**Imbalance 데이터셋이 모델 성능이 미칠 수 있는 악영향**

**1. Imbalance 데이터셋으로 학습 -> Imbalance 데이터셋 예측**

문제점

1. Biased classification on major class
2. Overfitting on major class
3. Precision-recall imbalance : 평가 지표를 바꾸면 된다.

**2. [씨젠 현재 상태] Balance 데이터셋으로 모델 학습 -> Imbalance 데이터셋 예측**

문제점

1. Biased classification on major class : 해결
2. Overfitting on major class : 해결
3. Precision-recall imbalance : 평가 지표를 바꾸면 된다.
4. [추가] Prior probability shift

=> 충분한 양의 Balance dataset을 마련했을 경우, 4) 문제 외에는 대다수 해결된다.

* Prior probability shift 가 발생했을 때 성능이 떨어진다는 다수의 연구가 존재한다.

**[단계별 시나리오]**

**Step 1 : 2 Step Framework 형성**

**기대효과 : 1) Prior probability shift 완화, 2) One-Classification 도입을 통한 정확도 향상**

=> Check) 성능이 더 향상되냐?

**추가 문제점 발생**

- Step1 을 통해 필터링 된 데이터의 분포가 변동적이다.

- 즉, Prior probability shift 문제가 완전히 해소되진 않는다.

**Step 2 : Step1을 통해 필터링 된 데이터에 대해 Prior probability shift 문제를 해결한 Classification 방법을 적용하자!**

**기대 효과 : 1) Prior probability shift 문제 해소를 통한 모델 정확도 향상**

**추가 의문점 발생**

* 1) 2 Step Framework를 거친 후 Prior probability shift 문제를 해소하는 게 좋을지
* 2) 전체 데이터에 대해서 Prior probability shift 를 해소한 Classification을 하는 게 좋을지

**Step 3 : 전체 데이터에 대해서 Prior probability shift 한 것과, 2 Step Framework를 적용한 후 prior probability shift 한 것의 결과를 비교하자!**

* 2 Step Frame work 구조가 정말 효과가 좋은가?