다. 주성분 1,2,3,4는 전체 변동의 ()%를 설명하고 있으며 각 주성분의 변수명은 job1, job2, job3, job4이다.
- ‘MSKA: Social Class A’ ~ ‘MSKD: Social class D’

위 6개의 설명변수는 3개의 주성분으로 차원축소 되었다. 주성분 1,2,3은 전체 변동의 ()%를 설명하고 있으며 각 주성분의 변수명은 job1, job2, job3, job4이다.

- ‘MHHUUR: Rented house’ ~ ‘MHKOOP: Home owners’

위 2개의 설명변수는 1개의 주성분으로 차원축소 되었다. 주성분 1은 전체 변동의 ()%를 설명하고 있으며 주성분의 변수명은 rent1이다.

- ‘MAUT1: 1 car’ ~ ‘MAUT0: No car’

위 3개의 설명변수는 2개의 주성분으로 차원축소 되었다. 주성분 1,2는 전체 변동의 ()%를 설명하고 있으며 각 주성분의 변수명은 car1, car2이다.

- ‘MZFONDS: National Health Service’ ~ ‘MZPART: Private health insurance’

위 2개의 설명변수는 1개의 주성분으로 차원축소 되었다. 주성분 1은 전체 변동의 ()%를 설명하고 있으며 주성분의 변수명은 insurance1이다.

- ‘MINKM30: Income < 30,000’ ~ ‘MINK123M: Income > 123,000’

위 5개의 설명변수는 3개의 주성분으로 차원축소 되었다. 주성분 1,2,3은 전체 변동의 ()%를 설명하고 있으며 각 주성분의 변수명은 income1, income2, income3이다.

(4) Feature scaling - min-max normalization

앞서 우리는 이상치 탐색 과정에서 설명변수들의 범위가 각각 다름을 볼 수 있었다. 우리 설명변수의 특성상 표준화(Standardiztion)보다는 ‘min-max normalization’이 적합하다고 판단하여 설명변수들의 범위를 [0, 1]로 변환해주었다.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

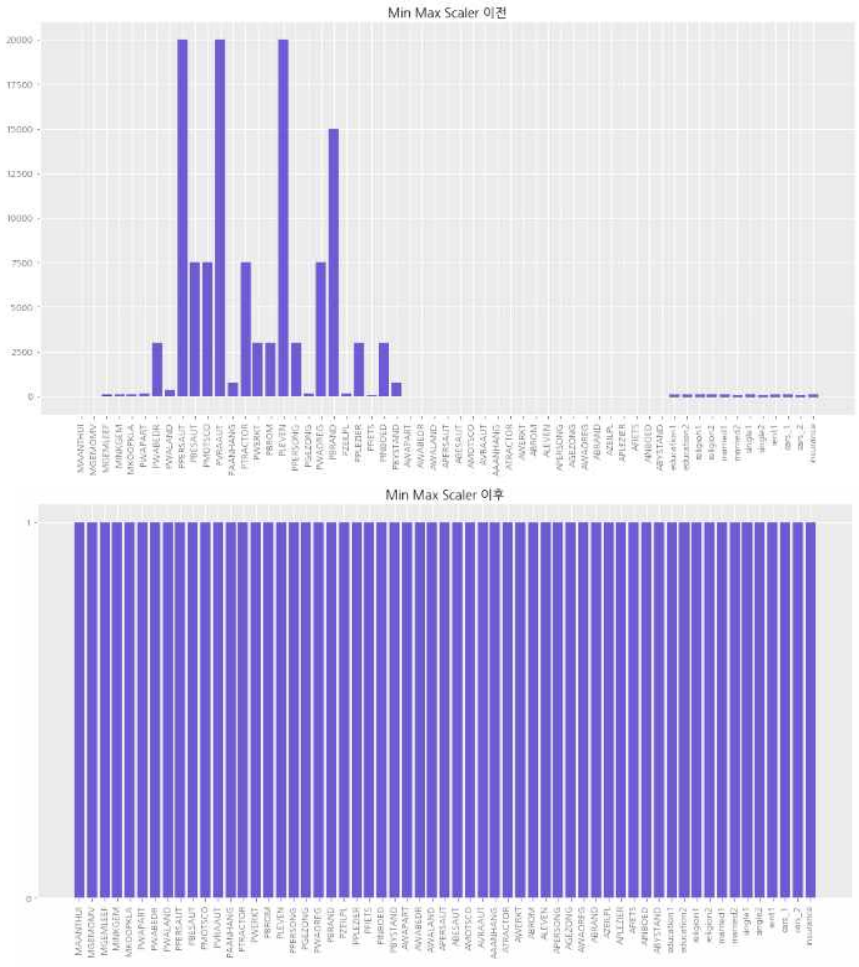


그림 21

iii. Under Sampling

우리는 앞서 앞으로 분석할 데이터가 심한 불균형 데이터임을 확인했다. 심한 불균형 데이터는 소수로 분류된 관찰값에 대해서 전혀 학습이 되지 않거나 더디게 학습된다. 때문에 불균형 데이터 처리는 필수적이다. 우리는 불균형 데이터를 다루기 위해 교수님이 제시해주신 Under Sampling 방법을 사용해 분석하고자 한다. 여기서 작은 관측값의 데이터는 모두 사용하고 큰 관측값의 데이터는 작은 관측값의 데이터 개수와 똑같이 맞춰준다. 즉, 데이터를 1:1로 만들어준다.

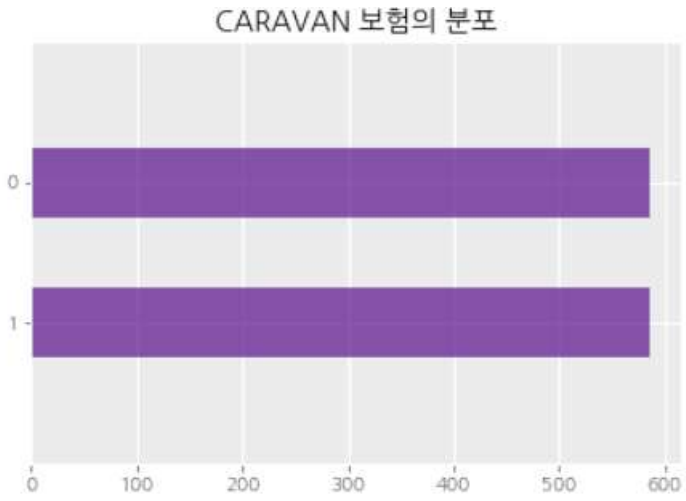


그림 22

새로 만든 Ticdata의 반응변수의 비율을 보면 1대1 로 동일 함을 볼 수 있다.

그리고 우리는 원래 데이터이 분포에 대한 정보를 잃지 않기위해 따로 sample weight값도 변수로 만들어 저장해 두었다.

```
3 print(sampleweight_0) ###반응변수가 0인것의 sample weight
4 print(sampleweight_1) ###반응변수가 1인것의 sample weight
```

```
15.761092150170649
1.0
```

그림 23

iv. **중요변수 선택**

데이터 변환 및 전처리가 끝난 지금 우리는 79개의 설명변수를 가지고 있다. 이는 1172개의 데이터를 학습시키는데 너무 많은 변수이다. 앞서 살펴본 것처럼 반응변수와 직접적인 상관관계를 가지는 변수는 많아 보이나 그와 동시에 변수들 사이에 공통적인 요인이 있어 상관관계가 있음을 유추해 볼 수 있다. 우리는 결정에 있어서 변수의 중요성과 변수와 변수간의 상관성을 같이 고려하면서 차원축소를 했다. 이를 위한 기법으로 randomf forest를 통한 변수중요도 평가, vif 분산팽창계수를 통한 변수 선택, 그리고 AIC 규제 기법을 통한 변수선택을 하고 최종적으로 모델에 사용할 변수들을 선택하였다.

(1) 1차 RandomForest 변수 중요도 평가

우리는 Random Forest 모델의 변수의 중요도 평가를 통해 먼저 1차적으로 변수를 선택하기로 결정했다. Random Forest 중요 변수 평가에서 전체 변수의 중요도의 합은 1이고 변수들은 그에 따른 상대적인 비율로 나타난다. 중요도가 0인 것들은 모두 제거하였으며 다음은 최소한 0.005이상인 변수들이다.

15	PBROM	0.01
1	MGEMOMV	0.01
66	rank_1	0.01
64	job_3	0.01
57	cars_2	0.01
58	insurance	0.01
60	income_2	0.01
36	ABROM	0.01
73	MOSHOOFD_5	0.02
70	MOSHOOFD_2	0.02
67	rank_2	0.02
63	job_2	0.02
41	ABRAND	0.02
53	single1	0.02
52	married2	0.02
78	MOSHOOFD_10	0.02
55	rent1	0.03
51	married1	0.03
3	MINKGEM	0.04
47	education1	0.04
59	income_1	0.04
56	cars_1	0.05
26	AWAPART	0.06
4	MKOOPKLA	0.06
29	APERSAUT	0.08
5	PWAPART	0.08
20	PBRAND	0.09
8	PPERSAUT	0.12

그림 24

위 결과 79개의 설명변수에서 28개의 설명변수로 줄었다.

(2) 다중공선성 제거

앞서 다중공선성을 의심해봐야한다고 언급했다. 1차적으로 변수제거한 후 남은 설명변수들 간의 상관관계가 순위형 변수 변환 및 정규화를 통해 상관관계 조금은 줄었지만 여전히 강한 상관관계를 보이는 변수들이 있다. 따라서 분산팽창지수 측정법을 이용하여 분산팽창계수가 10 이상인 경우, 다중공선성이 존재한다고 판단하여 이 변수들을 제거하고자 한다.

[남은 변수]

	VIF Factor	features
0	3.411410	PBROM
1	6.671478	MGEMOMV
2	8.909051	rank_1
3	6.247957	job_3
4	4.847936	insurance
5	3.454182	ABROM
6	1.384133	MOSHOOFD_5
7	1.613451	MOSHOOFD_2
8	5.545659	rank_2
9	8.751763	job_2
10	4.335823	ABRAND
11	4.452644	married2
12	1.170118	MOSHOOFD_10
13	3.597038	rent1
14	5.505323	married1
15	5.764081	cars_1
16	7.352016	MKOOPKLA
17	8.726838	APERSAUT
18	2.845487	PWAPART
19	1.635778	PBRAND
20	8.770715	PPERSAUT

그림 25

[사라진 변수]

```
1  delcolumns ###사라진 특성들
```

```
{'AWAPART': 33,80084356848073,
 'income_1': 16,95391343749816,
 'education1': 13,3343575969246 99,
 'single1': 12,447583827314713,
 'MINKGEM': 12,2542818254920 22,
 'income_2': 10,800775830087957,
 'cars_2': 10,472816215970743}
```

그림 26

(3) AIC, BIC 기법을 통한 변수 최종선택

위 두 과정을 거친 후 남은 21개의 변수들에 대해 최종적으로 AIC 규제기법을 적용해 backward selection을 통해 최종 변수들을 선택했다.

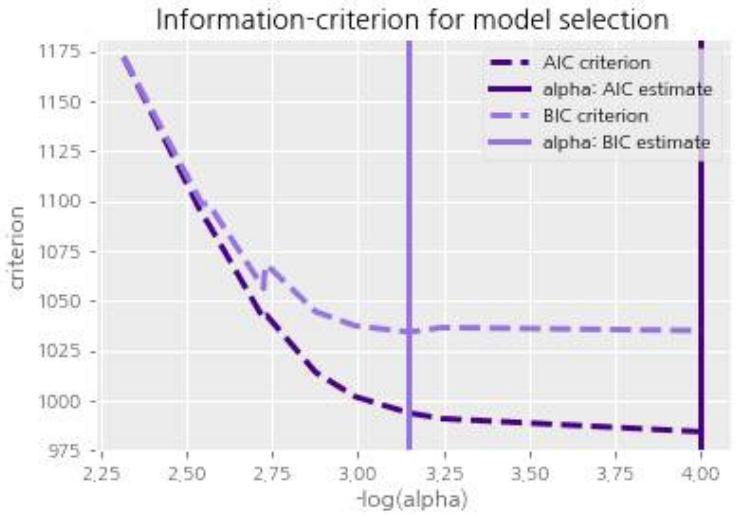


그림 27

```
In [1]: aic_var
Out[1]: Index(['ABROM', 'MOSHOOFD_10', 'MGEMOMV', 'job_2', 'married2', 'insurance',
               'MOSHOOFD_5', 'MOSHOOFD_2', 'rent1', 'ABRAND', 'married1', 'cars_1',
               'PWAPART', 'MKOOPKLA', 'PBRAND', 'PPERSAUT'],
              dtype='object')
```

```
In [1]: bic_var
Out[1]: Index(['ABROM', 'MOSHOOFD_10', 'MOSHOOFD_5', 'MOSHOOFD_2', 'rent1', 'married1',
               'cars_1', 'PWAPART', 'MKOOPKLA', 'PPERSAUT'],
              dtype='object')
```

그림 28

우리는 이중 덜 강력한 규제기법을 적용해 남은 aic_var들을 모델을 훈련시키는데 사용했다. aic 기법을 적용할 때 변수들을 표준화 시켜서 평가 하였으므로 그 계수의 절대값이 그 변수의 영향력을 의미한다고 할 수 있다. 그러므로 우리가 뽑은 변수들의 중요도를 다음과 같이 나타낼 수 있다.

AIC 기법을 통해 사라진 변수들로는 다음과 같다.

```
rank_2          0.000000
APERSAUT        0.000000
job_3           0.000000
rank_1          0.000000
PBROM           0.000000
```

그림 29

그리고 남은 변수들은 다음과 같다.

```
PPERSAUT        0.150990
PWAPART         0.053764
MOSHOOFD_5      0.039220
cars_1          0.035388
MOSHOOFD_10     0.033724
rent1           0.027027
ABROM           0.019995
married1        0.017679
MOSHOOFD_2      0.014854
married2        0.012536
job_2           0.010457
ABRAND          0.007201
MGEMOMV         0.006995
insurance       0.006193
PBRAND          0.005105
MKOOPKLA        0.002960
```

그림 30

IV. 본론2

1. 모형적합 준비단계

먼저 모델들에 대한 소개를 하기 전에 우리가 분석에 사용한 변수를 선택해주자.

```
사용할 변수들
Index(['ABROM', 'MOSHOOFD_10', 'MGEMOMV', 'job_2', 'married2', 'insurance',
       'MOSHOOFD_5', 'MOSHOOFD_2', 'rent1', 'ABRAND', 'married1', 'cars_1',
       'PWAPART', 'MKOOPKLA', 'PBRAND', 'PPERSAUT'],
      dtype='object')
사용할 변수개수 : 16
```

그림 31

그리고 우리는 전체 데이터의 70%를 Training Set으로 그리고 30%를 Validation Set 으로 구분하였다. 이로 인해 나누어진 test set과 validation set 의 크기는 다음과 같다.

```
X and y Input Data:     (1172, 16) (1172,)
Training Set Shape:     (820, 16) (820,)
Validation Set Shape:   (352, 16) (352,)
```

그림 32

그리고 훈련할때나 test할 때 샘플웨이트 값을 담아주기 위해서 sample weight를 담아주는 배열을 만들어준다.

Train Set: VAlidation set: Test set 의 sample_weight 를 담아두는 array 를 만들어준다

```
def weight_array(y):
    a=[]
    for i in range(len(y)):
        if np.hstack(y)[i]==0:
            a.append(sampleweight_0)
        else:
            a.append(sampleweight_1)
    return a
```

```
train_sampleweight=weight_array(y_train)
val_sampleweight=weight_array(y_val)
```

그림 33

모델 적합후 평가에 사용할 이익도표를 그려주기 위한 함수도 따로 정의해 주었다.

이익도표를 만들기 위한 함수를 만들어준다

```
def calc_lift(x,y,clf,bins=10):
    #Actual Value of y
    y_actual = np.hstack(y)
    #Predicted Probability that y = 1
    y_prob = clf.predict_proba(x)
    #Predicted Value of Y
    y_pred = clf.predict(x)
    cols = ['ACTUAL','PROB_POSITIVE','PREDICTED']
    data = [y_actual,y_prob[:,1],y_pred]
    df = pd.DataFrame(dict(zip(cols,data)))
    #Observations where y=1
    total_positive_n = df['ACTUAL'].sum()
    #Total Observations
    total_n = df.index.size
    natural_positive_prob = total_positive_n/float(total_n)
    df['상위구간'] = pd.qcut(df['PROB_POSITIVE'],bins,labels=False, duplicates='drop')
    pos_group_df = df.groupby('상위구간')
    #Percentage of Observations in each Bin where y = 1
    actual=pos_group_df['ACTUAL'].sum().sort_index(ascending=False)
    bin_count=pos_group_df['ACTUAL'].count().sort_index(ascending=False)
    cumsum=np.cumsum(bin_count)
    cumsum_percentage=np.cumsum(bin_count)/np.sum(bin_count)
    lift_positive = pos_group_df['ACTUAL'].sum().sort_index(ascending=False)/pos_group_df['ACTUAL'].count().sort_index(ascending=False
    cum_active=np.cumsum(pos_group_df['ACTUAL'].sum().sort_index(ascending=False))/np.cumsum(pos_group_df['ACTUAL'].count().sort_index
    cum_lift_positive=np.cumsum(pos_group_df['ACTUAL'].sum().sort_index(ascending=False))/np.cumsum(pos_group_df['ACTUAL'].count().sor
    cum_lift_positive=cum_lift_positive/natural_positive_prob
    lift_index_positive = (lift_positive/natural_positive_prob)

    #Consolidate Results into Output Dataframe
    lift_df = pd.DataFrame({ '구간 관측치 개수':bin_count
                            ,'구간 활성화 비율(%)':lift_positive*100 ,
                              '구간 LIFT':lift_index_positive,
                             '구간 실제 구매한 사람 수':actual,
                             '누적 관측치 개수':np.round(cumsum,1),
                            '누적 메일 보낸 비율':np.round(cumsum_percentage,1),
                            '누적 활성화 비율 %':np.round(cum_active,2),
                          '누적 실제 구매한 사람수':np.cumsum(actual),
                          '누적 lift (%)':cum_lift_positive})
    lift_df.index=[1,2,3,4,5,6,7,8,9,10]
    lift_df.index.name='구간'
    return lift_df
```

그림 34

우리는 로지스틱 , 나이브 베이즈 분류기, 신경망 모델 적합, 랜덤포레스트, 아다부스트 , 그래디언트 부스트, 서포트 벡터 머신을 적합시키고 각 모델의 성능을 ROC 커브, PR 커브 그리고 이익도표를 통해서 시각화하고 다른 모델들과의 비교 지표로 AUC, AP score을 제시하겠다.

2. 모형적합

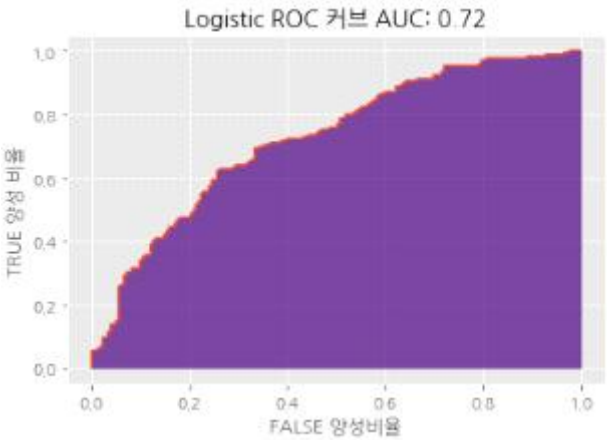
[Logistic Regression]

먼저 Logistic Regression 적합을 하겠다. logistic 모형에 적합한 이후 각 변수별 모형 계수를 나열하면 다음과 같다.

	0	1
	ABROM	-1,546908
	MOSHOOFD_10	-1,453082
	cars_1	-1,148155
	married1	-1,020027
	MOSHOOFD_5	-0,896549
	married2	-0,833053
	job_2	-0,751103
	rent1	-0,520202
	PBRAND	-0,516569
	MGEMOMV	-0,119983
	MKOOPKLA	0,071469
	ABRAND	0,085347
	insurance	0,135476
	MOSHOOFD_2	0,317492
	PWAPART	1,130559
	PPERSAUT	5,159477

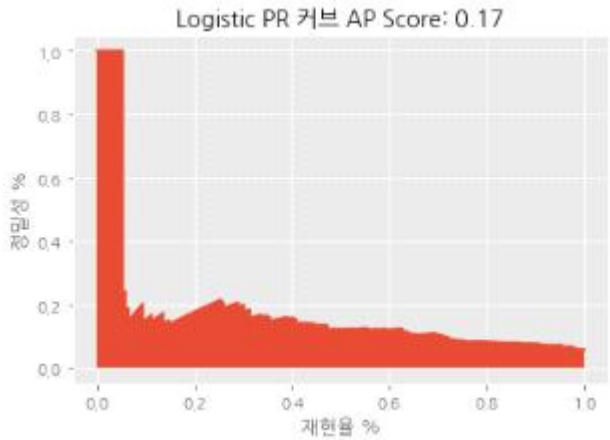
그림 35

ROC 커브



Logisitic Regression 모델의 ROC커브는 다음과 같다. AUC의 점수는 0.72이다.

PR 커브



Logisitic Regression 모델의 PR커브는 다음과 같다. AP의 점수는 0.17이다.

train 이익도표

구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	82	84.146341	1.654676	69	82	0.1	0.84	69	1.654676
2	81	75.308642	1.480889	61	163	0.2	0.80	130	1.568316
3	83	71.084337	1.397822	59	246	0.3	0.77	189	1.510791
4	82	60.975610	1.199041	50	328	0.4	0.73	239	1.432854
5	82	63.414634	1.247002	52	410	0.5	0.71	291	1.395683
6	82	48.780488	0.959233	40	492	0.6	0.67	331	1.322942
7	82	35.365854	0.695444	29	574	0.7	0.63	360	1.233299
8	82	36.585366	0.719424	30	656	0.8	0.59	390	1.169065
9	82	26.829268	0.527578	22	738	0.9	0.56	412	1.097788
10	82	6.097561	0.119904	5	820	1.0	0.51	417	1.000000

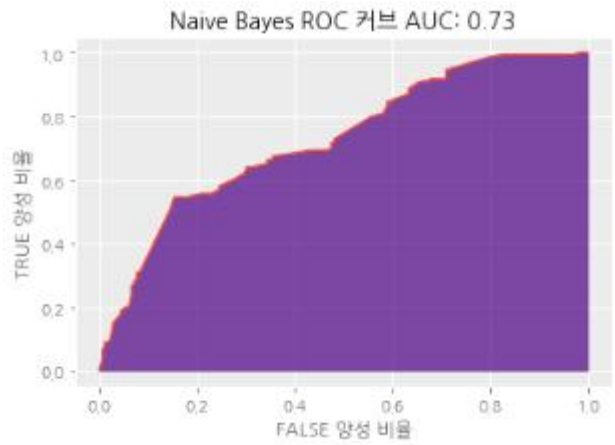
test이익도표

구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	36	72.222222	1.504274	26	36	0.1	0.72	26	1.504274
2	35	77.142857	1.606762	27	71	0.2	0.75	53	1.554796
3	35	62.857143	1.309214	22	106	0.3	0.71	75	1.473708
4	35	62.857143	1.309214	22	141	0.4	0.69	97	1.432876
5	35	51.428571	1.071175	18	176	0.5	0.65	115	1.360947
6	35	31.428571	0.654607	11	211	0.6	0.60	126	1.243781
7	35	42.857143	0.892646	15	246	0.7	0.57	141	1.193823
8	35	37.142857	0.773626	13	281	0.8	0.55	154	1.141485
9	35	31.428571	0.654607	11	316	0.9	0.52	165	1.087559
10	36	11.111111	0.231427	4	352	1.0	0.48	169	1.000000

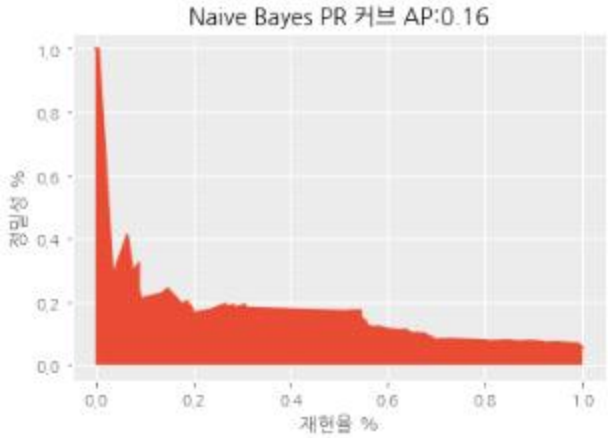
validation set의 가장 좋은 십분위(1)의 구간 LIFT 값(1.5043)은 가장 좋지 않은 십분위(10)의 구간 LIFR 값(0.2314)보다 약 6.5008배정도 크다.

[Naive BAYES CLASSIFIER 분류기]

ROC 커브



PR 커브



train 이익도표

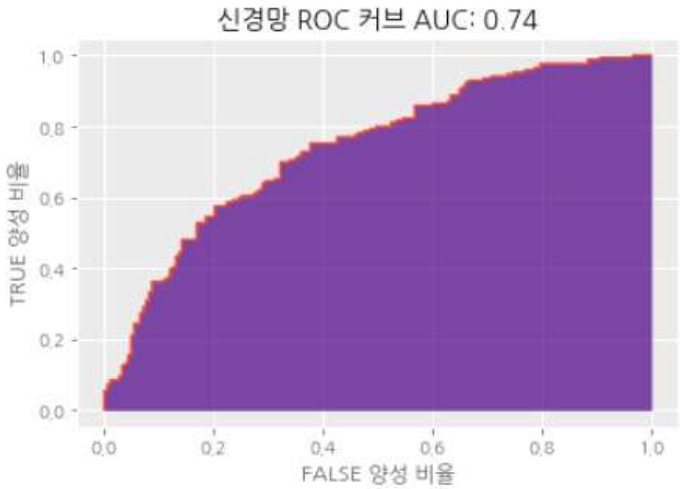
구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	81	79.012346	1.553720	64	81	0.1	0.79	64	1.553720
2	83	77.108434	1.516281	64	164	0.2	0.78	128	1.534772
3	12	58.333333	1.147082	7	176	0.2	0.77	135	1.508339
4	149	65.771812	1.293355	98	325	0.4	0.72	233	1.409777
5	85	58.823529	1.156722	50	410	0.5	0.69	283	1.357314
6	82	40.243902	0.791367	33	492	0.6	0.64	316	1.262990
7	72	50.000000	0.983213	36	564	0.7	0.62	352	1.227274
8	89	25.842697	0.508178	23	653	0.8	0.57	375	1.129265
9	49	20.408163	0.401312	10	702	0.9	0.55	385	1.078453
10	118	27.118644	0.533268	32	820	1.0	0.51	417	1.000000

test 이익도표

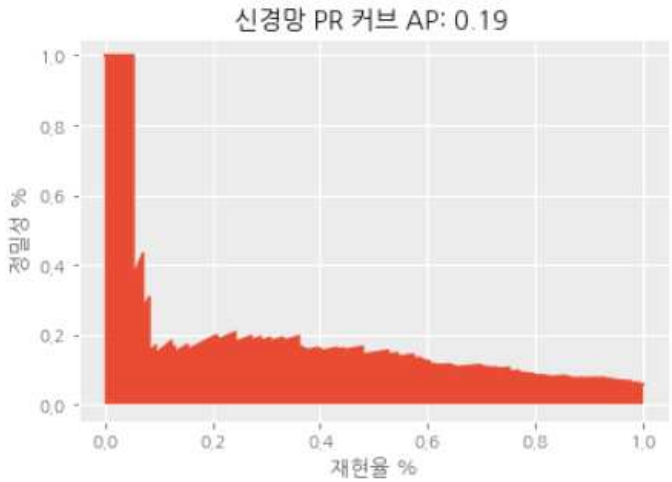
구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	36	72.222222	1.504274	26	36	0.1	0.72	26	1.504274
2	35	77.142857	1.606762	27	71	0.2	0.75	53	1.554796
3	35	62.857143	1.309214	22	106	0.3	0.71	75	1.473708
4	35	62.857143	1.309214	22	141	0.4	0.69	97	1.432876
5	35	51.428571	1.071175	18	176	0.5	0.65	115	1.360947
6	35	31.428571	0.654607	11	211	0.6	0.60	126	1.243781
7	35	42.857143	0.892646	15	246	0.7	0.57	141	1.193823
8	35	37.142857	0.773626	13	281	0.8	0.55	154	1.141485
9	35	31.428571	0.654607	11	316	0.9	0.52	165	1.087559
10	36	11.111111	0.231427	4	352	1.0	0.48	169	1.000000

[신경망 모델 적합]

ROC 커브



PR 커브



Train set 이익도표

구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	82	91.463415	1.798561	75	82	0.1	0.91	75	1.798561
2	82	79.268293	1.558753	65	164	0.2	0.85	140	1.678657
3	82	64.634146	1.270983	53	246	0.3	0.78	193	1.542766
4	82	69.512195	1.366906	57	328	0.4	0.76	250	1.498801
5	82	57.317073	1.127098	47	410	0.5	0.72	297	1.424460
6	82	54.878049	1.079137	45	492	0.6	0.70	342	1.366906
7	82	37.804878	0.743405	31	574	0.7	0.65	373	1.277835
8	82	31.707317	0.623501	26	656	0.8	0.61	399	1.196043
9	82	18.292683	0.359712	15	738	0.9	0.56	414	1.103118
10	82	3.658537	0.071942	3	820	1.0	0.51	417	1.000000

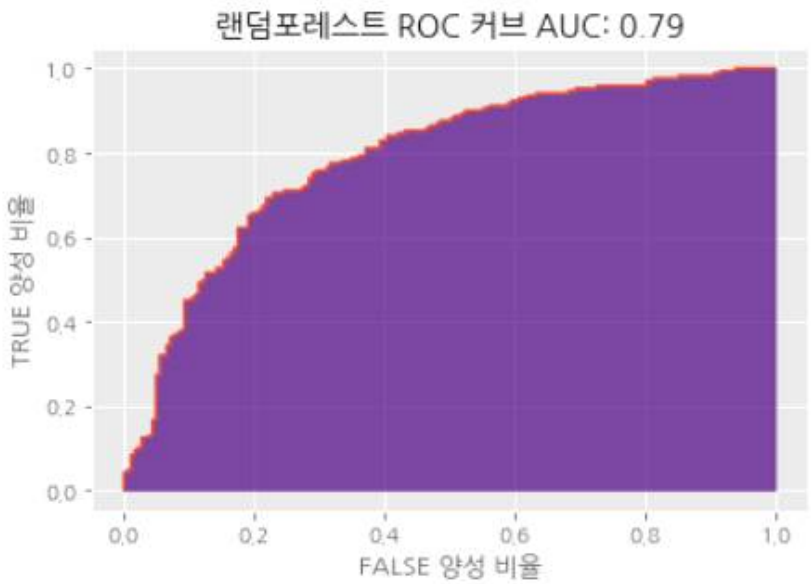
Test set 이익도표

```
calc_lift(X_val,y_val,clf_MLP)
```

구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	36	75.000000	1.562130	27	36	0.1	0.75	27	1.562130
2	35	82.857143	1.725782	29	71	0.2	0.79	56	1.642804
3	35	68.571429	1.428233	24	106	0.3	0.75	80	1.571955
4	35	54.285714	1.130685	19	141	0.4	0.70	99	1.462420
5	35	51.428571	1.071175	18	176	0.5	0.66	117	1.384615
6	35	37.142857	0.773626	13	211	0.6	0.62	130	1.283266
7	35	34.285714	0.714117	12	246	0.7	0.58	142	1.202290
8	35	42.857143	0.892646	15	281	0.8	0.56	157	1.163722
9	35	22.857143	0.476078	8	316	0.9	0.52	165	1.087559
10	36	11.111111	0.231427	4	352	1.0	0.48	169	1.000000

[랜덤포레스트]

ROC 커브



PR 커브

Train Set 이익도표

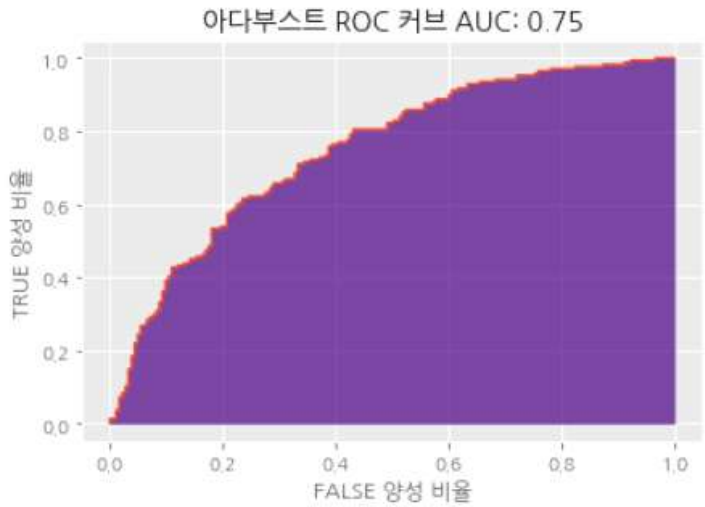
구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	82	89.024390	1.750600	73	82	0.1	0.89	73	1.750600
2	82	70.731707	1.390887	58	164	0.2	0.80	131	1.570743
3	82	71.951220	1.414868	59	246	0.3	0.77	190	1.518785
4	82	62.195122	1.223022	51	328	0.4	0.73	241	1.444844
5	82	53.658537	1.055156	44	410	0.5	0.70	285	1.366906
6	82	56.097561	1.103118	46	492	0.6	0.67	331	1.322942
7	82	40.243902	0.791367	33	574	0.7	0.63	364	1.247002
8	82	19.512195	0.383693	16	656	0.8	0.58	380	1.139089
9	82	21.951220	0.431655	18	738	0.9	0.54	398	1.060485
10	82	23.170732	0.455635	19	820	1.0	0.51	417	1.000000

Test Set 이익도표

구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	36	77.777778	1.619987	28	36	0.1	0.78	28	1.619987
2	35	85.714286	1.785292	30	71	0.2	0.82	58	1.701475
3	35	74.285714	1.547253	26	106	0.3	0.79	84	1.650553
4	35	62.857143	1.309214	22	141	0.4	0.75	106	1.565823
5	35	51.428571	1.071175	18	176	0.5	0.70	124	1.467456
6	35	42.857143	0.892646	15	211	0.6	0.66	139	1.372108
7	35	34.285714	0.714117	12	246	0.7	0.61	151	1.278491
8	35	22.857143	0.476078	8	281	0.8	0.57	159	1.178547
9	35	17.142857	0.357058	6	316	0.9	0.52	165	1.087559
10	36	11.111111	0.231427	4	352	1.0	0.48	169	1.000000

[아다부스트]

ROC 커브



PR 커브

Train 이익도표

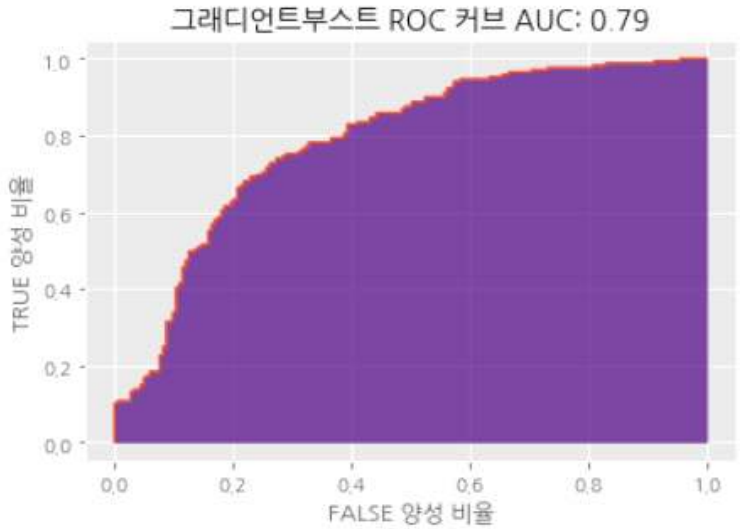
구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	82	92.682927	1.822542	76	82	0.1	0.93	76	1.822542
2	82	82.926829	1.630695	68	164	0.2	0.88	144	1.726619
3	82	71.951220	1.414868	59	246	0.3	0.83	203	1.622702
4	82	70.731707	1.390887	58	328	0.4	0.80	261	1.564748
5	82	58.536585	1.151079	48	410	0.5	0.75	309	1.482014
6	82	53.658537	1.055156	44	492	0.6	0.72	353	1.410871
7	82	40.243902	0.791367	33	574	0.7	0.67	386	1.322371
8	82	23.170732	0.455635	19	656	0.8	0.62	405	1.214029
9	82	14.634146	0.287770	12	738	0.9	0.57	417	1.111111
10	82	0.000000	0.000000	0	820	1.0	0.51	417	1.000000

Test 이익도표

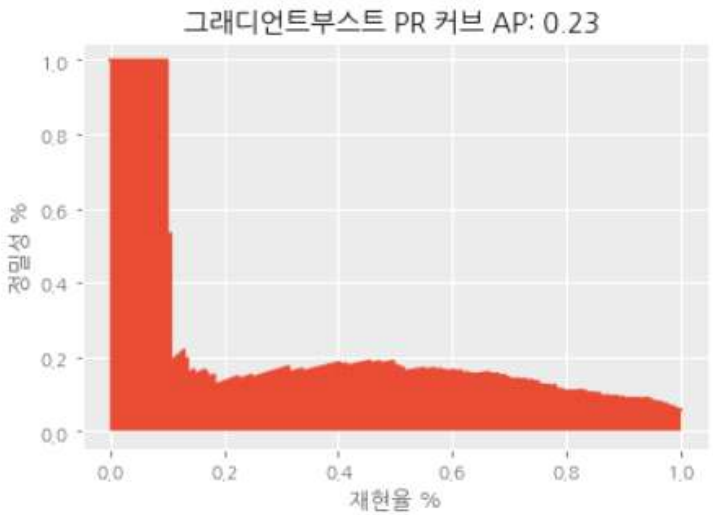
구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	36	80.555556	1.677844	29	36	0.1	0.81	29	1.677844
2	35	74.285714	1.547253	26	71	0.2	0.77	55	1.613468
3	35	62.857143	1.309214	22	106	0.3	0.73	77	1.513007
4	35	65.714286	1.368724	23	141	0.4	0.71	100	1.477192
5	35	45.714286	0.952156	16	176	0.5	0.66	116	1.372781
6	35	48.571429	1.011665	17	211	0.6	0.63	133	1.312880
7	35	34.285714	0.714117	12	246	0.7	0.59	145	1.227690
8	35	37.142857	0.773626	13	281	0.8	0.56	158	1.171134
9	35	20.000000	0.416568	7	316	0.9	0.52	165	1.087559
10	36	11.111111	0.231427	4	352	1.0	0.48	169	1.000000

[그래디언트 부스트]

ROC 커브



PR 커브



Train 이익도표

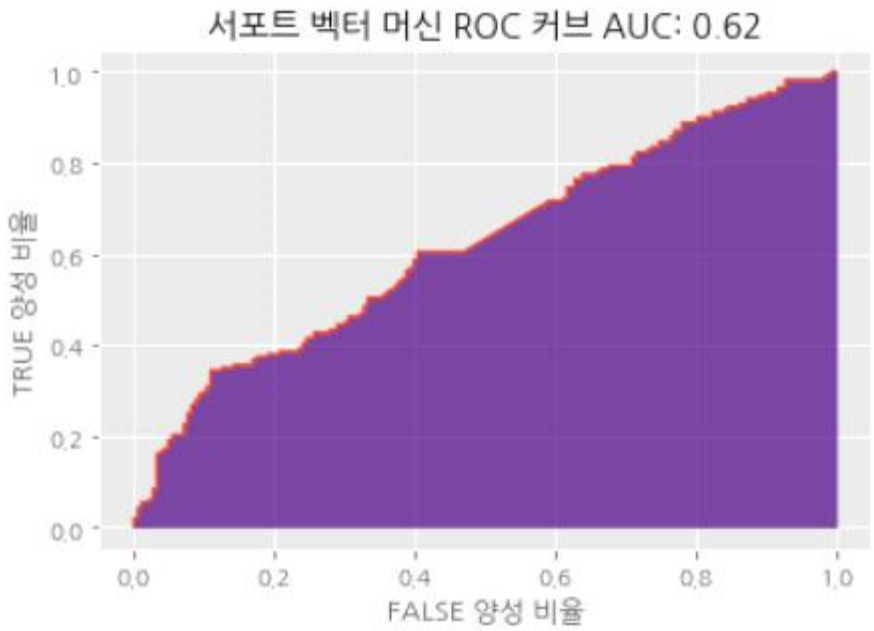
구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	82	100.000000	1.966427	82	82	0.1	1.00	82	1.966427
2	81	93.827160	1.845042	76	163	0.2	0.97	158	1.906107
3	83	81.927711	1.611049	68	246	0.3	0.92	226	1.806555
4	82	70.731707	1.390887	58	328	0.4	0.87	284	1.702638
5	82	56.097561	1.103118	46	410	0.5	0.80	330	1.582734
6	82	34.146341	0.671463	28	492	0.6	0.73	358	1.430855
7	82	30.487805	0.599520	25	574	0.7	0.67	383	1.312093
8	82	30.487805	0.599520	25	656	0.8	0.62	408	1.223022
9	82	9.756098	0.191847	8	738	0.9	0.56	416	1.108447
10	82	1.219512	0.023981	1	820	1.0	0.51	417	1.000000

Test 이익도표

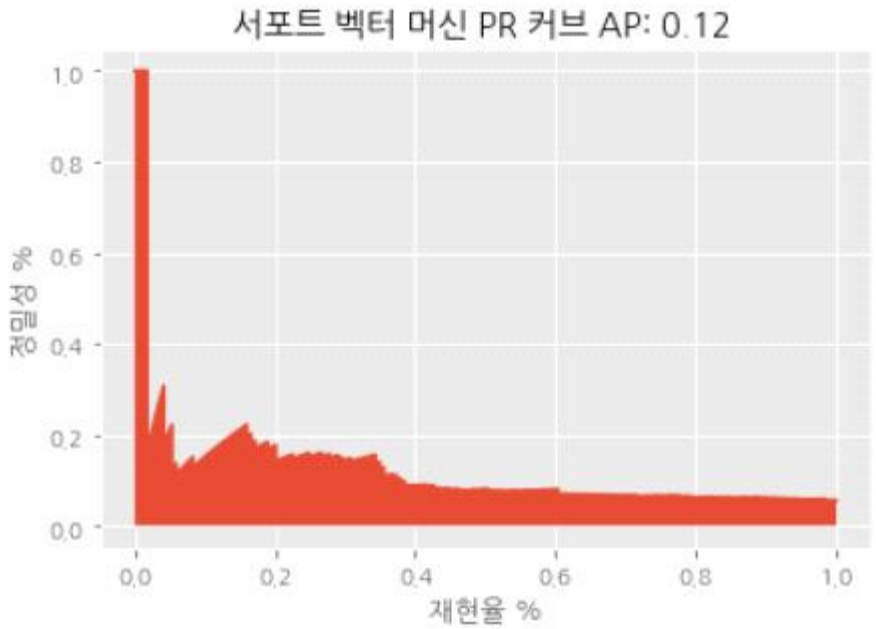
구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	36	75.000000	1.562130	27	36	0.1	0.75	27	1.562130
2	35	74.285714	1.547253	26	71	0.2	0.75	53	1.554796
3	35	85.714286	1.785292	30	106	0.3	0.78	83	1.630903
4	35	62.857143	1.309214	22	141	0.4	0.74	105	1.551051
5	34	58.823529	1.225200	20	175	0.5	0.71	125	1.487743
6	36	38.888889	0.809993	14	211	0.6	0.66	139	1.372108
7	35	31.428571	0.654607	11	246	0.7	0.61	150	1.270025
8	35	31.428571	0.654607	11	281	0.8	0.57	161	1.193371
9	35	14.285714	0.297549	5	316	0.9	0.53	166	1.094150
10	36	8.333333	0.173570	3	352	1.0	0.48	169	1.000000

[서포트 벡터 머신]

ROC 커브



PR 커브



Train 이익도표

구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	81	97.530864	1.917873	79	81	0.1	0.98	79	1.917873
2	83	77.108434	1.516281	64	164	0.2	0.87	143	1.714628
3	82	50.000000	0.983213	41	246	0.3	0.75	184	1.470823
4	82	43.902439	0.863309	36	328	0.4	0.67	220	1.318945
5	82	51.219512	1.007194	42	410	0.5	0.64	262	1.256595
6	21	42.857143	0.842754	9	431	0.5	0.63	271	1.236431
7	143	44.755245	0.880079	64	574	0.7	0.58	335	1.147653
8	82	37.804878	0.743405	31	656	0.8	0.56	366	1.097122
9	82	31.707317	0.623501	26	738	0.9	0.53	392	1.044498
10	82	30.487805	0.599520	25	820	1.0	0.51	417	1.000000

Testset 이익도표

구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	36	77.777778	1.619987	28	36	0.1	0.78	28	1.619987
2	35	68.571429	1.428233	24	71	0.2	0.73	52	1.525460
3	35	37.142857	0.773626	13	106	0.3	0.61	65	1.277213
4	35	45.714286	0.952156	16	141	0.4	0.57	81	1.196525
5	35	60.000000	1.249704	21	176	0.5	0.58	102	1.207101
6	12	0.000000	0.000000	0	188	0.5	0.54	102	1.130052
7	58	46.551724	0.969598	27	246	0.7	0.52	129	1.092221
8	35	40.000000	0.833136	14	281	0.8	0.51	143	1.059951
9	35	40.000000	0.833136	14	316	0.9	0.50	157	1.034829
10	36	33.333333	0.694280	12	352	1.0	0.48	169	1.000000

3. 모형평가의 기준

우리는 모형을 평가할 때 PR Curve의 면적을 나타내는 Average Precision 을 사용하였다. AP 점수를 쓴 이유는 불균형이 심한 데이터일수록 소수의 클래스를 맞추는 것이 쉽지않고 중요하므로 AUC 보다 PR 점수가 중요하다고 볼 수 있다. AUC는 불균형 데이터에 대해서도 점수를 후하게 주는 경향이 있다. 그리고 AP 점수는 이익도표 와도 연관이 깊어 AP 점수가 가장 높은 모형이 이익도표도 가장 좋은 경향이 있다. 그래서 모든 모형들의 AP점수를 나열해보면 다음과 같다.

```
print('     모델들의 Validation AP Score     ')
print('------------------------------')
print('Naive Bayes의                    AP  Score: %s '%ap_score_NB )
print('Neural Network의                 AP  Score: %s '%ap_score_MLP )
print('Logistic Regression의            AP  Score: %s '%ap_score_Log )
print('Random Forest의                  AP  Score: %s '%ap_score_RF )
print('AdaBoost의                       AP  Score: %s '%ap_score_ADA )
print('GradientBoost의                  AP  Score: %s '%ap_score_GB )
print('Support Vector Machine의         AP  Score: %s '%ap_score_SVM )
```

```
     모델들의 Validation AP Score
------------------------------
Naive Bayes의                    AP  Score: 0.16
Neural Network의                 AP  Score: 0.19
Logistic Regression의            AP  Score: 0.17
Random Forest의                  AP  Score: 0.2
AdaBoost의                       AP  Score: 0.15
GradientBoost의                  AP  Score: 0.23
Support Vector Machine의         AP  Score: 0.12
```

위의 그림을 보면 Gradient Boost 모델이 제일 좋은 걸 볼 수 있다. 그러므로 Gradient Boost 모델의 성능을 보기위해 Test Set 에 대한 이익도표를 살펴보면 다음과 같다.

구간	구간 관측치 개수	구간 활성화 비율(%)	구간 LIFT	구간 실제 구매한 사람 수	누적 관측치 개수	누적 메일 보낸 비율	누적 활성화 비율 %	누적 실제 구매한 사람수	누적 lift (%)
1	36	75.000000	1.562130	27	36	0.1	0.75	27	1.562130
2	35	74.285714	1.547253	26	71	0.2	0.75	53	1.554796
3	35	85.714286	1.785292	30	106	0.3	0.78	83	1.630903
4	35	62.857143	1.309214	22	141	0.4	0.74	105	1.551051
5	34	58.823529	1.225200	20	175	0.5	0.71	125	1.487743
6	36	38.888889	0.809993	14	211	0.6	0.66	139	1.372108
7	35	31.428571	0.654607	11	246	0.7	0.61	150	1.270025
8	35	31.428571	0.654607	11	281	0.8	0.57	161	1.193371
9	35	14.285714	0.297549	5	316	0.9	0.53	166	1.094150
10	36	8.333333	0.173570	3	352	1.0	0.48	169	1.000000

상위 10%의 활성화 비율은 75%로 높은 반면 하위 10% 값의 활성화 비율은 8% 로 전반적으로 하위 구간에서 상위 구간으로 갈수록 lift 꾸준히 증가하고 상위 40% 의 lift값이 1.55에 이를 정도로 분류를 잘 하고 있는 모습을 볼 수 있다. 다른 모든 모델이 테스트셋에 적합한것보다 결과가 좋으므로 Gradient Boost 모형을 제시한다.

Ⅴ. 결론

다음은 최종변수들 중요도를 알아본다. CARAVAN 보험을 구매한 고객을 예측하는데 있어 'PPERSAUT', 'MKOOPKLA', 'PWAPART', 'PBRAND', 'cars_1' 순으로 중요함을 알 수 있다.

0	ABROM	0.00
1	MOSHOOFD_10	0.00
6	MOSHOOFD_5	0.00
9	ABRAND	0.00
2	MGEMOMV	0.01
4	married2	0.02
5	insurance	0.02
3	job_2	0.04
8	rent1	0.05
7	MOSHOOFD_2	0.06
10	married1	0.06
11	cars_1	0.09
14	PBRAND	0.11
12	PWAPART	0.12
13	MKOOPKLA	0.16
15	PPERSAUT	0.26

다음은 표준화된 회귀 계수들이다. 표준화된 회귀계수는 단위에 무관하기 때문에 설명변수들의 상대적인 영향력을 직접 비교할 수 있으며 표준화 된 회귀계수가 클수록 설명별수의 관심변수에 대한 영향력이 크다고 말할 수 있다. 따라서 회귀계수들의 절대값 기준을 볼 때, ‘PPERSAUT’이 제일 기여도가 크고 다음으로 ‘ABROM’, ‘MOSHOOFD_10’, ‘PWAPART’, ‘cars_1’ 순이다.

Logistic 모형 계수

	0	1
0	ABROM	-1.546908
1	MOSHOOFD_10	-1.453082
11	cars_1	-1.148155
10	married1	-1.020027
6	MOSHOOFD_5	-0.896549
4	married2	-0.833053
3	job_2	-0.751103
8	rent1	-0.520202
14	PBRAND	-0.516569
2	MGEMOMV	-0.119983
13	MKOOPKLA	0.071469
9	ABRAND	0.085347
5	insurance	0.135476
7	MOSHOOFD_2	0.317492
12	PWAPART	1.130559
15	PPERSAUT	5.159477

[최종 평가]

다음은 모형에서 쓰이는 변수들에 대한 최종 평가다. 로지스틱회귀와 랜덤포레스트를 활용하면 실제로 CARAVAN 보험을 구입하는 사람들의 특성을 살펴 볼 수 있다.

먼저 가장 중요한 변수는 'PPERSAUT'이 변수는 고객이 지불한 차 보험금을 나타낸다. 차 보험금에 많은 비용을 지불하는 고객일수록 CARAVAN 보험을 구입할 확률이 높다.

그 다음 변수는 'MKOOPKLA' 이다. 이 변수는 고객이 속한지역에 사는 사람들의 구매력을 나타내는 변수로 더 잘사는 지역에 사는 사람일수록 더 CARAVAN 보험을 구입할 확률이 높다.

세 번째 변수는 'PWAPART' 이다. 이는 고객이 제 3자 보험에 지불하는 금액으로서 제 3자 보험에 많은 비용을 지불하는 고객일수록 CARAVAN 보험을 구입할 확률이 높다.

네 번째 변수는 'PBRAND'이다. 이 변수는 고객이 화재 보험에 지불하는 보험금을 의미한다. 화재 보험금을 많이 내는 사람일수록 CARABAN을 보험을 구입할 확률이 높다.

다섯 번째 변수는 'cars_1'이다. 이 변수는 우리가 pca를 통해 만들어 낸 변수로 이는 차를 가지고 있는 사람들을 의미한다. 차를 가지고 있는 사람들의 비율이 높은 지역일수록 그 지역에 사는 고객이 CARAVAN을 구입할 확률 낮아진다.

여섯 번째 변수는 결혼한 사람과 결혼하지 않은 사람의 대비를 나타내는 'married1' 변수이다. 결혼한사람의 비율이 높은 지역에 사는 사람일수록 CARAVAN을 구입할 확률이 높다.

이정도까지 변수를 종합하면 결국 CARAVAN을 구입한 사람들은 이동식 주택이므로 화재보험, 제 3자 보험, 그리고 차량보험을 같이 들 확률이 높은 것이다. 그리고 지역적으로 봤을 때 차를 가지고 있거나 결혼을 했거나 아님 구매력이 좋은 사람들이 모여 사는 곳에 사는 사람일수록 CARAVAN 보험을 구입해 왔다는 것이다. 이는 합리적인 결과이다. 이 변수들 말고는 이 변수들과 상관관계가 없으면서 유의한 변수는 없다고 봐도 된다.

[부록 1. 상세한 데이터 설명]

① 86개 변수 설명

No.	Name	Description	Type	Domain
1	MOSTYPE	Customer Subtype	명목형	L0
2	MAANTHUI	Number of houses	연속형	1 ~ 10
3	MGEMOMV	Avg size household	연속형	1 ~ 6
4	MGEMLEEF	Avg age	순위형	L1
5	MOSHOOFD	Customer main type	명목형	L2
6	MGODRK	Roman catholic	순위형	L3
7	MGODPR	Protestant	순위형	L3
8	MGODOV	Other religion	순위형	L3
9	MGODGE	No religion	순위형	L3
10	MRELGE	Married	순위형	L3
11	MRELSA	Living together	순위형	L3
12	MRELOV	Other relation	순위형	L3
13	MFALLEEN	Singles	순위형	L3
14	MFGEKIND	Household without children	순위형	L3
15	MFWEKIND	Household with children	순위형	L3
16	MOPLHOOG	High level education	순위형	L3
17	MOPLMIDD	Medium level education	순위형	L3
18	MOPLLAAG	Lower level education	순위형	L3
19	MBERHOOG	High status	순위형	L3
20	MBERZELF	Entrepreneur	순위형	L3
21	MBERBOER	Farmer	순위형	L3
22	MBERMIDD	Middle management	순위형	L3
23	MBERARBG	Skilled labourers	순위형	L3
24	MBERARBO	Unskilled labourers	순위형	L3

25	MSKA	Social class A	순위형	L3
26	MSKB1	Social class B1	순위형	L3
27	MSKB2	Social class B2	순위형	L3
28	MSKC	Social class C	순위형	L3
29	MSKD	Social class D	순위형	L3
30	MHHUUR	Rented house	순위형	L3
31	MHKOOP	Home owners	순위형	L3
32	MAUT1	1 car	순위형	L3
33	MAUT2	2 cars	순위형	L3
34	MAUT0	No car	순위형	L3
35	MZFONDS	National Health Service	순위형	L3
36	MZPART	Private health insurance	순위형	L3
37	MINKM30	Income < 30,000	순위형	L3
38	MINK3045	Income 30-45,000	순위형	L3
39	MINK4575	Income 45-75,000	순위형	L3
40	MINK7512	Income 75-122.000	순위형	L3
41	MINK123M	Income > 123.000	순위형	L3
42	MINKGEM	Average income	순위형	L3
43	MKOOPKLA	Purchasing power class	순위형	L4
44	PWAPART	Contribution private third party insurance	순위형	L4
45	PWABEDR	Contribution third party insurance (firms)	순위형	L4
46	PWALAND	Contribution third party insurane (agriculture)	순위형	L4
47	PPERSAUT	Contribution car policies	순위형	L4
48	PBESAUT	Contribution delivery van policies	순위형	L4
49	PMOTSCO	Contribution motorcycle/scooter policies	순위형	L4
50	PVRAAUT	Contribution lorry policies	순위형	L4

51	PAANHANG	Contribution trailer policies	순위형	L4
52	PTRACTOR	Contribution tractor policies	순위형	L4
53	PWERKT	Contribution agricultural machines policies	순위형	L4
54	PBROM	Contribution moped policies	순위형	L4
55	PLEVEN	Contribution life insurances	순위형	L4
56	PPERSONG	Contribution private accident insurance policies	순위형	L4
57	PGEZONG	Contribution family accidents insurance policies	순위형	L4
58	PWAOREG	Contribution disability insurance policies	순위형	L4
59	PBRAND	Contribution fire policies	순위형	L4
60	PZEILPL	Contribution surfboard policies	순위형	L4
61	PPLEZIER	Contribution boat policies	순위형	L4
62	PFIETS	Contribution bicycle policies	순위형	L4
63	PINBOED	Contribution property insurance policies	순위형	L4
64	PBYSTAND	Contribution social security insurance policies	순위형	L4
65	AWAPART	Number of private third party insurance	연속형	1 ~ 12
66	AWABEDR	Number of third party insurance (firms)	연속형	1 ~ 12
67	AWALAND	Number of third party insurane (agriculture)	연속형	1 ~ 12
68	APERSAUT	Number of car policies	연속형	1 ~ 12
69	ABESAUT	Number of delivery van policies	연속형	1 ~ 12
70	AMOTSCO	Number of motorcycle/scooter policies	연속형	1 ~ 12
71	AVRAAUT	Number of lorry policies	연속형	1 ~ 12
72	AAANHANG	Number of trailer policies	연속형	1 ~ 12
73	ATRACTOR	Number of tractor policies	연속형	1 ~ 12
74	AWERKT	Number of agricultural machines policies	연속형	1 ~ 12
75	ABROM	Number of moped policies	연속형	1 ~ 12
76	ALEVEN	Number of life insurances	연속형	1 ~ 12

77	APERSONG	Number of private accident insurance policies	연속형	1 ~ 12
78	AGEZONG	Number of family accidents insurance policies	연속형	1 ~ 12
79	AWAOREG	Number of disability insurance policies	연속형	1 ~ 12
80	ABRAND	Number of fire policies	연속형	1 ~ 12
81	AZEILPL	Number of surfboard policies	연속형	1 ~ 12
82	APLEZIER	Number of boat policies	연속형	1 ~ 12
83	AFIETS	Number of bicycle policies	연속형	1 ~ 12
84	AINBOED	Number of property insurance policies	연속형	1 ~ 12
85	ABYSTAND	Number of social security insurance policies	연속형	1 ~ 12
86	CARAVAN	Number of mobile home policies	명목형	0 / 1

< L0 >

Value	Label	Value	Label
1	High Income, expensive child	20	Etnically diverse
2	Very Important Provincials	21	Young urban have-nots
3	High status seniors	22	Mixed apartment dwellers
4	Affluentsenior apartments	23	Young and rising
5	Mixed seniors	24	Young, low educated
6	Carre and childcare	25	Young seniors in the city
7	Dinki's(double income no kids)	26	Own home elderly
8	Middle class families	27	Seniors in apartments
9	Modern, complete families	28	Residential elderly
10	Stable family	29	Porchless seniors: no front yard
11	Family starters	30	Religious elderly singles
12	Affluent young families	31	Low income catholics
13	Young all american family	32	Mixed seniors
14	Junior cosmopolitan	33	Lower class large families
15	Senior cosmopolitans	34	Large family, employed child
16	Students in apartments	35	Village families
17	Fresh masters in the city	36	Couples with teens 'Married with children'
18	Single youth	37	Mixed small town dwellers

19	Suburban youth	38	Traditional familes
		39	Large religious familes
		40	Larfe family farms
		41	Mixed rurals

< L1 >

Value	Label
1	20-30 years
2	30-40 years
3	40-50 years
4	50-60 years
5	60-70 years
6	70-80 years

< L2 >

Value	Label
1	Successful hodonists
2	Driven Growers
3	Average Family
4	Career Loners
5	Living well
6	Crusing Seniors
7	Retired and Religeous
8	Family with grown ups
9	Conservative families
10	Farmers

< L3 >

Value	Label
0	0 %
1	1 - 10 %
2	11 - 23 %
3	24 - 36 %
4	36 - 49 %
5	50 - 62 %
6	63 - 75 %

7	76 - 88 %
8	89 - 99 %
9	100 %

< L4 >

Value	Label
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 -

[부록 2. 이상치탐색]

각 변수에 존재하는 값이 정수값이 아닌 것이 있는지 확인한다.

MOSTYPE
[1 2 3 4 5 6 7 8 9 10 11 12 13 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41]
MAANTHUI
[1 2 3 4 5 6 7 8 10]
MGEMOMV
[1 2 3 4 5 6]
MGEMLEEF
[1 2 3 4 5 6]
MOSHOOFD
[1 2 3 4 5 6 7 8 9 10]
MGODRK
[0 1 2 3 4 5 6 7 8 9]
MGODPR
[0 1 2 3 4 5 6 7 8 9]
MGODOV
[0 1 2 3 4 5]
MGODGE
[0 1 2 3 4 5 6 7 8 9]
MRELGE
[0 1 2 3 4 5 6 7 8 9]
MRELSA
[0 1 2 3 4 5 6 7]
MRELOV
[0 1 2 3 4 5 6 7 8 9]
MFALLEEN
[0 1 2 3 4 5 6 7 8 9]
MFGEKIND
[0 1 2 3 4 5 6 7 8 9]
MFWEKIND

[0 1 2 3 4 5 6 7 8 9]
MOPLHOOG
[0 1 2 3 4 5 6 7 8 9]
MOPLMIDD
[0 1 2 3 4 5 6 7 8 9]
MOPLLAAG
[0 1 2 3 4 5 6 7 8 9]
MBERHOOG
[0 1 2 3 4 5 6 7 8 9]
MBERZELF
[0 1 2 3 4 5]
MBERBOER
[0 1 2 3 4 5 6 7 8 9]
MBERMIDD
[0 1 2 3 4 5 6 7 8 9]
MBERARBG
[0 1 2 3 4 5 6 7 8 9]
MBERARBO
[0 1 2 3 4 5 6 7 8 9]
MSKA
[0 1 2 3 4 5 6 7 8 9]
MSKB1
[0 1 2 3 4 5 6 7 8 9]
MSKB2
[0 1 2 3 4 5 6 7 8 9]
MSKC
[0 1 2 3 4 5 6 7 8 9]
MSKD
[0 1 2 3 4 5 6 7 8 9]
MHHUUR
[0 1 2 3 4 5 6 7 8 9]
MHKOOP
[0 1 2 3 4 5 6 7 8 9]
MAUT1
[0 1 2 3 4 5 6 7 8 9]
MAUT2
[0 1 2 3 4 5 6 7 9]
MAUT0
[0 1 2 3 4 5 6 7 8 9]
MZFONDS
[0 1 2 3 4 5 6 7 8 9]
MZPART
[0 1 2 3 4 5 6 7 8 9]
MINKM30
[0 1 2 3 4 5 6 7 8 9]
MINK3045
[0 1 2 3 4 5 6 7 8 9]
MINK4575
[0 1 2 3 4 5 6 7 8 9]
MINK7512
[0 1 2 3 4 5 6 7 8 9]
MINK123M
[0 1 2 3 4 5 6 7 9]
MINKGEM
[0 1 2 3 4 5 6 7 8 9]
MKOOPKLA
[1 2 3 4 5 6 7 8]
PWAPART
[0 1 2 3]
PWABEDR
[0 1 2 3 4 5 6]
PWALAND
[0 1 2 3 4]
PPERSAUT
[0 4 5 6 7 8 9]
PBESAUT
[0 5 6 7]
PMOTSCO
[0 3 4 5 6 7]
PVRAAUT
[0 4 6 7 9]
PAANHANG
[0 1 2 3 4 5]
PTRACTOR
[0 3 4 5 6 7]
PWERKT
[0 1 2 3 4 6]
PBROM
[0 2 3 4 5 6]
PLEVEN
[0 1 2 3 4 5 6 7 8 9]
PPERSONG

[0 1 2 3 4 5 6]
PGEZONG
[0 2 3]
PWAOREG
[0 4 5 6 7]
PBRAND
[0 1 2 3 4 5 6 7 8]
PZEILPL
[0 1 2 3]
PPLEZIER
[0 1 2 3 4 5 6]
PFIETS
[0 1]
PINBOED
[0 1 2 3 4 5 6]
PBYSTAND
[0 2 3 4 5]
AWAPART
[0 1 2]
AWABEDR
[0 1 5]
AWALAND
[0 1]
APERSAUT
[0 1 2 3 4 5 6 7 12]
ABESAUT
[0 1 2 3 4 5]
AMOTSCO
[0 1 2 3 8]
AVRAAUT
[0 1 2 3 4]
AAANHANG
[0 1 2 3]
ATRACTOR
[0 1 2 3 4 5 6]
AWERKT
[0 1 2 3 4 6]
ABROM
[0 1 2 3]
ALEVEN
[0 1 2 3 4 5 8]
APERSONG
[0 1]
AGEZONG
[0 1]
AWAOREG
[0 1 2]
ABRAND
[0 1 2 3 4 5 6 7]
AZEILPL
[0 1]
APLEZIER
[0 1 2]
AFIETS
[0 1 2 3 4]
AINBOED
[0 1 2]
ABYSTAND
[0 1 2]
CARAVAN
[0 1]

[부록 3. 관측값 개수]

① MOSTYPE

value	1	2	3	4	5	6	7	8	9	10
num	218	148	433	90	70	209	72	546	460	271
value	11	12	13	15	16	17	18	19	20	21
num	286	194	302	7	25	13	27	7	42	29
value	22	23	24	25	26	27	28	29	30	31
num	169	376	324	129	79	77	41	139	190	318
value	32	33	34	35	36	37	38	39	40	41
num	234	1401	325	362	373	233	569	542	137	355

* 14번 관측값 개수는 0이다.

② MOSHOOFD

value	1	2	3	4	5
num	959	827	1513	79	940
value	6	7	8	9	10
num	326	881	2694	1111	492

[부록 4. PCA]

(1) 'MGODRK: Roman catholic' ~ 'MGODGE: No religion'

	Prin1	Prin2	Prin3	Prin4
proportion of variance	0.4265	0.2899	0.2196	0.0640
cumulative proportion	0.4265	0.7164	0.9360	1.0000

	Prin1	Prin2	Prin3	Prin4
MGODRK	0.6564	-0.2971	0.6857	-0.1040
MGODPR	-0.4343	-0.5807	0.0601	-0.6859
MGODOV	0.5542	-0.4139	-0.7194	-0.0932
MGODGE	-0.1040	-0.6859	-0.0634	-0.7174

주성분1은 ' ____ '로 해석할 수 있으며 변수명은 'religion_1'로 사용하겠다.

(2) 'MRELGE: Married' ~ 'MRELOV: Other relatioin'

	Prin1	Prin2	Prin3
proportion of variance	0.5253	0.3683	0.1064
cumulative proportion	0.5253	0.8936	1.0000

	Prin1	Prin2	Prin3
MRELGE	-0.6322	-0.3480	-0.6922
MRELSA	0.4886	-0.8725	-0.0077
MRELOV	0.6013	0.3431	-0.7216

(3) 'MFALLEN: Singles' ~ 'MFWEKIND: Household with children'

	Prin1	Prin2	Prin3
proportion of variance	0.4984	0.3678	0.1338
cumulative proportion	0.4984	0.8662	1.0000

	Prin1	Prin2	Prin3
MFALLEEN	0.5819	0.7321	0.3541
MFGEKIND	0.2989	-0.5975	0.7441
MFWEKIND	-0.7564	0.3271	0.5665

(4) 'MOPLHOOG: High level education' ~ 'MOPLLAAG: Lower level education'

	Prin1	Prin2	Prin3
proportion of variance	0.5790	0.3060	0.1150
cumulative proportion	0.5790	0.8850	1.0000

	Prin1	Prin2	Prin3
MOPLHOOG	0.4635	0.8443	0.2689
MOPLMIDD	0.4174	-0.4758	0.7742
MOPLLAAG	-0.7816	0.2466	0.5729

(5) 'MBERHOOG: High status' ~ 'MBERARBO: Unskilled labourers'

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
proportion of variance	0.2263	0.1935	0.1717	0.1605	0.1345	0.1135
cumulative proportion	0.2263	0.4198	0.5915	0.7520	0.8865	1.0000

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
MBERHOOG	0.6609	-0.4703	-0.4088	-0.3681	-0.1322	0.1477
MBERZELF	0.5256	0.3689	0.4692	0.0008	-0.5080	-0.3309
MBERBOER	0.4454	0.3565	0.1845	0.1931	0.5073	0.5881
MBERMIDD	-0.1864	-0.4610	0.6262	-0.1339	-0.2832	0.5124
MBERARBG	-0.2308	0.5487	-0.2866	-0.4914	-0.3802	0.4213
MBERARBO	0.0246	-0.0456	-0.3232	0.7535	-0.4923	0.2876

(6) 'MSKA: Social Class A' ~ 'MSKD: Social class D'

	Prin1	Prin2	Prin3	Prin4	Prin5
proportion of variance	0.2856	0.2353	0.1995	0.1542	0.1254
cumulative proportion		0.5209	0.7204	0.8746	1.0000

	Prin1	Prin2	Prin3	Prin4	Prin5
MSKA	0.7920	-0.0777	-0.3751	-0.3763	-0.2904
MSKB1	0.1983	0.2570	0.6839	-0.5674	0.3240
MSKB2	0.1134	0.0282	0.5547	0.2782	-0.7754
MSKC	-0.5317	0.3184	-0.2430	-0.5931	-0.4529
MSKD	0.1944	0.9087	-0.1575	0.3275	0.0662

(7) 'MHHUUR: Rented house' ~ 'MHKOOP: Home owners;

	Prin1	Prin2

proportion of variance	0.9208	0.0792
cumulative proportion	0.9208	1.0000

	Prin1	Prin2
MHHUUR	0.7236	0.6902
MHKOOP	-0.6902	0.7236

(8) 'MAUT1: 1 car' ~ 'MAUT0: No car'

	Prin1	Prin2	Prin3
proportion of variance	0.4756	0.3762	0.1482
cumulative proportion	0.4756	0.8518	1.0000

	Prin1	Prin2	Prin3
MAUT1	-0.6046	0.1111	-0.7887
MAUT2	0.4414	0.8710	-0.2157
MAUT0	0.6630	-0.4786	-0.5767

(9) 'MZFONDS: National Health Service' ~ 'MZPART: Private health insurance'

	Prin1	Prin2
proportion of variance	0.8738	0.1262
cumulative proportion	0.8738	1.0000

	Prin1	Prin2
MZFONDS	-0.7107	0.7034
MZPART	0.7034	0.7107

(10) 'MINKM30: Income < 30,000' ~ 'MINK123M: Income > 123,000'

	Prin1	Prin2	Prin3	Prin4	Prin5
proportion of variance	0.2843	0.2666	0.2254	0.1264	0.0973
cumulative proportion	0.2843	0.5509	0.7763	0.9027	1.0000

	Prin1	Prin2	Prin3	Prin4	Prin5
MINK30	-0.6337	0.6260	-0.0079	-0.1249	0.4369
MINK3045	-0.1790	-0.6571	0.2886	-0.0587	0.6703
MINK4575	0.6195	0.1994	-0.4886	-0.1651	0.5570
MINK7512	0.3883	0.3021	0.7786	-0.3904	0.0299
MINK123M	0.1826	0.2128	0.2672	0.8952	0.2206