241122 수상작리뷰

[심장 질환 예측 AI 해커톤톤]

☑️**대회**: [머신러닝 입문 트랙 시즌3]심장 질환 예측 AI 해커톤 - DACON
https://dacon.io/competitions/official/236333/overview/description

☑️**참고 코드**: Private 1위, LGBM 모델, 순서형(라벨) 인코딩 전처리 기법 활용 - DACON
https://dacon.io/competitions/official/236333/codeshare/11599

📗 데이터

a. train.csv:
- id
- age
- sex: 여자=0, 남자1
- cp: 가슴통증 종류(0=무증상, 1=일반적이지 않은 협심증, 2=협심증이 아닌 통증, 3=일반적인 협심증)
- trestbps: 휴식중 혈압(mmHg)
- chol: 혈중 콜레스테롤(mg/dl)
- fbs: 공복 중 혈당 (120mg/dl 이하일 시 0, 초과일 시 1)
- restecg: 휴식 중 심전도 결과
(0=showing probable or definite left ventricular hypertrophy by Estes' criteria, 1= 정상, 2=having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV))
- thalach: 최대 심박수
- exang: 활동으로 인한 협심증 여부(0=없음, 1=있음)
- oldpeak: 휴식 대비 운동으로 인한 ST 하강
- slope: 활동 ST분절 피크의 기울기(0=하강, 1=평탄, 2=상승)
- ca: 형광 투시로 확인된 주요 혈관 수(0~3개, Null은 4로 인코딩)
- thal: 지중해빈혈 여부(0=Null, 1=정상, 2=고정 결함, 3=가역 결함)
- **target**: 심장 질환 진단 여부(0: <50% diameter narrowing, 1: >50% diameter narrowing)

b. test.csv:
target 칼럼 제외 동일

c. sample_submission.csv:
- id
- **target**: 심장 질환 진단 여부

📗 코드흐름

1. **라이브러리 로드**
2. **EDA**
    - 결측치 처리: Ca 결측치는 모델 학습으로 예측, Thal 결측치는 다른 피처로 간접적 처리
    - 전체 피처 순서형 범주로 인코딩: 순서형 범주화하여 학습에 반영
      - 다중 카테고리형 피처는 원핫 인코딩
      - 숫자형 피처는 StandardScaler() fit_transform()
      - slope의 downsloping, flat, upsloping 0,1,2 -> 2,1,0
      - age는 young, middle, old 3개로 범주화
      - risk, chol, thalach, oldpeak Low, middle, High risk로 3개 범주화

    - 피처 선택: LGBM 기반으로 피처 중요도 계산하여 모델 기반 피처 선택
      - Id 드롭

3. **모델링: LGBM**
    - LGBM으로 모델링 > accuracy: 0.78
    - feature_importance() 로 중요도 계산, 중요하지 않은 피처는 드롭
    - 다시 LGBM > accuracy: 0.81
    - XGBRegressor, GridSearchCV로 튜닝 후 모델링 > accuracy:0.76
    - LGBM으로 최종 모델 선정

4. **제출**

📗 배운점 / 특이점

- 이미 숫자형인 피처들도 필요에 따라 적은 위험도 ~ 높은 위험도로 중요도를 차등 적용해 범주형 피처로 만든 후 원핫 인코딩을 진행할 수 있다.
- 결측치를 XGBRegressor()이용한 학습 예측치로 채우는 방법이 있음
- 0,1,2 > 2,1,0으로 값을 바꾸며 피처를 전처리 할 수 있음
- 대부분 피처를 범주화 하여 학습시킴
- 피처 중요도 확인 후 중요 피처만 선택하니 중요도가 높아짐을 확인