난임환자 임신 성공 여부 예측 AI 해커톤

순풍산부인과 팀

대회 개요

난임은 증가하는 의료 문제로, AI 기반 예측 모델이 환자의 부담을 줄이고 맞춤형 치료를 지원하고자 함

이번 해커톤은 난임 환자 데이터를 활용한 임신 성공 예측 AI 모델 개발을 목표로 하며, 효과적인 치료 방안을 모색

대회 주제

난임 환자 대상 임신 성공 여부 예측 AI모델 개발

Project Definition

Dataset

Train.csv
256,351개
ID : 샘플별 고유 ID
난임 환자 시술 데이터 (67개의 컬럼)

Train.csv
90,067개
ID: 샘플별 고유 ID
난임 환자 시술 데이터 (67개의 컬럼)

Features

기본정보 변수
시술관련 변수
불임원인 변수
배아 및 난자 생성 관련 변수
난자 및 정자 출처 관련 변수
경과일 관련 변수
시술결과 변수
등 총 67개

Target

임신 성공 여부

1: 성공

0: 실패

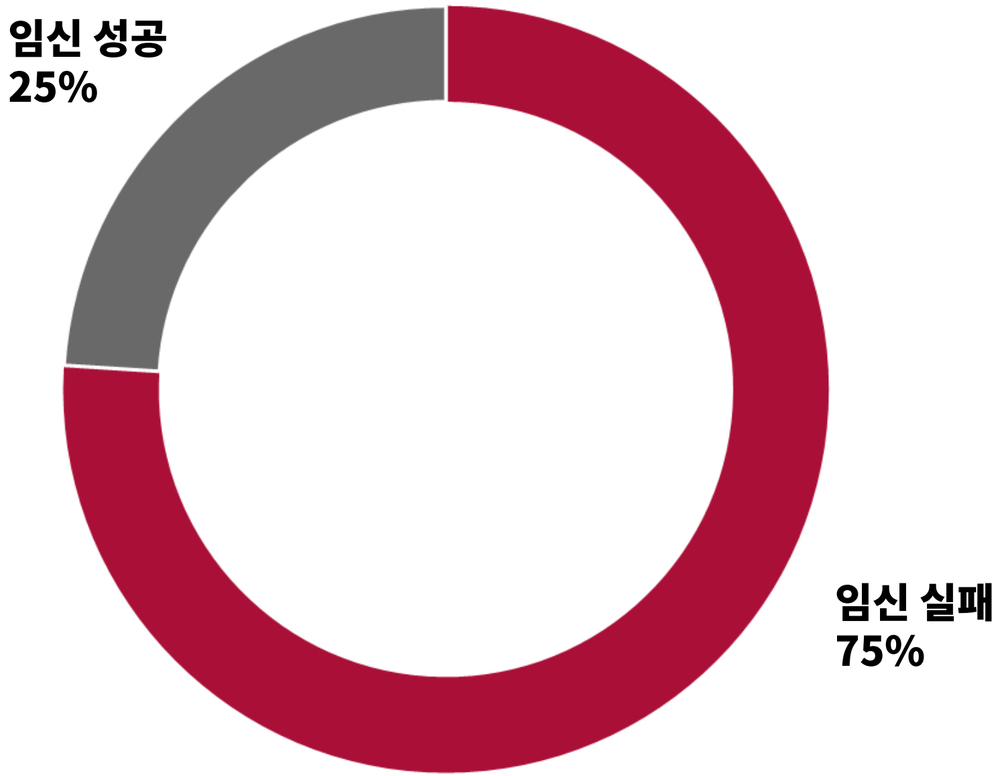
Target Distribution

Class Imbalance

임신 실패 : 75%

임신 성공 : 25%

이러한 불균형은 임신 성공 예측 성능 영향을 줄 가능성



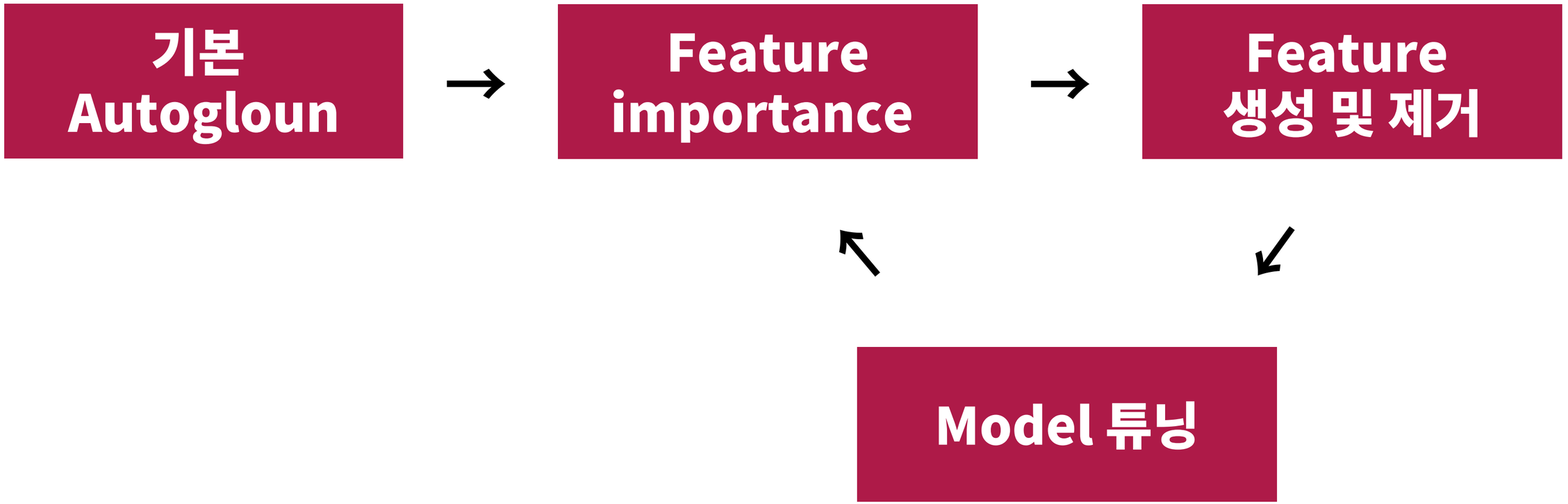
Model

AutoGluon

모델 사용 이유

1. **다양한 모델(XGBoost, LightGBM, Neural Network 등) 자동 학습**
2. **스택 앙상블을 통해 성능 극대화**
3. **하이퍼파라미터 튜닝, 데이터 전처리 자동 수행**

Model Flow



전처리

Column 삭제

Feature Importance 음수 및 하위권, p value 0.05 이상인 컬럼들 제거

기대효과

예측 성능 향상 및 과적합 방지

데이터 균형성 확보

전처리

삭제 Column

불임원인

불임원인 - 여성요인
불임원인 - 자궁경부
불임원인 - 정자 면역학
불임원인 - 정자 운동성
불임원인 - 정자 농도
부부 주 불임원인
여성 불임 원인

시술 배아

시술 유형
난자 해동 경과일
배아 해동 경과일
미세주입 배아 이식 수
착상 전 유전 진단 사용 여부

정자 수정

저장된 신선 난자 수
PGD 시술 여부
파트너 정자와 혼합된 난자 수
기증자 정자와 혼합된 난자 수

임신 출산

DI 출산 횟수
DI 임신 횟수
임신 시도 또는 마지막 임신 경과 연수
대리모 여부

총 22개

전처리

Feature 생성

이식된 배아 수의 비선형성 학습을 위한 제곱 피처 생성

```
# ✅ 피처 엔지니어링 적용
df["배아 수 제곱"] = df["이식된 배아 수"] ** 2
```

전처리

Feature 생성

시술 당시 나이 중앙값 생성 및 “알 수 없음“ 값 -1 처리

```
# ✅ 연령대를 중앙값으로 변환하여 숫자로 변환
age_mapping = {
    "만18-34세": 26,
    "만35-37세": 36,
    "만38-39세": 38.5,
    "만40-42세": 41,
    "만43-44세": 43.5,
    "만45-50세": 47.5,
    "알 수 없음": -1  # "알 수 없음"을 특별한 값으로 처리
}
```

전처리

Feature 생성

Feature importance 상위 feature 활용 파생 변수 생성

[총 생성 배아수, 저장된 배아 수, 이식된 배아 수, 배아 이식 경과일, 시술 당시 나이]

↓

파생변수

전처리

feature 생성

1. 배아 보존율

배아 보존율 = 저장된 배아 수 / (총 생성 배아 수 + 1)

2. 나이와 배아 수의 비율 변수

배아 수 대비 나이 = 시술 당시 나이 숫자 / (이식된 배아 수 + 1)

3. 특정 그룹화 변수 생성

1~2개 이식 여부 = (이식 배아 수 1~2개 = 1, not 0)

나이 26~36 여부 = (시술 나이 26~36세 = 1, not 0)

4. 추가 비율 Feature 생성

배아 이식 경과일 대비 나이 = 배아 이식 경과일 /
(시술 당시 나이 숫자 + 1)

전처리

상호작용 feature

1. 이식된 배아 수 × 시술 당시 나이

→ 시술 당시 연령과 배아 수가 이식 성공률에 미치는 영향

2. 이식된 배아 수 × 배아 이식 경과일 대비 나이

→ 배아 수와 이식 후 시간이 결합된 영향을 분석

3. 배아 이식 경과일 대비 나이 × 배아 수 대비 나이

→ 두 개의 비율 변수를 결합하여 복합적인 영향 분석

4. 시술 당시 나이 숫자 / 총 생성 배아 수

→ 연령과 배아 수의 관계를 정량적으로 반영

5. 배아 수 대비 나이 / 총 생성 배아 수

→ 전체 배아 수 대비 개별 변수가 미치는 영향을 조정

Model 튜닝

Model 선택

모델 성능 평가

하위권 모델 제거

상위권 모델 6개 선택

```
hyperparameters = {
    "GBM": {},
    "CAT": {},
    "XGB": {},
    "RF": {},
    "FASTAI": {},
    "XT": {},
    "NN_TORCH": {}
}
```

Model 튜닝

```
predictor.fit(
    train_data=train,
    presets="best_quality",
    time_limit=3600*12,
    num_bag_folds=8,
    num_stack_levels=0,
    dynamic_stacking=False,
    save_space=True,
    hyperparameters = hyperparameters,
    hyperparameter_tune_kwargs = {
        "num_trials": 20,
        "scheduler": "local",
        "searcher": "random"
    }
```

→ 최상의 성능을 목표

→ 학습시간 제한 = 12시간

→ Bagging 기법 적용 8개 폴드 사용해 훈련

→ stacking = 0 모델 복잡도 줄여 과적합 방지

→ 동적 스태킹 비활성화 과적합 방지

→ 모델 6개 선택 사용

→ 각 모델당 20번의 하이퍼파라미터 랜덤 탐색 수행

결과

모델 생성

WeightedEnsemble_L2 최종 앙상블 모델 생성

→ **총 13개 모델 앙상블**

모델 구성

LightGBM_BAG_L1/T3: 0.083
LightGBM_BAG_L1/T13: 0.0416
LightGBM_BAG_L1/T19: 0.0416
CatBoost_BAG_L1/T1: 0.208
CatBoost_BAG_L1/T2: 0.08
CatBoost_BAG_L1/T3: 0.08
NeuralNetFastAI_BAG_L1: 0.17
XGBoost_BAG_L1/T4: 0.08
XGBoost_BAG_L1/T5: 0.042
XGBoost_BAG_L1/T9: 0.042
NeuralNetTorch_BAG_L1/cbc41_00003: 0.0417
NeuralNetTorch_BAG_L1/cbc41_00005: 0.0417
NeuralNetTorch_BAG_L1/cbc41_00006: 0.0417

결과

예측 수행

ROC_AUC Validation score = 0.7404

Public score = 0.74203

Private score = 0.74221

→

Private score > Public score

모델의 일반화 성능이 안정적

과적합 X

감사합니다!

순풍산부인과 팀