09. 분류 실습 - 캐글 산탄데르 고객 만족 예측

- 데이터 전처리
    - 사이킷런 래퍼를 이용해 필요한 모듈 로딩하고 학습 데이터를 DataFrame으로 로딩
    - 클래스/피처 데이터 세트 분리
    - 학습/데이터 세트 분리
- XGBoost 모델 학습과 하이퍼 파라미터 튜닝
    - 분리한 학습/테스트 데이터 세트로 학습을 진행한 뒤 평가 진행
    - HyperOpt를 이용해 베이지안 최적화 기반 XGBoost 하이퍼 파리미터 튜닝
    - 목적 함수를 만들어 최대 ROC-AUC 값을 최소값으로 반환
    - fmin() 함수 호출해 반복하며 최적의 파라미터 도출
    - 도출된 최적의 하이퍼 파라미터를 기반으로 재학습 및 재측정
    - 수행시간이 많이 요구된다는 단점
- LightGBM 모델 학습과 하이퍼 파리터 튜닝
    - 분리한 학습/검증 데이터 세트 이용해 학습 및 평가 진행
    - XGBoost에 비해 단축된 학습 시간
    - 목적 함수 생성
    - fmin()호출해 최적 하이퍼 파라미터 도출
    - 최적의 하이퍼 파라미터를 이용해 재학습 후 재평가

10. 분류 실습 - 캐글 신용카드 사기 검출

- 언더 샘플링과 오버 샘플링의 이해
    - 이상 레이블 데이터 건수가 적어 다양한 유형을 학습하지 못하는 문제 발생
    - 언더 샘플링: 많은 데이터 세트를 적은 데이터 세트 수준으로 감소
    - 오버 샘플링: 적은 데이터 세트를 증식하여 학습을 위한 충분한 데이터 확보
- 데이터 일차 가공 및 모델 학습/예측/평가
    - DataFrame으로 데이터 로딩
    - get_preprocessed_df() 함수와 get_train_test_df()함수 이용해 데이터 가공 빛 학습/데이터 세트 반환
    - 학습/예측/평가를 위한 별도의 함수 생성
    - 테스트 데이터 세트에서 예측 및 평가 수행
- 데이터 분포도 변환 후 모델 학습/예측/평가
    - get_processed_df()함수를 이용해 정규 분포 형태로 변환
    - get_train_test_df()함수를 호출해 학습/데이터 세트 생성
    - get_model_train_eval() 이용해 로지스틱 회귀와 모델 학습/예측/평가
    - get_preprocessed_df() 사용해 로그 변환 로직으로 변경
- 이상피 데이터 제거 후 모델 학습/예측/평가
    - 이상치 데이터: 전체 데이터 패턴에서 벗어난 이상 값을 가진 데이터, 머신러닝 모델 성능에 영향
    - IQR 방식을 통해 찾아낼 수 있음
    - IQR 방식: 4분위 값의 편차를 이용해 최댓값과 최솟값을 결정한 뒤, 그 범위에서 벗어난 데이터를 이상치로 간주
    - get_outlier()함수로 이상치 검출한 칼럼 입력받은 후 넘파이의 percentile() 이용해 IQR 계산

- SMOTE 오버 샘플링 적용 후 모델 학습/예측
    - SMOTE 기법으로 오버 샘플링 적용한 뒤 로지스틱 회귀와 LightGBM 모델의 예측 성능 평가
    - imbalanced-learn 패키지의 SMOTE 클래스를 이용해 구현
    - 반드시 학습 데이터 세트만 오버 샘플링 해야 함
    - fit_resample() 메서드를 이용해 증식하여 학습