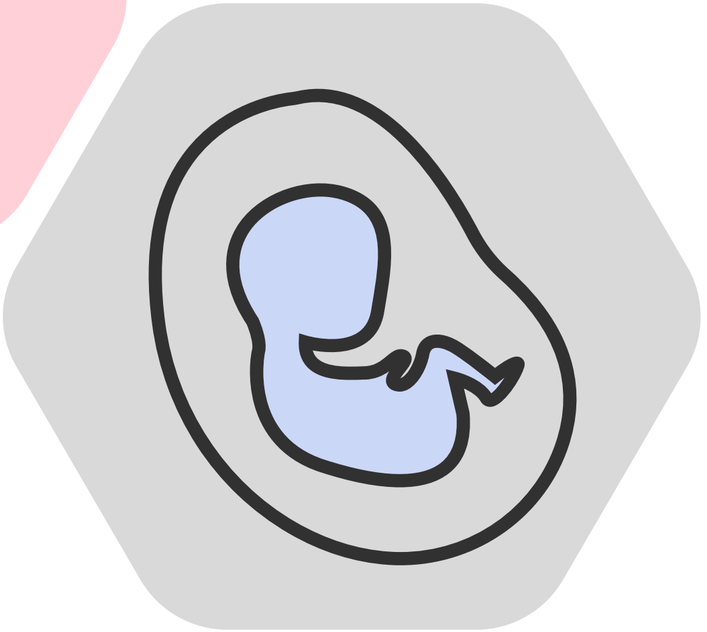
LG

Aimers



난임 환자 대상 임신 성공 여부 예측 AI 온라인 해커톤

code7monkey 이승현

Problem Definition

LG화학 생명과학사업본부

hfea 데이터 셋 → 호르몬 수치, 신장, 체중 같은 정보가 제한적

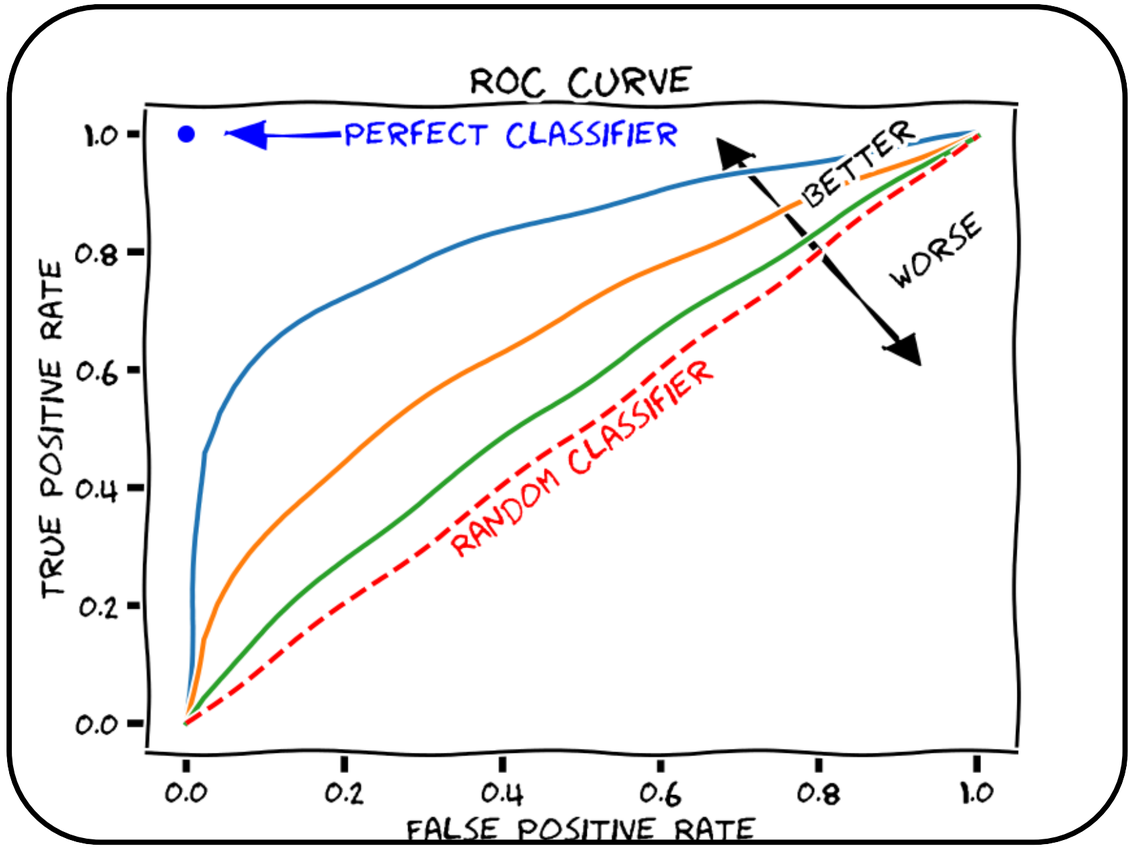
임신X, 출산O → 임신 기간 동안의 변수도 고려해야 함

AUC → 임계값 조절 등의 후처리가 의미가 없음

AUC?

LG화학 생명과학사업본부

- Area Under the Curve
  : 임계값을 변화시키며 TPR / FPR를 그린 그래프 아래 면적



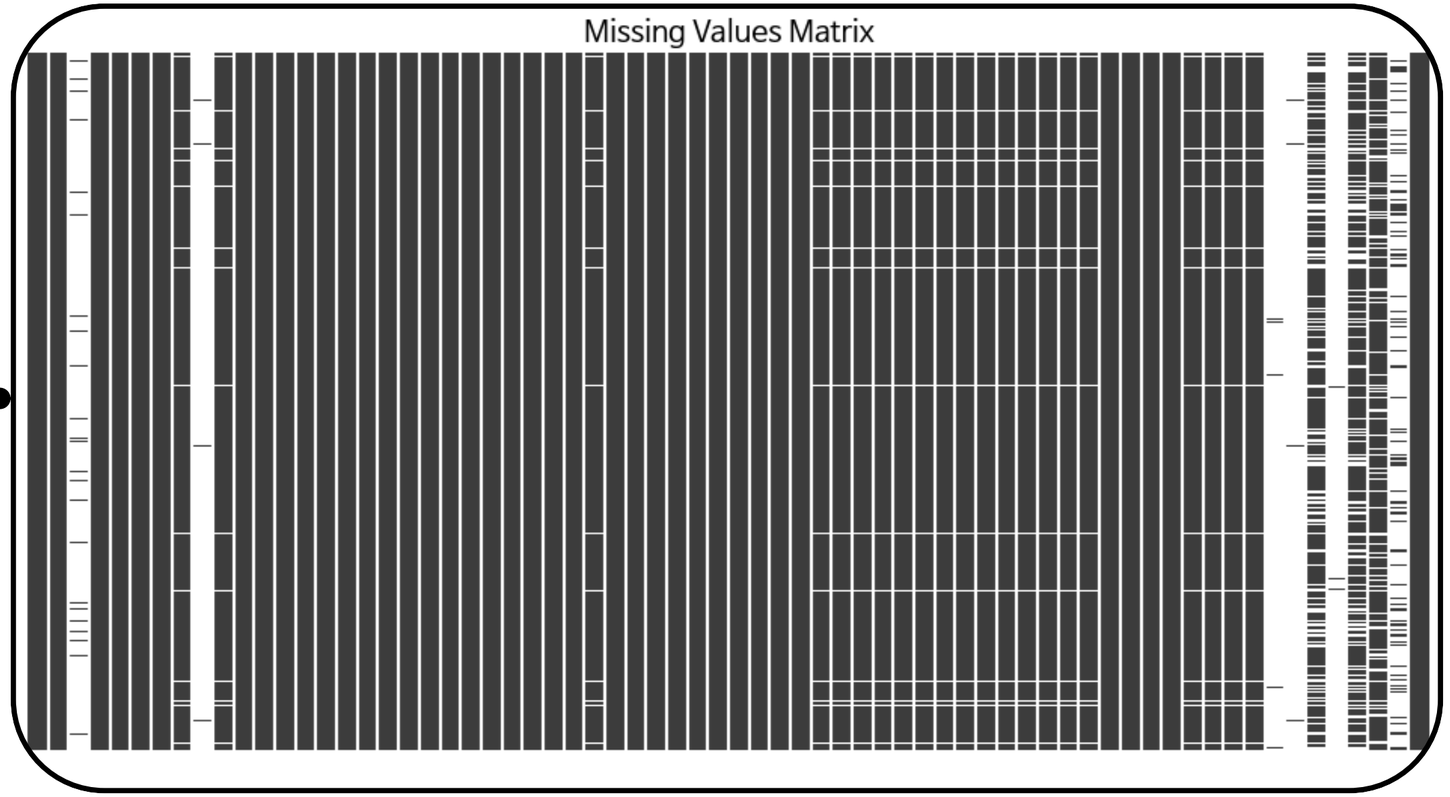
1. 순위 기반 평가

2. 임계값 독립성

3. 클래스 비율에 무관

전처리 - 결측치

LG화학 생명과학사업본부



동일한 결측치 패턴(이식된 배아 수, 저장된 배아 수.. 약 21개)

=> 전부 시술유형이 'DI'

- 숫자형: fillna(0)
- 문자형: 'Not Answer(DI)'

나머지

- 특정 시술 유형
  - Unkown 대체
- 난자 출처
  - 알 수 없음 -> 본인 제공
- PGD/ PGS
  - fillna(0)
- 경과일 컬럼
  - 난자 채취 경과일
    - fillna(1)
  - 혼합/이식/해동
    - fillna(999)

기타 전처리

LG화학 생명과학사업본부

난자 기증자 나이

조건) 난자 출처 == '본인 제공'
: 난자 기증자 나이 -> 시술 당시 나이

컬럼 drop

1. '임신 시도 또는 마지막 임신 경과 연수'
2. '난자 해동 경과일'
3. '불임 원인 - 정자 형태'
4. '불임 원인 - 정자 운동성'
5. '불임 원인 - 정자 면역학적 요인'
6. '불임 원인 - 정자 농도'
7. '불임 원인 - 자궁경부 문제'
8. '불임 원인 - 여성 요인'
9. '대리모 여부'
10. '부부 부 불임 원인'
11. PGS 검사를 받고도 안한 사람

🎯 drop 기준

1. 결측치 특징이 없고, 비율이 높음
2. 거의 단일 값

추가 drop

PGS 검사를 받고도 안한 사람
-> XGB/CAT에서만 drop

Keywords Extraction

LG화학 생명과학사업본부

키워드 추출하여 각각의 컬럼으로 만들어 줌

배아 생성 주요 이유
기증용, 현재 시술용

→

기증용	현재 시술용
1	1

배아 생성 주요 이유

- 기증용
- 현재 시술용
- 난자 저장용
- 배아 저장용

특정 시술 유형

- ICSI
- IVF
- UnKnown
- IUI
- BLASTOCYST
- AH

파생 변수

LG화학 생명과학사업본부

난임 여부

총 시술 횟수 - 총 임신 횟수

PGS 검사를 받고도 안한 사람

착상 전 유전 검사 사용 여부
- PGS 시술 여부

미세주입이 아닌 배아 이식/생성/저장 수

이식된 배아 수 - 미세주입 배아 이식 수
총 생성 배아 수 - 미세주입에서 생성된 배아 수
저장된 배아 수 - 미세주입 후 저장된 배아 수

유산 여부

총 임신 횟수 - 총 출산 횟수
& 이진 분류로 변환

클리닉 외 총 시술 횟수

총 시술 횟수 -
클리닉 내 총 시술 횟수

이식된 배아 수 / 나이

나이는 범주형이므로, 1~7로 매칭
단일 배아 이식 여부가 1인 경우, 이식된 배아 수를 1.5로 취급 (가중치)

인코딩? 범주형!

LG화학
생명과학사업본부

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

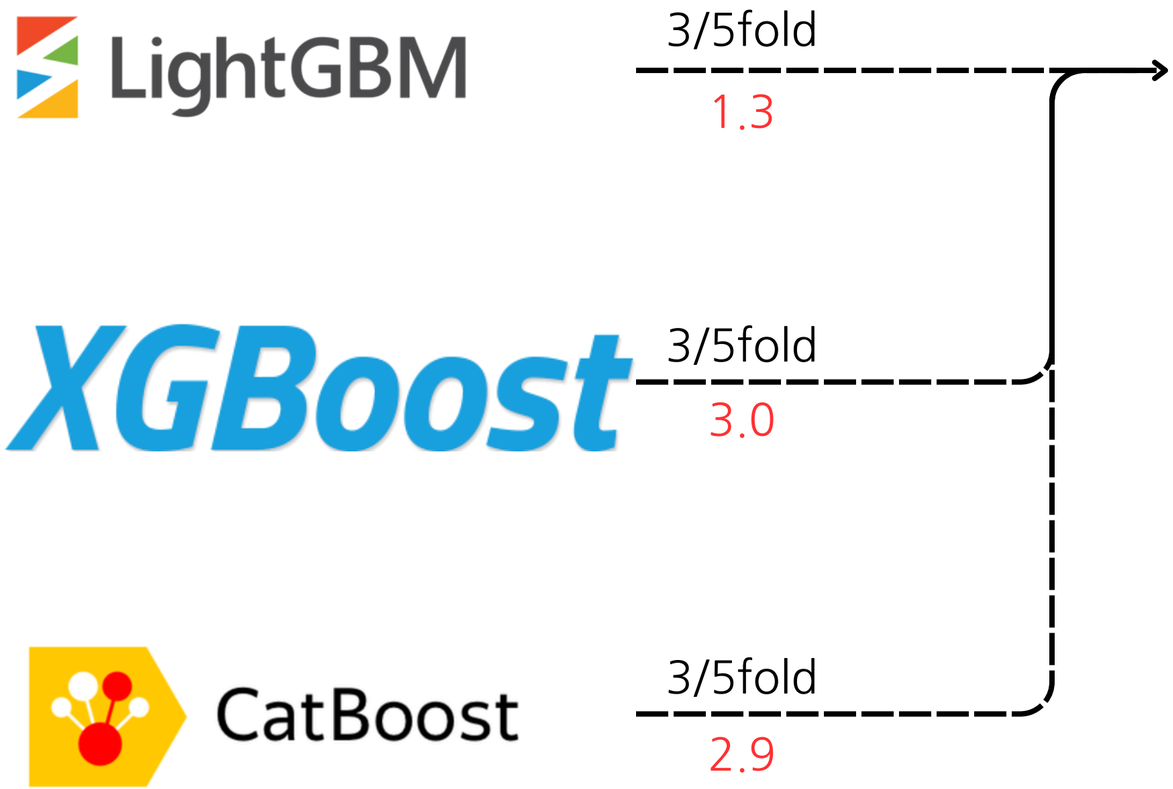
Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

GDBT 계열의 모델은 범주형으로 지정해주면,
별도의 인코딩 과정이 필요 없음!
ex) catboost의 cat_features

Modeling

LG화학 생명과학사업본부

빨간색: 모델 별 가중치



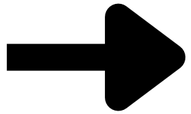
RANK 앙상블

- soft voting, sigmoid ensemble에 비해 성능이 좋았음
- AUC: 절대적 크기 < 상대적 순서
  - 순위 정보를 이용하는 rank 앙상블이 적합
- 각 모델 별 가중치는 수동으로 조정

Model Performance

LG화학 생명과학사업본부

fold	lgbm(cv)	xgb(cv)	cat(cv)
2	0.74155	0.74159	0.74222
3	0.73985	0.73859	0.74021
4	0.73955	0.73844	0.73942
mean	0.73944	0.73790	0.73952



public : 0.74201
private: 0.74228

1,5 fold는 성능이 낮아 앙상블에서 배제

Thank You!