데이터 마이닝을 활용한
보험 사기 적발 모형

응용통계학과 201352004 최문석
응용통계학과 201452024 박상희
응용통계학과 201452041 이인풍

목차

목차

01 데이터 설명

빅콘테스트 소개

한국정보화진흥원(NIA)과 한국빅데이터연합회는 빅데이터 우수인재 발굴과 취업 연계를 지원하는 '빅콘테스트 2016'를 연다고 29일 밝혔다.

빅콘테스트 2016은 빅데이터 관련 기업의 인턴 채용 기회가 주어지는 빅데이터 분석 경진대회다. 올해는 한화생명, SK텔레콤, 한국정보통신진흥협회가 대회를 공동 주관하며 미래창조과학부, 네이버, 다음소프트, 한국빅데이터포럼 등이 후원한다.

데이터 소개

이번 대회는 퓨처스 리그, 챌린지 리그로 나뉘어 진행된다. 퓨처스 리그는 개봉 영화의 관객 수를 예측하는 문제로 매출액과 점유율, 관객 수, 증감률, 상영 횟수 등의 데이터가 제공된다. 챌린지 리그는 보험사기를 예측하는 문제로 보험 고객정보와 계약정보, 지급정보, 설계사 정보 등의 데이터가 주어진다.

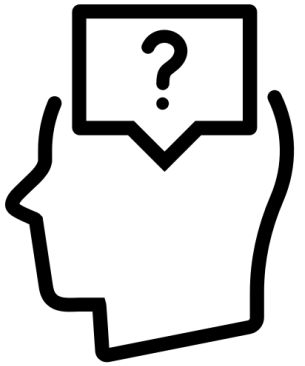
- ☑ 한화생명에서 제공한 보험과 관련된 데이터를 사용
- ☑ 보험 사기를 적발하는 분류 모형 개발

2016 빅콘테스트 홍보 포스터

01 데이터 설명

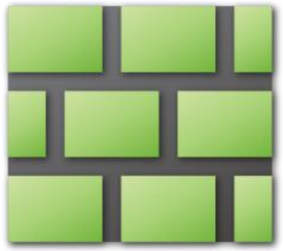
제공된 데이터의 3가지 특징

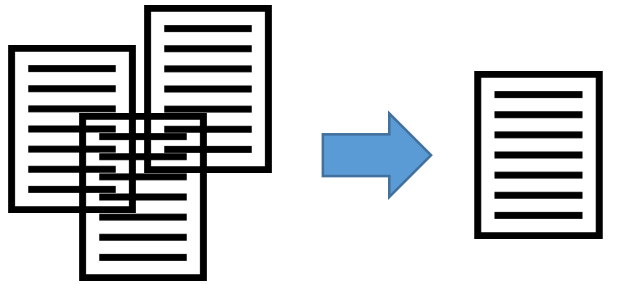
- 실제 Data를 기반으로 한 현실세계의 기업 내부 Data
- 누구나 쉽게 분석을 시도할 수 있는 Data-Set 구성
- 누구나 한번쯤은 좌절할 수 있는 Data의 한계 존재



분석시 유의사항

- 잘못 입력된 값, 결측값 등의 불친절한 데이터들의 전처리
- 데이터의 특성에 따라 여러 파일로 나누어진 데이터의 병합
- 보험 사기의 "원인" 보다는 "패턴"을 파악하여 보험 사기를 사전에 탐지





01 데이터 설명

제공된 5가지의 데이터

총 5개의 파일이 CSV 형태로 존재, 데이터의 특성에 따라 구분되어 있으며, 총 90 Columns 이상의 개인 작업 컴퓨터에서 작업할 수 있는 최선의 빅데이터를 제공

CUST DATA

- 고객의 특성을 나타내는 데이터
- SIU_CUST_YN 이라는 최종 보험사기 구분 Factor 포함(학습용 Set에만 결과가 있음)
- 고객의 성/연령/거주지/직업/배우자/소득 및 신용등급 정보 등 포함

CNTT DATA

- 고객들의 계약 속성을 나타내는 데이터
- 고객과 연관된 계약들의 상품종류 및 상태변화 및 보험료 수준 등
- 고객테이블과 CUST_ID 값을 Key값으로 하여 Join 가능

CLAIM DATA

- 고객들을 대상으로 한 지급 속성을 나타내는 데이터
- 언제 / 어떠한 사유로 / 얼마의 보험금이 지급 되었는지에 대한 정보를 포함함
- CNTT_DATA와 POLY_NO 값을 key로 하여 Join 가능

01 데이터 설명

제공된 5가지의 데이터

총 5개의 파일이 CSV 형태로 존재, 데이터의 특성에 따라 구분되어 있으며, 총 90 Columns 이상의 개인 작업 컴퓨터에서 작업할 수 있는 최선의 빅데이터를 제공

FMLY DATA

- 고객간의 가족 여부를 알 수 있는 데이터
- 보험 사기의 경우 다수가 연계하여 발생하는 경우가 많으므로 Network 분석 등으로 접근하는 참가자를 위하여 정보를 제공

FPINFO DATA

- 보험설계사 정보, 보험설계사의 재직기간 등을 알 수 있는 데이터
- 고객 대비 보험에 대한 이해도가 높음. Network 분석을 위한 데이터

	CLLT_FP_PRNO	INCB_DVSN	ETRS_YM	FIRE_YM	BEFO_JOB	EDGB	BRCH_CODE
0	6200	R	200306	200407	기타	고졸	25.0
1	2316	R	200104	200904	보험관계인	대학원졸	11.0
2	6207	R	200306	200311	무직	고졸	32.0
3	20797	R	198605	200804	기타	고졸	59.0
4	26245	R	199707	201508	기타	고졸	207.0

실제 FPINFO DATA

	CUST_ID	SUB_CUST_ID	FMLY_RELN_CODE
0	21564	20967	13
1	20672	6462	12
2	1342	12929	17
3	19239	18556	13
4	1649	13961	13

실제 FMLY DATA

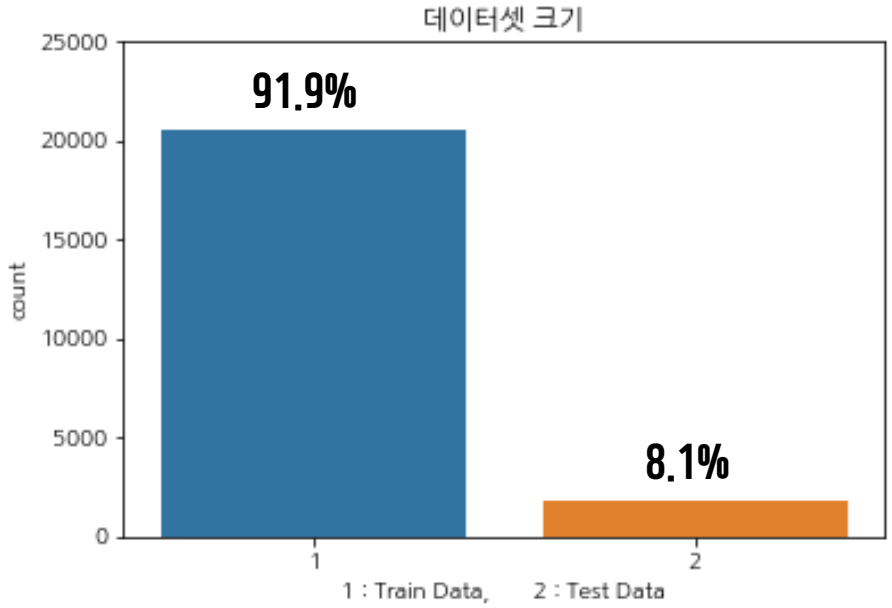
목차

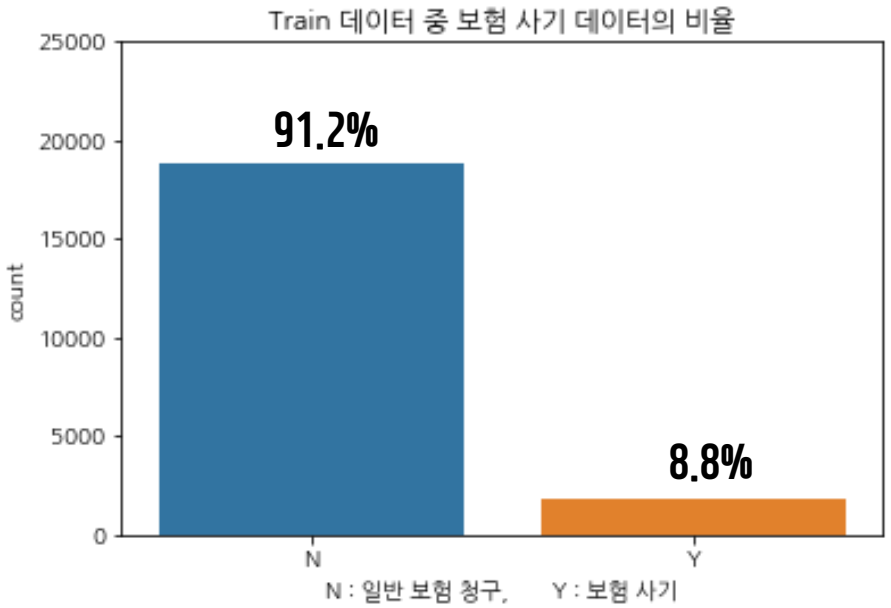
02 모델링을 위한 변수 선택

CUST_DATA

- 22,400개의 데이터, Train 20,607(91.9%) 과 Test 1,793(8.1%) 으로 구분.
- 20,607개의 Train 데이터 중 일반 18,801명(91.2%) 과 사기 1,806명(8.8%) 로 구분.
- 일반과 사기의 비율이 9:1로 매우 극단적. 모델의 정분류율이 91.9% 이상 되어야 함.
- 1,806 개의 보험 사기 데이터를 깊게 살펴볼 필요가 있음.

	일반	사기	Total
Train	18,801	1,806	20,607
Test	?	?	1,793

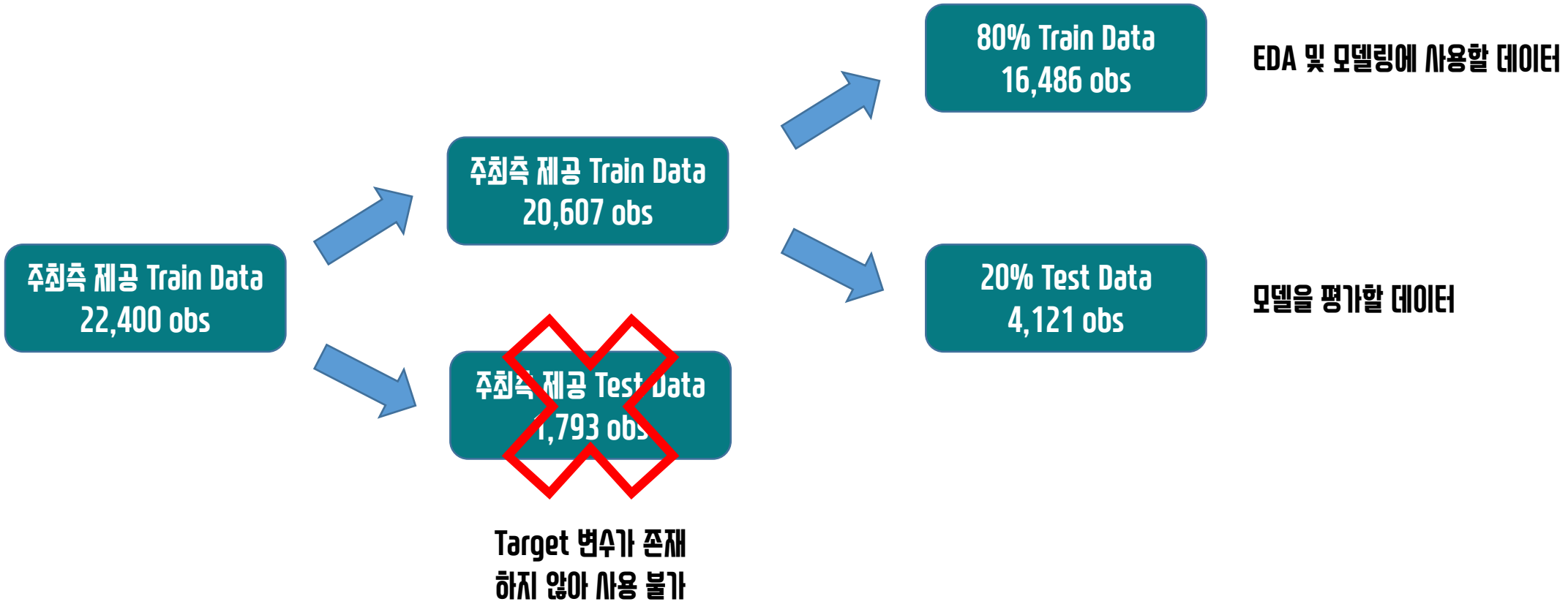




02 모델링을 위한 변수 선택

Train Test Split

- Target 변수가 존재하는 20,607개의 데이터를 8 : 2 로 분리
- 80%의 데이터로 EDA와 모델링 진행, 나머지 20%의 데이터로 모델 평가



02 모델링을 위한 변수 선택

CONTRACT_COUNT

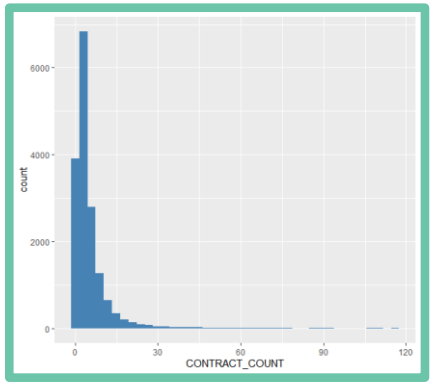
고객별 계약 건수를 나타내는 변수

Processing

CUST_ID를 기준으로 CNTT_DATA의 계약 건수를 계산

CUST_ID	COUNT
1	34
2	38
3	2
4	108
5	5
6	1
. . .	. . .

연속형 변수의 분포 확인



왼쪽으로 치우친 모양

범주형 변수로 변환

CUST_ID	COUNT
1	20 ~
2	20 ~
3	2
4	20 ~
5	5
6	1
. . .	. . .

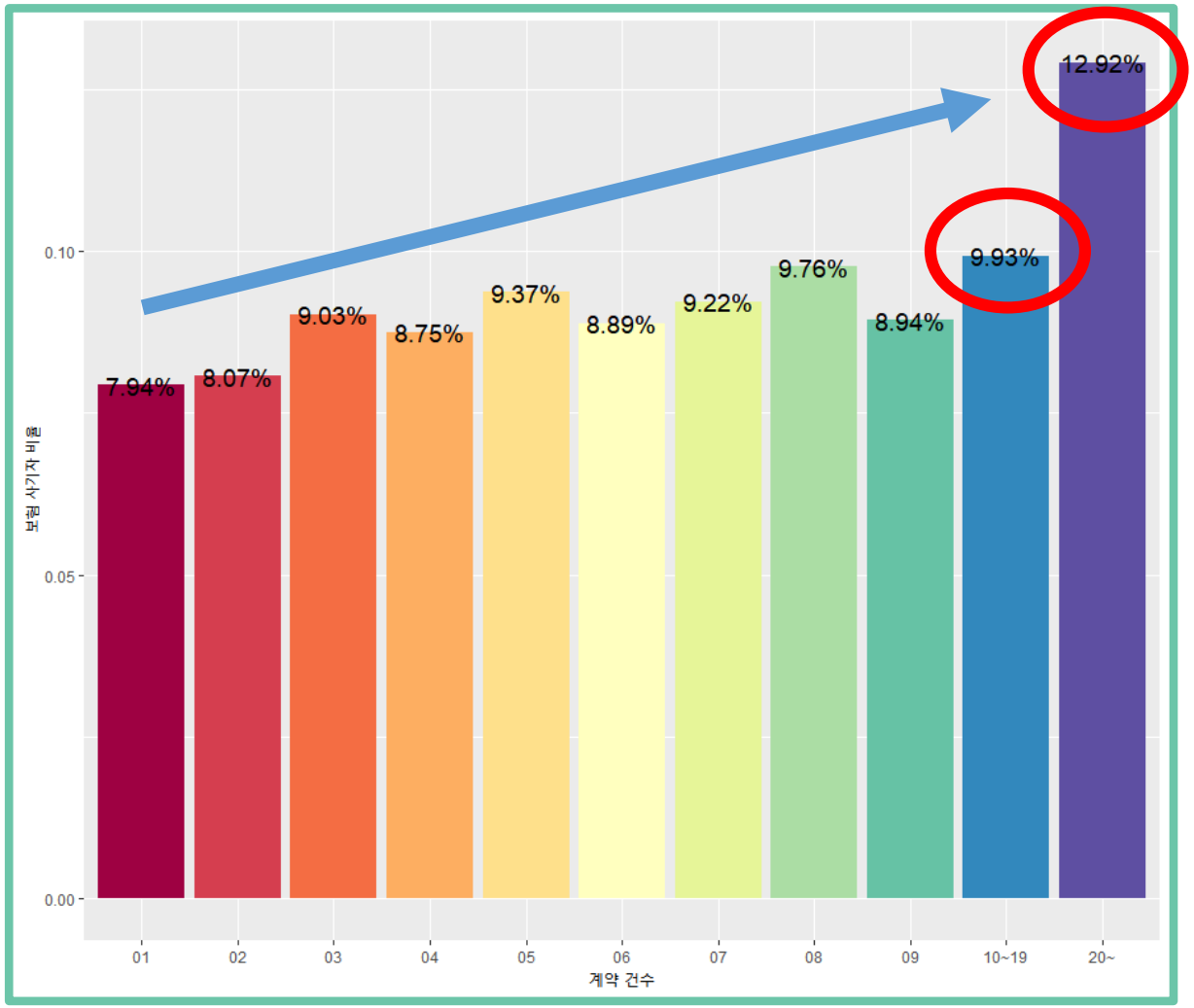
02 모델링을 위한 변수 선택

CONTRACT_COUNT

· 고객의 계약 건수가 증가할수록
보험 사기자의 비율도 증가하는 추세를 보였다.

보험 사기 여부와 연령대의 분할표

	전체 고객	일반 고객	보험 사기자
01	3,903	3,593	310
02	2,948	2,710	238
03	2,225	2,024	201
04	1,657	1,512	145
05	1,259	1,141	118
06	866	789	77
07	672	610	62
08	543	490	53
09	380	346	34
10~19	1,530	1,370	152
20 ~	503	438	65



02 모델링을 위한 변수 선택

CLAIM_COUNT

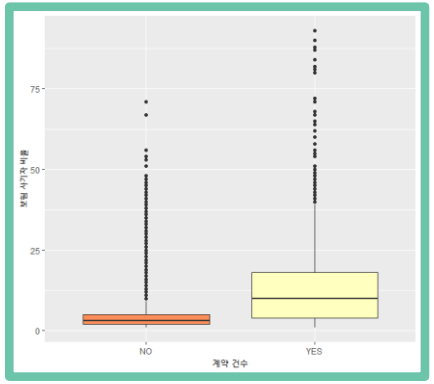
고객별 보험 청구 건수를 나타내는 변수

Processing

CUST_ID를 기준으로 CLIAM_DATA의 청구 건수를 계산

CUST_ID	COUNT
1	4
2	3
3	1
4	9
5	1
6	3
. . .	. . .

연속형 변수의 분포 확인



보험 사기 여부에 따른
분포의 확연한 차이가 드러남

전체 청구 건수와 평균 청구 건수

보험 사기 여부	전체 청구 건수
YES	19,941
NO	68,219

보험 사기 여부	평균 청구 건수
YES	13.7
NO	4.54

보험 사기자는 인원은 적지만
상당한 양의 보험을 청구함

02 모델링을 위한 변수 선택

CLAIM_COUNT

· 정규화(Normalization)

연속형 데이터의 전체 구간을 0 ~ 1으로 설정하여 데이터를 관찰하는 방법이다. 정규화를 하면 데이터의 범위를 일치시켜 스케일의 영향을 줄일 수 있다.

· 정규화 수식

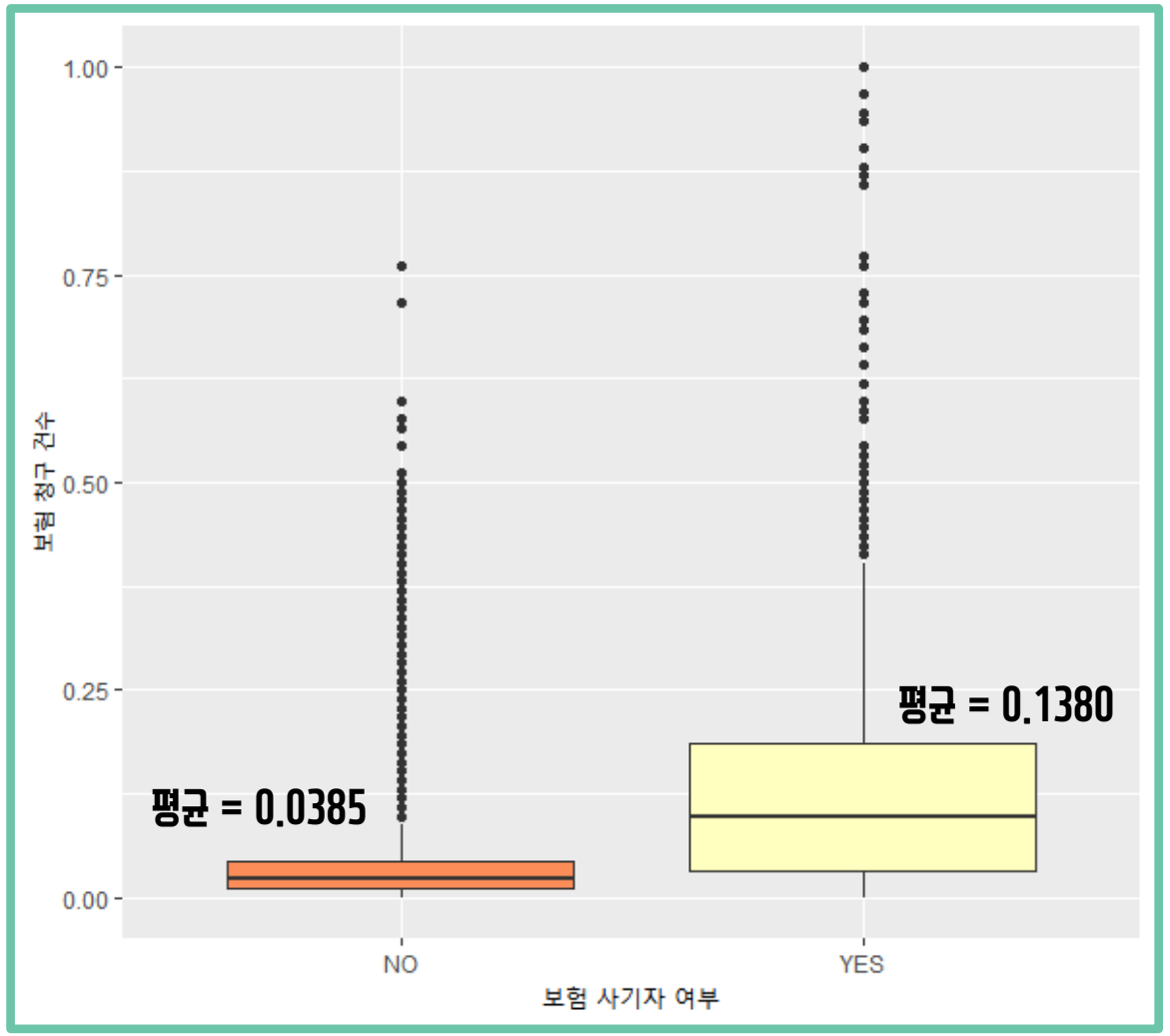
$$X_{NEW} = \frac{X - X_{MIN}}{X_{MAX} - X_{MIN}}$$

· 정규화 이전 CONTRACT_COUNT의 분포

Min	1st QU	Median	Mean	3rd QU	Max
1.000	2.000	3.000	5.348	6.000	93.000

· 정규화 이후 CONTRACT_COUNT의 분포

Min	1st QU	Median	Mean	3rd QU	Max
0.000	0.011	0.022	0.047	0.054	1.000



02 모델링을 위한 변수 선택

DANGER_FP

· FP는 보험 계약을 담당하는 보험사의 직원으로, 보험 사기 고객을 도와주거나, 일을 시작한지 얼마 되지 않아 보험 사기를 많이 당하거나, 본인이 담당하는 고객을 이용해 직접 보험 사기를 저지르는 경우가 있을 것이라 판단. 각 FP의 담당 계약 건수 대비 보험 사기자의 계약 비율을 체크

Processing

보험 사기자의 CUST_ID를 추출

13, 14, 26, 54, 56, 63, ···
9442, 9451, ···, 15244

113,010개 계약 건에 대하여 보험사기자의 계약인지를 표시

POLY_NO	CLLT_FP_PRNO	SIU
1	4671	0
2	5900	0
3	5778	1
···	···	···

FP별로 담당한 사기자의 계약 수를 카운트

CLLT_FP_PRNO	SIU_COUNT
1	0
2	0
3	0
···	···

FP별 담당 계약 건 대비 보험 사기자의 계약 건 비율

CLLT_FP_PRNO	SIU_COUNT
1	0
2	0
···	···

113,010개 계약 건에 대하여 FP의 위험도를 표시

POLY_NO	CUST_ID	SIU_COUNT
88997	4980	0.0000
99104	2341	0.1475
···	···	···

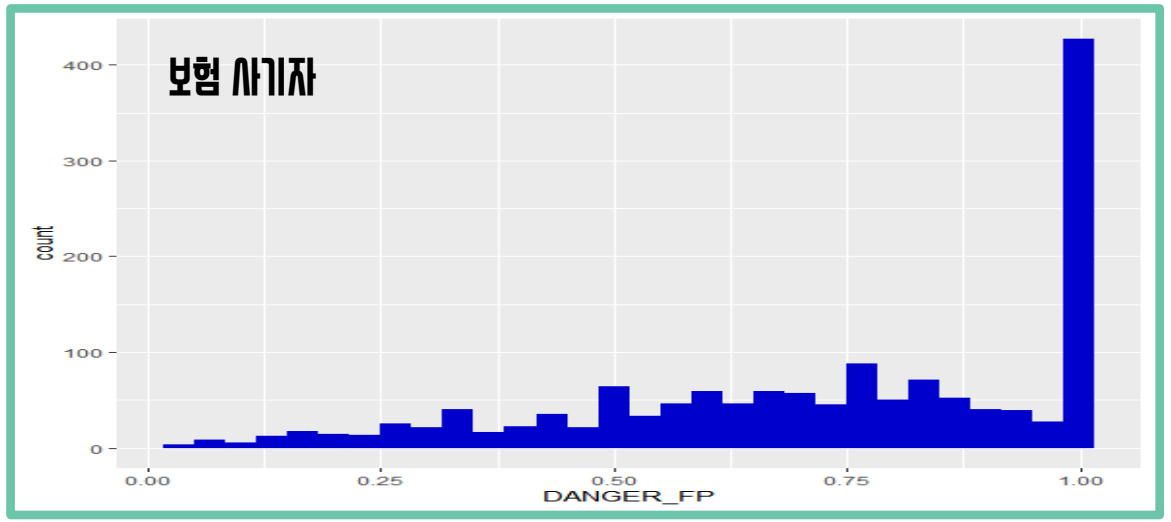
고객 별 계약 건들의 FP 위험도의 평균을 계산

CUST_ID	FP_DANGER
1	0.0455
2	0.0568
···	···

02 모델링을 위한 변수 선택

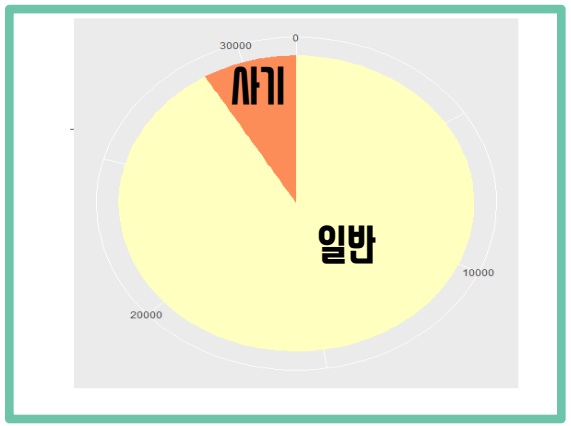
DANGER_FP

- 일반 고객 15,031명의 평균 FP 위험도는 0.0183으로 나타났다.
- 보험 사기자 1,455명의 평균 FP 위험도는 0.733 으로 나타났다.
- 일반 고객들 중 11,732명(78.05%)은 FP 위험도가 0.00이었다.
- 보험 사기자 중 426명(29.27%)은 FP 위험도가 1.00 이었다.

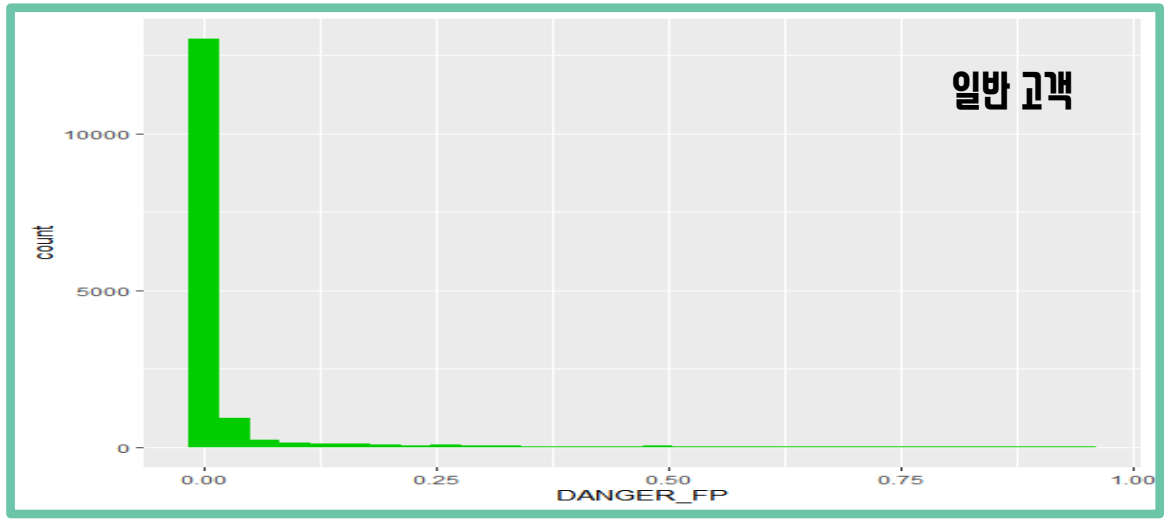


전체 FP 담당자 중 보험 사기에 관련된 FP

- 보험 사기에 관련되어 있다는 말은 FP의 위험도가 0.000이 아닌 FP를 뜻함



- 전체 FP 수 : 31,523
- 사기와 관련 : 2,707
- 9.39%가 사기와 관련



DANGER_DOCTOR

· 의사 역시 고객의 보험 사기를 도와주는 등의 방법으로 특정 의사들이 보험 사기에 연관되어 있을 것이라 판단. 의사의 면허 번호를 활용하여 FP_DANGER와 같은 방식으로 의사의 담당 보험 청구 건수 대비 보험 사기자의 청구 건수의 비율을 계산

Processing

119,020개 청구 건에 대하여 보험사기자의 청구인지를 표시

보험 사기자의 CUS

13, 14, 26, 54,
9442, 9451,

담당 의사별로 담당한 사기자의 계약 수를 카운트

SIU_COUNT
0
0
1
…

DANGER_FP와 같은 방식으로
Processing

고객 별 청구 건들의 의사

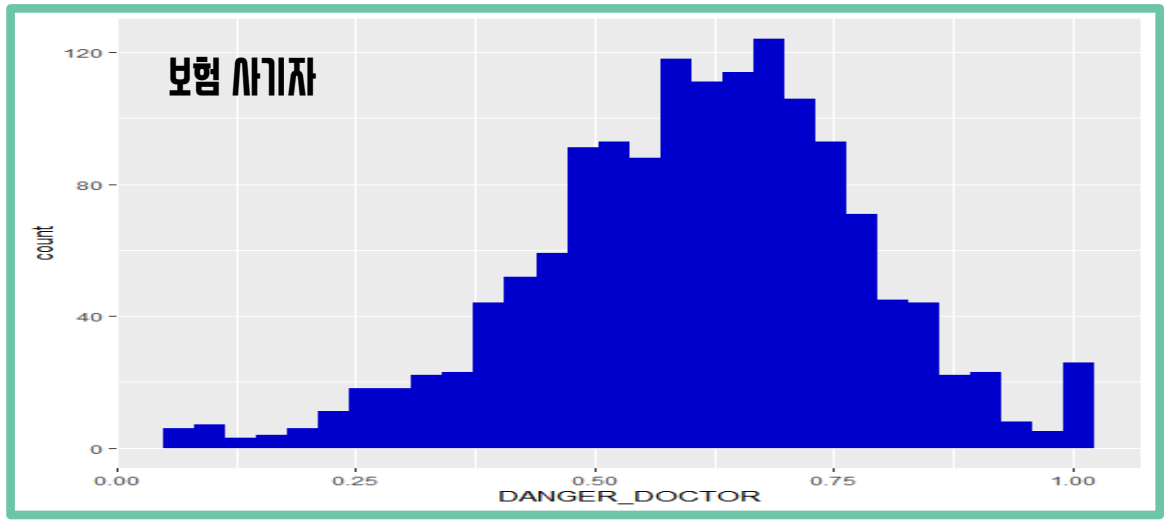
험 사기자의 계약 건 비율

CUST_ID						SIU_COUNT
1						0
2						1
…	…		…	…	…	…	…

02 모델링을 위한 변수 선택

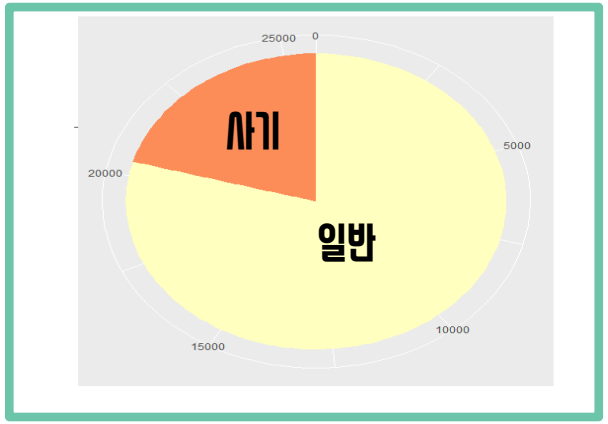
DANGER_DOCTOR

- 일반 고객 15,031명의 평균 의사 위험도는 0.061 으로 나타났다.
- 보험 사기자 1,455명의 평균 의사 위험도는 0.610 으로 나타났다.
- 일반 고객들 중 8,290명(55.15%)은 의사 위험도가 0.00 이었다.
- 보험 사기자 중 26명(1.78%)은 의사 위험도가 1.00 이었다.

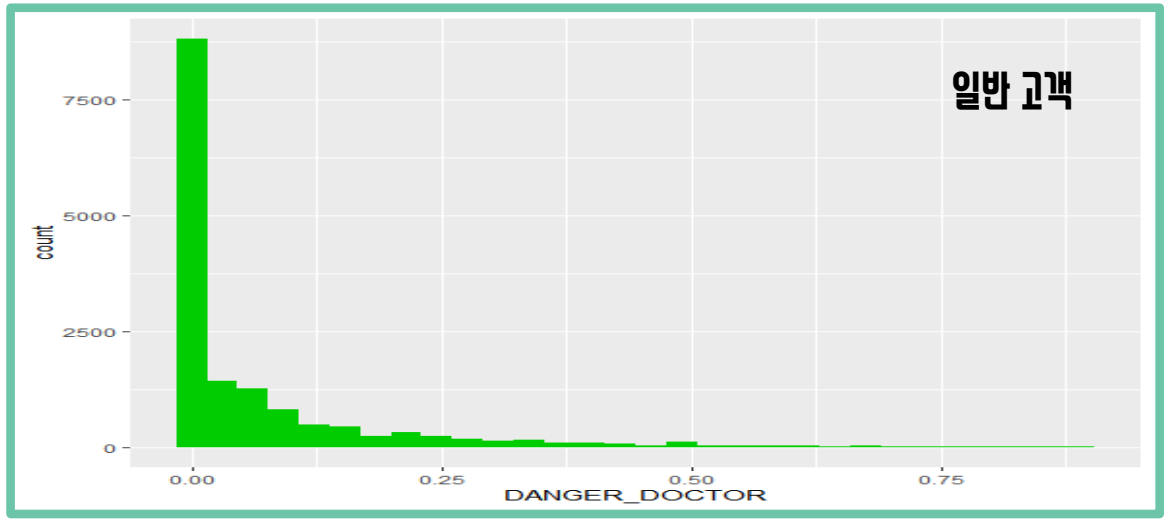


 # 전체 담당 의사 중 보험 사기에 관련된 의사

- 보험 사기에 관련되어 있다는 말은 의사의 위험도가 0.000이 아닌 FP를 뜻함



- 전체 의사 수 : 25,702
- 사기와 관련 : 5,306
- 20.64%가 사기와 관련



DANGER_HOSP

· 병원 역시 고객의 보험 사기를 도와주는 등의 방법으로 특정 병원들이 보험 사기에 연관되어 있을 것이라 판단. 병원 고유 번호를 활용하여 FP_DANGER와 같은 방식으로 병원의 담당 보험 청구 건수 대비 보험 사기자의 청구 건수의 비율을 계산

Processing

119,020개 청구 건에 대하여 보험사기자의 청구인지를 표시

보험 사기자의 CUS

13, 14, 26, 54,
9442, 9451,

담당 병원별로 담당한 사기자의 계약 수를 카운트

SIU_COUNT
0
0
1
…

DANGER_FP와 같은 방식으로
Processing

고객 별 청구 건들의 병원

CUST_ID					
1					
2					
…	…	…	…	…	…

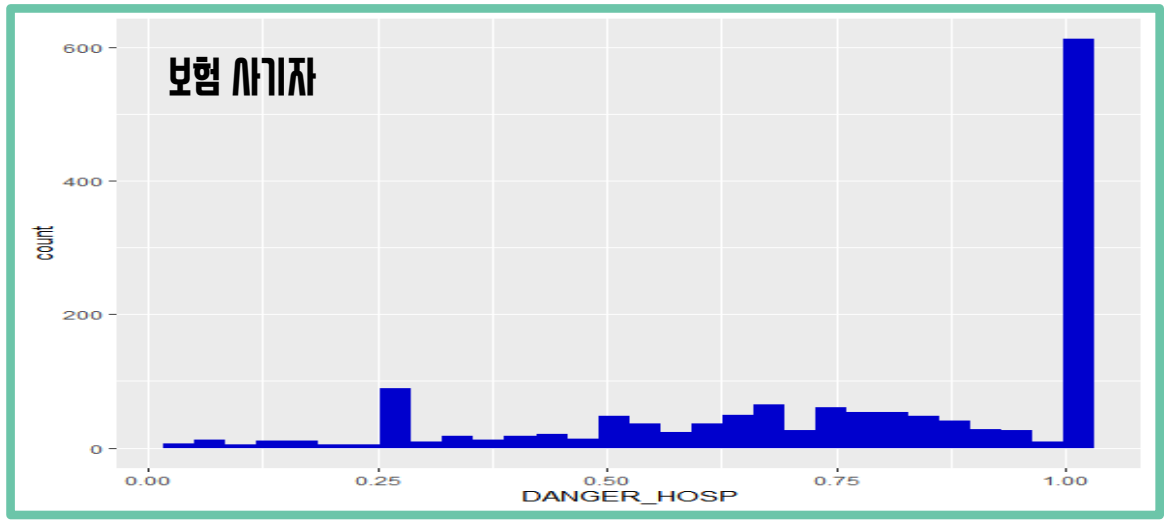
험 사기자의 계약 건 비율

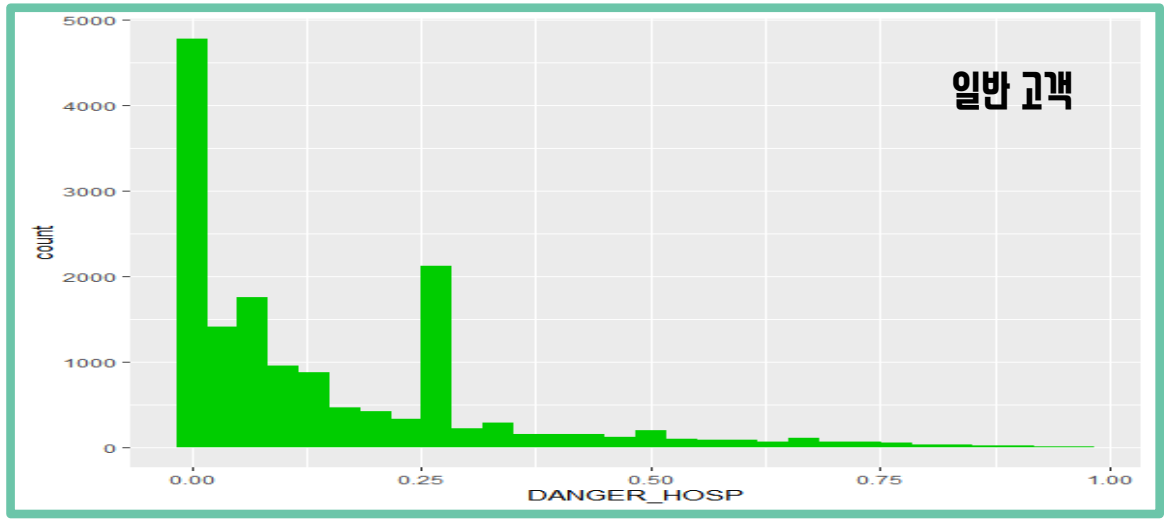
	SIU_COUNT
…	…

02 모델링을 위한 변수 선택

DANGER_HOSP

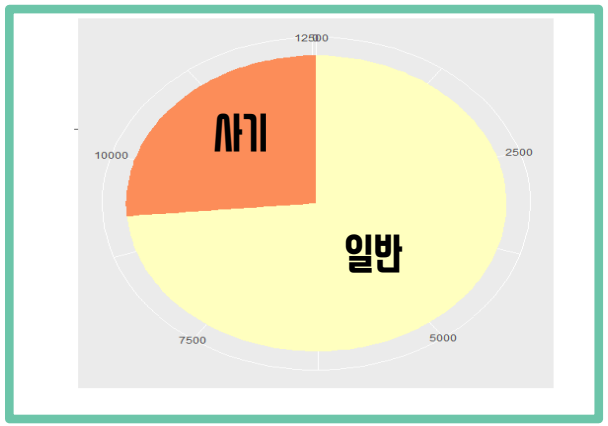
- 일반 고객 15,031명의 평균 병원 위험도는 0.144 으로 나타났다.
- 보험 사기자 1,455명의 평균 병원 위험도는 0.770 으로 나타났다.
- 일반 고객들 중 4,123명(27.42%)은 병원 위험도가 0.00 이었다.
- 보험 사기자 중 613명(42.13%)은 병원 위험도가 1.00 이었다.





전체 담당 병원 중 보험 사기에 관련된 병원

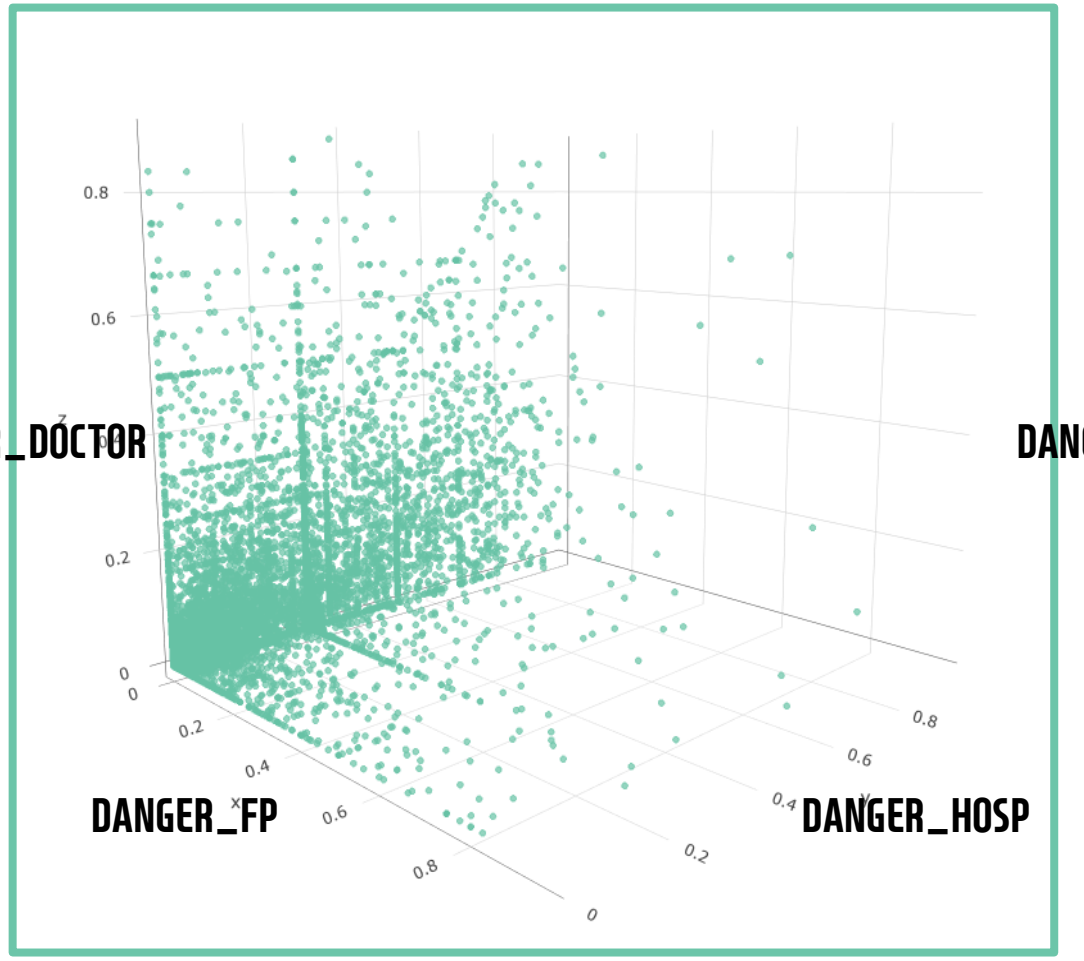
- 보험 사기에 관련되어 있다는 말은 병원의 위험도가 0.000이 아닌 FP를 뜻함

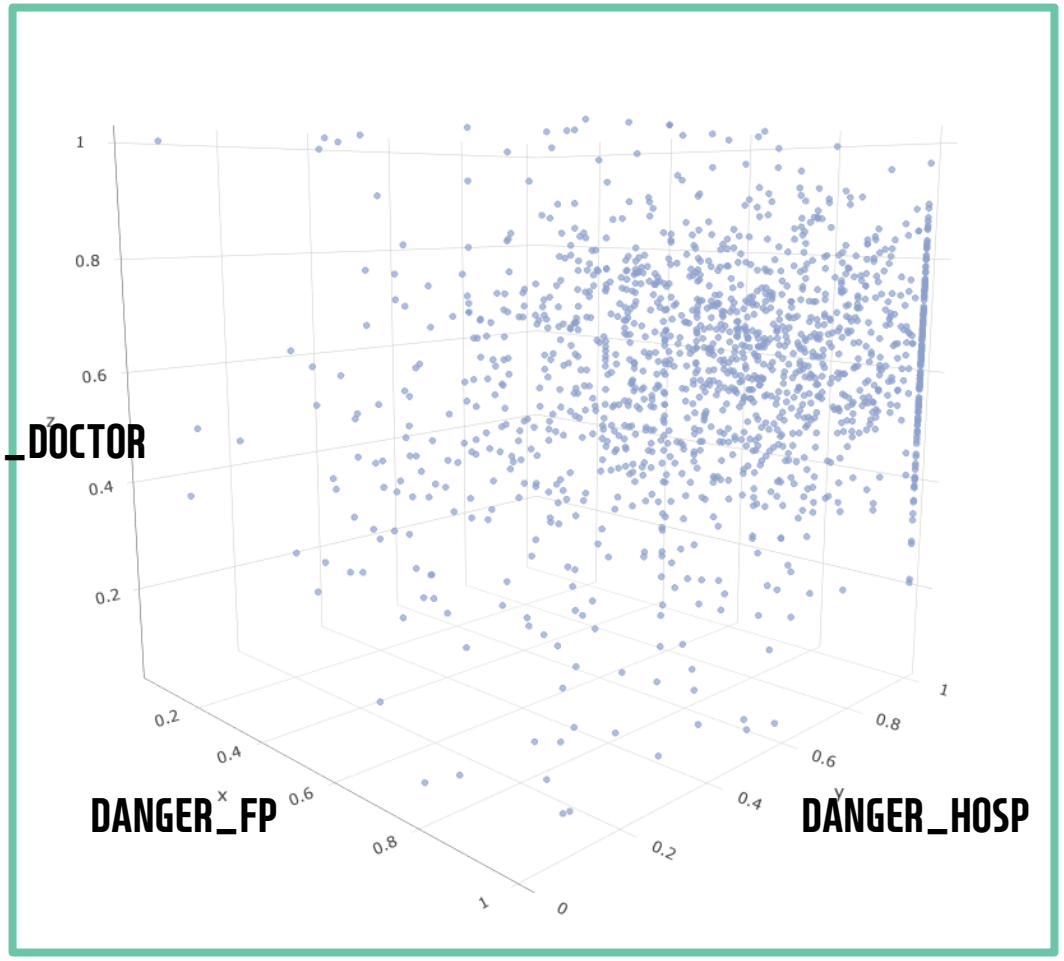


- 전체 병원 수 : 12,538
- 사기와 관련 : 3,307
- 26.37%가 사기와 관련

02 모델링을 위한 변수 선택

FP_DOCTOR_HOSP

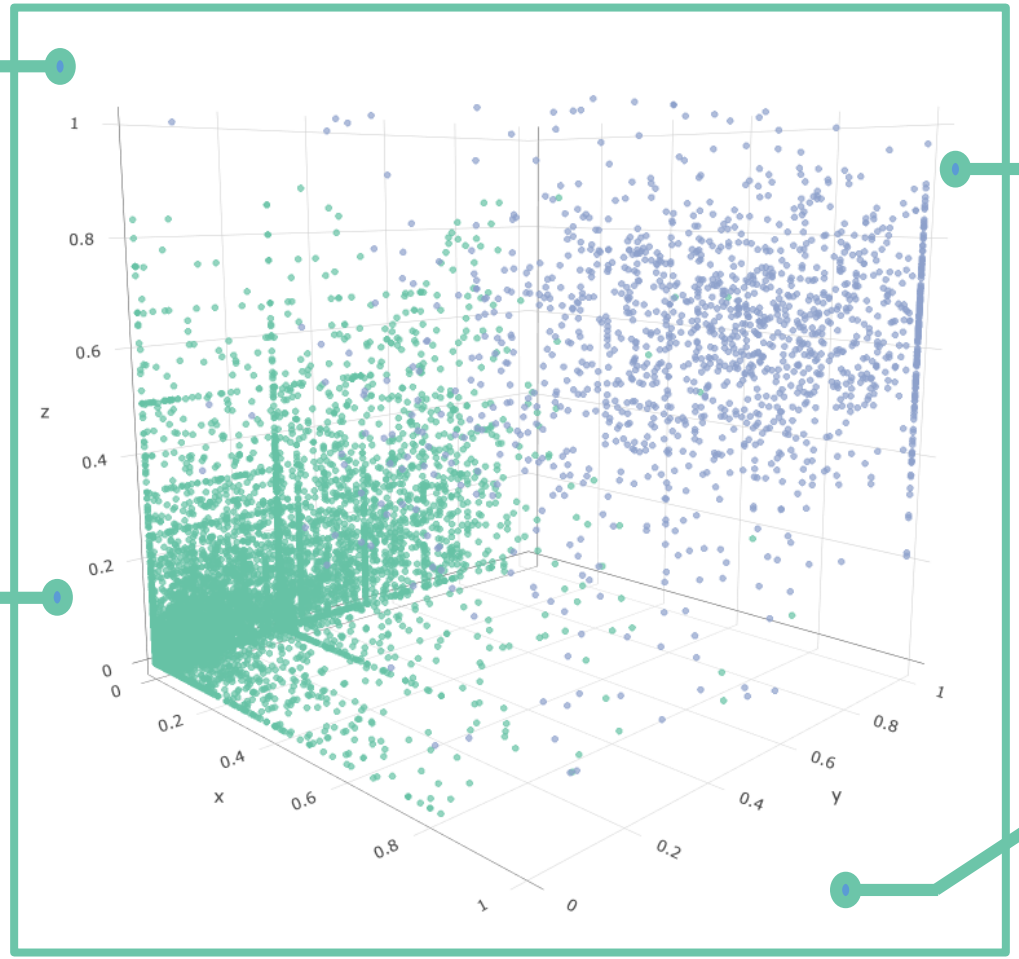




02 모델링을 위한 변수 선택

FP_DOCTOR_HOSP

각 위험도의 평균		
	일반 고객	보험 사기자
FP	0.018	0.733
DOCTOR	0.061	0.610
HOSP	0.144	0.770



보험 사기자의 경우
FP, DOCTOR, HOSP의 위험도가
모두 "1" 에 가까움

일반 고객의 경우
FP, DOCTOR, HOSP의 위험도가
모두 "0" 에 가까움

위험도 간의 상관관계			
	FP	DOCTOR	HOSP
FP	-	0.63	0.71
DOCTOR	0.63	-	0.73
HOSP	0.71	0.73	-

02 모델링을 위한 변수 선택

CAUSE_CODE

· 보험 청구시 지급 청구의 원인이 되는 사유 코드

· 사망(01), 입원(02), 통원(03), 장해(04), 수술(05), 진단(06), 치료(07), 해지/무효(09)의 8개의 범주로 구성

· 보험 사기자들이 주로 사용한 청구 사유 코드에 대하여 가중치를 계산

전체 담당 병원 중 보험 사기에 관련된 병원

CAUSE_CODE	청구 건수	보험 사기자의 청구 건수	보험 사기자들의 사용 비율	사유 코드 별 가중치
01	589	31	0.15%	0.0015
02	62,016	15,533	77.90%	0.7790
03	24,433	1,352	6.78%	0.0678
04	902	223	1.11%	0.0111
05	20,511	2,024	10.15%	0.1015
06	10,395	774	3.89%	0.0389
07	169	4	0.02%	0.0002
09	5	0	0.00%	0.0000
Total	119,020	19,941	100%	1.0000

CAUSE_CODE

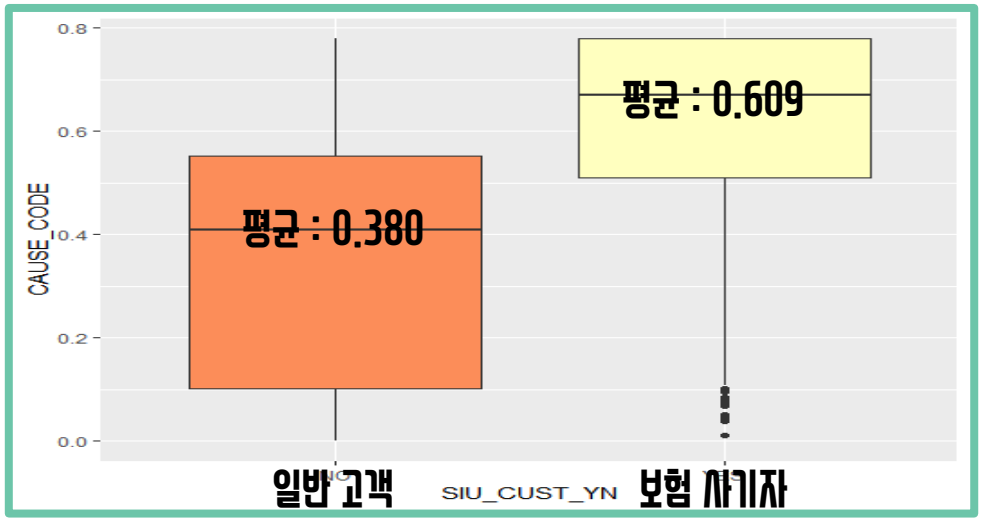
· 계산된 청구 사유 코드의 가중치를 모든 청구 건수(119,020 건)에 대하여 부여

POLY_NO	CUST_ID	DMND_RESN_CODE	CAUSE_CODE
1365	5936	3	0.0678
1247	1043	2	0.7790
…	…	…	…

· 각 고객들에 대하여 평균 청구 사유 가중치를 계산

CUST_ID	CAUSE_CODE
1	0.0678
2	0.0678
3	0.7790
4	0.0389
5	0.7790
6	0.3060
…	…

· 보험 사기 여부에 따른 청구 사유 가중치



CNTT_ROLE

- 계약 당시 고객의 역할을 구분하는 코드

- 계약자이나 주피보험자는 아님(0), 계약자이며 주피보험자임(1), 계약자는 아니지만 주피보험자임(2), 주피보험자는 아니지만 피보험자임(21)
  피보험자는 아니지만 생존수익자(3), 피보험자 및 생존수익자는 아니지만 입원장해수익자(4), 피보험자는 생존급부의 수익자는 아니지만 사망수익자(5)

- 보험 사기자들이 주로 사용한 역할 코드에 대하여 가중치를 계산

전체 담당 병원 중 보험 사기에 관련된 병원

CNTT_ROLE	총 계약 건수	보험 사기자의 계약 건수	보험 사기자들의 사용 비율	역할 코드 별 가중치
0	23,712	2122	25.4%	0.2540
1	48,059	3799	45.4%	0.4540
2	27,802	1706	20.4%	0.2040
3	616	60	0.717%	0.0072
4	436	69	0.825%	0.0083
5	9,216	502	6.00%	0.0600
21	3,169	110	1.31%	0.0131
Total	113,010	8,368	100%	1.0000

CNTT_ROLE

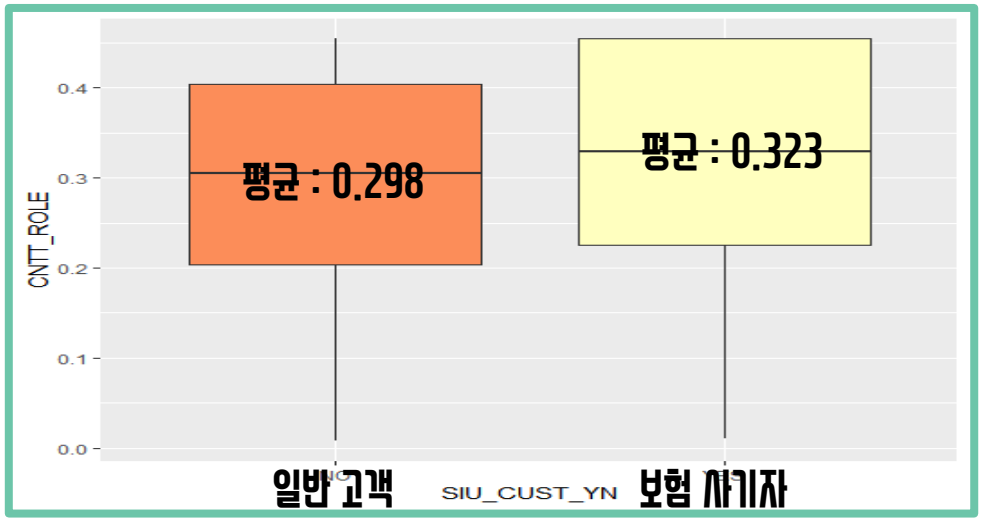
· 계산된 계약 역할 코드의 가중치를 모든 계약 건수(113,010 건)에 대하여 부여

POLY_NO	CUST_ID	CNTT_ROLE	CAUSE_CODE
1	2805	1	0.4540
2	5658	1	0.4540
…	…	…	…

· 각 고객들에 대하여 평균 계약 역할 코드 가중치를 계산

CUST_ID	CNTT_CODE
1	0.2570
2	0.2340
3	0.4540
4	0.2960
5	0.4140
6	0.4540
…	…

· 보험 사기 여부에 따른 계약 역할 코드 가중치



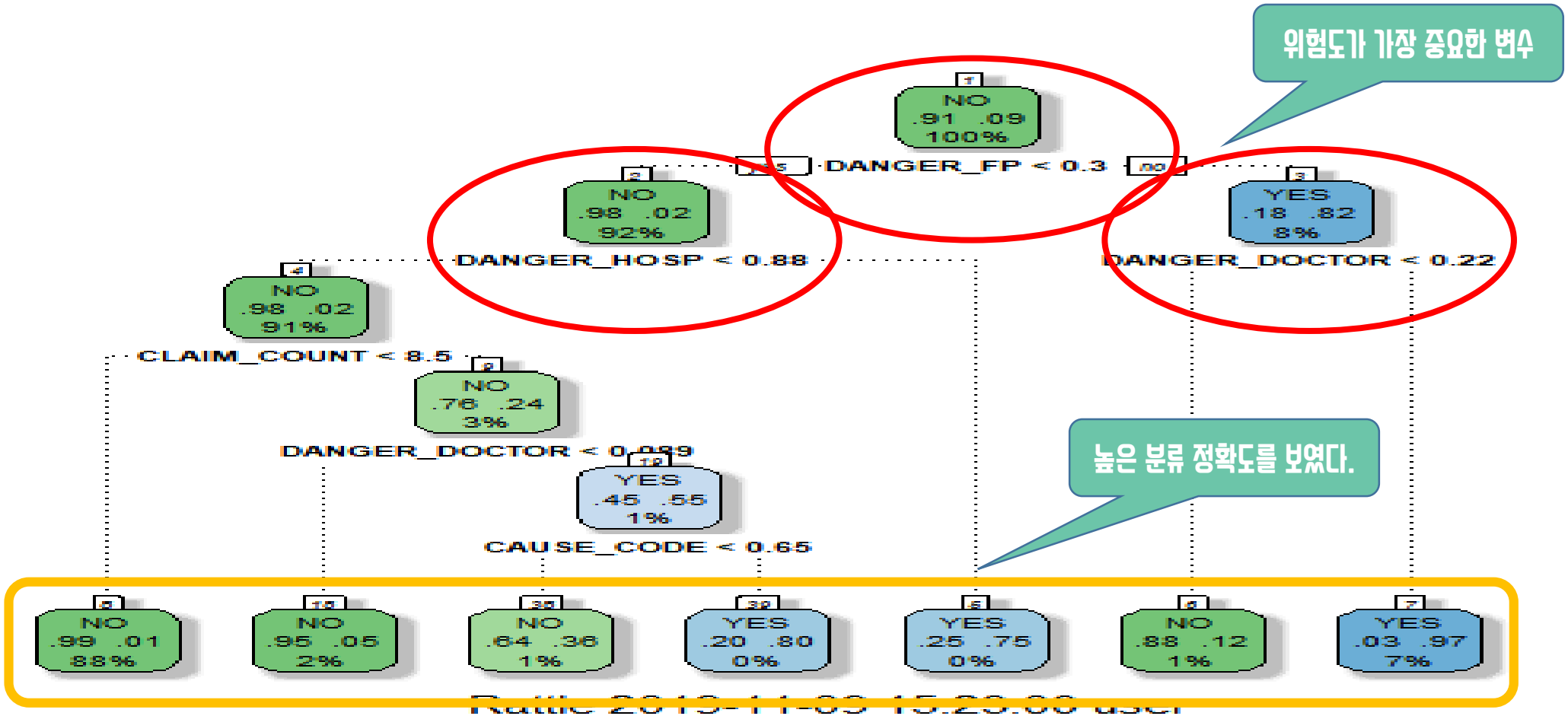
목차

분석에 사용할 변수

	변수명	설명	수준	비고
INPUT	CONTRACT_COUNT	계약 건수	Binary	“01”, “02”, “03”, “04”, “05”, “06”, “07”, “08”, “09”, “10~19”, “20~”
	CLAIM_COUNT	청구 건수	Numeric	0.0000 ~ 1.0000
	DANGER_FP	FP 위험도	Numeric	0.0000 ~ 1.0000
	DANGER_DOCTOR	의사 위험도	Numeric	0.0000 ~ 1.0000
	DANGER_HOSP	병원 위험도	Numeric	0.0000 ~ 1.0000
	CAUSE_CODE	보험 청구 사유	Numeric	0.0000 ~ 1.0000
	CNTT_ROLE	계약 역할	Numeric	0.0000 ~ 1.0000
TARGET	SIU_CUST_YN	보험 사기자 유무	Binary	“YES”, “NO”

Decision Tree

· 의사결정 나무는 데이터를 분석하여 이들 사이에 존재하는 패턴을 예측 가능한 규칙들의 조합으로 나타내며, 그 모양이 '나무' 와 같다고 해서 의사결정나무라 불린다.



Error Matrix

TRAIN DATA ERROR MATRIX

16,486 obs		Predicted	
		NO	YES
Actual	NO	15,025	46
	YES	223	1,191

TEST DATA ERROR MATRIX

4,212 obs		Predicted	
		NO	YES
Actual	NO	3,715	15
	YES	63	329

Test Accuracy and Score

· 정확도 = 전체 고객 중 올바르게 예측한 경우 = $\frac{3,715+329}{4,212} = 0.9601$

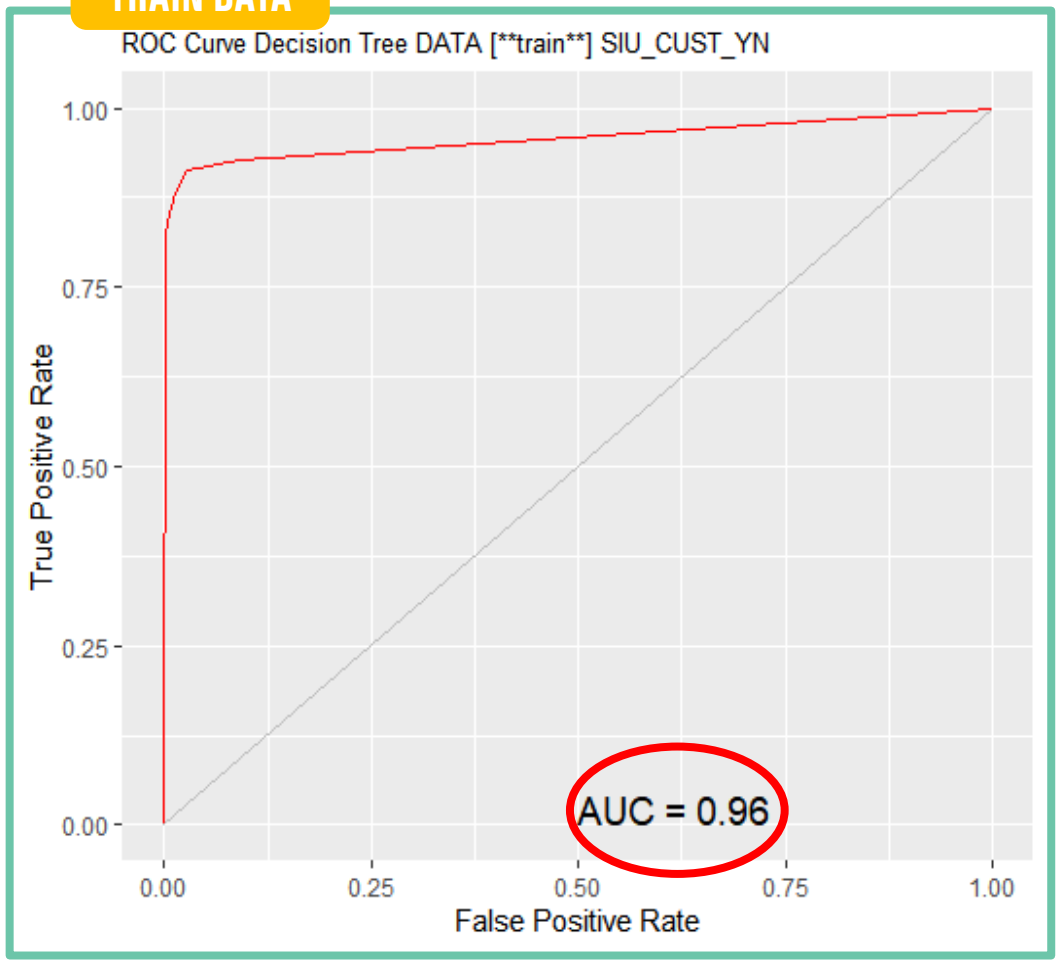
· Precision(정밀도) = 보험 사기자라고 예측한 경우 중 실제 보험 사기자의 비율 = $\frac{329}{15+329} = 0.9564$

· Recall(재현율) = 실제 보험 사기자 중 올바르게 예측해낸 비율 = $\frac{329}{63+329} = 0.8393$

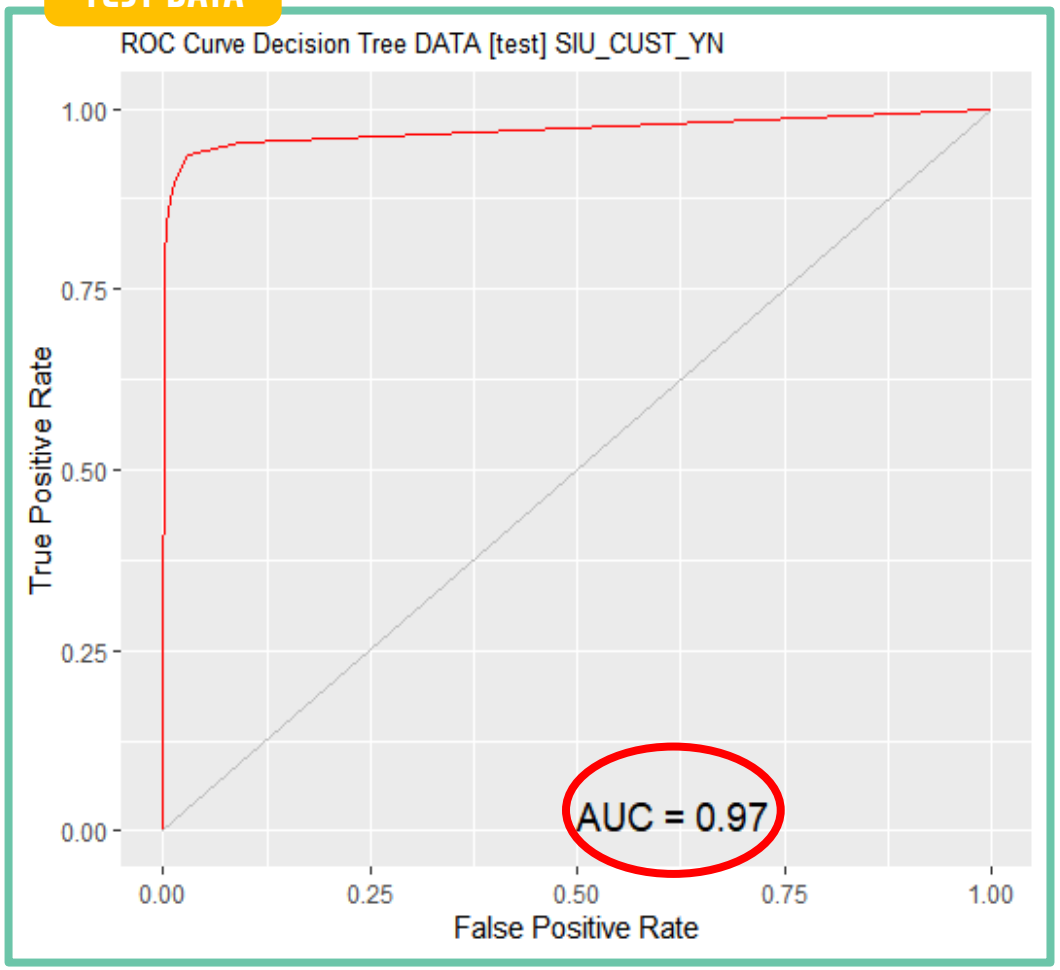
· F1-score = Precision과 Recall의 조화 평균 = $\frac{2 \times Precision \times Recall}{Precision+Recall} = 0.8940$

ROC Curve and AUC

TRAIN DATA



TEST DATA



목차

04 결론 및 제언

정리

Feature Selecting

- 보험 사기자를 예측(이항 분류)하기 위해 1개의 범주형 변수와 6개의 연속형 변수를 사용하여 분석 진행
- 6개의 연속형 범주의 스케일은 0.0000 ~ 1.000으로 조정
- 보험 사기에 있어서 FP, 의사, 병원이 매우 밀접한 관계를 가지고 있음

Modeling

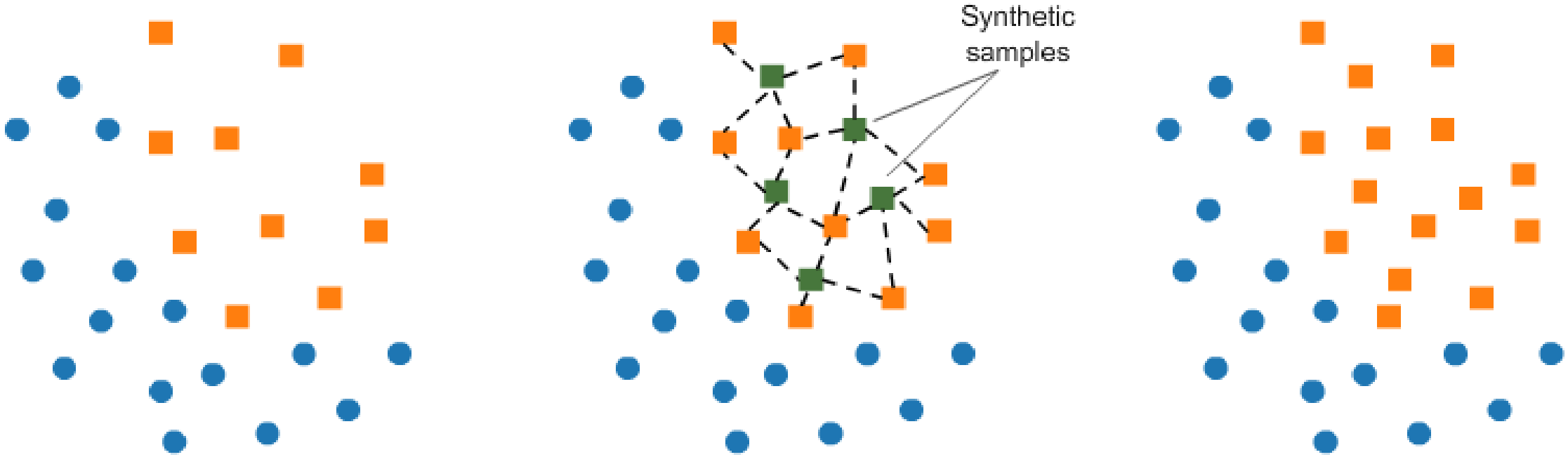
- 비록 적은 변수를 사용함에도 불구하고 비교적 높은 정확도를 보여줌
- 분석 진행 시에 사전에 Train과 Test를 분리하여 Test에 대해서는 전혀 사용하지 않았기 때문에 어느정도 Overfitting의 문제점을 해소

추후 분석 방향

Target 변수의 불균형 (YES : NO = 9 : 1)을 해소 하기 위해 오버 샘플링 진행

SMOTE 오버 샘플링 (Synthetic Minority Over-sampling Technique)

SMOTE 기법은 오버 샘플링 기법 중 하나로 합성 데이터를 사용하는 가장 많이 사용되고 있는 방식이다. SMOTE란 합성 소수 샘플링 기술로 다수 클래스를 샘플링하고, 기존 소수 샘플을 보간하여 새로운 소수 인스턴스를 합성해낸다.

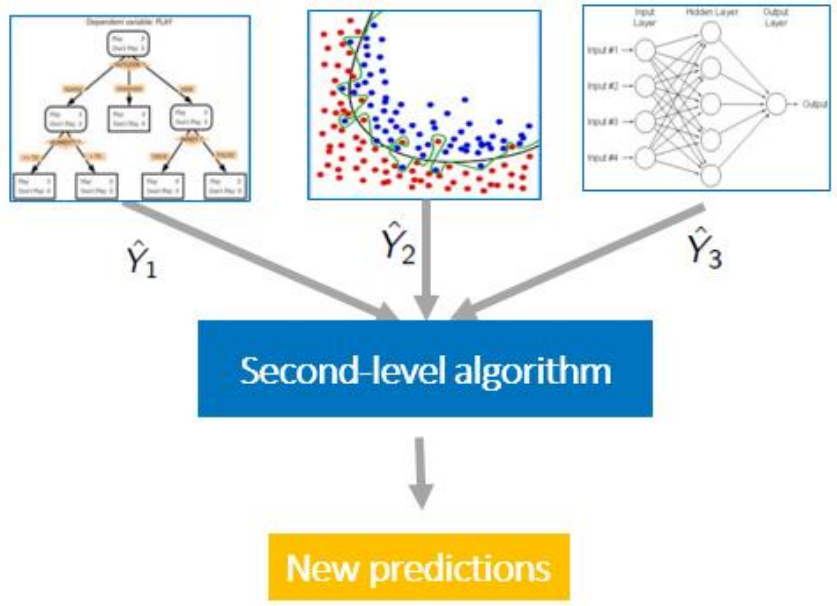


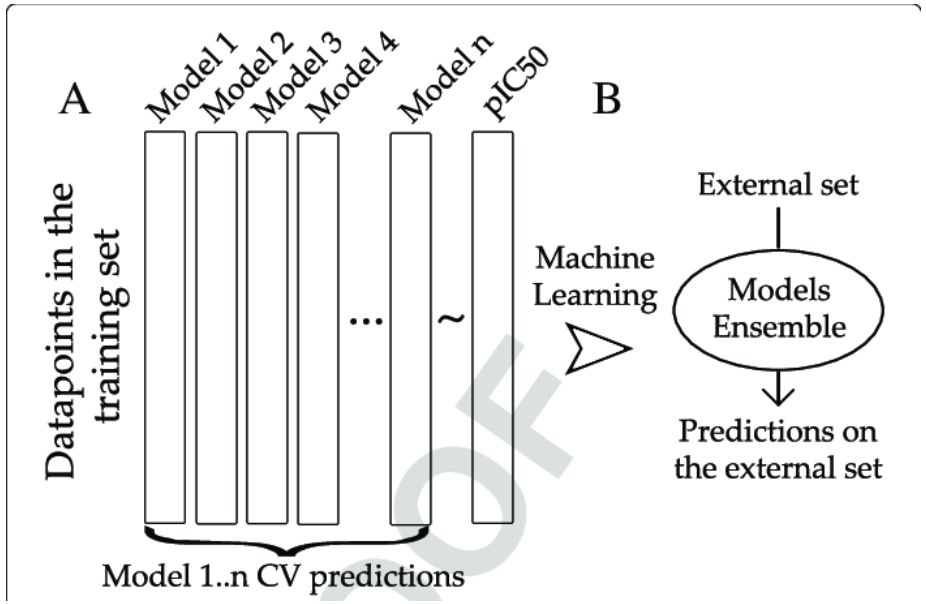
추후 분석 방향

Model Stacking을 통한 Ensemble 을 통한 Overfitting 해소

Model Stacking

이 기법은 서로 다른 여러 알고리즘들을 Level-0 모델로 사용하여 그 Output을 다시 Level-1의 Input으로 사용하여 전체적인 모델 구조를 2중 구조로 구현하는 방법이다. 이 Model Stacking을 사용하면 서로 다른 모델을 사용하게 되므로 한 모델에 대하여 Overfitting을 방지할 수 있다는 장점이 있다.





THANK YOU!