수상작리뷰
[Private 3위] 전통적인 앙상블 기법 활용(ExtraTree + CatBoost + LightGBM) - DACON

[축구선수의 유망 여부 예측 AI 해커톤]

◈ 주제

👉 축구선수 관련 데이터를 활용해 선수의 유망 여부 예측(이진 분류)하는 AI 알고리즘 개발

◈ 데이터

- train.csv : ID , 축구선수 관련 정보(나이, 키, 몸무게, 포지션, 선호하는발, 공격력, 수비력, 슈팅, 패스, 드리블, 정확도, 롱패스 등…), 유망 여부
- test.csv : ID , 축구선수 관련 정보(나이, 키, 몸무게, 포지션, 선호하는발, 공격력, 수비력, 슈팅, 패스, 드리블, 정확도, 롱패스 등 …)
- sample_submission.csv : ID , 유망 여부

◈ 코드 리뷰

(1) 라이브러리 설치, 불러오기, 패키지 임포트

- numpy, pandas, matplotlib, seaborn / LabelEncoder, StandardScaler / LR, RF …

(2) 사용할 함수 생성 *(교차 검증 함수)*

🔷 모델 적용 기본 베이스라인 함수 basic_model_evaluation
🔷 target 불균형 문제를 반영한 함수 skf_fold_evaluation: 예측 확률이 0.36 이상 → 1로 예측
🔷 EDA 용 함수: 데이터 정보 확인-data_info / 칼럼명 비율 확인-search_columns, search_target_ratio / 연속형 변수 plot 그리기-numerical_plot, numerical_group_plot, numerical_target_group_ratio / 상관계수 히트맵 그리기-correlation_matrix, describe_matrix

(3) EDA 수행

(4) 데이터 전처리 및 스케일링

- 포지션 개수 너무 많음 → attack/midfield/defend/goalkeeper 4개로 축소
- 범주형 변수 → 라벨인코딩
- 연속형 변수 → 표준화
- 학습데이터, 테스트데이터 비율 나누기
- ID 변수 제거
- 파생변수 생성

- 불필요한 컬럼 제거 (근거: 단일 모델 수행 시 설명력이 낮은 컬럼 제거)

(5) 단일 모델 성능 파악 *(↔ (4) 불필요한 컬럼 제거에 사용)*

(6) 하이퍼 파라미터 튜닝
- optuna 사용
- acc, f1, 정밀도, 재현율 roc 출력

(7) 알고리즘 학습 및 예측 수행
- 소프트 보팅 수행

◈ 배울점

- 피처개수가 너무 많은 점을 고려해 반복함수를 생성해 EDA 수행한 점
- 전반적으로 자신만의 함수를 만들어 편리하게 작업을 수행함.
- 함수 생성하는 능력을 길러야겠다고 다짐함.
- 단일모델 성능 파악 후에 불필요한 컬럼을 제거하는 것처럼 수행 과정 중에 즉각적인 피드백을 수행함.
- 하이퍼 파라미터를 자동으로 탐색하는 optuna 을 실제로 어떻게 사용할 수 있는지 알게됨.