보험사기 예측 모델

12214227 박윤서

목차

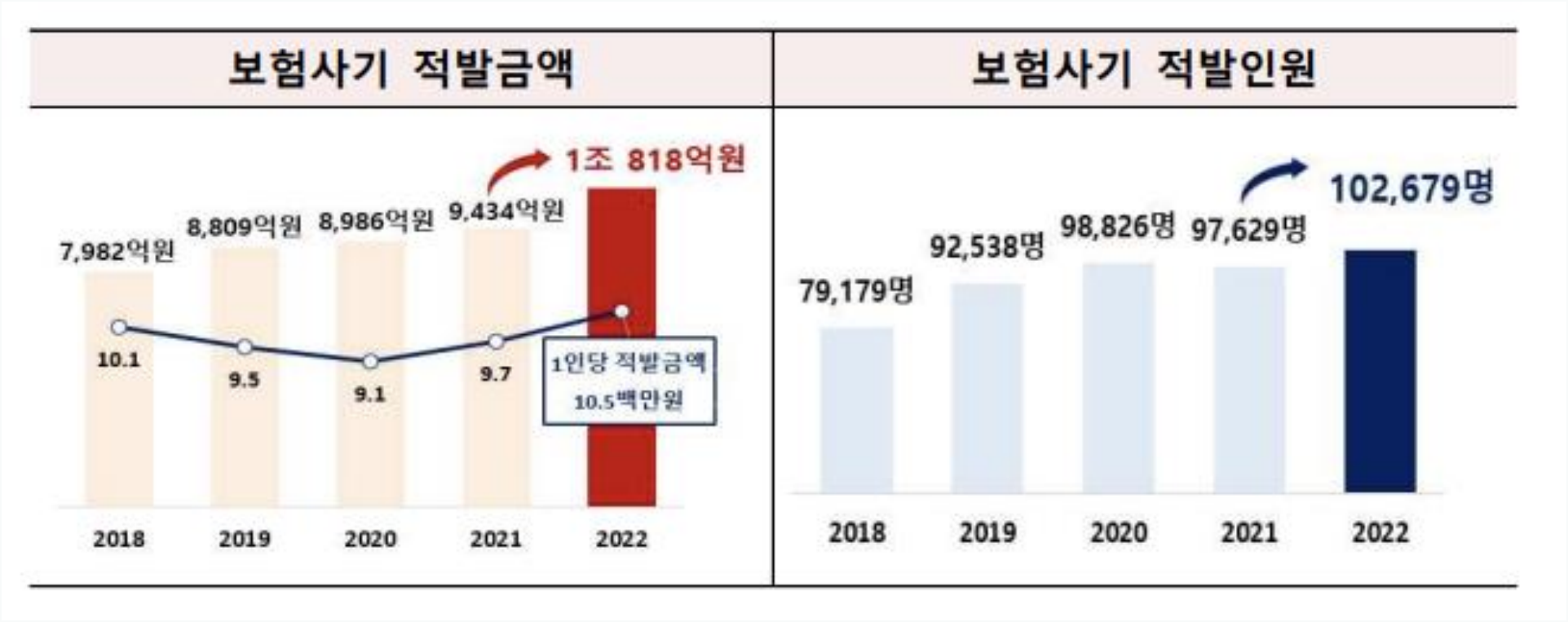
1. INTRO

INTRO

「보험사기방지 특별법」

보험사기행위란 보험사고의 발생, 원인 또는 내용에 관하여 보험자를 기망하여 보험금을 청구하는 행위.

INTRO

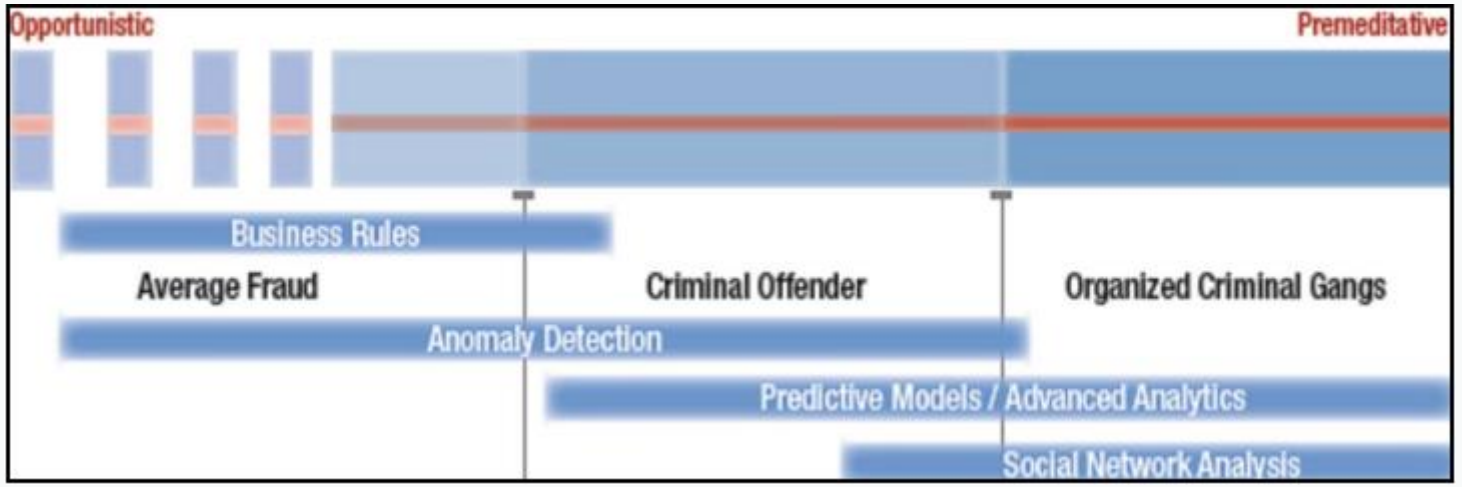


① 보험 회사 손해
② 보험사의 성실한 계약자들의 보험료 인상을 초래
③ 보험 제도 자체의 존립 기반을 위협

INTRO

과거: 경험에 의존한 사기 적발

현재: 정확한 정보를 얻기 위해 과학적인 알고리즘 활용



DATA

보험 가입자

- 인구통계적 정보
- 신용등급
- 최초 고객 등록일
- 소득
- 총 보험 납입 금액 등

보험 계약

- 보험의 종류
- 보험상품 구입 채널
- 보험 만기일
- 주보험금
- 합계보험금 등

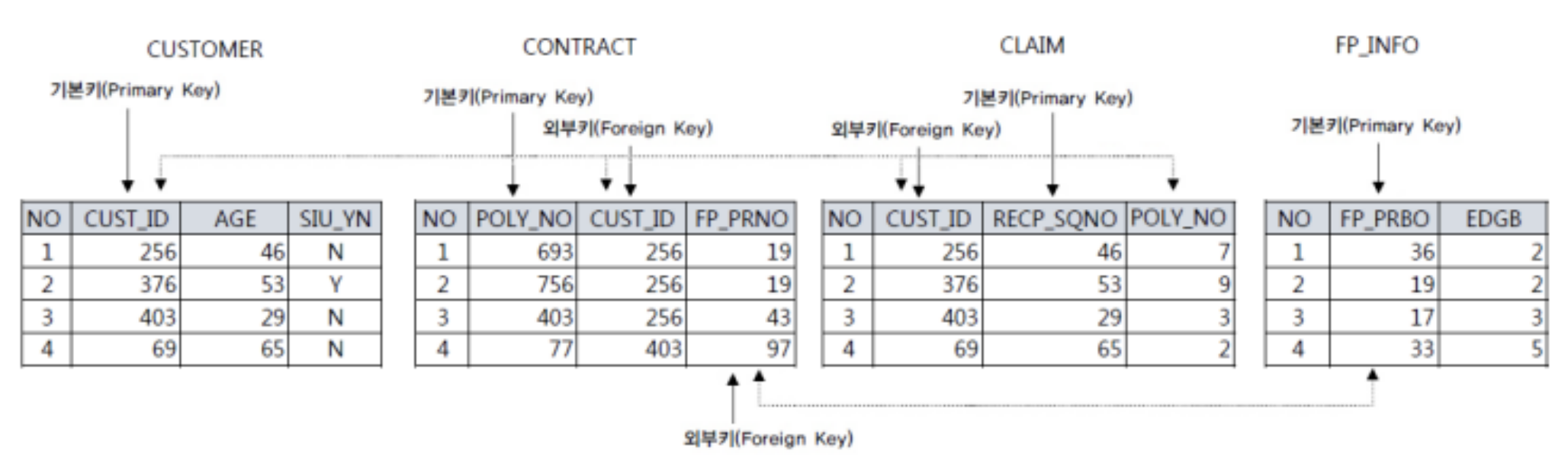
보험 청구

- 보험금 청구 일자
- 사고 원인
- 입원 기간
- 치료병원
- 실손 처리 여부
- 청구금액
- 지급금액 등

보험 설계사

- 재직 여부
- 입사년월
- 퇴사년월
- 학력
- 설계사 이전 직업

DATA



NO	CUST_ID	AGE	SIU_YN
1	256	46	N
2	376	53	Y
3	403	29	N
4	69	65	N

NO	POLY_NO	CUST_ID	FP_PRNO
1	693	256	19
2	756	256	19
3	403	256	43
4	77	403	97

NO	CUST_ID	RECP_SQNO	POLY_NO
1	256	46	7
2	376	53	9
3	403	29	3
4	69	65	2

NO	FP_PRBO	EDGB
1	36	2
2	19	2
3	17	3
4	33	5

2. EDA & FEATURE SELECTION

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 62 (2012) 989 – 994

Procedia
Social and Behavioral Sciences

WC-BEM 2012

A fraud detection approach with data mining in health insurance

Melih Kirlidog[a,b*], Cuneyt Asuk[b]

[a]*North-West University, Vanderbijlpark, South Africa*
[b]*Marmara University, Istanbul, Turkey*

변수 선택 및 가공에 참고

보험 사기자가 보이는 경향

1) 과도한 의료비용

2) 짧은 기간 안에 많은 보험청구

3) 보험청구일자와 만기일의 차이가 많지 않은 경우

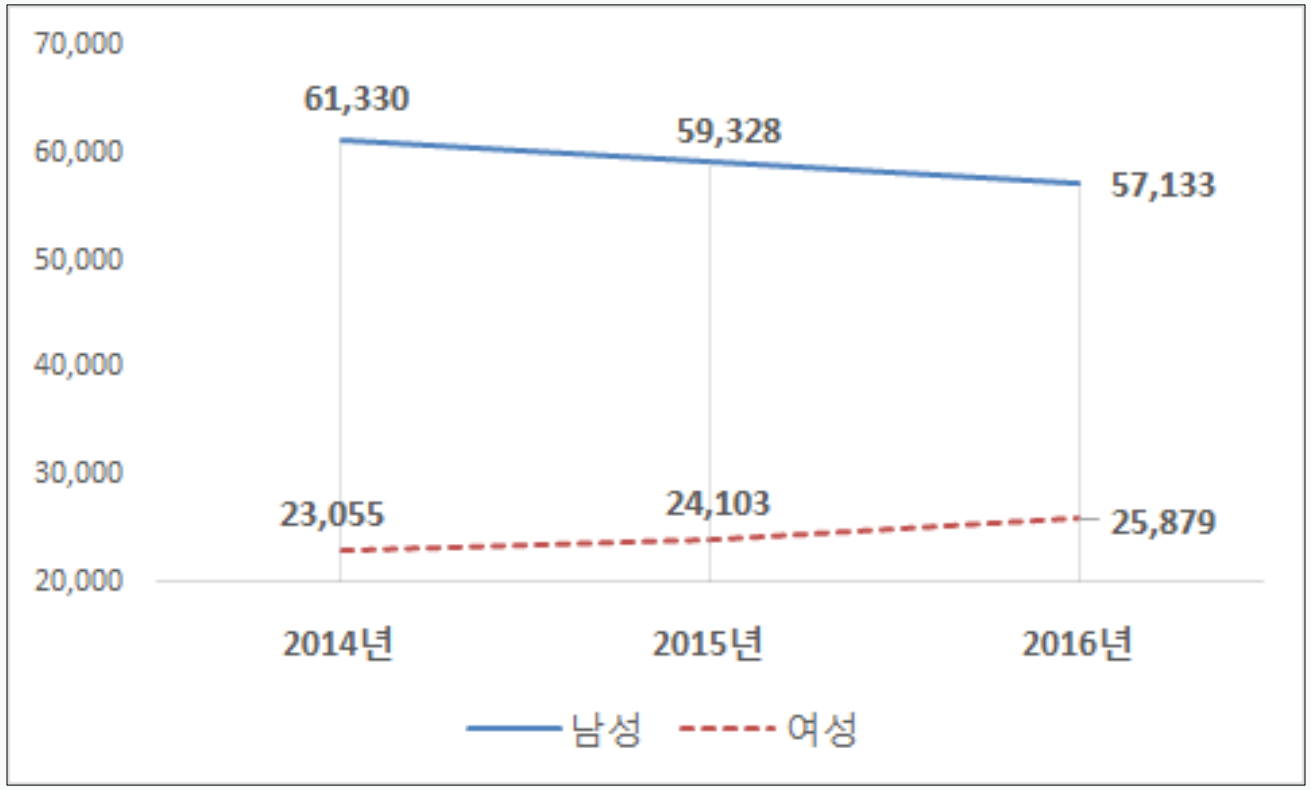
4) 많은 횟수의 보험청구

5) 과도한 치료기간

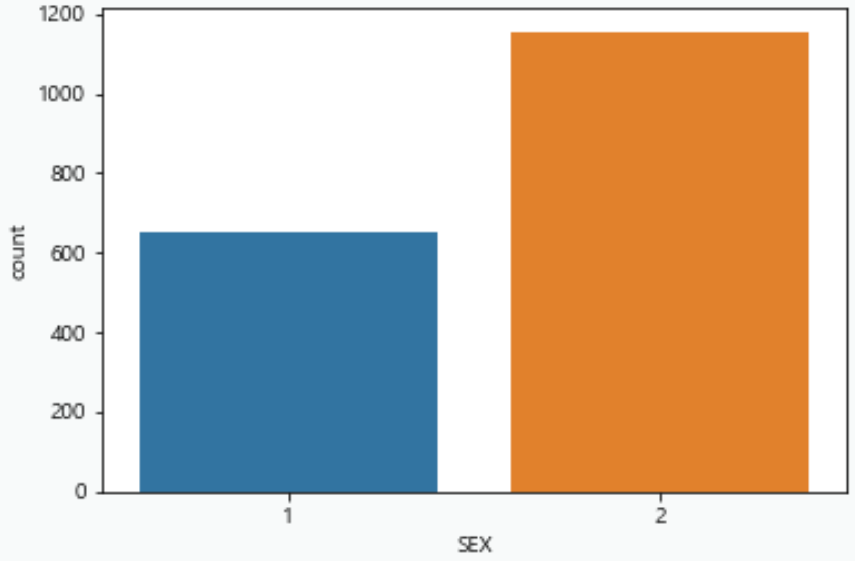
6) 보험사기와 관련이 많은 병원을 이용

7) 추가적으로 유의미한 결과를 보이는 변수 확인

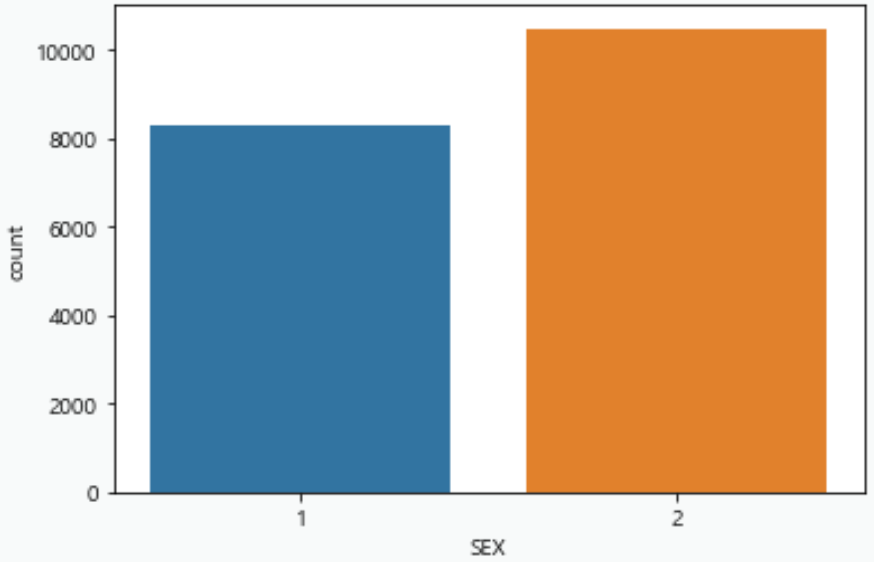
성별


70,000
60,000
50,000
40,000
30,000
20,000
61,330
59,328
57,133
23,055
24,103
25,879
2014년
2015년
2016년
남성
여성

성별



사기



사기 아님

여성의 경우 유의!

2.9064414184147852e-11
귀무가설을 기각합니다. 여자와 보험 사기 여부 간에 유의한 관련성이 있습니다.

→ 카이제곱검정, T-test

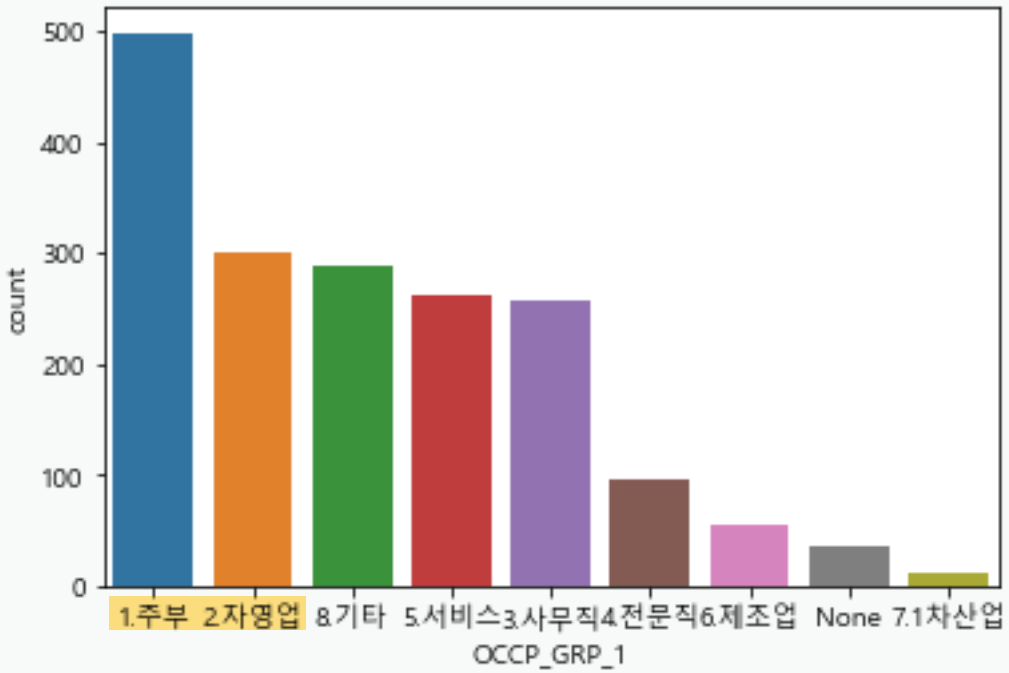
직업

(직업별) 보험사기 적발자의 직업은 **회사원**(19.1%), **무직 · 일용직**(11.1%), **전업주부**(10.6%), **학생**(4.9%) 순

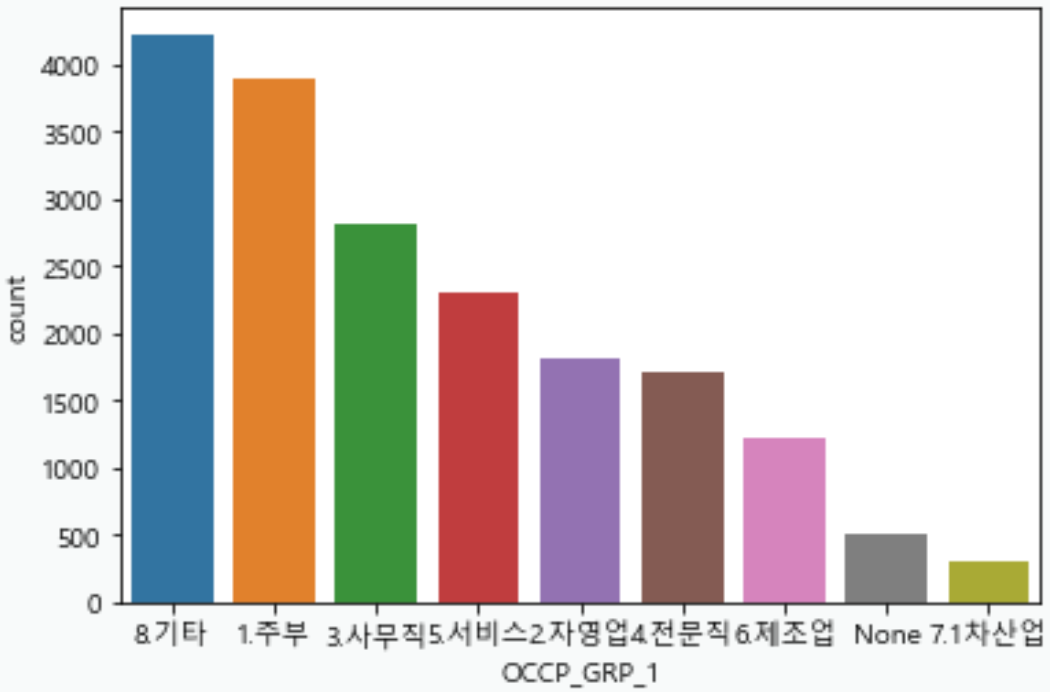
보험설계사, 의료인, 자동차정비업자 등 관련 **전문종사자**의 **비중**은 **4.3%**(4,428명) 수준

(직업) 직업별 보험사기자 비중은 무직·일용직(24.7%), 회사원(18.5%), 자영업(7.7%) 순으로 그 구성비는 **전년과 유사한 수준을 유지**

직업



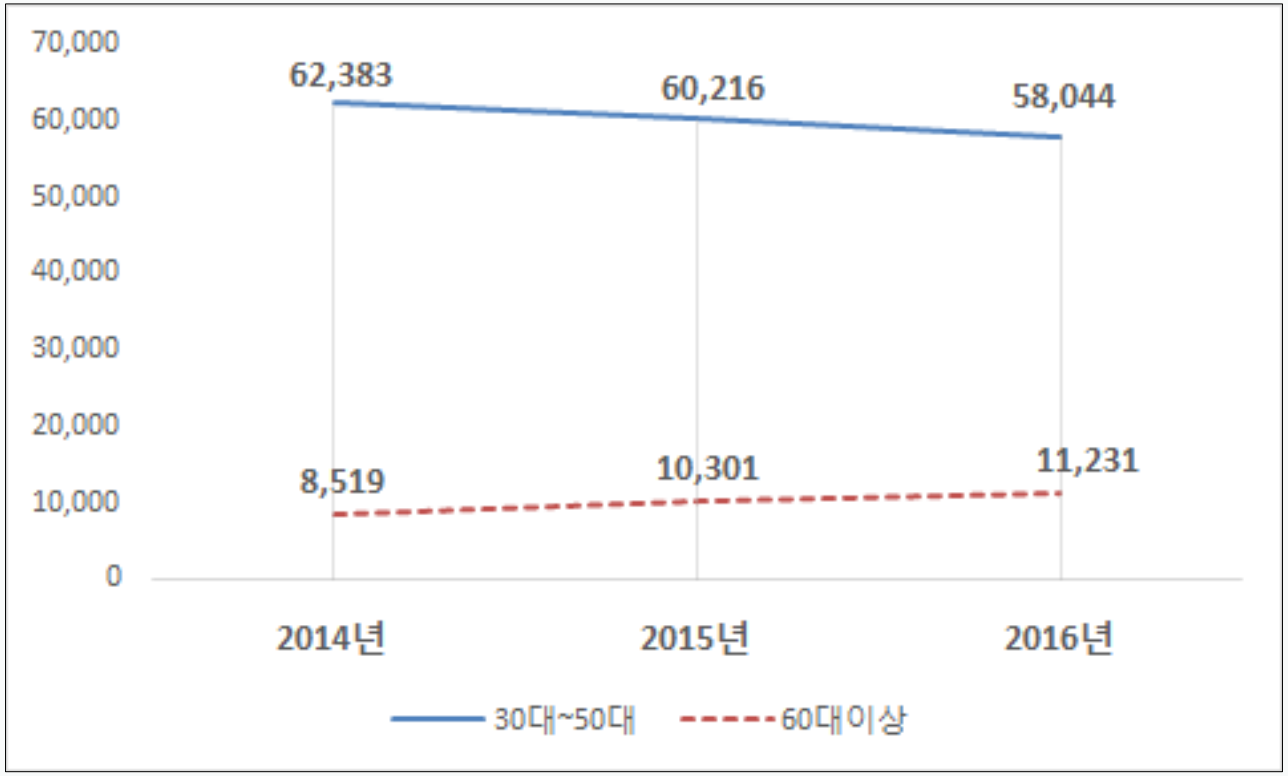
사기



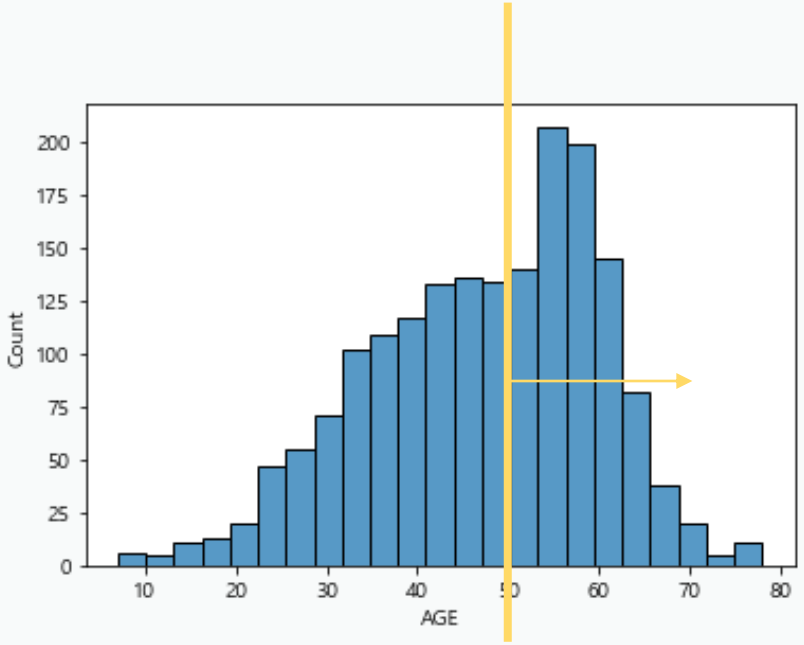
사기 아님

주부와 자영업자에게서 유의한 결과(6%씩 높음)
서비스업 (3%)

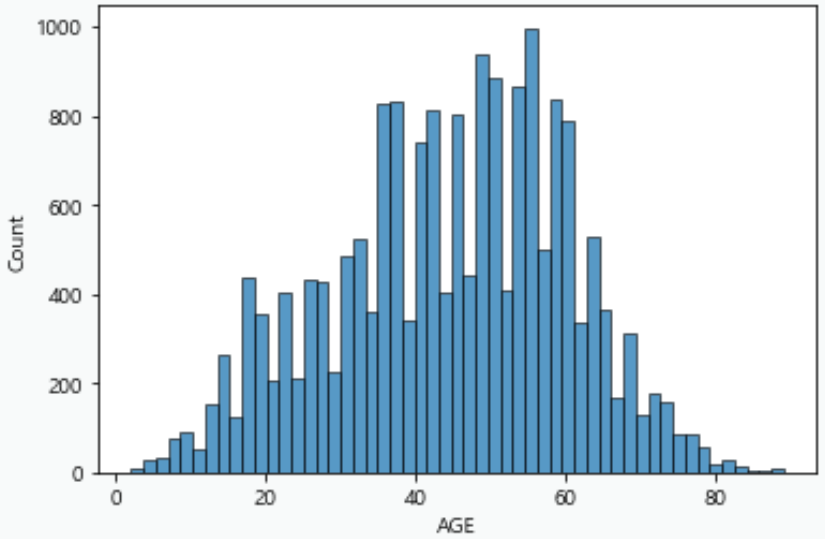
연령


70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
62,383
60,216
58,044
8,519
10,301
11,231
2014년
2015년
2016년
30대~50대
60대이상

연령

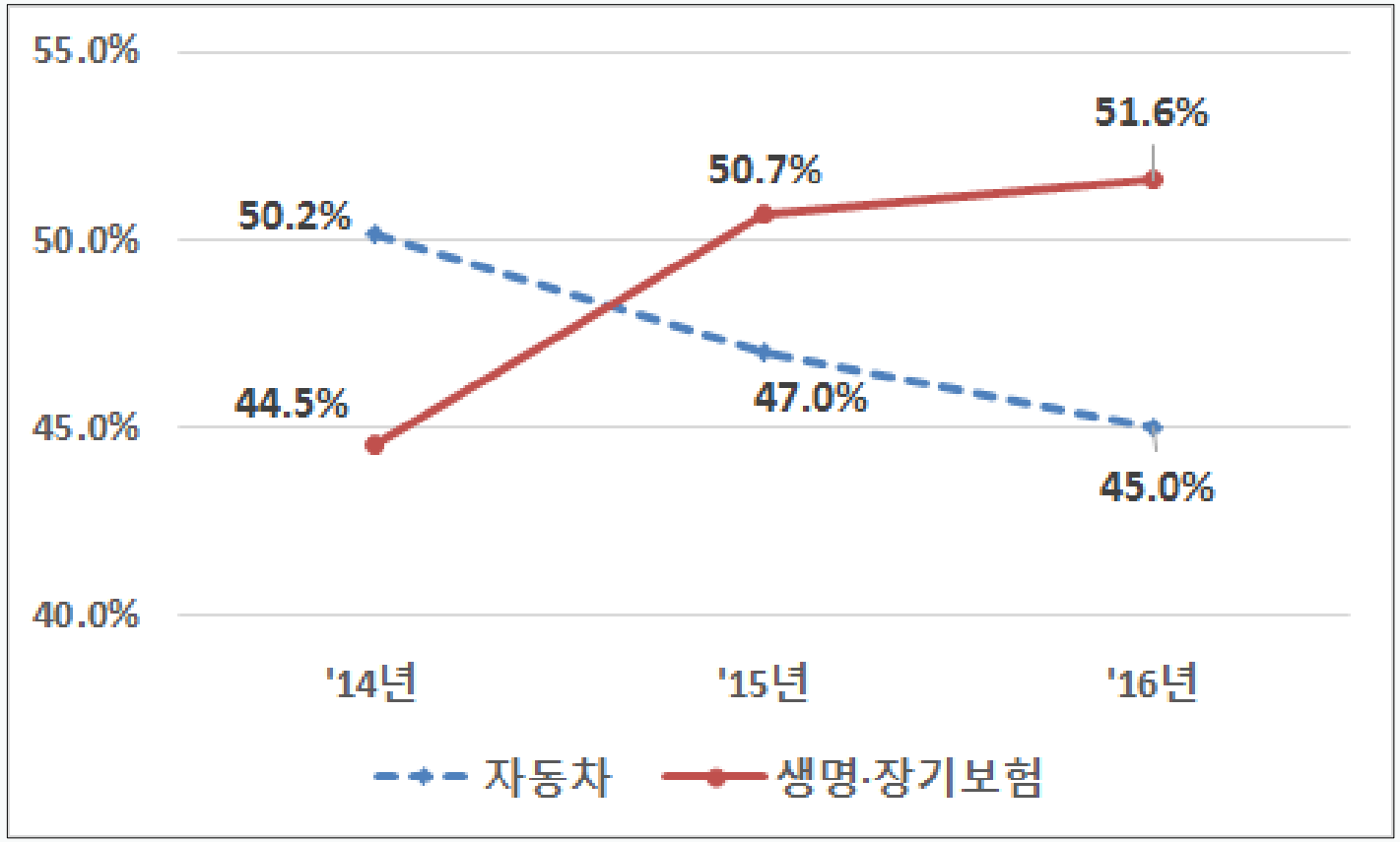

Count
AGE
0
25
50
75
100
125
150
175
200
10
20
30
40
50
60
70
80

사기


Count
AGE
0
200
400
600
800
1000
0
20
40
60
80

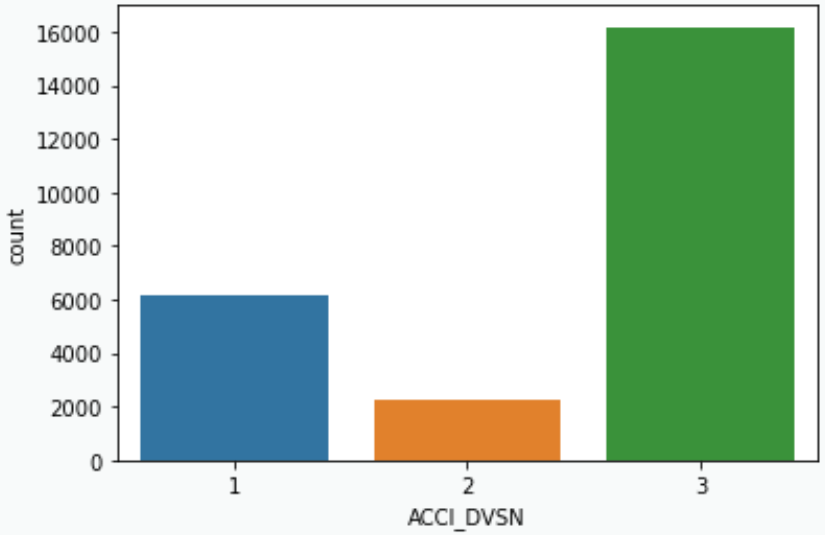
사기 아님

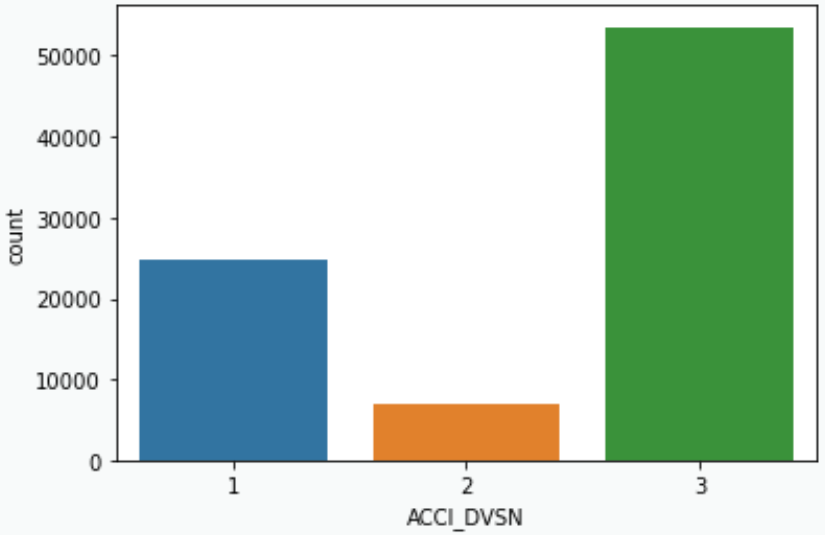
사고구분



사고구분

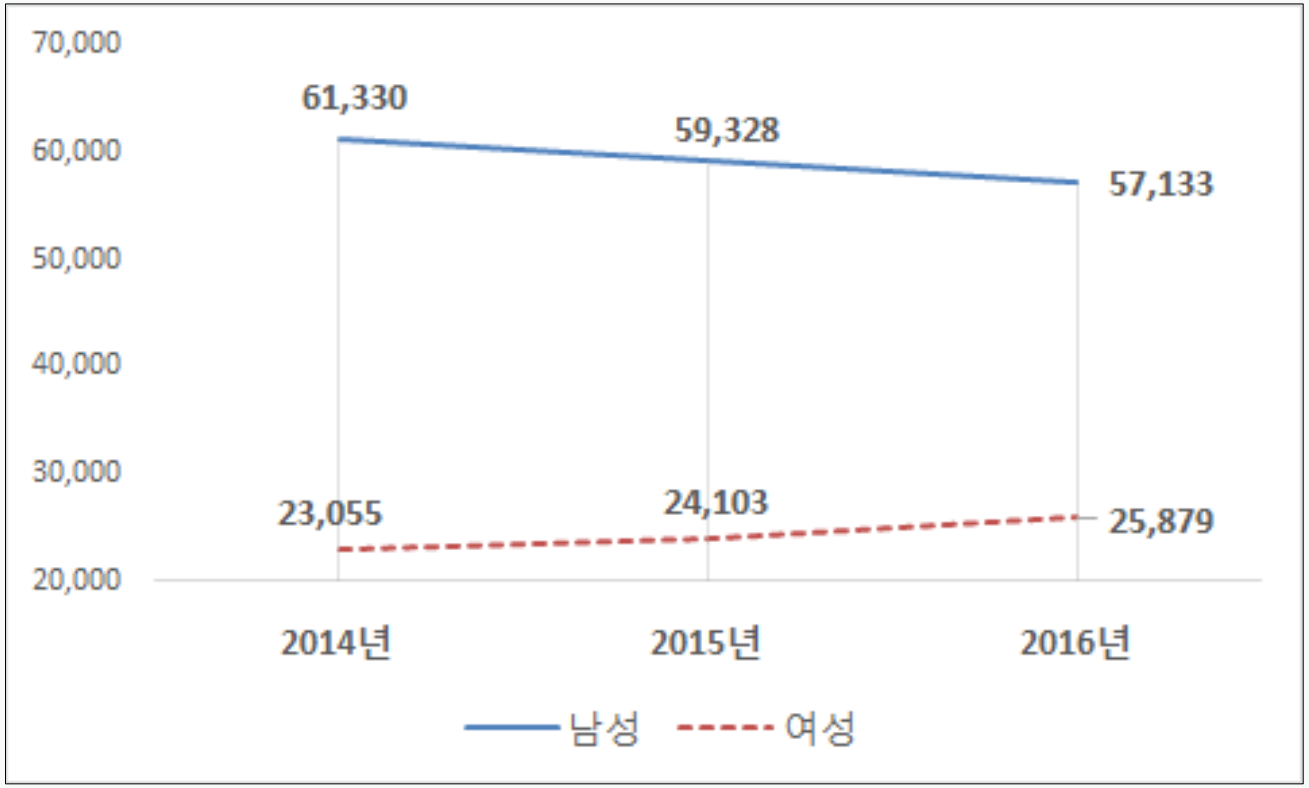
재해 교통재해 질병





질병이 유의

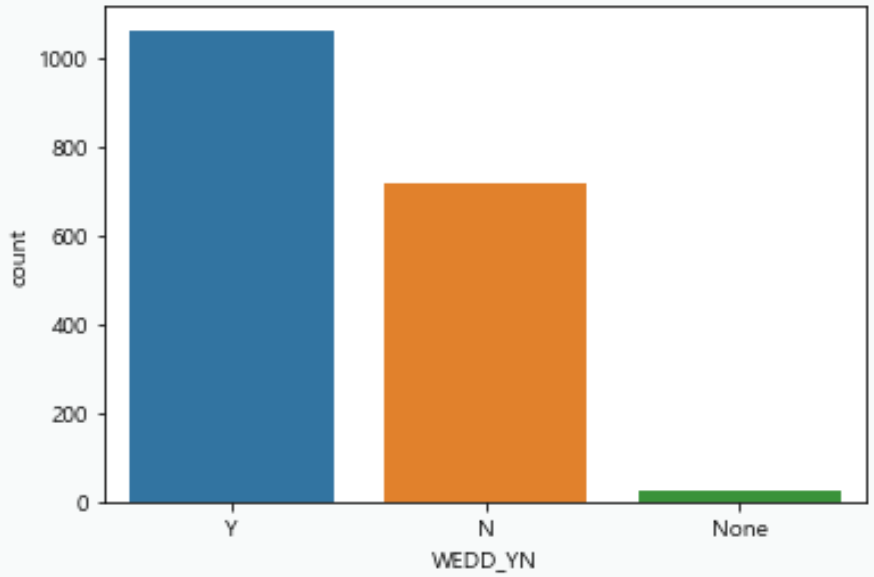
성별



자동차 보험사기 감소

미용·건강 목적 시술을
질병으로 조작 가능성

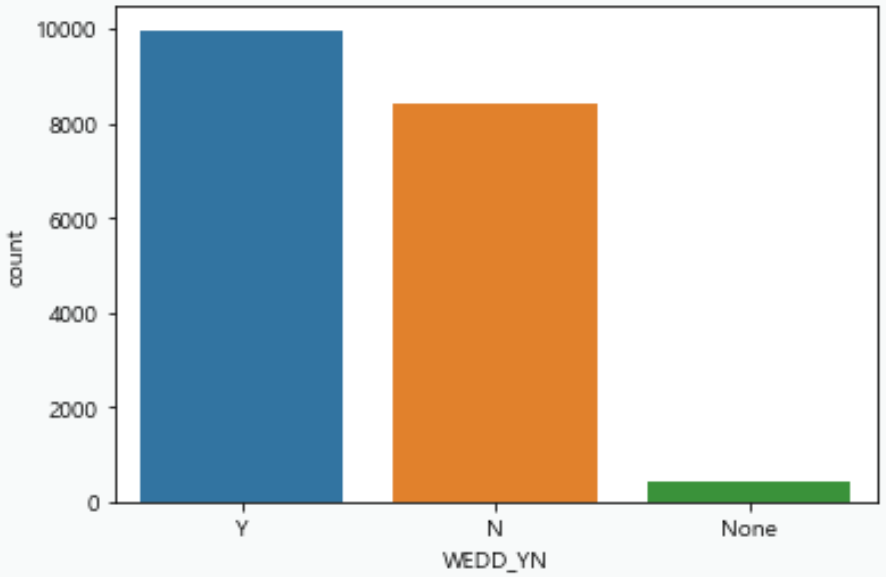
결혼 여부



사기

4%

>



사기 아님

결혼한 경우가 유의

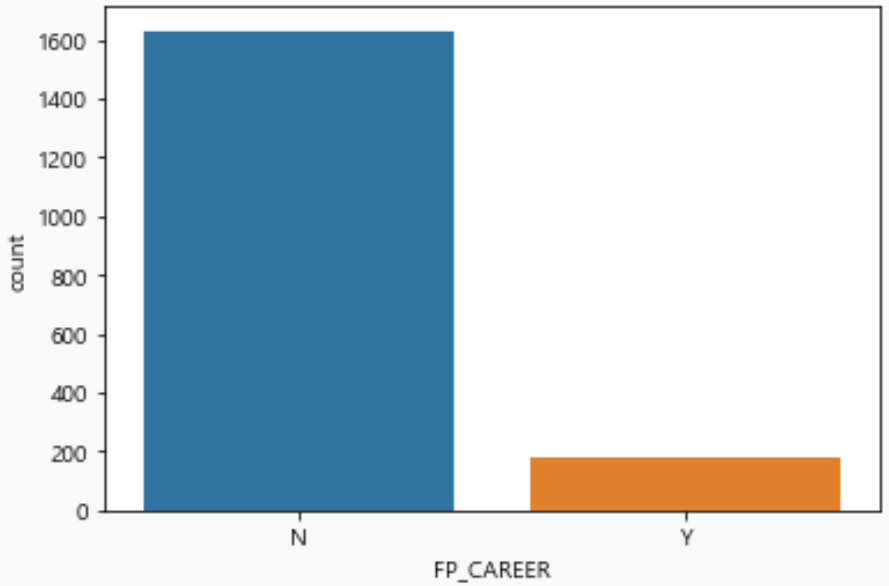
보험 설계사 경력

3 전직 보험설계사가 진단서를 위조한 사례

□ **(개요) 전직 보험설계사** 'A'는 자녀 2명과 함께 13개 보험사에 **63건의 보험을 가입한 후,** 입원 사실이 없음에도 **입원확인서, 진단서 등을 위조**하여 보험금 **1억 3천만원**을 편취 ('16.8월 검찰 송치)

□ **(특이사항)** 보험설계사 근무 경험을 토대로 보험사기로 의심받지 않도록 **보험금 청구 시기** 등을 **지능적으로 조절하여 청구**

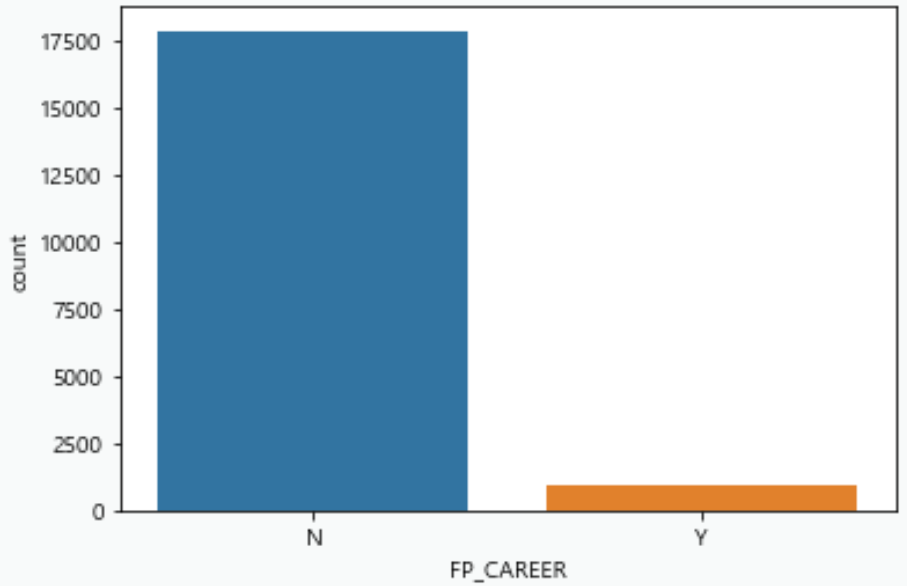
보험 설계사 경력



사기

5%

>



사기 아님

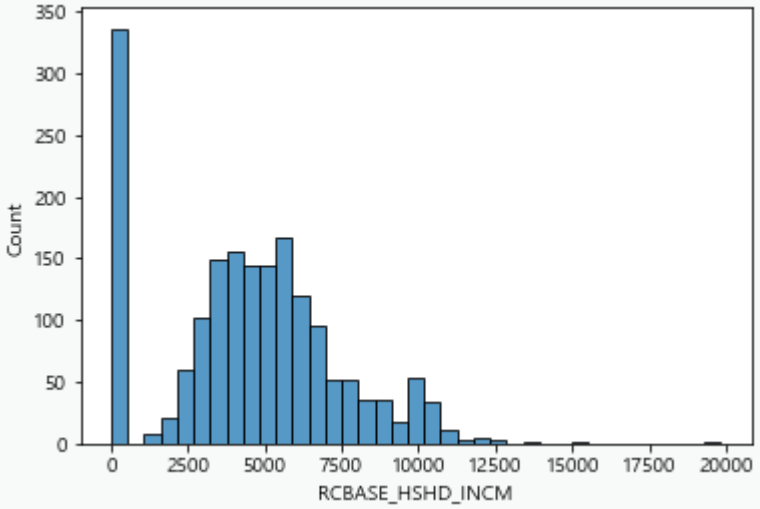
경력 있는 경우가 유의

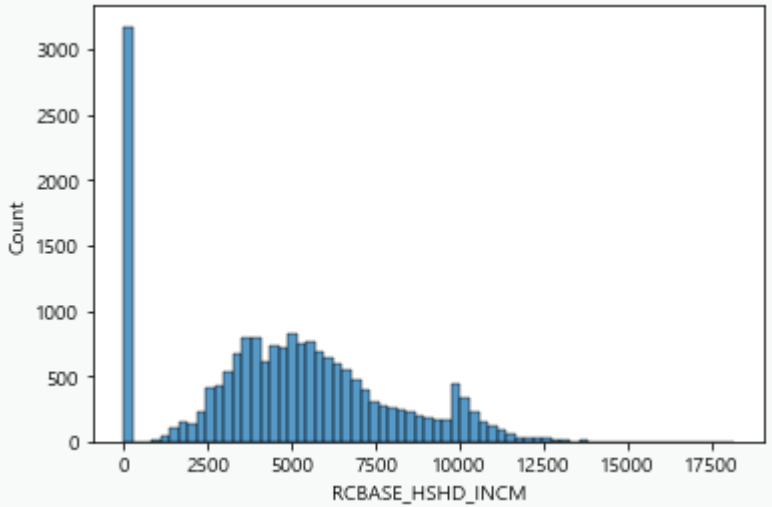
소득

✓고객 추정 소득

✓추정가구소득1(주택가격 우선)

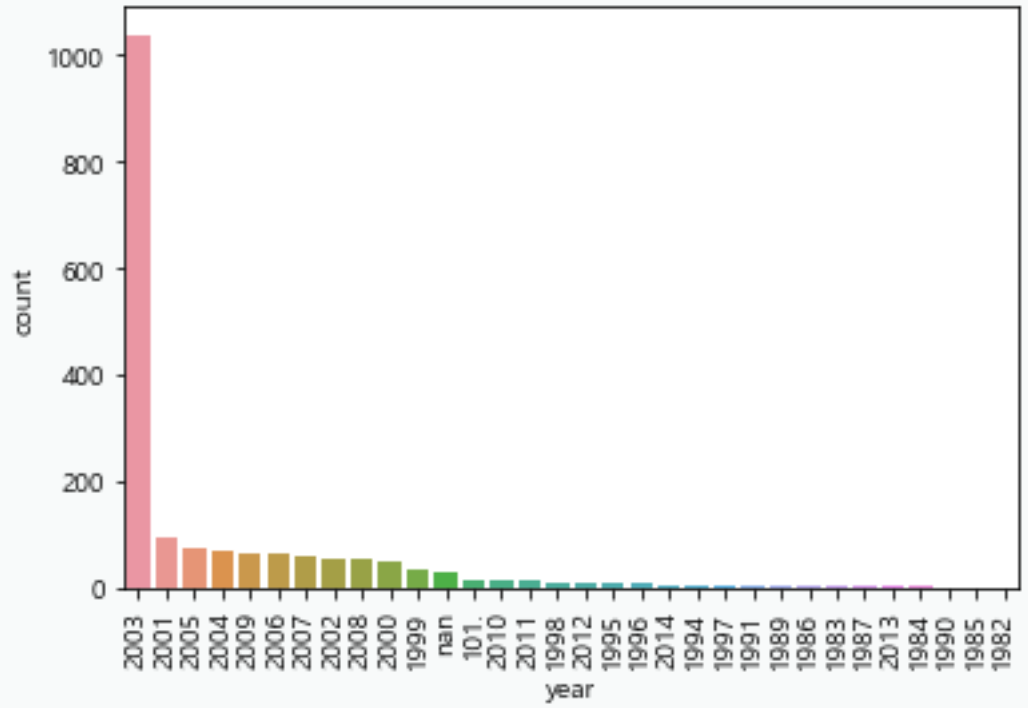
✓추정가구소득2(직업, 납입보험료 수준 우선하여)

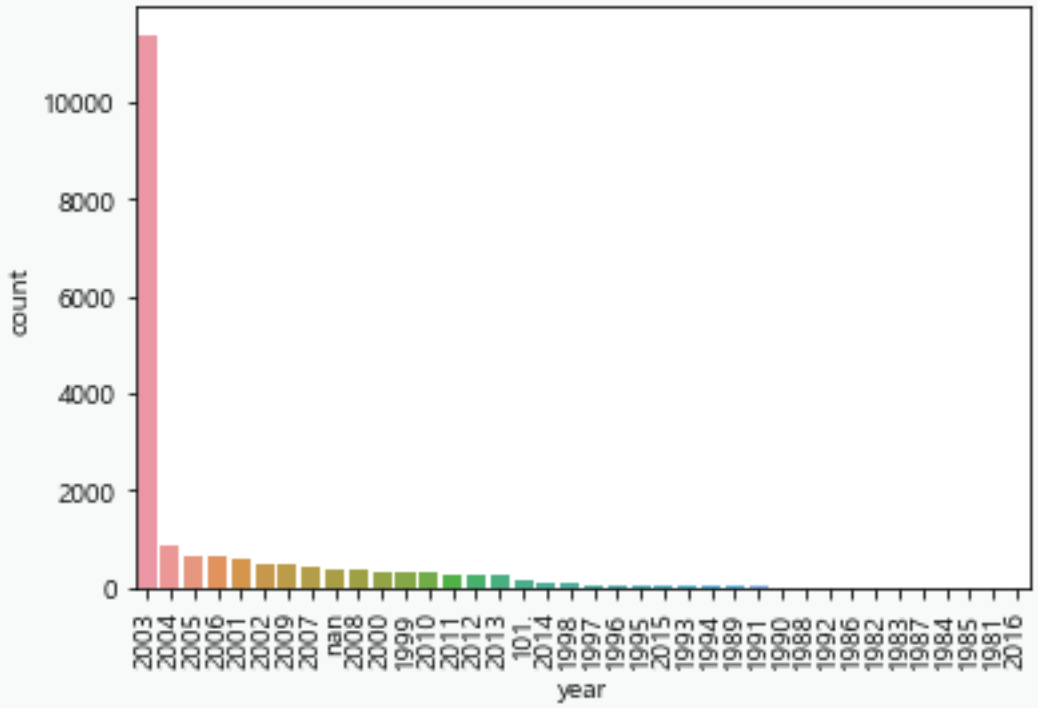




사기를 친 경우의 소득이 더 낮은 편

고객등록년월

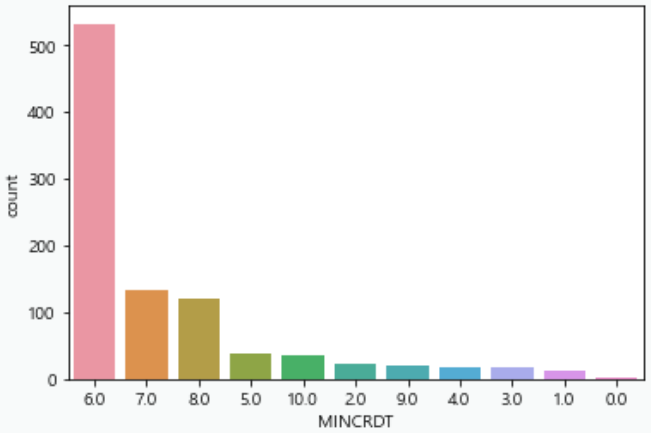


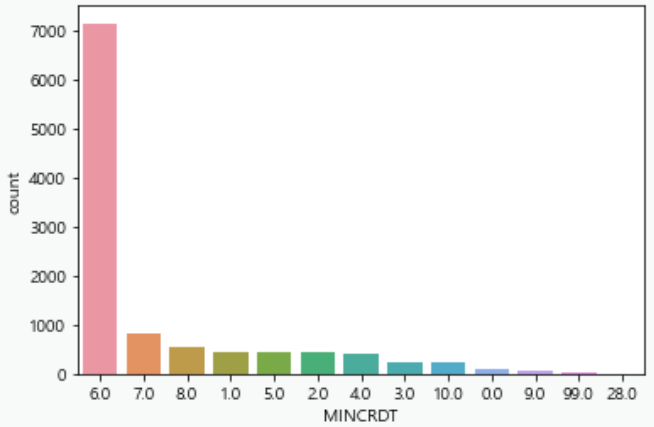


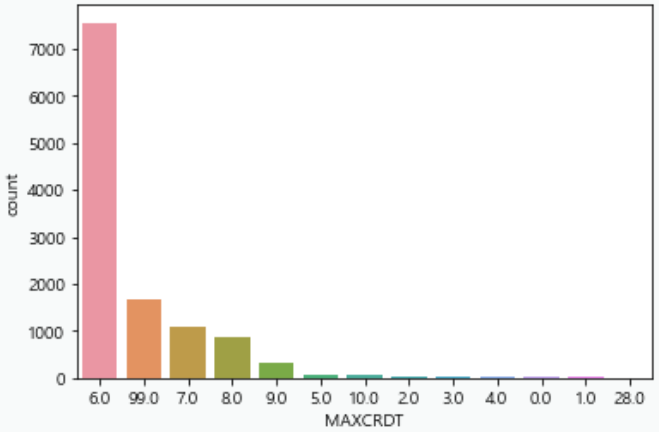
2003년, 2001년과 같은 2000년대 초반이 유의

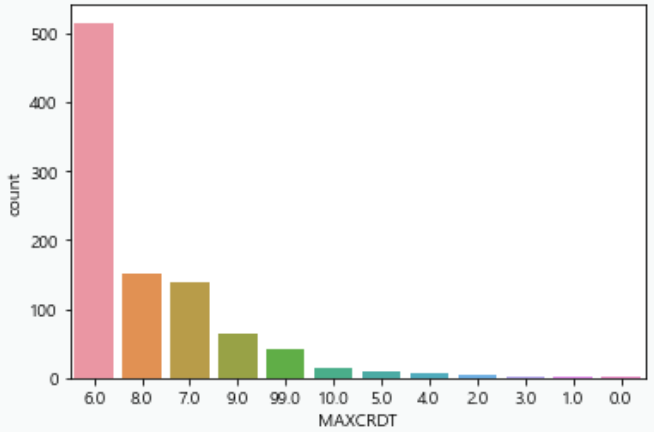
⟶ 데이터상 날짜를 기준으로 가입한지 며칠인지 계산하여 변수로 저장

신용등급



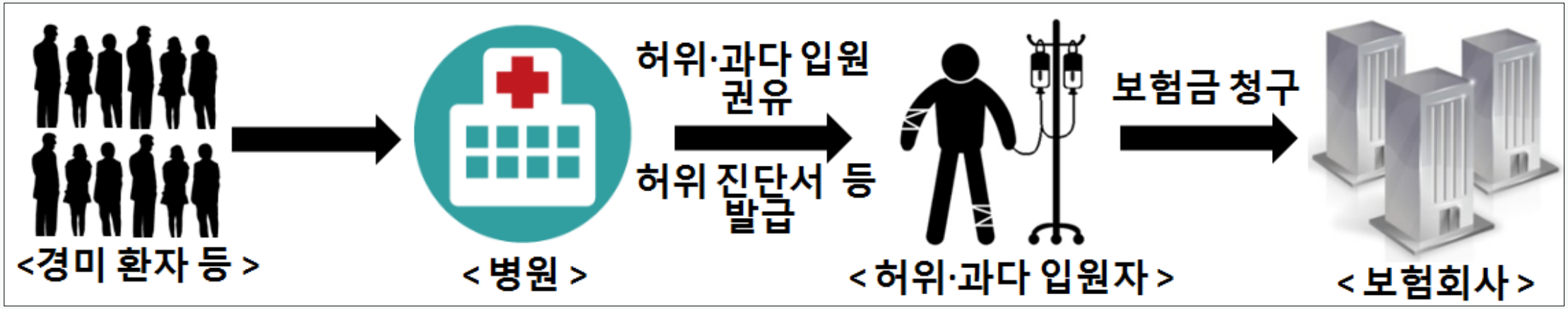




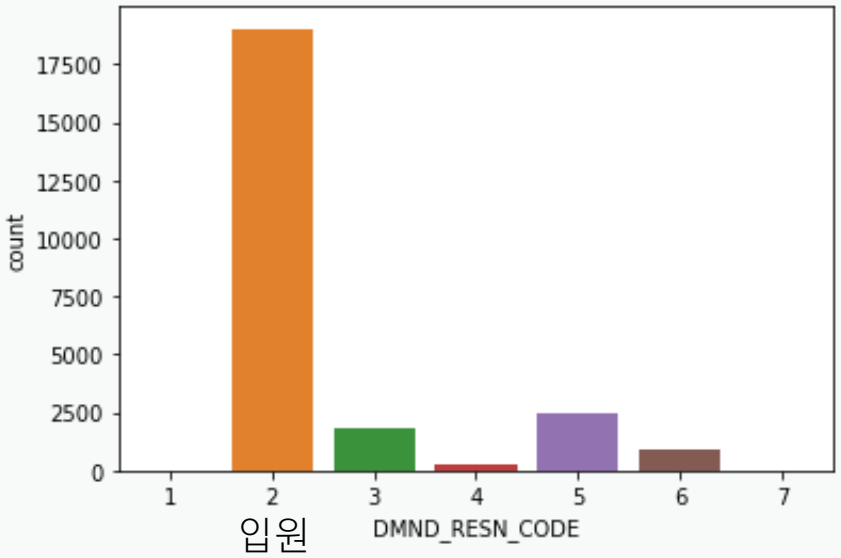


사기인 경우가 신용등급이 낮은 편

청구 사유



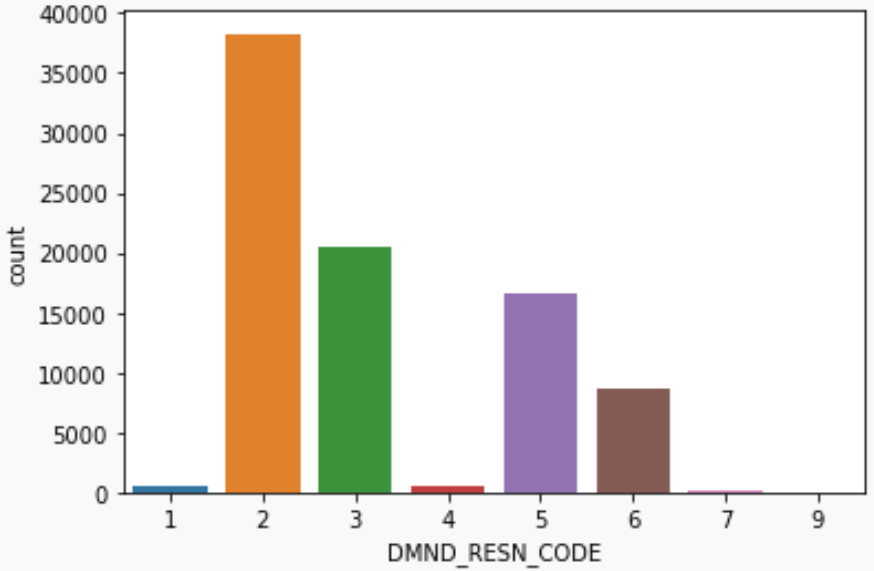
청구 사유



사기

23%

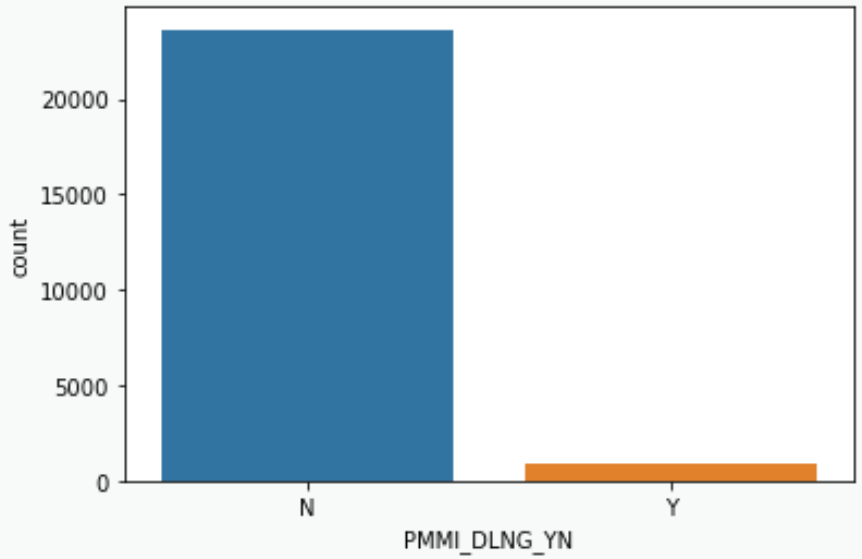
>



사기 아님

입원의 경우가 매우 유의

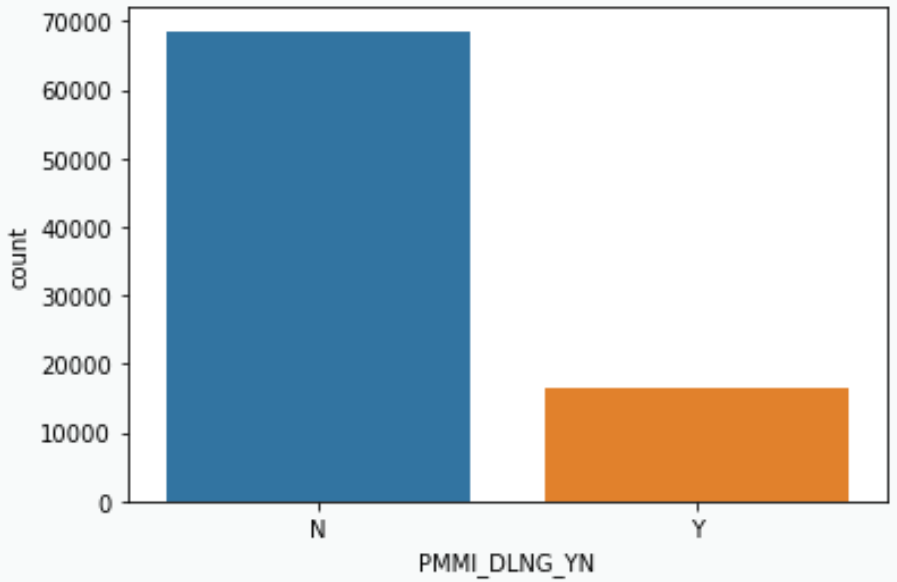
실손 처리 여부



사기

15%

<



사기 아님

실손 처리를 해준 경우가 유의하게 적음

보험 사기자가 보이는 경향

1) 과도한 의료비용

2) 짧은 기간 안에 많은 보험청구

3) 보험청구일자와 만기일의 차이가 많지 않은 경우

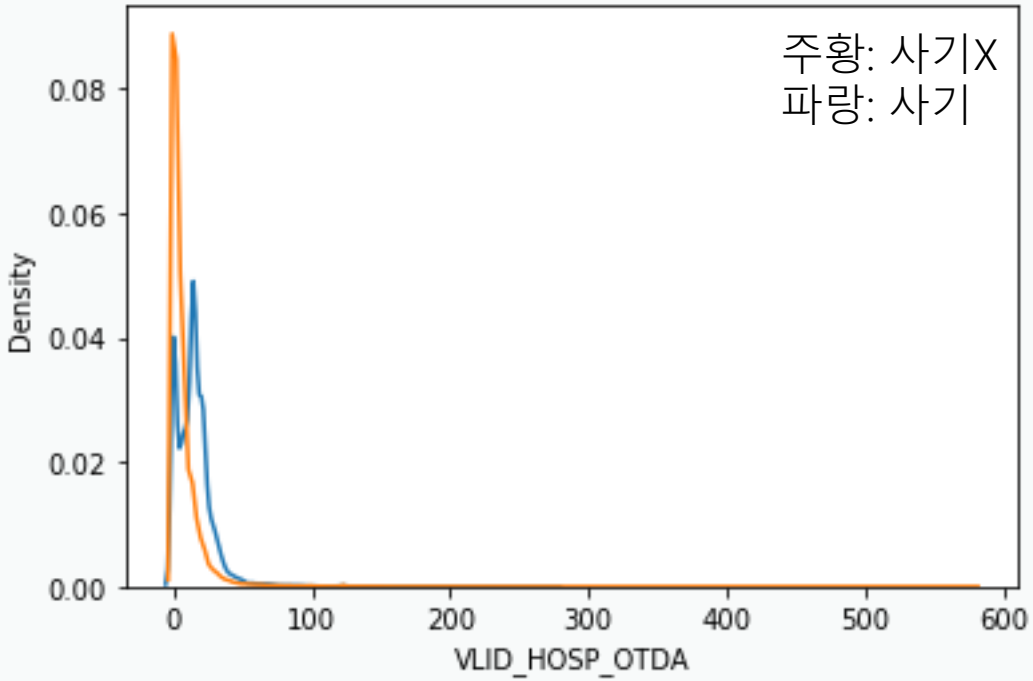
4) 많은 횟수의 보험청구

5) 과도한 치료기간

6) 보험사기와 관련이 많은 병원을 이용

7) 추가적으로 유의미한 결과를 보이는 변수 확인

유효입원/통원일수



사기인 경우가 일수가 길게 나타남.

보험 사기자가 보이는 경향

1) 과도한 의료비용

2) 짧은 기간 안에 많은 보험청구

3) 보험청구일자와 만기일의 차이가 많지 않은 경우

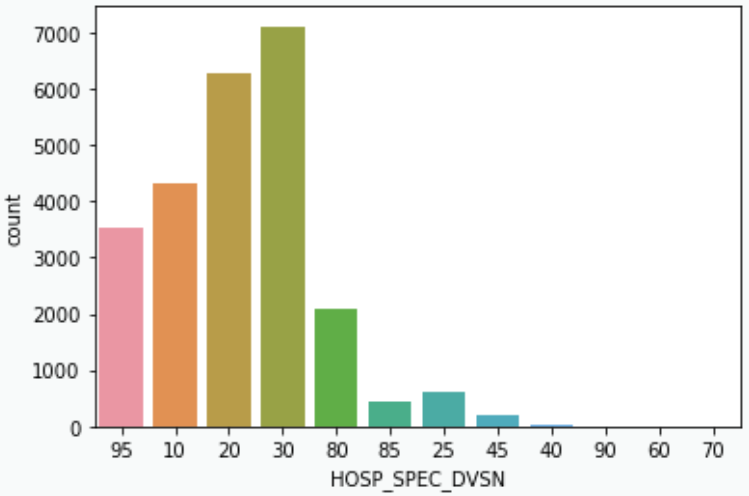
4) 많은 횟수의 보험청구

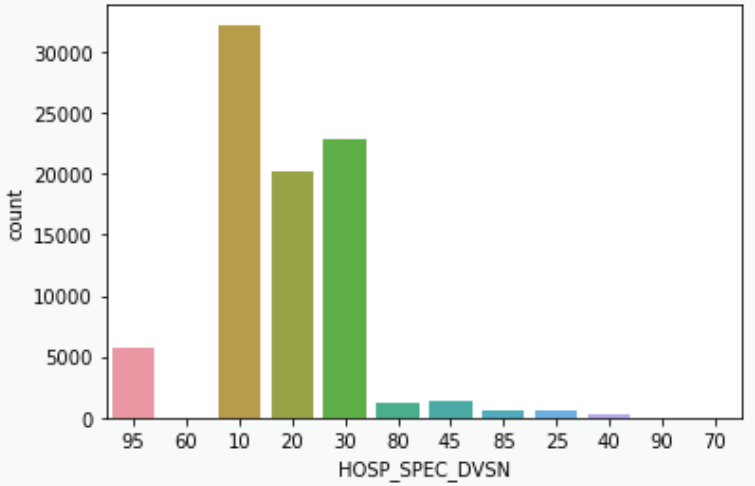
5) 과도한 치료기간

6) 보험사기와 관련이 많은 병원을 이용

7) 추가적으로 유의미한 결과를 보이는 변수 확인

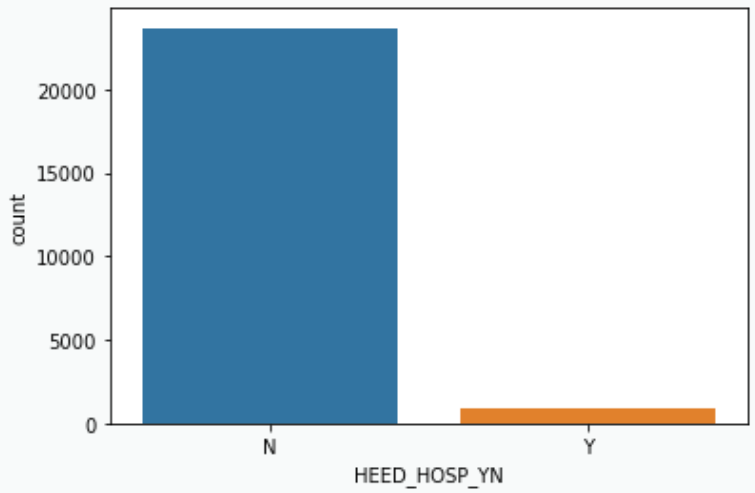
병원 종류





25(요양병원), 80(한방병원), 85(한의원), 95(의료기관이외)에서 확연한 차이를 보임.

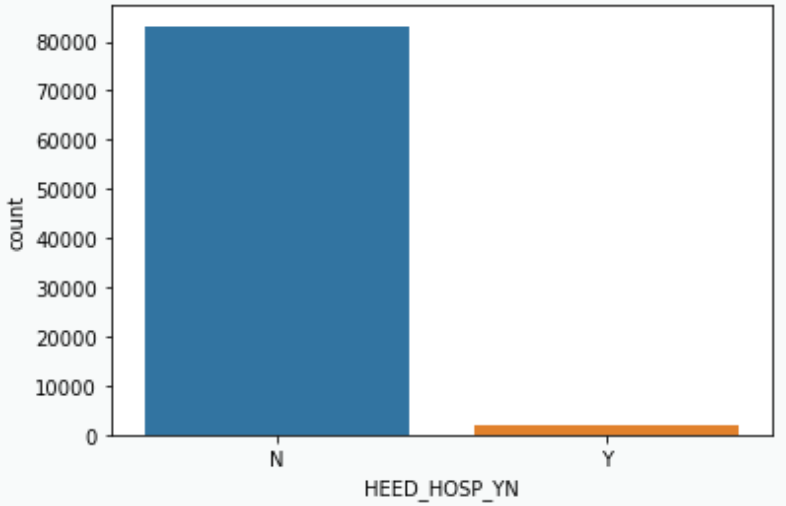
유의 병원 여부



사기

2%

>



사기 아님

사기 유의 병원인 경우가 조금 더 높음

실제 사례

Ⅰ 허위 진료비영수증 등의 발급을 통한 보험사기

□ **A병원**은 **실제 수술비를 초과한 금액**으로 비급여 진료비를 부풀려서 **허위 진료비영수증**을 **발급***하여 **환자**들로 하여금 **보험금**을 **편취토록 함**

* (예시) 실손보험이 있는 환자들에게 시술비로 약 300만원 상당의 허위 진료비영수증을 발급해주고, 지급보험금 중 200만원을 병원 관계자에게 이체하도록 제안

○ **브로커**들은 **A병원**에 **환자**를 **소개**하면 **알선수수료**를 지급받기로 병원과 **공모**하고, 실손의료보험 가입자를 병원에 **소개·알선**

○ 위와 같은 수법에 가담한 **환자**들은 허위 청구를 통해 **부당이득*** **편취**

* (예시) 240만원(300만원 * 0.8 <실손 보상비율 80% 가정>) - 200만원 = 40만원

실제 사례

② 한의원의 허위 진료기록부 발급을 통한 보험사기

□ **B한의원**은 **실손보험으로 보장되지 않는 보신제** 등을 **처방**하고 보험금 청구가 가능한 치료제로 **허위의 진료기록부**를 **교부**

- ○ **브로커들**은 B한의원에 **환자**를 **소개**하고 **매출액의 일부** 또는 **매월 수천만원**을 **알선수수료**로 **수취**
- ○ **다수의 보험소비자**가 **허위 청구서류**를 이용하여 **보험금 부당 편취**

3. FEATURE ENGINEERING

보험 사기자가 보이는 경향

1) 과도한 의료비용

2) 짧은 기간 안에 많은 보험청구

3) 보험청구일자와 만기일의 차이가 많지 않은 경우

4) 많은 횟수의 보험청구

5) 과도한 치료기간

6) 보험사기와 관련이 많은 병원을 이용

7) 추가적으로 유의미한 결과를 보이는 변수 확인

변수 생성

1) 보험 계약 개수

2) 보험 청구 개수

3) FP의 risk 계산

➡ 사기로 처리된 경우 주선한 FP에게 가중치 부여

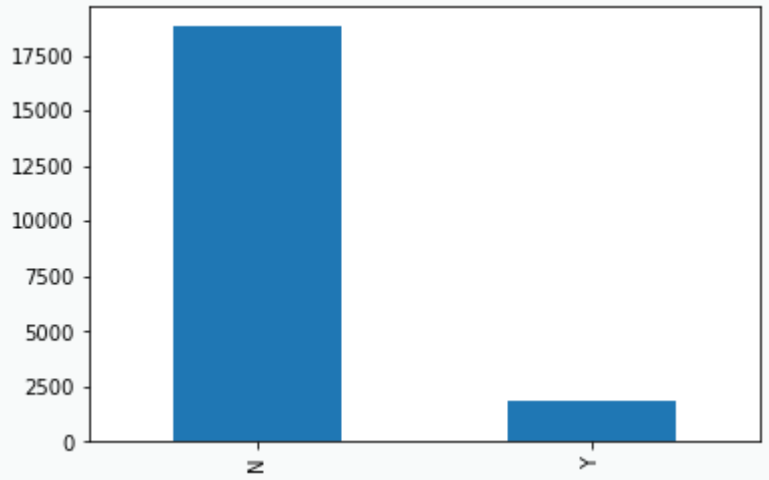
join	risk	claim
34	0.0	4
34	0.0	4
34	0.0	4
34	0.0	4
34	0.0	4
...	...	...
1	0.0	2
2	0.0	2
2	0.0	2
2	0.0	2
2	0.0	2

결측치 처리

1) 주택거주코드, 병원 종류 : 새로운 범주로 저장
2) 총납입액, 자녀 수, 등록일수 : 평균값 처리
3) 최소/ 최대 신용등급: 미확인시 6으로 처리
4) FP risk: 0으로 처리

4. MODEL 1

모델 선택



1. 변수가 많다.
2. 범주형과 연속형 변수가 섞여 있다. ⟶ 랜덤 포레스트, XGBoostClassifier
3. 불균형 데이터

평가 지표

1. F1-score

불균형 데이터셋에서 효과적으로 사용되는 지표.

분류 문제에서 주로 평가 지표로 활용됨.

➡ 1에 가까울수록 좋은 분류 모형으로 판단.

2. 특이도

보험 사기자를 옳게 예측하는 것이 중요하여 참-거짓 비율이 중요

➡ 1-특이도: FPR(오분류율) 최소화 위해 높은 특이도 필요

MODEL1 결과

- Gridsearch CV를 사용하여 최적 하이퍼 파라미터 조정후의 결과

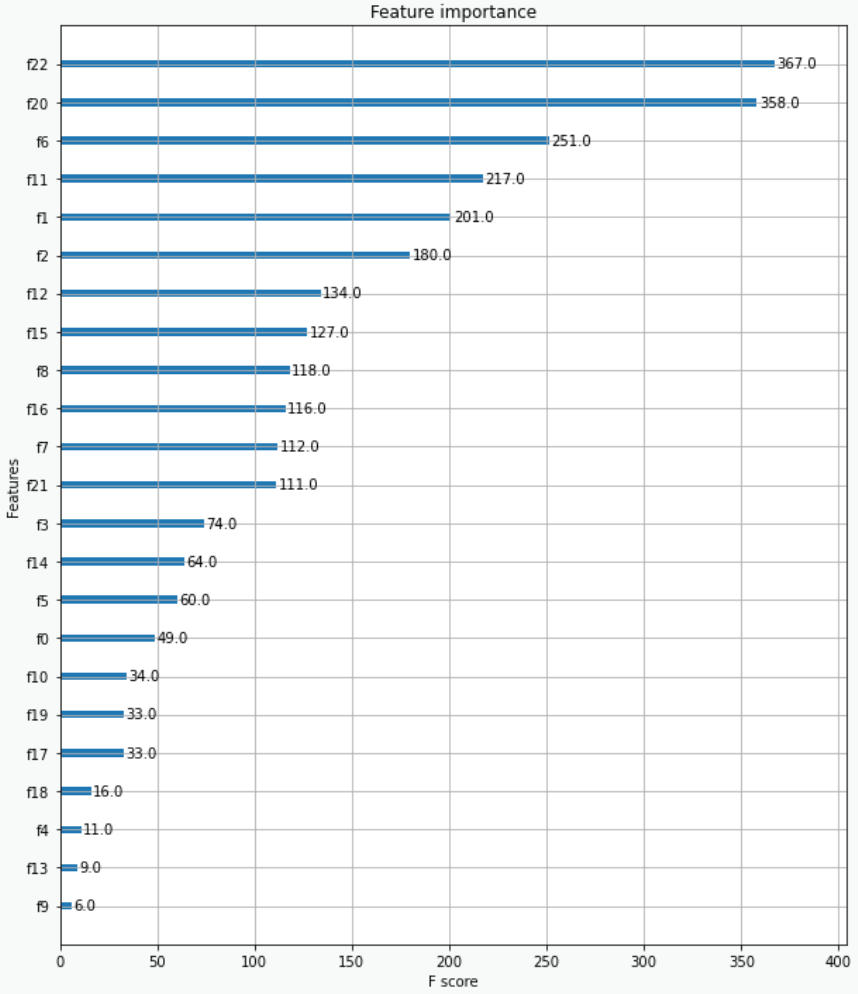
- 랜덤 포레스트 (SMOTE 적용 결과)

```
오차 행렬
[[99575  2594]
 [ 1186 28940]]
특이도: 0.9746, 정밀도: 0.9177, 재현율: 0.9606,    F1: 0.9387
```

- XGBoost (SMOTE 적용 안 함)

```
오차 행렬
[[102006    163]
 [  1683  28443]]
특이도: 0.9984, 정밀도: 0.9943, 재현율: 0.9441,    F1: 0.9686,
```

MODEL1 결과



청구 개수, 계약 개수, 총납입액, 추정가구소득, 연령이 분류에 중요함을 확인.

MODEL 2

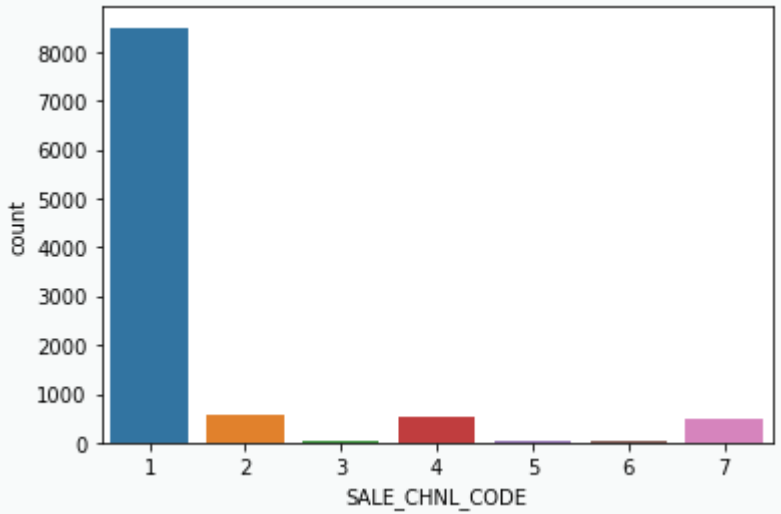
MODEL2 수정 사항

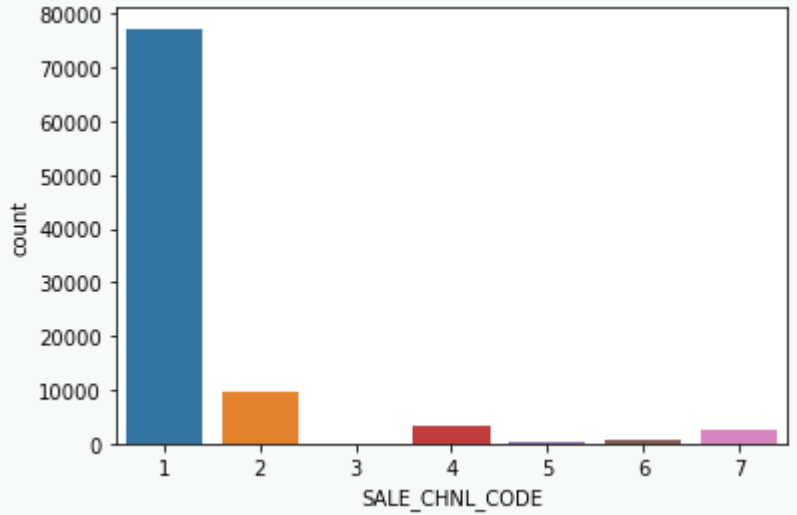
1) 청구 사유 코드: 입원 유의(Y)

2) 병원 종류: 요양병원, 한방병원, 한의원, 그외 기간에서 유의(Y)

3) 사고 원인에 대한 결과 코드 추가

4) 고객이 상품을 구매한 경로: 법인인 경우 사기가 낮은편(N)





MODEL2 결과

- XGBoost

```
오차 행렬
[[102043    126]
 [  1436  28690]]
특이도: 0.9988, 정밀도: 0.9956, 재현율: 0.9523,    F1: 0.9735
```

더 향상된 결과를 보임

5. DISSCUSION

DISCUSSION

1. FP risk가 중요 변수로 여겨졌으나 결측치가 너무 많아 모델에 큰 영향을 주지 못하였다. 보험 회사측에서 이와 관련한 양질의 데이터를 수집하는 것이 중요하다.

2. 교차 검증 수행시 F1-score가 특히 하락하는 경향을 보임. 과적합 가능성이 있어 이것을 조금 더 해결해야 할 필요가 있다.(조기종료)

THANK YOU