

난임 환자 대상 임신 성공 여부 예측 AI 오프라인 해커톤

엄마손은 AI손

* CONTENTS

- 1. 문제 및 데이터 소개**
- 2. EDA & PROCESSING**
- 3. MODELING**
- 4. 최종 선정**

* 문제 및 데이터 소개

주제 및 배경

난임 시술을 진행하는 환자들에게 시술 부담을 줄이는 동시에, 의료 기관이 차별화된 서비스를 제공할 수 있도록 난임 환자 데이터를 활용하여 **임신 성공 확률을 예측**하고 임신을 결정짓는 최적의 특성을 탐색하는 AI 모델 개발

데이터 소개

train data : 샘플별 고유 ID (126,244행)의 임신 성공 여부를 포함한 난임 환자 시술 데이터 (34개 컬럼)

test data : 샘플별 고유 ID (54,412행)의 난임 환자 시술 데이터 (33개 컬럼)

환자 시술 당시 나이, 시술 유형, 이전 총 임신 성공 횟수,
이식된 배아 수, 정자 출처 등..

* EDA & PREPROCESSING

변수 유형 변환

환자 시술 당시 나이, 난자 기증자 나이, 정자 기증자 나이

범주형에서 수치형으로 변환, 각 구간의 중앙값으로 매핑, '알 수 없음'은 -1로 처리하여 식별 가능하게 함.

난자 수, 배아 수 관련 칼럼

범위형 값을 수치형 최댓값으로 변환 (예: '1-5' → 5),
동일한 수준의 임신 성공률과 적은 양의 데이터로, 분리의 유의미성 낮다 판단하여 마지막 두 구간 병합

횟수 관련 칼럼

변수의 의미가 수치이므로 수치형으로 처리

* EDA & PREPROCESSING

결측치 처리

이전 총 임신 횟수

‘이전 총 임신 횟수’ 변수가 결측, ‘이전 총 성공 임신 횟수’ 변수가 결측이 아닌 경우,
최소한 성공 횟수만큼 임신했을 것이므로 ‘이전 총 임신 횟수’의 값으로 대체

이전 총 임신 횟수 결측	이전 총 성공 임신 횟수 결측	count
False	False	114541
True	False	6622
True	True	5081

그 외 결측치

나머지 결측은 모두 -1로 대체

* EDA & PREPROCESSING

변수 제거

신선 난자 사용 여부

의미상 연결되는 변수인 ‘채취된 신선 난자 수’ 변수와 분포가 일치하지 않는 것 확인
⇒ 정보량이 더 많은 ‘채취된 신선 난자 수’를 남기고 변수 제거

신선 난자 사용 여부	채취된 신선 난자 수	개수
0	0	39876
0	1	85963
1	1	405

* EDA & PREPROCESSING

세부 시술 유형	count	mean_success_rate
ICSI / AH:Unknown	1	1.000000
ICSI / BLASTOCYST:ICSI	1	1.000000
IVF / BLASTOCYST	1062	0.349509
ICSI / BLASTOCYST	1356	0.348777
FER	3	0.333333
IVF:ICSI	350	0.309524
ICSI / BLASTOCYST:IVF / BLASTOCYST	10	0.300000
IVF	43912	0.270535
ICSI	56714	0.269160
Unknown	18380	0.257319
ICSI:IVF	863	0.252646
ICSI / AH	684	0.209969
IVF / AH	303	0.202420
IVF:Unknown	96	0.187500
ICSI:Unknown	199	0.132328
DI	847	0.124224
ICSI:ICSI	910	0.025275
IVF:IVF	550	0.021818
IVF / AH:ICSI / AH	1	0.000000

범주형 변수의 고유값

BLASTOCYST 가 포함되는 경우 임신 성공률 평균이 전체적으로 **높은** 경향

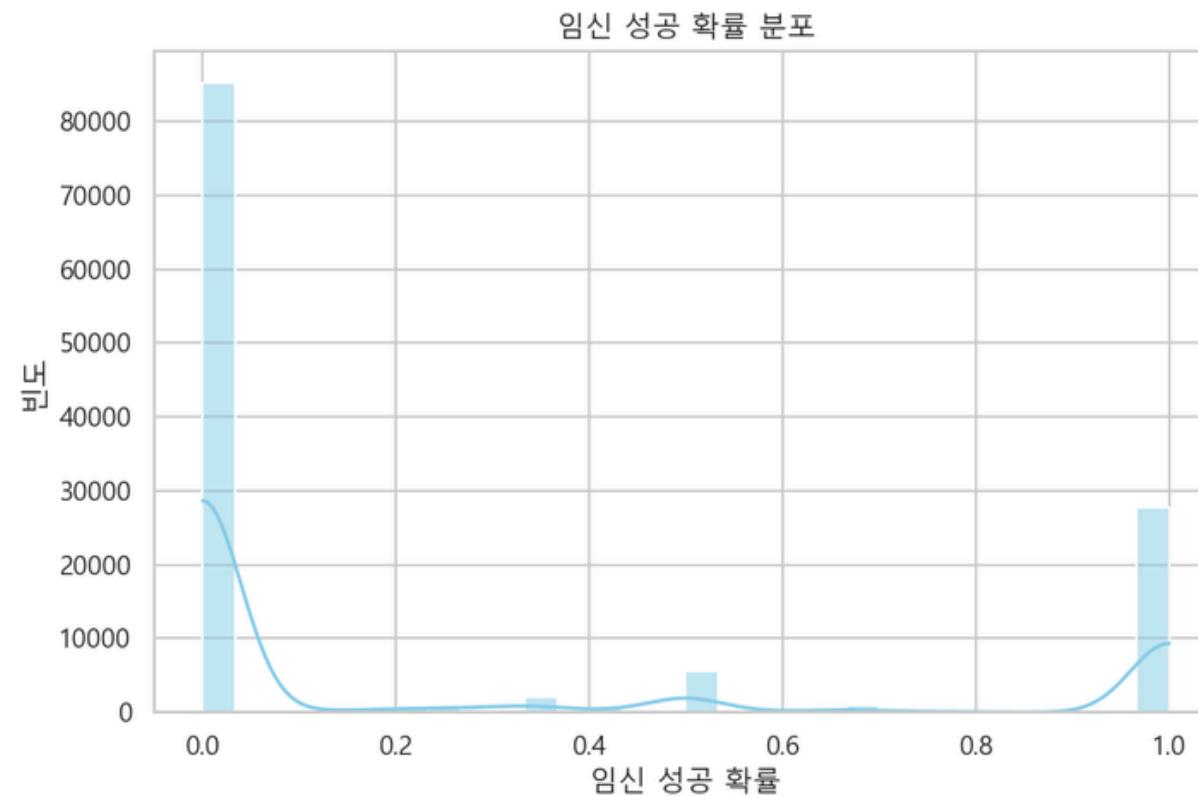
1. BLASTOCYST 가 포함되는 경우 **BLASTOCYST**
2. AH가 포함되는 경우 **AH**
3. 같은 시술 유형이 :로 연결되는 경우 **DOUBLE**
4. 다른 시술 유형이 :로 연결되는 경우 **DOUBLE_CROSS**
5. Unknown 와 다른 시술이 포함되는 경우 **DOUBLE_UNKNOWN**
6. ICSI, IVF, Unknown 은 그대로
7. 그 외는 ETC

=> 고유값 개수 19개 → 10개

같은 시술 유형이 :로 연결되는 경우 임신 성공률 평균이 전체적으로 **낮은** 경향

* EDA & PREPROCESSING

Target 분포



임신 성공 확률	Count
(0.0, 0.1]	178
(0.1, 0.2]	1178
(0.2, 0.3]	1468
(0.3, 0.4]	2724
(0.4, 0.5]	5911
(0.5, 0.6]	376
(0.6, 0.7]	990
(0.7, 0.8]	356
(0.8, 0.9]	46
(0.9, 1.0]	27794

본선에서 Target(임신 성공 확률)은 0~1 사이의 확률 값. (Regression)
대부분 0 또는 1의 값이며, 사이 값은 다양하게 분포함

* MODELING

문제 정의

이진 분류 모델

- Validation Set 구성에 **Target의 불균형**을 반영
- output을 범위에 맞는 **확률 값**으로 추출

회귀 모델

- Validation Set 구성에 Target 비율 반영에 어려움
- output을 확률 값으로 Clipping 필요



* MODELING

StratifiedKFold + class weight

K-fold cross validation을 통해 모델 학습 및 실험 진행

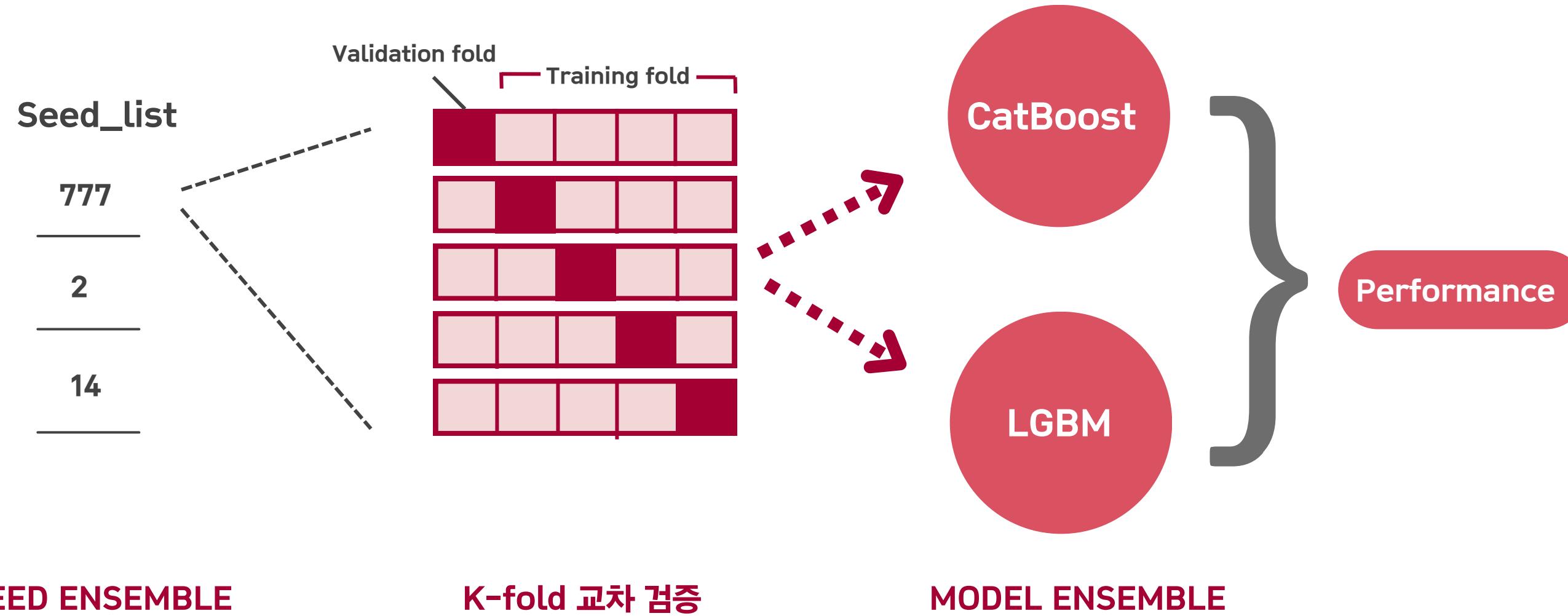
target imbalance를 고려하여 **Stratified K-fold**로 fold별 target 분포 동일하게 데이터 분할

5-fold 교차검증을 통해 모델 성능 일반화와 과적합 방지

각 **fold** 내의 **train data**의 target 비율을 활용해 **모델의 class weight 적용**

ValidationSet에 **Target**의 불균형을 반영

* MODELING _SEED ENSEMBLE & MODEL ENSEMBLE



모델의 일반화, 안정성 증가를 통해 과적합을 방지하기 위해 **Seed Ensemble**과 **Model Ensemble** 같이 활용

CatBoost: 순차적 통계 방식의 자체 인코딩 방식을 통한 Categorical 변수 처리

LGBM: 비대칭 구조를 활용한 유연한 분기로 데이터의 다양한 특징 학습

두 가지 모델을 양상불하여 각 모델의 장점을 반영

* MODELING

Hyperparameter Tuning

Optuna를 활용하여 **LightGBM**, **CatBoost** 하이퍼파라미터 튜닝을 진행

⇒ 트리 개수, 학습률, 최대 깊이, 샘플링 비율, L1/L2 정규화, 감마 등의 하이퍼파라미터를 고려하여 범위를 좁혀가며 최적의 하이퍼파라미터 리스트를 출력받아, 이를 모델에 적용

* 최종 모델의 실용성 및 활용 가능성

```
=====
Seed Number : 777
=====
<Model : LGBM>
[Valid] Fold #1 Score: 0.6638
[Valid] Fold #2 Score: 0.6577
[Valid] Fold #3 Score: 0.6551
[Valid] Fold #4 Score: 0.6552
[Valid] Fold #5 Score: 0.6576
Avg. Score of validset: 0.6578974152274235
Std. Score of validset: 0.0031739081839709158

<Model : CB>
[Valid] Fold #1 Score: 0.6619
[Valid] Fold #2 Score: 0.6577
[Valid] Fold #3 Score: 0.6544
[Valid] Fold #4 Score: 0.6559
[Valid] Fold #5 Score: 0.6582
Avg. Score of validset: 0.657637377928958
Std. Score of validset: 0.0025436854621755265

=====
```

```
=====
Seed Number : 2
=====
<Model : LGBM>
[Valid] Fold #1 Score: 0.6564
[Valid] Fold #2 Score: 0.6549
[Valid] Fold #3 Score: 0.6560
[Valid] Fold #4 Score: 0.6580
[Valid] Fold #5 Score: 0.6597
Avg. Score of validset: 0.6570215218873903
Std. Score of validset: 0.0016813447877224383

<Model : CB>
[Valid] Fold #1 Score: 0.6569
[Valid] Fold #2 Score: 0.6556
[Valid] Fold #3 Score: 0.6563
[Valid] Fold #4 Score: 0.6580
[Valid] Fold #5 Score: 0.6592
Avg. Score of validset: 0.6572010668291671
Std. Score of validset: 0.0012703523155506793

총 실행 시간: 8분 19.28초
```

```
=====
Seed Number : 1234
=====
<Model : LGBM>
[Valid] Fold #1 Score: 0.6571
[Valid] Fold #2 Score: 0.6581
[Valid] Fold #3 Score: 0.6548
[Valid] Fold #4 Score: 0.6593
[Valid] Fold #5 Score: 0.6565
Avg. Score of validset: 0.6571594308876059
Std. Score of validset: 0.0015107519394656212
```

다양한 모델의 양상블, 하이퍼파라미터 튜닝, 시드 양상블, SMOTE 기반 오버샘플링 등의 여러 조합의 실험 끝에,

최종적으로 Optuna 기반 하이퍼파라미터 최적화가 적용된

LightGBM, CatBoost 모델의

3개 시드 양상블이 최적 모델로 채택

#	팀	팀 멤버	최종점수	제출수
3	엄마손은 AI손	고강 ye ss yo le	0.6648	42

신속성 : 학습(8분) 및 추론 시간(1초) 단축 → 실시간 적용 가능

일반화 성능 : 높은 Private Score → 실제 환경에서도 안정적인 성능

* 현업 적용 가능성 부분

💡 배란 중심 변수 활용을 통한 LG화학 제품 효과 증명



COMPANY

PRODUCT

SUSTAINABILITY

CAREERS

난임(5)

IVF-C™ 주 IVF-C™ Inj. IVF-C™ Inj.	IVF-M HP™ 멀티도... IVF-M HP Multidose™ Inj. 다음과 같은 여성의 불임증 치료 1. 클로 미펜구연산염으로 치료되지 않는 여성...	IVF-M HP™ 주 IVF-M HP™ Inj. 다음과 같은 여성의 불임증 치료 1. 클로 미펜구연산염으로 치료되지 않는 여성...	가니레버 프리필드 시... Ganilever™ Prefilled syring... LG화학 생명과학사업본부의 전문의약품 제품 페이지입니다.	폴리트롭 프리필드 시... Follitropin™ Prefilled syrin... Follitropin™ Prefilled syringe Inj.
---	---	--	--	---

배란 성공 여부, 배란 시기, 시술 당시 난포 개수 등과 같은 배란 관련 데이터와 LG 화학에서 제공하는 자극제나 치료제의 환자의 약물 종류, 약물 투여 용량 및 기간 등과 같은 데이터를 추가로 사용하여 환자의 약물 반응 패턴을 정밀하게 모델링 하고자 함.

감사합니다.

LG Almers 6th

엄마손은 AI손

고경준 김예진 나영은 이수민 정서윤