머신러닝 프로젝트 주제 고안 – 신용카드 사기 감지

https://www.kaggle.com/mlg-ulb/creditcardfraud

신용 카드 회사가 사기 신용 카드 거래를 인식할 수 있는 모델을 만들어 고객이 구매하지 않은 항목에 대해 비용이 청구되지 않도록 만들기(=> 경향성 필요. 2종 오류(Miss Alarm) 줄여야 함.)

31개 열로 이루어져 있음. 단, 보안을 위해 이미 PCA 차원축소가 이루어진 데이터임 ㅠㅠ 그래서 우리가 각 feature가 뭔지는 알 수 없음.

샘플 수는 약 28만 5천 개. 144MB의 꽤 큰 데이터임!

Target 은 0 또는 1임. 단, 음성(0) 데이터보다 양성(1) 데이터가 훨씬 적음.

즉, 사기 건보다 정상 건이 훨씬 많음. nba때랑 비슷하다면 오히려 모델링 성능 높이는 것은 어렵지 않을 것으로 예상됨.(사기 건의 특성이 뚜렷하다면)

다른 사람들의 작업을 볼 수 있는데, 모델링 부문에서 가장 인기있는 작업물은 로지스틱 회귀를 사용함. 기타 SVM, 랜덤 포레스트, 단순신경망을 활용한 작업들도 있는 것으로 보임.

로지스틱 회귀 활용한 작업: https://www.kaggle.com/gauravduttakiit/creditcard-fraud-detection-by-logistic-regression

Nba때와 같이 인사이트 도출 등을 기대할 수는 없음.(특성이 뭔지 가려져 있기 때문에ㅜㅠ)

하지만 데이터가 이미 잘 정제되어 있는 것으로 보이고, 이 작업 참여 코드가 3천개가 넘는 데다 참고할 만한 좋은 작업들이 많고, 딥러닝 등이 아니라 머신러닝으로 되어 있음. 이미 다른 사람들이 성능이 꽤 나온다는 것을 검증해 줬기 때문에 나름…(??) 쉬운 작업이 되지 않을까 생각함.