2024-11-11 수상작 리뷰

[머신러닝 입문 트랙 시즌3]심장 질환 예측 AI 해커톤

https://dacon.io/competitions/official/236333/codeshare/11599

데이터

- id: 데이터 고유 id
- age: 나이
- sex: 성별 (여자 = 0, 남자 = 1)
- **cp: 가슴 통증(chest pain) 종류**
- **0 : asymptomatic 무증상**
- **1 : atypical angina 일반적이지 않은 협심증**
- **2 : non-anginal pain 협심증이 아닌 통증**
- **3 : typical angina 일반적인 협심증**
- trestbps: (resting blood pressure) 휴식 중 혈압(mmHg)
- chol: (serum cholestoral) 혈중 콜레스테롤 (mg/dl)
- fbs: (fasting blood sugar) 공복 중 혈당 (120 mg/dl 이하일 시 = 0, 초과일 시 = 1)
- **restecg: (resting electrocardiographic) 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: (maximum heart rate achieved) 최대 심박수
- exang: (exercise induced angina) 활동으로 인한 협심증 여부 (없음 = 0, 있음 = 1)
- oldpeak: (ST depression induced by exercise relative to rest) 휴식 대비 운동으로 인한 ST 하강
- **slope: (the slope of the peak exercise ST segment) 활동 ST 분절 피크의 기울기**
- **0: downsloping 하강**
- **1: flat 평탄**
- **2: upsloping 상승**
- ca: number of major vessels colored by flouroscopy 형광 투시로 확인된 주요 혈관 수 (0~3 개)
- **Null 값은 숫자 4로 인코딩됨**
- thal: thalassemia 지중해빈혈 여부
- **0 = Null**
- 1 = normal 정상
- 2 = fixed defect 고정 결함
- 3 = reversable defect 가역 결함
- target: 심장 질환 진단 여부

- 0: < 50% diameter narrowing
- 1: > 50% diameter narrowing

코드 흐름

1) **EDA**

   결측치 처리: Ca 결측치는 모델 학습을 통해 예측,Thal 결측치는 다른 피처를 활용하여 간접적으로 처리

   전체 피처를 순서형 범주로 인코딩 : 학습의 복잡도를 고려하여 전체 피처를 순서형 범주화하여 학습에 반영

   Corr() 확인.

2) **Feature Engineering & Processing**

   결측치 처리: Ca 결측치는 모델 학습을 통해 예측,Thal 결측치는 다른 피처들을 순서형 범주로 인코딩 후 추론하기로 결정

   전체 피처를 순서형 범주로 인코딩 : 학습의 복잡도를 고려하여 전체 피처를 순서형 범주화하여 학습에 반영

   피처 선택: LGBM 모델을 기반으로 피처 중요도를 계산하여 모델 기반 피처 선택을 활용

   모델링

3) **모델 학습(Linear Regression)**

   최종 모델 선정: 성능 평가에서 LGBM 모델이 다른 모델(XGBRegressor)보다 앞서는 경향을 보여 최종 모델로 선정

   하이퍼파라미터 튜닝: 모델 성능을 극대화하기 위해 하이퍼파라미터를 조정하며 최적의 성능을 도출

4) **후처리**

   예측값 할당

차별점, 배울점

이 수상작에서는 전체 피처를 순서형 범주화하였다. 이러한 과정을 학습에 반영했을 때의 장점은 모델의 복잡도가 감소된다는 것과, 연산의 효율성이 증가한다는 것이다. 그 외에도 해석도 쉬워지고 노이즈를 줄이는 효과도 있다고 한다.

또한, 이 수상작에서는 Feature Importance를 시각화해서 살펴보며 중요하지 않은 피처들을 제외시켜 모델에 적용해 여러 번 학습을 시도하였다. 이러한 과정을 학습에 반영했을 때의 장점은 모델이 단순화된다는 것과, 과적합이 방지된다는 장점이 있다.