# 건강 정보를 통한 흡연 여부 예측 및 해석

404-NOT-FOUND

모두의 연구소 4기 데이터톤

(2025.06.04)

# 목차

1. 팀원 및 팀명 소개
2. 데이터 소개 및 전처리
3. 특성 공학(Feature Engineering)
   3-1) 변수 변환
   3-2) 신규 Feature 정의
   3-3) 신규 Feature 탐색 결과
4. 모델 선택
   4-1) Tree 기반 아닌 모델
   4-2) Tree 기반 모델
5. 모델 해석 및 의의
6. Q&A
7. 참고자료

# 1. 조원 및 팀 소개

## 조원 소개

**손영조**

- (팀장) 분석 과제 기획, 분석 방향 설정
- 코드 정리, 발표자료 작성, 발표 수행
- 모델 해석 및 관련 이론 검토

**박병기**

- 데이터 전처리, 시각화, 모델별 성능 검토
- Tree 기반 모델 파라미터 튜닝 작업
- 팀명 선정 및 팀 로고 디자인

**윤경애**

- 데이터 전처리, 시각화, 모델별 성능 검토
- Tree 기반 모델 파라미터 튜닝 작업
- 코드 작성 단계별 검토

**최신애**

- Feature 해석 및 관련 자료 조사
- 파생 변수 관련 성능 확인
- 코드 작성 단계별 검토

## 팀명 '404-NOT-FOUND' 소개

- '더 많은 데이터'를 확보하는 것이 반드시 더 좋은 결과로 이어지지 않는 다는 점을 경험했기에, 이에 착안하여, 인터넷에서 원하는 정보를 찾을 수 없는 상황과 결부하여 팀명 결정

## 팀 목적

- 현재까지 익힌 데이터 전처리, 특성 공학, 머신러닝 기법 등을 적용해보고, '흡연 여부'에 영향 주는 주요 변수에 대해 탐색해보고 해석해보기

# 2. 데이터 소개 및 전처리

## ① 데이터 소개(1) – 데이터 출처

- 데이터 출처
  - https://www.Kaggle.com/competitions/playground-series-s3e24
- 데이터 구성
  - 식별자인 id 와, 흡연 여부(흡연자 = 1, 비흡연자 = 0) 제외한 22개의 feature로 구성됨
  - 원본 데이터로부터 '생성된'(GAN*으로 생성) 가짜 데이터
  - '원본 데이터' 중 흡연/비흡연자 구분이 되어있는 train-set은 38,984개 (중복 5,517개 포함)
  - 초기 159,256건 데이터에서 192,723 건의 데이터 분석하기로 목표 변경
- 데이터 구조 요약

중복 제거 전

| 데이터 출처 | 흡연 여부 | | 합계 |
|---|---|---|---|
| | 흡연(1) | 비흡연(0) | |
| 합성(train) | 89,603 | 69,653 | 159,256 |
| 원본(train) | 24,666 | 14,318 | 38,984 |

합성 데이터 흡연율 43.74%,
원본 데이터 흡연율 36.73%

중복 제거 후

| 데이터 출처 | 흡연 여부 | | 합계 |
|---|---|---|---|
| | 흡연(1) | 비흡연(0) | |
| 합성(train) | 89,603 | 69,653 | 159,256 |
| 원본(train) | 21,209 | 12,258 | 33,467 |

합성 데이터 흡연율 43.74%,
원본 데이터 흡연율 36.63%

GAN(Generative Adversarial Network, 적대적 생성 네트워크)는 데이터를 생성하는 generator와 데이터를 구별하는 discriminator가 경쟁하는 과정을 통해 학습하는 딥러닝 모델을 지칭함

# 2. 데이터 소개 및 전처리

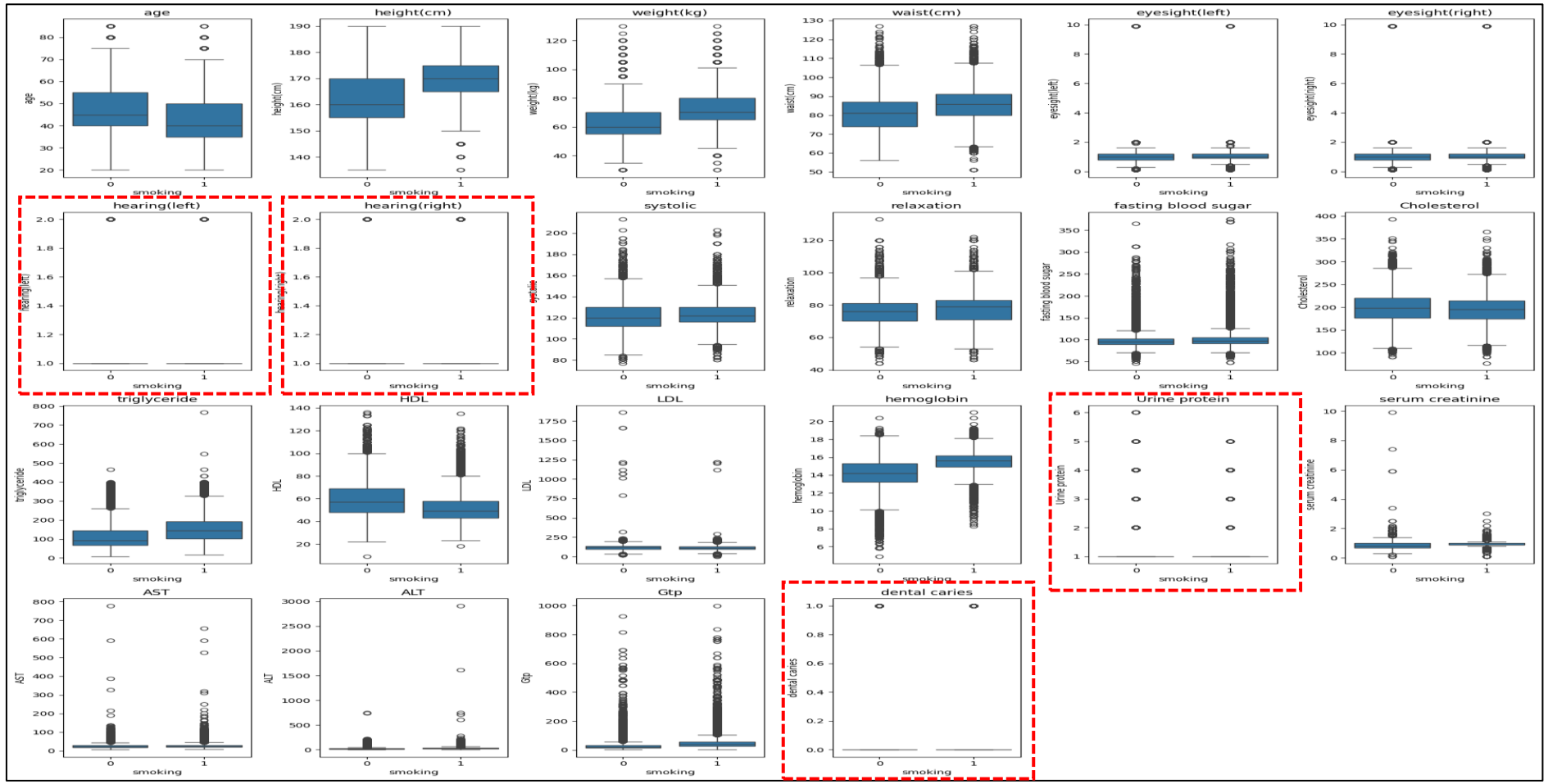
① 데이터 소개(2) – feature 소개

## Feature 상세 소개(22개)

| 변수명 | 설명 | 속성 | 변수명 | 설명 | 속성 |
|---|---|---|---|---|---|
| age | 나이 | 수치형 | Cholesterol | 콜레스테롤 | 수치형 |
| height(cm) | 키 | 수치형 | triglyceride | 중성지방 수치 | 수치형 |
| weight(kg) | 몸무게 | 수치형 | HDL | 이로운 콜레스테롤 | 수치형 |
| waist(cm) | 허리둘레 | 수치형 | LDL | 해로운 콜레스테롤 | 수치형 |
| eyesight(left) | 좌안 시력 | 수치형 | hemoglobin | 헤모글로빈 | 수치형 |
| eyesight(right) | 우안 시력 | 수치형 | Urine protein | 신장 이상 지표 | 수치형(범주) |
| hearing(left) | 왼쪽 청력 | 수치형(범주) | serum creatinine | 신장 기능 | 수치형 |
| hearing(right) | 오른쪽 청력 | 수치형(범주) | AST | 간 손상 지표 | 수치형 |
| systolic | 수축기 혈압 | 수치형 | ALT | 간 손상 지표 | 수치형 |
| relaxation | 이완기 혈압 | 수치형 | Gtp | 간 기능(담즙관련) | 수치형 |
| fasting blood sugar | 공복기 혈당 | 수치형 | dental caries | 충치 여부 | 수치형(범주) |

# 2. 데이터 소개 및 전처리

## ② 데이터 전처리(1) – 시각화




- hearing(left), hearing(right), dental caries, Urine protein 등은 '수치 ' 형태지만 범주로 구분되어야 함
- 특히 Urine protei은 추후 'one-hot encoding' 필수적임

# 2. 데이터 소개 및 전처리

## ② 데이터 전처리(2) – 결측치, Encoding 등

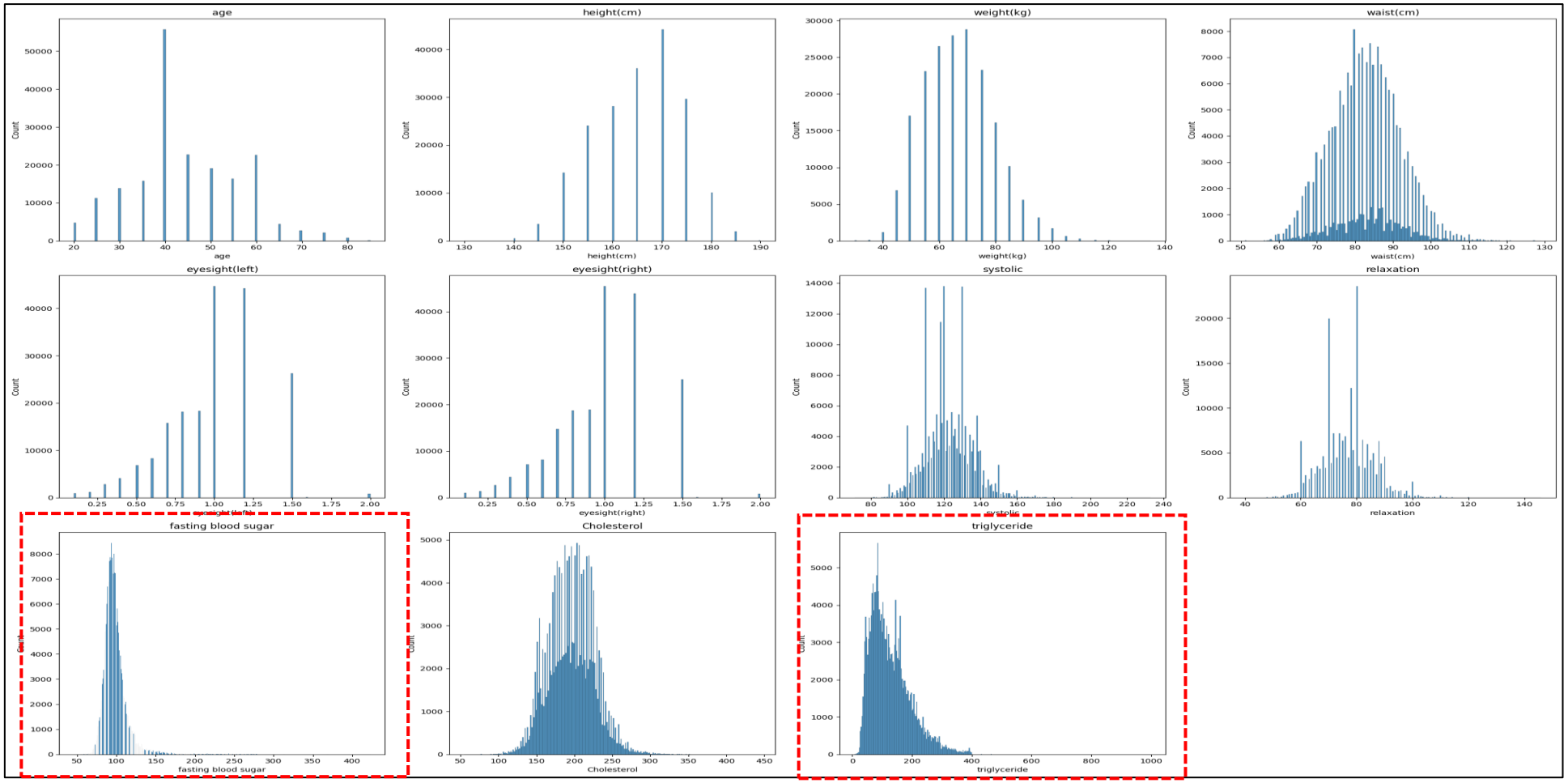
- 결측치 처리
  - eyesight(left), eyesight(right)는 9.9로 '똑같은 값'이 채워져 있음. (약 0.1% 미만)
  - 중앙값을 채워넣거나, 다른 feature로 예측해서 넣는대신, 9.9인 경우 제거하는 것으로 결정.
- Encoding 문제
  - hearing(left), hearing(right)는 정상 1, 이상 2로 되어 있는데, 정상 0, 이상 1로 변경
  - 정상 1, 이상 2로 모델 학습할 경우, 특히 tree기반 모델이 아닐 경우 왜곡 발생할 수 있음
- 이상치(outleir) 문제
  - 4~6% 데이터가 통계에서의 이상치*에 해당하여, x의 분포를 확인하고 변수 변환하는 방향으로 진행

| 변수명 | 이상치 비율(%) |
|---|---|
| fasting blood sugar | 5.52% |
| serum creatinine | 4.07% |
| ASL | 4.36% |
| ALT | 4.62% |
| Gtp | 6.10% |

이상치(Outlier)*는 다음과 같이 정의됨 (IQR = Q3 – Q1)
- 이상치 하한 : Q1 – 1.5 * IQR
- 이상치 상한 : Q3 + 1.5 * IQR

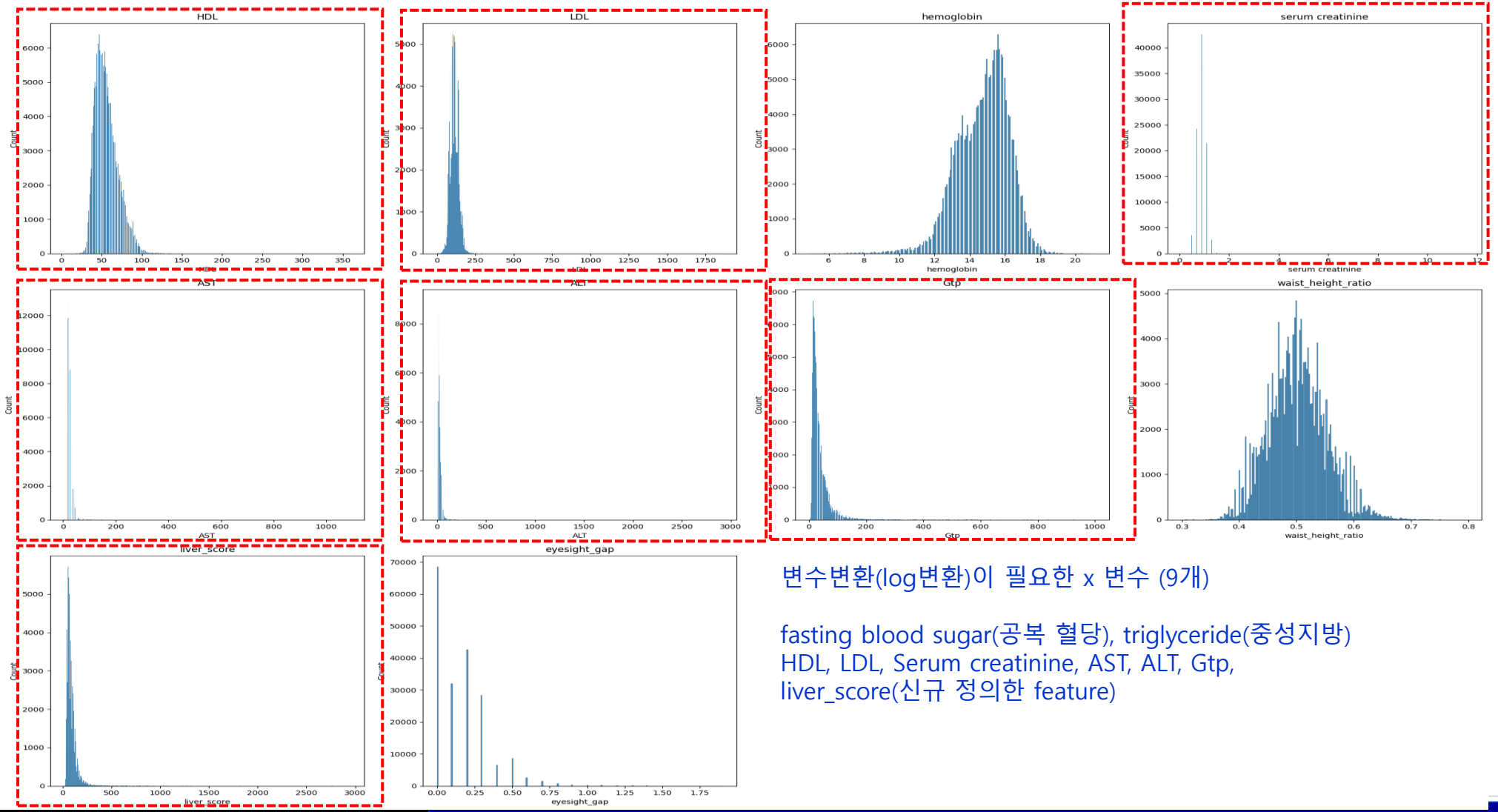
# 3. 특성 공학(Feature Engineering)

## ① 변수 변환(1)



# 3. 특성 공학(Feature Engineering)

## ① 변수 변환(2)




변수변환(log변환)이 필요한 x 변수 (9개)

fasting blood sugar(공복 혈당), triglyceride(중성지방)
HDL, LDL, Serum creatinine, AST, ALT, Gtp,
liver_score(신규 정의한 feature)

# 3. 특성 공학(Feature Engineering)

② 신규 Feature 정의

- 신규 Feature 정의
  - 모델 성능 높이기 위해 신규 Feature 정의

| N | 신규 Feature 후보군 | Feature 생성 방법 | 근거 |
|---|---|---|---|
| ① | 데이터출처(source) | 원래대상(0) / 추가된대상(1) | 두 데이터의 특성이 다르면, 모델 성능에 영향 |
| ② | BMI | 몸무게(kg)/키(m단위)의 제곱 | 흡연이 비만에 악영향 줄 가능성 |
| ③ | 허리둘레/키 비율 | 허리둘레(cm)/키(cm) | 흡연이 복부비만(허리둘레)에 악영향 줄 가능성 |
| ④ | 간점수(liver_score) | AST + ALT + Gtp | AST + ALT + Gtp 등의 점수를 활용(의학, 검색) |
| ⑤ | HDL/LDL비율 | HDL / (LDL + 1) | 해로운 콜레스테롤 비율 |
| ⑥ | 중성지방/HDL비율 | 중성지방 / (HDL + 1) | 중성지방이 HDL대비 높으면 건강 악영향 |
| ⑦ | 양안시력차 | 양쪽 시력 절대값 차이 | 양안 시력차 큰 경우도 건강 안 좋은 상태 |

- 신규 feature 효과 확인 방법
  - 전체 데이터 셋 중 7 대 3으로 Train, Test 분할(random_state = 42)
  - Train 내에서 7 대 3으로 분할하여 feature 를 넣으면서 모델 성능 비교
  - Feature 반영 시 사용 모델 : XG-boost(여러 모델 중 가장 우수한 성능을 지닌 편)

# 3. 특성 공학(Feature Engineering)

## ③ 신규 Feature 탐색 결과

- 신규 Feature 탐색 결과
  - 모델 성능 높이기 위해 신규 Feature 정의

| 구분 | 성능(ROC-AUC) |
|---|---|
| Base-line(Feature추가 없음) | 0.86125 |
| ①만 추가 | 0.86192 |
| ⑥만 추가 | 0.86185 |
| ⑦만 추가 | 0.86186 |
| ③, ④, ⑦ 추가 | 0.86236 |
| ①, ③, ④, ⑦ 추가 | 0.86241 |
| ②~⑦, 6개 추가 | 0.86059 |

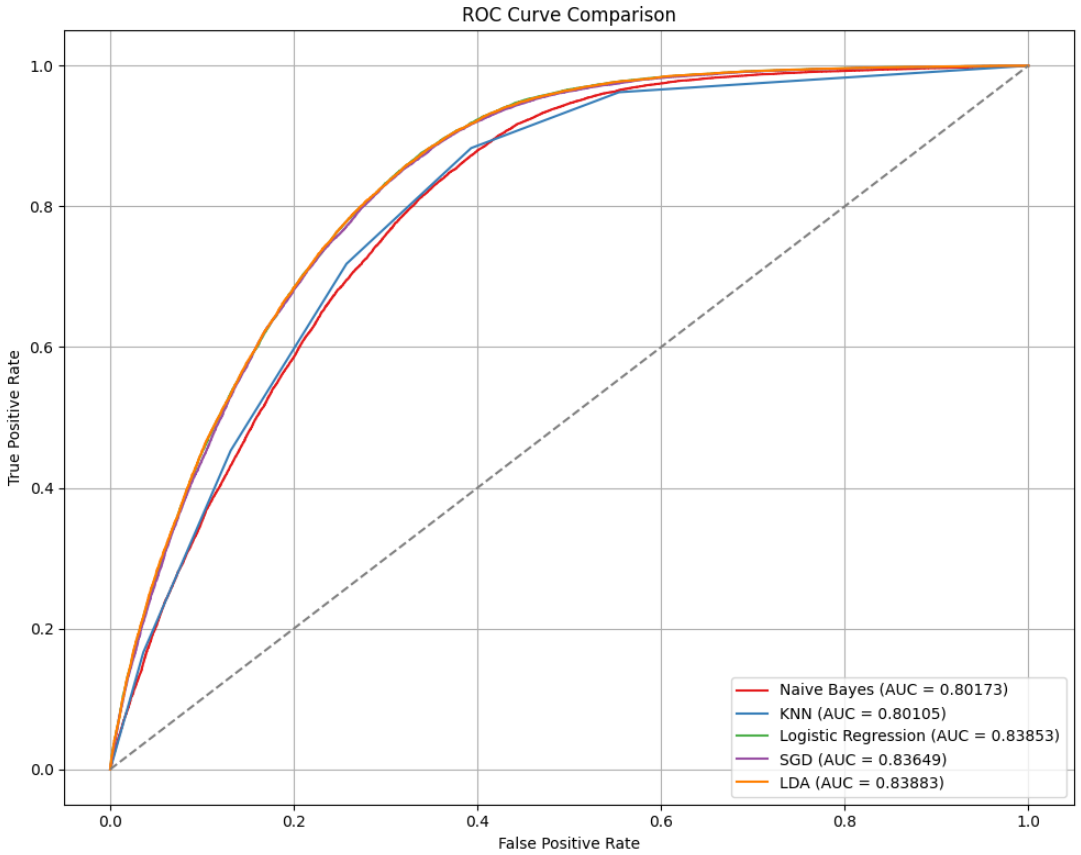
| N | 신규 Feature 후보군 |
|---|---|
| ① | 데이터출처(source) |
| ② | BMI |
| ③ | 허리둘레/키 비율 |
| ④ | 간점수(liver_score) |
| ⑤ | HDL/LDL비율 |
| ⑥ | 중성지방/HDL비율 |
| ⑦ | 양안시력차 |

- 결과 해석
  - '데이터 출처' 구분하는 것만으로도 ROC-AUC에 변동이 생김(0.078%)
  - 허리둘레/키 비율, 간 점수, 양안시력차가 최적 조합
  - 최적 조합에 '데이터 출처' 구분 feature가 추가되면, ROC-AUC 변동폭은 상대적으로 적어짐(0.006%)
    -> 상기 세 feature 추가로, 성격 상이한 데이터 결합으로 인한 영향이 거의 설명 가능함

# 4. 모델 선택

## ① Tree기반이 아닌 모델

- 모델 탐색 결과(parameter Tuning 미적용된 기본 모델)
  - Naive bayes, KNN, Logistic, SGD, LDA 확인 결과 LDA(0.83883), Logistic(0.83853) 등이 성능 준수함



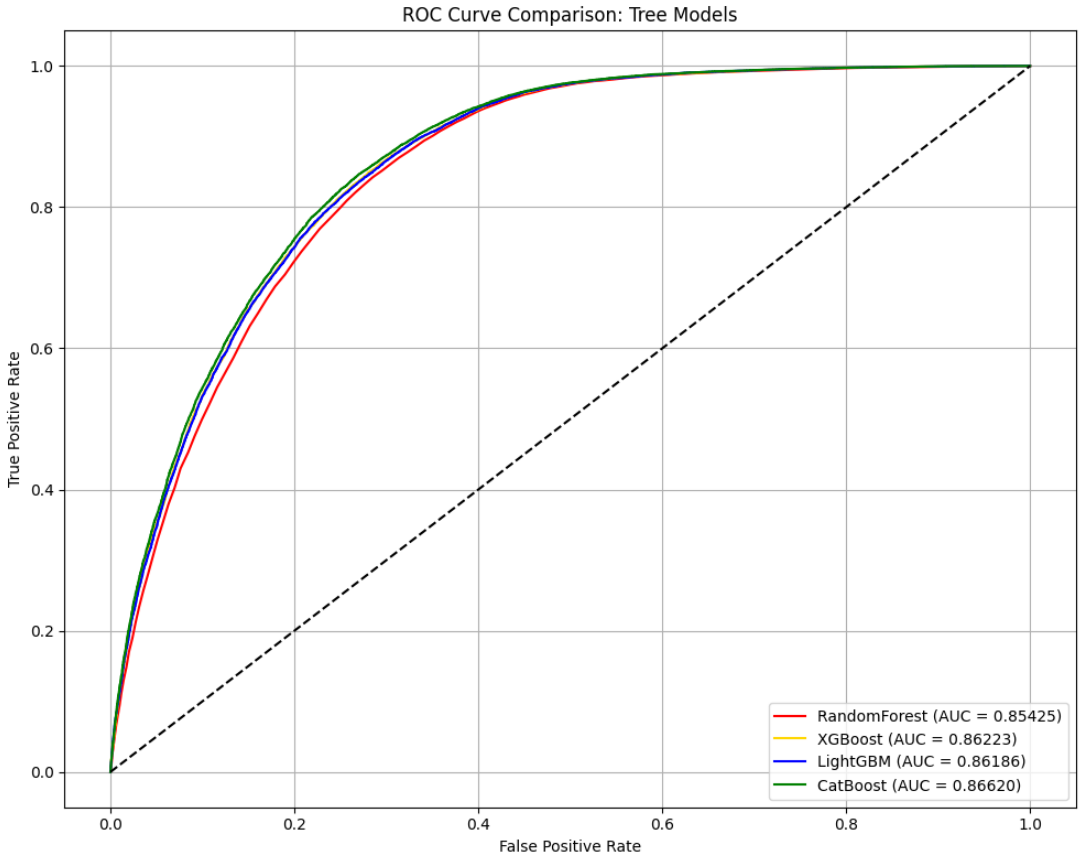

LDA가 Logistic 대비 성능이 근소하게 높지만,
LDA의 경우 데이터의 분포가 '정규성'을 가정하기 때문에

일반화된 성능을 위해서는 Logistic을 활용하는 것이 더 바람직할 것으로 예상됨.
(특히, voting이나 stacking 목적으로 tree기반 모델과 결합해서 활용하고자 하는 경우)

# 4. 모델 선택

## ② Tree기반 모델

- 모델 탐색 결과(parameter Tuning 미적용된 기본 모델)
  - CatBoost(0.86620), Xgboost(0.86223), LightGBM(0.86186) 순서로 성능이 우수함




CatBoost가 성능이 가장 우수하긴 하지만,

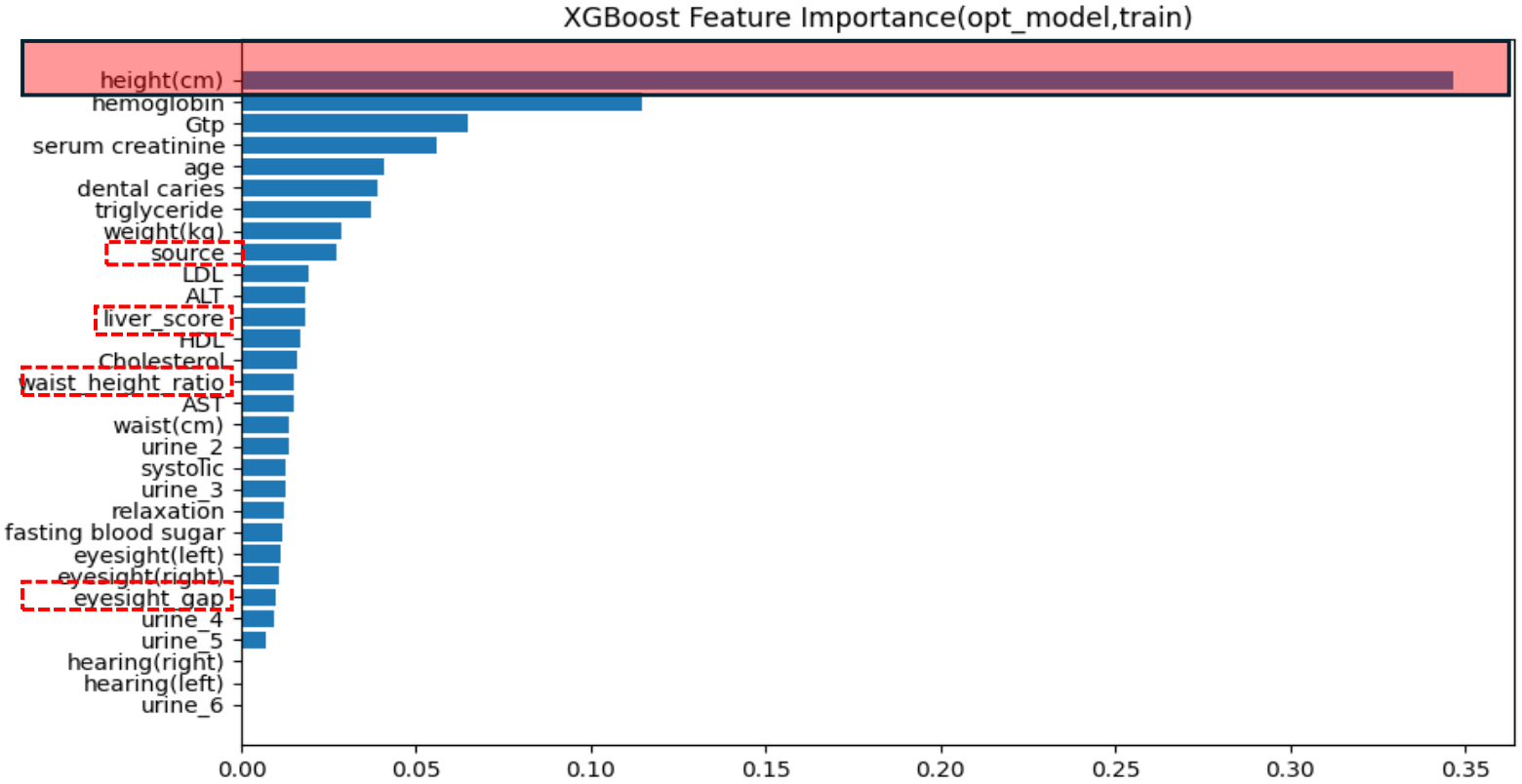
Optuna, hyperopt, RandomSearch 등으로 파라미터 튜닝을 할 경우, CatBoost는 성능 향상이 잘 이뤄지지 않음.
(파라미터 튜닝까지 고려할 경우, XG-boost가 가장 우수한 성능의 모델로 판단됨.)

특히 optuna의 경우, XGboost는 colab에서 10분 정도 소요
CatBoost는 거의 1시간 정도 소요

# 5. 모델 해석 및 의의

## ① Importance 기준 해석

- XGboost 모델 해석(*optuna로 최적화)
  - Xgboost(0.86223->0.86848)성능 향상, 4개의 신규 feature중 source, liver_score, waist_height_ratio는 상위권에 진입(단, source는 데이터 구분자이므로 두 데이터의 차이를 완벽히 잡아내지 못했다는 의미)

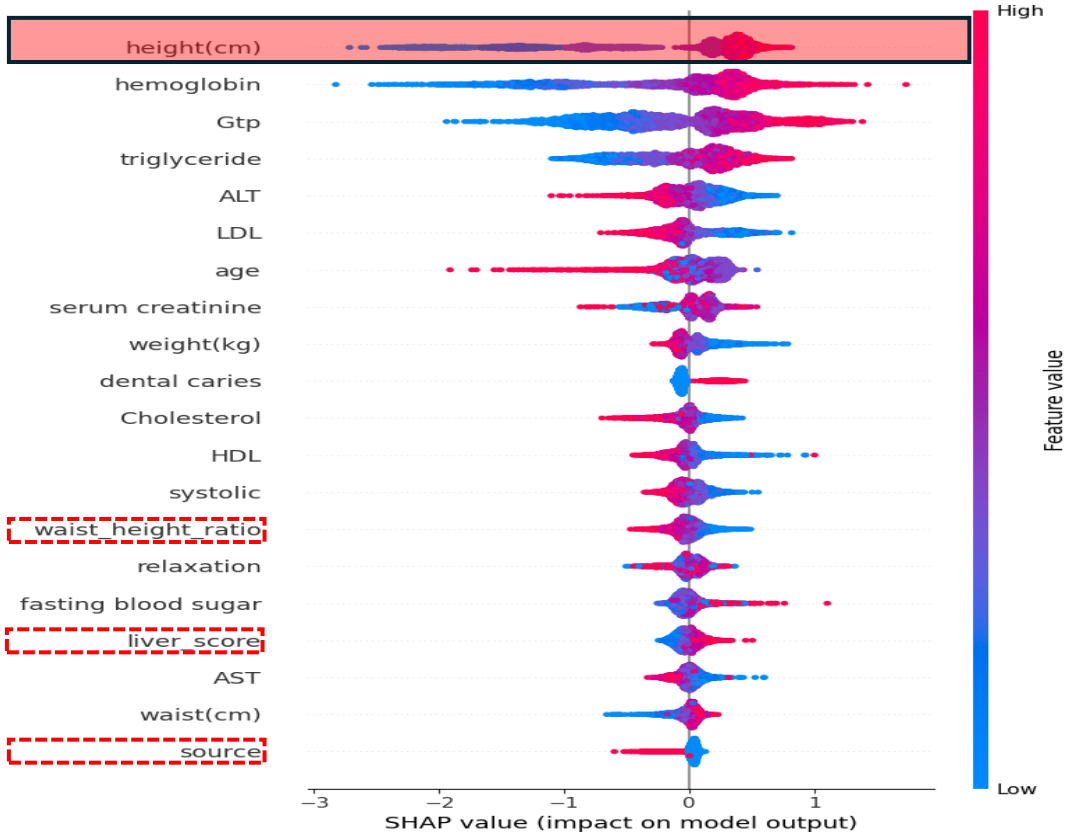


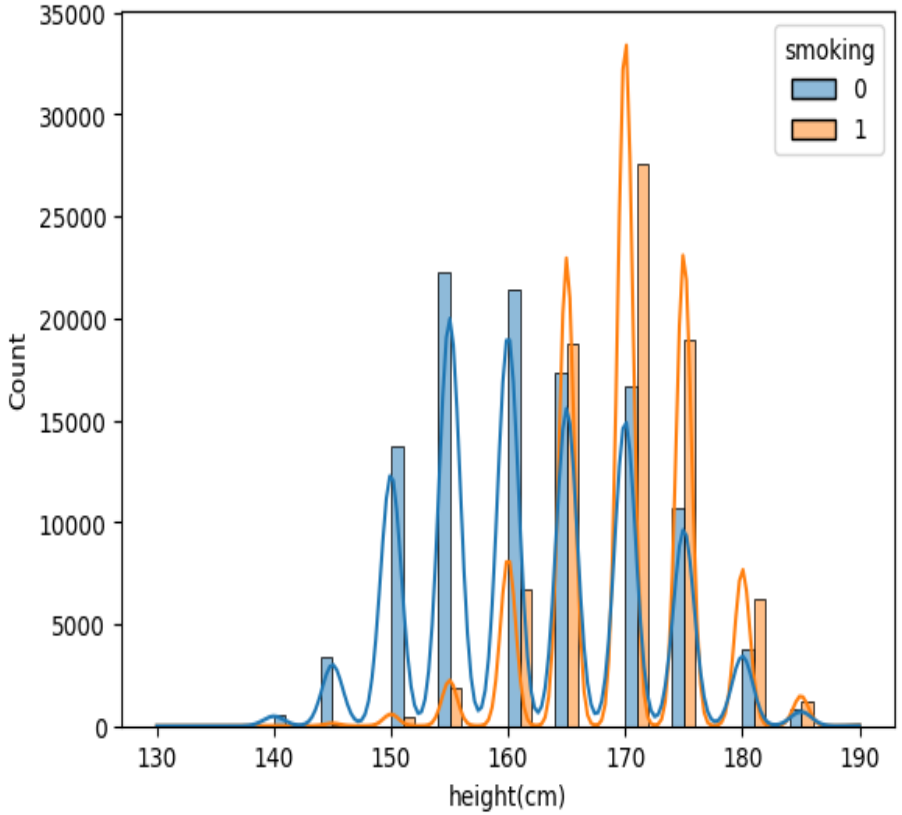

그런데, 다른 특성도 아니고 키(height)가 압도적으로 중요하다고 나옴.

# 5. 모델 해석 및 의의

## ② Shap_value 기준 해석

- XGboost 모델 Shap-value해석(*optuna로 최적화)
  - 역시 이 경우에도 키(height)가 독보적으로 중요한 영향을 끼쳤다는 것을 확인할 수 있음!
  - 단순히 시각화 해도 키에 따른 차이가 있음



# 5. 모델 해석 및 의의

## ③ Importance vs Shap_value 비교

- Importance vs Shap_value
  - Shap_value에 대해서는, 계산 시간으로 인해 약 10% 데이터만 sample추출 해서 그림을 그렸음

| 항목 | Importance | Shap_value |
|---|---|---|
| 기반 원리 | 트리 분할 기준 | 게임이론의 Shapley value |
| 중요도 계산 기준 | 손실 감소량(gain), 빈도(weight)등 | 예측 값 변화에 대한 기여도(샘플별) |
| 산출 형태 | 전체 데이터 기준 단일 값 | 샘플마다 기여도 존재, 평균값으로 정리 |
| 해석 방식 | 모델이 해당 feature를 얼마나 자주 사용? | 해당 feature가 예측값에 얼마나 영향? |
| 비선형성/상호작용 | 제한적으로 반영 | 반영함(비선형, 상호작용 모두) |
| 속도/계산 | 매우 빠름 | 매우 오래 걸림 |
| 사용 목적 | 대략적으로 빠른 판단, 전체적 경향 확인 | 상세한 분석, 모델 설명, 변수 해석 |

# 5. 모델 해석 및 의의

## ④ 모델 Feature 관련 제안

- 모델 성능 향상 목적의 feature 제안

| 구분 | 항목 | 제안 사유 |
|---|---|---|
| 과제 외부적 Feature | 성별 | - 미국 내의 흡연률은 인종, 성별 막론하고 꾸준히 감소 중이나, 남성 흡연률이 더 높음<br>- 남성이 여성 대비 평균적으로 키가 크기 때문에, '성별' 변수가 있으면 '키 ' 변수의<br>과도한 설명력을 억제하고, 의학적으로 알려진 간, 신장 기능 관련 변수들의<br>모델에 대한 설명력이나, 예측에 대한 영향도 증가할 수 있음 |
| | 인종 | - 미국 사회는 인종별로 소득, 교육, 음주, 흡연, 비만 등의 지표가 상이한 것으로 알려져 있음 |
| 과제 내부적 Feature | Triglyceride | - 위험도 따른 구간 나누기(150미만/150~199/200~499/500이상) |
| | Gtp | - 위험도 따른 구간 나누기(50미만/50~100/100~200/200이상) |
| | AST | - 위험도 따른 구간 나누기(40미만/40~80/80~160/160이상) |
| | ALT | - 위험도 따른 구간 나누기(40미만/40~80/80~160/160이상) |

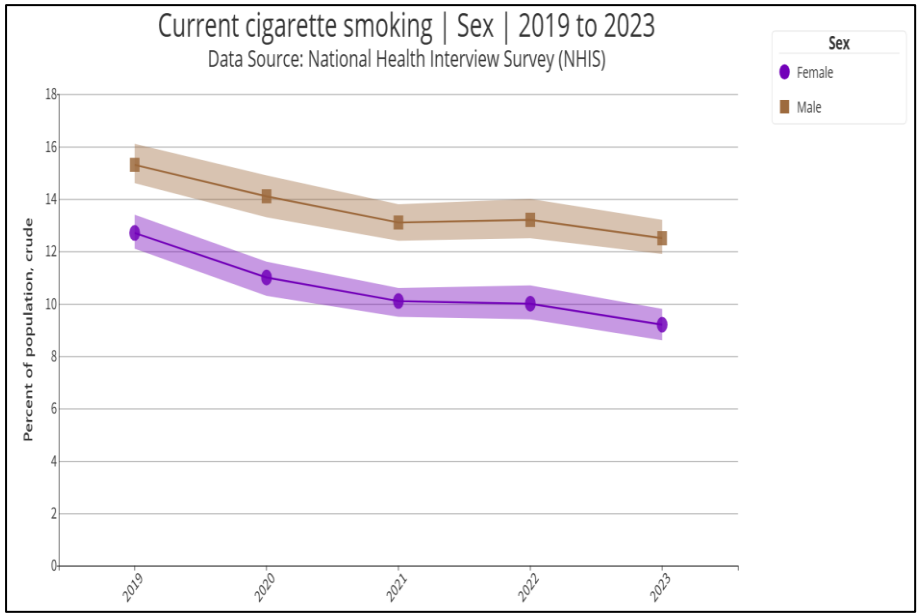
- Triglyceride(중성 지방) 등의 지표 다룰 시, 해당 국가 의료적인 임상 기준을 활용하는 것도 도움이 됨 (단순히 비율로 나누는 것은 지양)

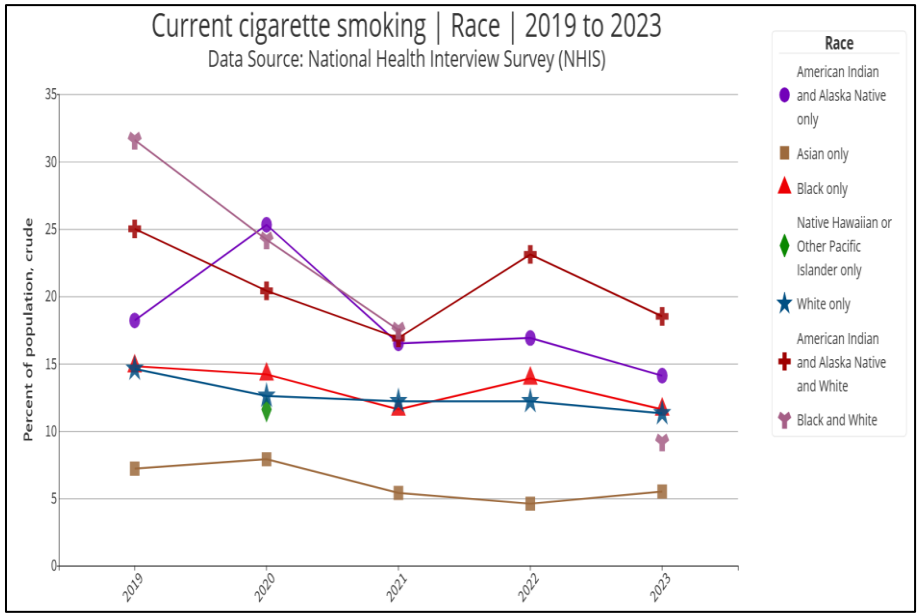
# 6. Q&A

# 7. 참고 자료

## ① 미국 성별, 인종별 흡연율

- 미국 성별, 인종별 흡연율
  - 항시 남성이, 여성 대비 흡연율이 높음
  - 미국은 인종별 소득 수준, 교육 수준, 범죄율, 키, 비만율 등 차이가 있으며 흡연율도 예외는 아님



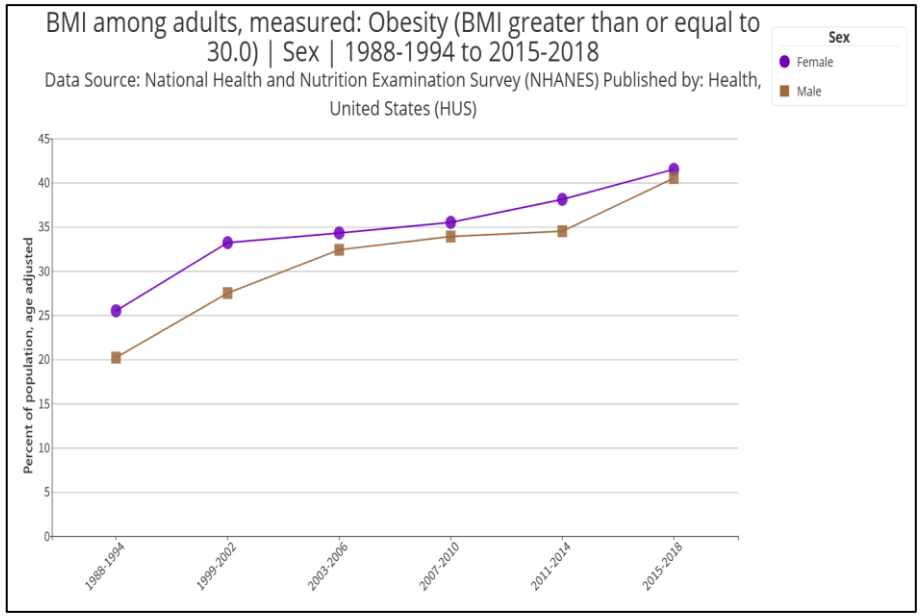



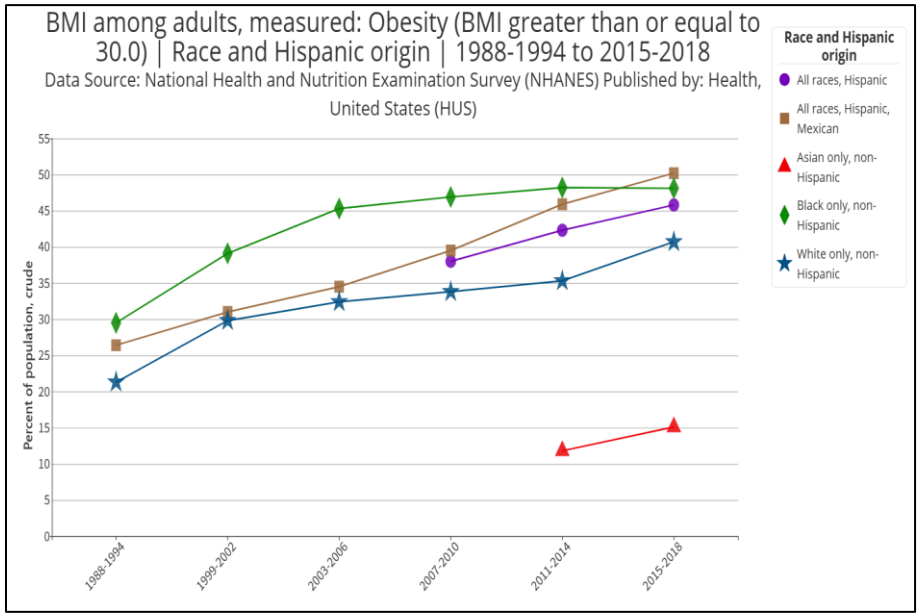

- 이 자료는 설문 자료로서 Confidence Interval도 표시 가능 (응답자들이 '흡연 여부'에 대해 거짓으로 응답할 가능성)
- 자료 출처 : https://nchsdata.cdc.gov/DQS

# 7. 참고 자료

## ② (참고) 미국 성별, 인종별 비만율

- 미국 성별, 인종별 비만율
  - 한국은 BMI 25기준, 미국은 BMI 30 기준(예시, 키 174cm면 한국 75kg, 미국 91kg)
  - 건강에 악영향 주는 '비만(중성지방, LDL, HDL 유관)' 역시 인종별 차이가 뚜렷함




- 성별, 인종별 정보가 있다면 인구 집단별 키, 비만 관련 지표에 대한 차이를 모델이 더 잘 학습할 가능성↑
- 키, 비만 지표 제외하고 흡연이 직접적으로 영향 끼친다고 알려진 특성(간, 신장 등)의 중요도, 설명력 상승 가능성↑