

LG Aimers 6th 본선 발표

팀

간지포s

목차

01 데이터 전처리

02 주요 기여 및 방법론

03 결론

01

데이터 EDA 및 분석 목적

1. 분석 목적

데이터 분석 목적

- 난임 환자 시술 데이터를 기반으로 임신 성공 확률을 예측
- 정밀한 예측 결과를 통해 의료진의 임상 의사결정 및 신약 개발 보조

핵심 인사이트 도출

- 상관관계 및 분포 분석을 통해 임신 성공에 유의미한 핵심 변수 선별
- 배아·난자 상태, 연령, 시술 이력 등 중심으로 예측에 중요한 특징 도출
- 도출된 변수 기반 설명력 높은 파생 변수 생성 및 적용

최종 기대 효과

- 의료진 및 신약 개발자에게 신뢰 가능한 보조 지표 제공
- 환자에게 예측 기반 상담 자료 제공
- 임상적 활용을 고려한 설명 가능하고 일반화된 ML 모델 개발 기여

2. 평가 지표 분석

$$Score = 0.5 \times \left(1 - \frac{\sum w_i (y_i - \hat{y}_i)^2}{\sum w_i} \right) + 0.5 \times \frac{2TP}{2TP + FP + FN}$$

$$w_i = 1 + 4y_i + (0.5 - y_i)^2$$

$$\hat{y}_i^{(bin)} = \begin{cases} 1, & \text{if } \hat{y}_i > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

모델 설계에 반영된 전략

- 단순 정밀도만이 아닌, 고정밀 + 민감도 균형을 고려한 모델 튜닝 필요
- 가중치 기반 학습을 통해 임상적으로 중요한 케이스에 더 민감하게 반응하도록 유도
- 예측값을 확률로 출력 가능하도록 Calibration 친화적 모델(CatBoost 등) 활용

02

데이터 전처리

1. 데이터 형 변환 및 문자열 값 파싱

- 목적:

- 모델 학습에 적합한 형태로 데이터를 정제하고, 수치형·범주형 변수로 변환

- 문자열 값 파싱 및 수치형 변환:

- 범위 값 파싱:

- '0회', '1-5회', '>20', '6회 이상' 등의 문자열 값을 수치로 변환

- 중앙값 추정 or 기준값 + 1 방식 적용

- 수치형 변환 :

- 원본 컬럼 복사 → _num (숫자형)

- 문자열 변환 → _cate (범주형)

- 적용 대상 예시 : '이식된 배아 수', '미세주입(ICSI) 배아 이식 수', '배아 이식 후 경과일'

2. 파생 변수 생성 및 예측 특징 도출 (Correlation 분석 기반)

특징 도출 배경:

- 환자의 연령, 과거 시술 경험, 배아와 난자의 상태 등 다양한 요소가 임신 성공률에 큰 영향을 미침.
- Correlation Matrix 분석:
 - 각 변수 간 상관관계를 확인하여 예측에 중요한 변수들을 선별 및 파생 변수로 생성

임신 성공 예측 파생 변수:

- 배아 관련:
 - 배아 이식 비율, ICSI 배아 비율, 배아 저장 비율, 배아 품질 지표
- 난자 관련:
 - 난자 수정 시도 비율, 난자 배아 생성 성공률, 신선/해동 난자 여부 결합
- 시술 경험 및 연령 관련:
 - 이전 시술 총 횟수, 이전 임신 성공률, 시술 경험 점수, **고위험 연령군**, **연령 가중치**,
기증자 나이 차이, 이식 후 걱정기간

3. 전처리 결과 및 기대 효과

기대효과

- 데이터 품질 향상:
 - 불완전한 정보 보완 및 정제
 - 모델 학습에 최적화된 변수 생성
- 모델 성능 개선:
 - 단순 변수 대신 비율 및 상호작용 변수 활용
 - 예측력 강화 및 오버피팅 완화
- 해석 가능성 증가:
 - 의미 있는 변수 도출로 인사이트 제공
 - 단순 숫자보다 해석에 유용한 비율/상호작용 변수 추가

03

주요 기여 및 방법론

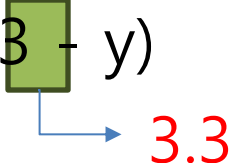
1. 평가지표 기반 가중치 최적화 전략

배경:

- 평가 기준에 따라 임신 성공 사례에 더 높은 중요도 부여 필요
- 예측 확률의 정확도와 이진 분류 성능 모두를 고려하기 위해 가중치(weight) 적용 필요
- 가중치 함수 (Python 코드):

```
def compute_weights(y):
```

```
    return (1 + 4 * y + (0.5 - y) ** 2) * (3 - y)
```



적용 목적 :

- 예측 확률 정밀도 + 이진 분류 균형 있게 최적화
- 희귀 클래스(임신 성공)에 대한 민감도 향상

기대 효과 :

- Weight Brier Score, F1 score 개선 · 향상
- 임상적으로 중요한 샘플에 모델 집중도 증가

2. Stacking 기반 예측 모델 구조

구성 :

- **Stage1** : CatBoost -> 예측값 및 잔차 계산
- **Stage2** : CatBoost 출력값 + 잔차 -> LightGBM의 입력값으로 활용

의의 :

- CatBoost의 범주형 처리 및 일반화 능력 활용
- LightGBM이 예측 오류 영역을 보완하여 전체 성능 개선

기대 효과 :

- 개별 모델 한계를 상호 보완
- 예측 성능 및 안전성 동시 확보
- 앙상블 구조로 과적합 위험 완화 및 강건성 강화

3. Ensemble 비율 최적화 실험

실험 배경 :

- Stacking 구조 : 복합적 비선형 구조 학습에 강점, 과적합 위험 있음

의의 :

- CatBoost 단일 모델: 범주형 처리에 특화, 해석 가능하고 일반화 성능 우수
- 서로 다른 학습 특성을 반영하여 예측 성능과 일반화 성능 상호 보완

조합 목적 :

- Stacking의 고성능 예측력을 CatBoost의 안정성으로 보정
- Ensemble 구조를 적용하여 예측 결과의 불확실성을 줄이고, 강건한 성능 도출
- 실험적으로 **최적 비율(n=3)** 도출 → Stability Accuracy 균형 달성

$$\hat{y}_{final} = \frac{n \cdot \hat{y}_{stacking} + \hat{y}_{catboost}}{n + 1}$$

03

결론

결론

- 난임 환자 시술 데이터를 기반으로 임신 성공 확률을 정량 예측
- 가중치 기반 손실 함수 설계로 평가 지표 최적화 및 민감 사례 대응
- CatBoost + LGBM Stacking 기반 예측, Ensemble 비율 조정을 통해 성능 극대화
- 의학적 맥락에 맞춘 파생 변수 구성으로 해석 가능한 예측 구조 설계

감사합니다.