준비한 데이터셋 : openml의 credit-g, adult, higgs, covertype

데이터셋 크기 : credit-g<adult<higgs<covertype

Credit-g, adult, higgs 모두 이진분류하는 문제, covertype은 7개의 category 존재

Categorical한 value는 adult에 많고, higgs 는 모두 numeric한 value들. 자료 조사 중 Catboost와 lgbm의 차이가 categorical한 value들이 많은 데이터셋에서 catboost의 성능이 더 좋다는 언급이 있어서 비교하기 위해 이렇게 데이터를 선정했다.

Catboost, lgbm은 categorical한 value도 알아서 처리해주지만 xgboost는 그러지 못한다. 하지만 세 모델을 한 코드 내에서 같이 사용했으므로 일괄적으로 데이터셋을 전처리해주고 모델에 사용하였다.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

모델 구현텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

이런 식으로 4개의 데이터셋에 lgbm, catboost, xgb 모델들 다 적용해서 구현했다. 정확한 성능 비교를 위해 하이퍼파라미터 중 트리 개수, 학습률, 트리 최대 깊이는 통일했다.

걸린 시간,메모리 비교

위와 같이 모델이 작동되는 시간과 사용한 메모리량을 기록하여 배열에 저장한 다음, 마지막에 걸린 시간과 메모리를 비교하였음.

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

위 사진은 credit-g 데이터의 결과, 순서대로 lgbm, catboost, xgboost

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

adult에서 time과 memory

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

higgs에서 time과 memory

텍스트, 스크린샷, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

covertype에서 time과 memory

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

그래프로 시간 소요량 그려봄

데이터셋의 크기가 커질수록 xgb는 대폭 느려졌고, catboost도 다소 느려지는 모습을 볼 수 있음

반면 lgbm은 데이터의 크기가 굉장히 커져도 안정적인 속도를 보여줌 🡪 매우 빠르다는 것을 알 수 있음

메모리 사용량에 대해선 세 모델 다 비슷한 양상을 보였음

모델 성능 비교

텍스트, 도표, 라인, 그래프이(가) 표시된 사진

자동 생성된 설명

텍스트, 라인, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

텍스트, 라인, 스크린샷, 그래프이(가) 표시된 사진

자동 생성된 설명

Covertype 제외하고 credit-g, adult, higgs 순서대로 ROC Curve

성능에서 유의미한 차이가 있다고 보기는 힘들었음

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명  
텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

자동 생성된 설명

이런 식으로 precision, recall, accuracy, f1-score를 확인하고 confusion-matrix로 나타내었음

Logistic과의 비교

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

Logistic regression과 성능을 비교하여 Tree-based의 gradient boosting 모델이 얼마나 좋은 성능을 보이는지 비교해봄.

텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 번호, 도표이(가) 표시된 사진

자동 생성된 설명 텍스트, 스크린샷, 번호, 도표이(가) 표시된 사진

자동 생성된 설명

위는 higgs, 아래는 covertype 데이터셋,

왼쪽은 lgbm, 오른쪽은 logistic regression의 confusion matrix

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

Adult 데이터셋, 위는 lgbm, 아래는 logistic의 성능

성능 비교한 결과 확실히 gradient-boosting 류 트리 모델들이 logistic regression 보다 좋은 성능을 내고 있음을 확인할 수 있었음. F1-score도 boosting류 모델들이 훨씬 더 높음. 예측이 더 정확함.

결론

자료조사 결과 catboost류가 category value가 많은 데이터셋에선 유리하다고 조사됐지만 모델을 적용해보니 딱히 그렇지 않았음. 또한 데이터의 크기가 커진다고 해서, feature가 많아진다고 해서 세 모델 (lgbm, cat, xgb) 의 성능 차이가 뚜렷히 벌어지진 않았음. 이론과 실제는 약간의 차이가 존재하는 듯.

본 프로젝트에선 몇몇 하이퍼파라미터를 모든 경우에서 다 통일시켰으나 실제론 다르게 변경해나가면서 성능을 더 높일 수 있을 것으로 보임. 어떤 모델을 사용한다기보단, 하이퍼파라미터 세팅과 Cross-validation 등으로 모델들을 다양하게 테스트해보고 최선의 모델을 찾아나가는 것이 중요한 것임을 알게 되었음.