**POSCO 청년 AI Big Data 아카데미 9기**

**진행 현황 보고서**

A반 2조

강지영 김범수 김효진 배향운 양혜지 정지성 최지영

**데이터 전처리**

1. 보험가입 사전 승인 검진 정보 (insu\_pre\_review.csv)

· 결측 데이터 처리

- bust(가슴둘레) : 1개

- waist(허리둘레) : 1개

- pulse\_count(맥박) : 2개

→ 건강검진 데이터에서 나이, 키, 허리둘레 등 비슷한 특성을 가진 사람들의 평균치로 대체

· 파생 변수 생성

- BMI(체질량 지수)

→ 키와 몸무게를 이용한 비만 측정법 (몸무게/키\*2)

- Whtr(Waist to height ratio)

→ 허리둘레와 신장을 이용한 비만 측정법 (허리둘레/신장)

· 이상치 처리

- bp\_검사항목 (검사 항목당 판정결과)

→ 오판된 판정결과 데이터를 항목당 정상범주를 활용하여 재판정한 데이터로 수정

- bp\_judge (판정결과 점수 합)

→ 기존 검사 항목당 판정결과의 오류로 인해 올바르게 합산된 새로운 판정결과 점수의 합 설정

2. 보험 청구/지급 정보 (insu\_request.csv)

· 결측 데이터 처리

- insu\_pay\_amount(보험지급금액) : 52개

→ 보험상품, 질병당 다른 지급 금액으로 인해 지급액 대체 불가하므로 삭제 처리

-sick\_2nd(상병코드2) : 37,949개

-sick\_3rd(상병코드3) : 44,605개

→ 상병코드2, 상병코드3 변수 삭제 처리

· 파생 변수 생성

- dif\_req\_pay(청구-지급금액)

→ 손익 계산을 하기 위해 청구 대비 지급 금액이라는 새로운 변수 생성

3. 보험 가입 사전 승인 검진 정보 + 보험 청구/지급 정보

→ customer\_id(고객ID)를 기준으로 해당 고객의 사전 검진 정보와 보험 이용 내역을 연결한 데이터셋 생성

**데이터 모델링**

⠂목표변수 설정

1. 청구-지급
2. 위험도 산정
3. 지급액/납입액
4. 청구액

데이터 셋 : [insu\_request data.csv] + [insu\_pre\_review.csv]

→ 결측치 일반검진은 대체, 혈액검사는 제거

① 목표변수 (Y) : 청구-지급

１) 1분류 : 위험군과 비위험군

(청구 - 지급 > 0) 이거나 (청구 - 지급 == 0) 이면 0(비위험군) / 총 15119개

(청구 – 지급 < 0) 이면 1(위험군) / 총 3996개

⠂ 모델링 결과(Best 기준)

Train set : 0.4 / Validation set : 0.3 / Test set : 0.3

→ 모델 정확도 : 0.81

변수 중요도 : sick\_main\_ZZZ > sick\_main\_S13 > age 순

２) 2분류 : 위험군 내 저위험군, 고위험군 분류

위험군 1(청구-지급<0,총 3966개)을 위험도에 따라 다시 분류

(청구 – 지급)의 값 중 **중앙값(-20만원) 이용**

(A 청구 – 지급 >= -200,000) 을 위험군 1(저위험군)로 대체 / 총 2,793개

(B 청구 – 지급 < -200,000) 을 위험군 2(고위험군)로 대체 / 총 1,203개

⠂ 모델링 결과(Best 기준)

Train set : 0.4 / Validation set : 0.3 / Test set : 0.3

→ 모델 정확도 : 0.69

변수 중요도 : sick\_main\_zzz > 평생 건강 보장1 > age 순

３) 3분류 : 비위험군 내 수익성(-1), 비수익성(0) 분류

비위험군 0(청구 – 지급 > 0 이나 청구 – 지급 == 0, 총 15,119개)을 다시 분류(0과 -1)

(청구 – 지급)의 값 중 0을 기준으로 분류

A : (청구 – 지급 == 0)을 0(비수익성)으로 대체 / 총 12,003개

B : (청구 – 지급 > 0)을 -1(수익성)로 대체 / 총 3,116개

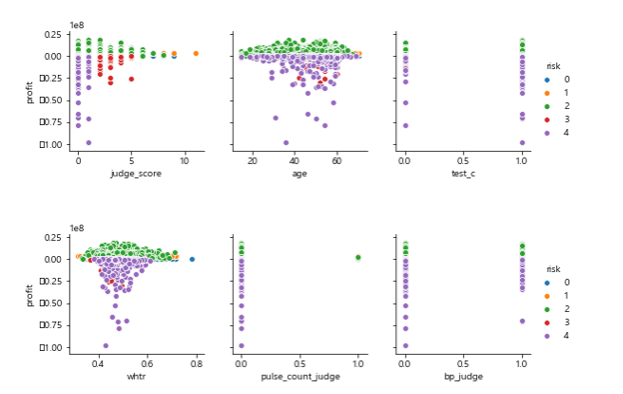
⠂ 모델링 결과(Best 기준)

Train set : 0.4 / Validation set : 0.3 / Test set : 0.3

→ 모델 정확도 : 0.69

변수 중요도 : sick\_main\_zzz > 평생건강보장1(보험명) > age 순

② 목표 변수 (Y) : 위험 정도 (risk라는 새로운 변수 생성)



⠂risk변수 생성 기준

|  |  |  |
| --- | --- | --- |
| risk | 대 상 | 설 명 |
| 0 | 거절 고객 | 가입 X, 위험 우려 X |
| 1 | 청구 이력 X | 수익 |
| 2 | 청구 이력 O, (납부액 > 지급액) | 보통 |
| 3 | 청구 이력 O, 저위험군 | 평균 손실이 적음 |
| 4 | 청구 이력 O, 고위험군 | 평균 손실이 큼 |

→ 3, 4 군집은 K-means으로 분류, 유사 특성의 2 집단을 형성,

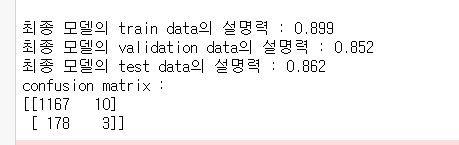
평균 손실 정도에 따라 손실 정도가 큰 집단은 고위험군, 적은 집단은 저위험군으로 분류

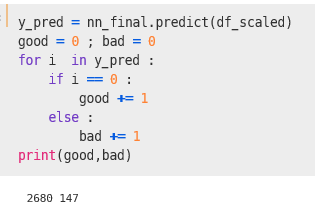
1 ) 1분류 : 청구 유무, 청구이력이 없는 고객(위험도1,2)과 있는 고객(위험도3,4)

⠂ 모델링 결과(Best 기준)

Train set : 0.4 / Validation set : 0.3 / Test set : 0.3

→ 유의 변수만 사용하여 모델 설명력 : 0.862





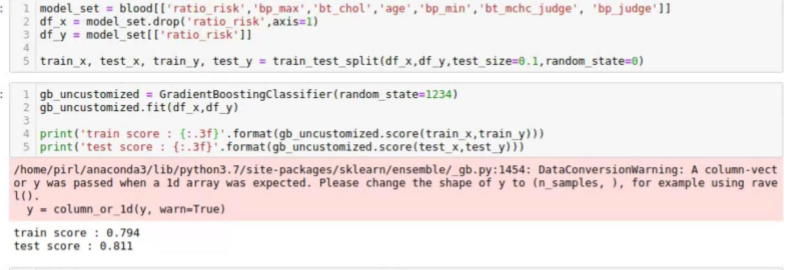
거절 고객 데이터 대입 : 위험군에 147명만 속하고, 나머지는 수익이라 판정

→ 거절 고객을 위험으로 판단한 기존 모델의 정확도가 너무 낮음 -> 현실적 X

2 ) 저위험군(3), 고위험군(4) -> 거절 고객 분류 결과가 147 -> 분류 현실성 X -> 기준 변경

③ 목표 변수2 (Y) = 지급액/납입액

1) Y의 평균보다 크거나 같은 경우 RISK = 1 (손해), 작은 경우 RISK = 0 (수익)



⠂ 모델링 결과(Best 기준)

→ 유의 변수만 사용하여 초기 모델 설명력 : 0.811

2) 거절 고객의 정보를 Gradient Boosting 모델에 적용

수익을 낼 것이라고 판단된 거절 고객 : 2272

손실을 낼 것이라고 판단된 거절 고객 : 72

→ 즉, 기존 모델이 거절 고객의 0.03만 정확하게 분류하므로 현실적이지 못함

④ 목표 변수 (Y) = 청구액

**< 질병에 따른 위험 분류 >**

1) 질병과 청구액 간 관계

- 가정 : 병에 따른 청구액이 있을 것이다.

- 병이 동일하여도 청구액이 다르다.

2) 보험에 따른 질병 청구액 관계

- 가정 : 보험 종류별로 동일 병에 따른 청구액이 다를 것이다.

- 동일 보험이여도 동일 병에 따른 청구액이 다르다.

**< 고려사항 >**

- 병 마다 데이터 수에 차이가 있다.

→ 기록된 데이터가 소수인 것은 모델을 생성하기 어려울 것으로 판단

- request 데이터 셋에 동일 request\_id로 여러번 청구하는 경우가 있다.

- 동일 질병이여도 청구마다 금액이 다르므로 금액을 기준으로 하기 어렵다.

- 청구 빈도를 고려한다.

→ 일부 높은 빈도의 질병을 선택하여 모델을 만든다.

**< 질병예측 모델 >**

고객 검사 데이터(pre\_review) 로 특정 질병을 예측하는 모델을 만든다.

INPUT : 혈액검사 데이터, 체중, 신장 등 / OUTPUT : 질병 유무

- 모든 질병을 고려하면 output 차원이 높아서 모델 생성이 어렵다.

- 높은 빈도의 질병들만 고려하여 차원을 한정한다.

- 주상병 뿐 아니라 2,3 상병도 label 로 사용한다.

ex) output label : [ 감기, 골절, 암, 타박상, 화상 ]으로 가정했을 때,

고객1(주상병 : 감기, 2상병 : 타박상) 인 경우 [1,0,0,1,0]으로 설정 한다.

결측치 제거 후 혈액검사를 받은 사람들의 데이터로 분석 시도

(청구금액 - 지급금액)을 산출해 0 미만일 경우 지급금액 많은 위험고객으로 분류

목표변수 : risk(0: 위험하지 않은 고객, 1: 위험한 고객)

설명변수 : 혈액검사 판정결과들

('bp\_judge', 'pulse\_count\_judge', 'bt\_chol\_judge', 'bt\_crea\_judge',

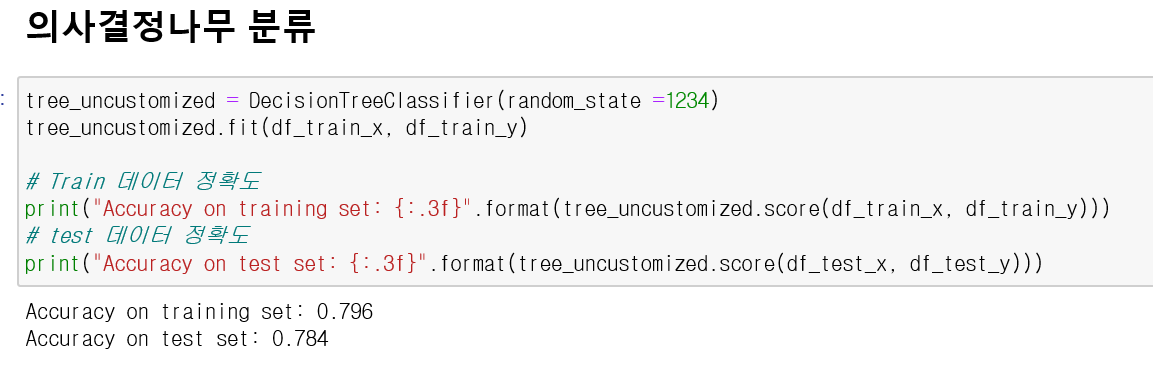
'bt\_gluc\_judge', 'bt\_hb\_judge', 'bt\_hbsa\_judge', 'bt\_hct\_judge',

'bt\_mch\_judge', 'bt\_mchc\_judge', 'bt\_mvc\_judge', 'bt\_plat\_judge',

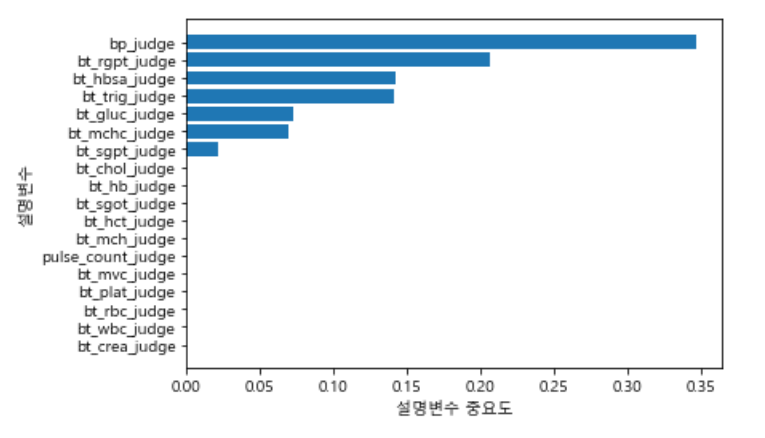
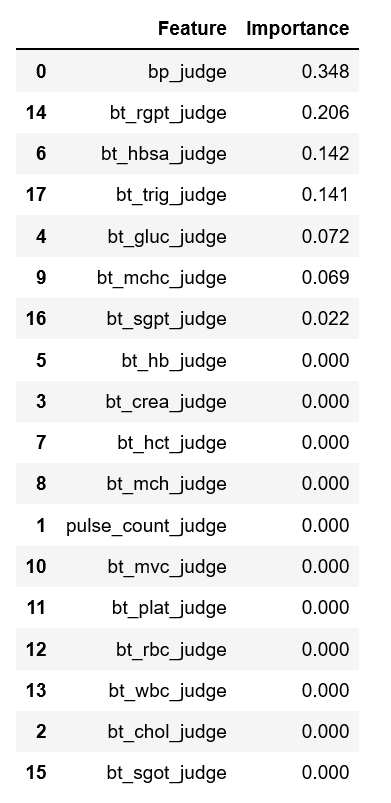
'bt\_rbc\_judge', 'bt\_wbc\_judge', 'bt\_rgpt\_judge', 'bt\_sgot\_judge',

'bt\_sgpt\_judge', 'bt\_trig\_judge')

**< 모델링 결과 >**

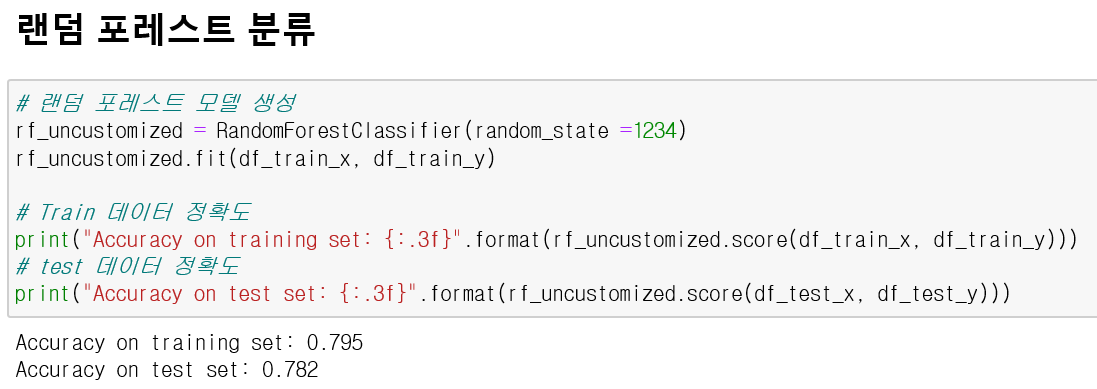


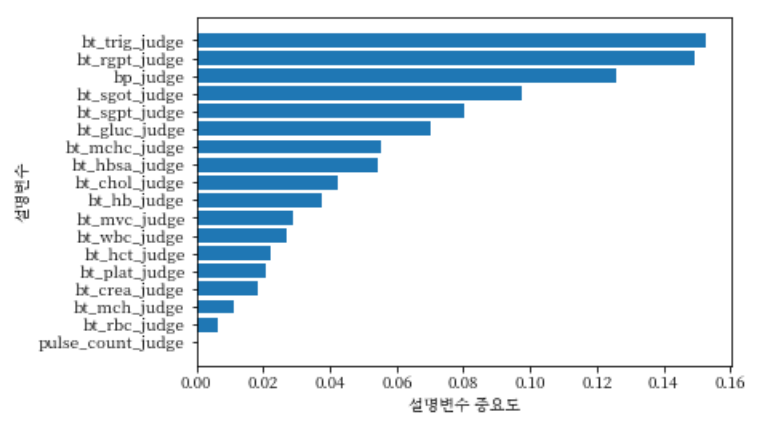
의사결정나무 설명변수 중요도 파악 결과



분석결과 고객의 위험도를 분류하는 데 중요도 순서

1. bp\_judge(혈압),
2. bt\_rgpt\_judge(감마),
3. bp\_hbsa\_judge(B형 간염 항원)
4. bt\_trig\_judge(중성지방)
5. bt\_gluc\_judge(공복 혈당)

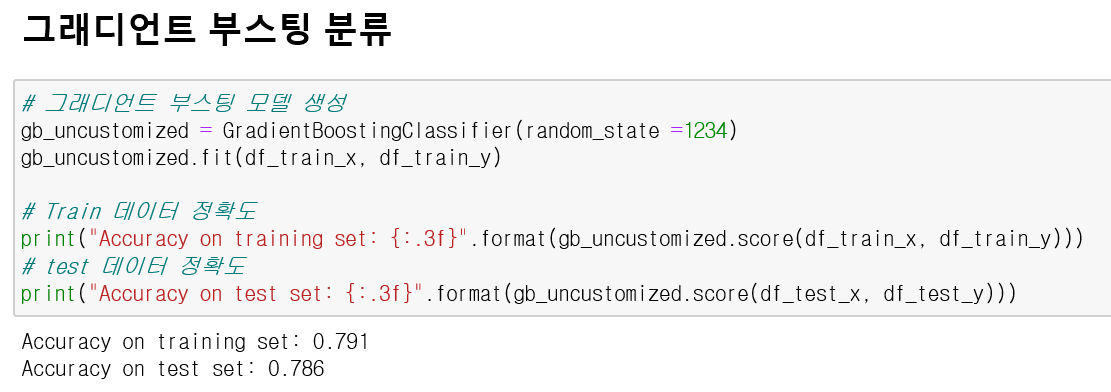




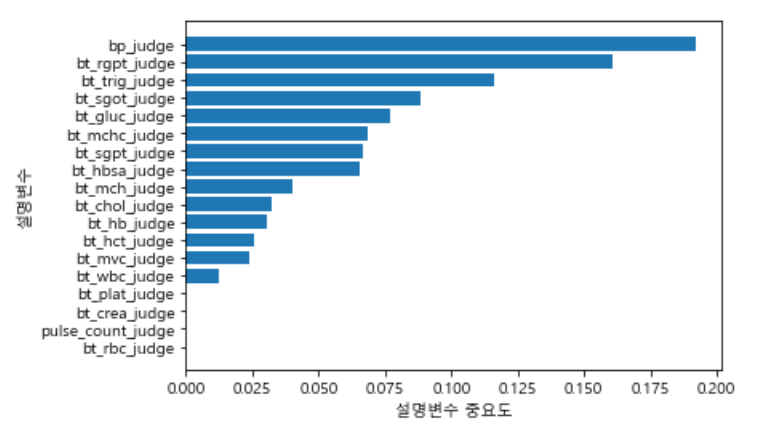
분석결과 고객의 위험도를 분류하는 데 중요도 순서

1. bt\_trig\_judge(중성지방),
2. bt\_rgpt\_judge(감마)
3. bp\_judge(혈압)
4. bp\_sgot\_judge(간세포)
5. bp\_sgpt\_judge(간세포)

순서로 중요함을 알 수 있음



그래디언트 부스팅 설명변수 중요도 파악 결과



분석결과 고객의 위험도를 분류하는 데 중요도 순서

1. bp\_judge(혈압)
2. bt\_rgpt\_judge(감마)
3. bt\_trig\_judge(중성지방)
4. bp\_sgot\_judge(간세포)
5. bp\_gluc\_judge(공복 혈당)