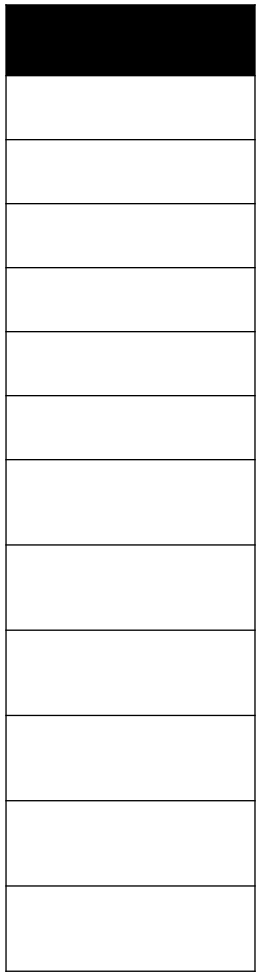
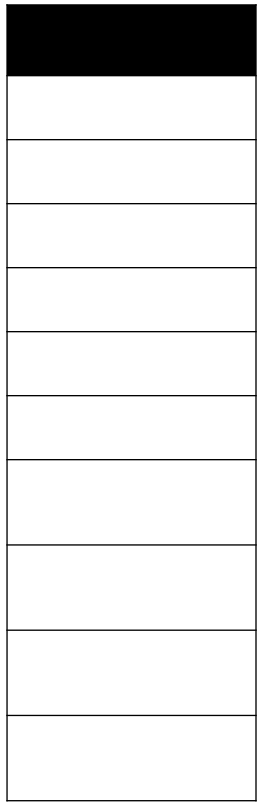
2018/11/1 신호결과 데이터 추출과정

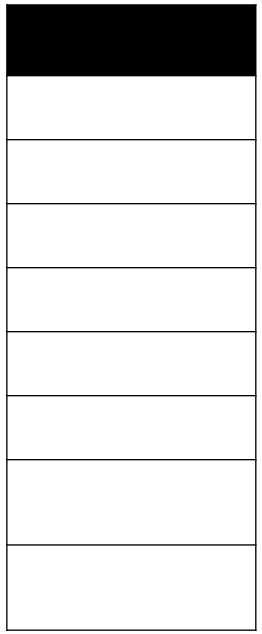
1.
11월 전체데이터
(22,687,614건)



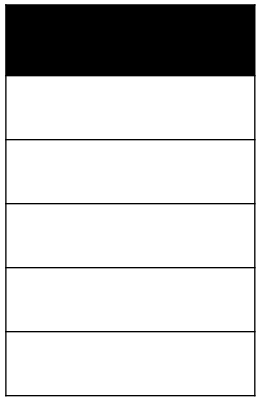
2.
신호명 '침입' 데이터
(860966건)



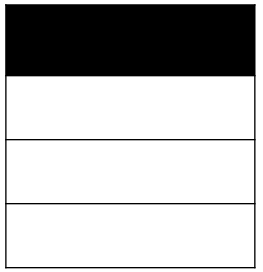
3.
2번을 충족하며
침입데이터 중
오전 6시 이전 데이터
(4313건), 상호수 624



4.
3번을 충족하며
첫번째 신호가 '침입'인
상호수 336



5.
4번을 충족하는 상호가
10월 이전 데이터 9건을
갖고 있는 **케이스 312**



학습시 비율이 맞아야 하는 이유

- 인공지능 학습의 최종목표 = loss를 줄이는 일
- Loss는 잘맞추면 줄어든다(강아지를 강아지로, 고양이를 고양이로)
- 만약 1000건의 강아지사진과 1건의 고양이사진으로 학습을 진행할 경우

  - 학습을 진행하며 강아지사진은 1000번 검토할 동안,
    고양이사진은 1번 검토하게 된다.
  - 그러므로 학습과정에서 강아지의 특징위주로 학습을 하게된다.
  - 또한, 모두 강아지사진으로 예측해도 정확도 99.9%가 되고,
    loss를 최소화 시킬수있으므로 학습 방향을 고수함.
  - 결국 모두 강아지로 예측하는 결과초래.

관련 논문

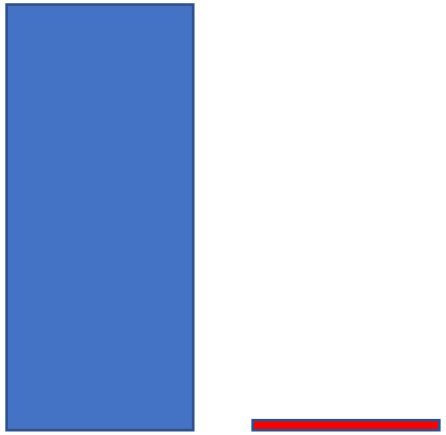
Fithria Siti Hanifah , Hari Wijayanto , Anang Kurnia "SMOTE Bagging Algorithm for Imbalanced Data Set in Logistic Regression Analysis". Applied Mathematical Sciences, Vol. 9, 2015
Lina Guzman, DIRECTV "Data sampling improvement by developing SMOTE technique in SAS" .Paper 3483-2015

학습시 비율이 맞아야 하는 이유

불필요출동
180,000,000건

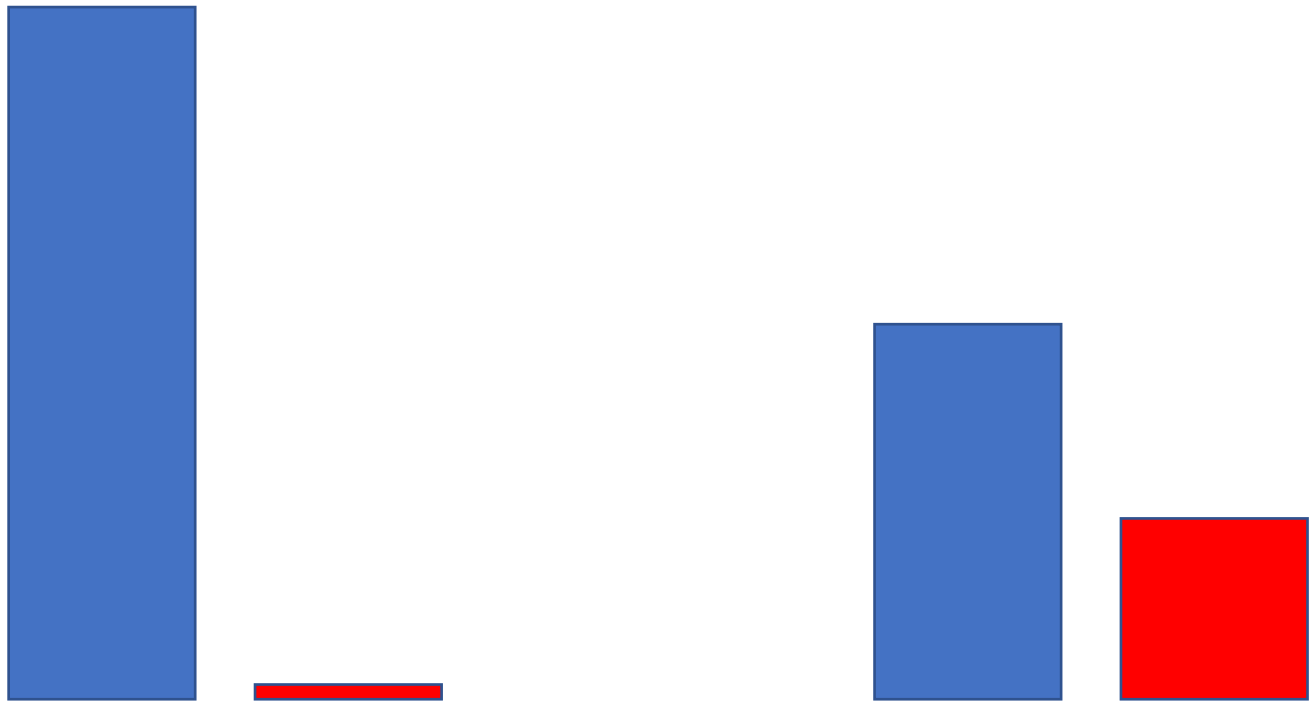
필요출동
약 600 건

불필요출동 대비 0.000003의 비율
-> 무조건 불필요출동으로 예측하게됨



단순히 온콜, ATM, 국선체크신호를 지우면 달라질까?
-> 설령 80%가 제거되어 36,000,000 건으로 줄었다 하더라도, 데이터 비율이 10%미만으로 떨어질 경우 학습이 정확하게 이루어 지지 않음

학습시 비율이 맞아야 하는 이유



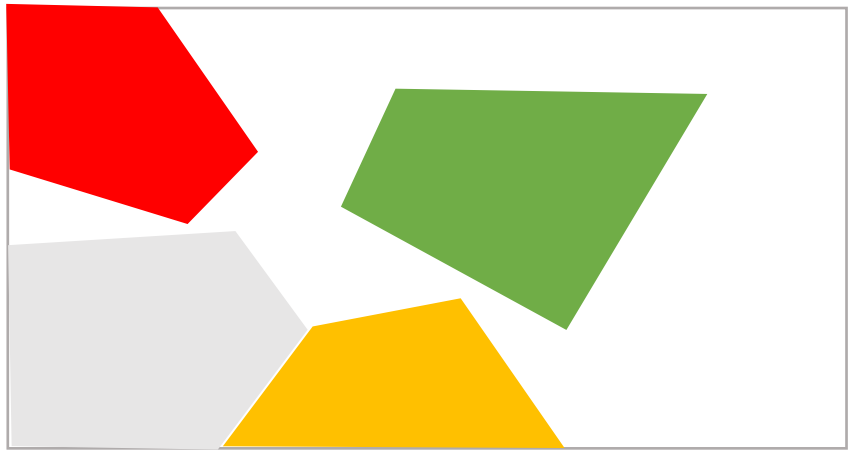
결론 : 많은 데이터는 줄이고(UNDERSAMPLING), 적은데이터는 늘려서(OVERSAMPLING) 비율을 맞춰야한다

단순히 비율만 맞추면 안되는 이유

보다 정확한 예측을 위해선 다양한 정보를 습득하는게 유리하다

목표 : 초록색 데이터 판별하기

3:1 비율로 학습 시도

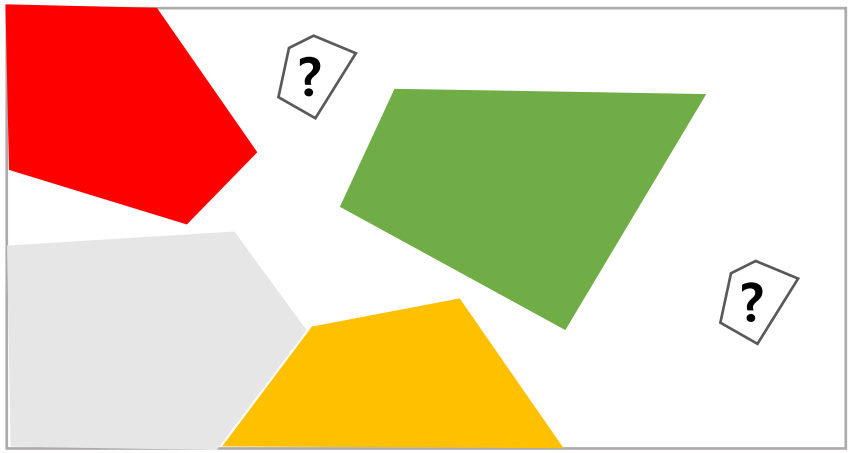


학습한 정보를 바탕으로 판별영역을 시각화한 예시 -> 해당 영역 안에 있을 경우 정확한 예측

단순히 비율만 맞추면 안되는 이유

보다 정확한 예측을 위해선 다양한 정보를 습득하는게 유리하다

목표 : 초록색 데이터 판별하기



물음표 영역은 어떻게 판단할 것인가?
->정확도가 떨어짐

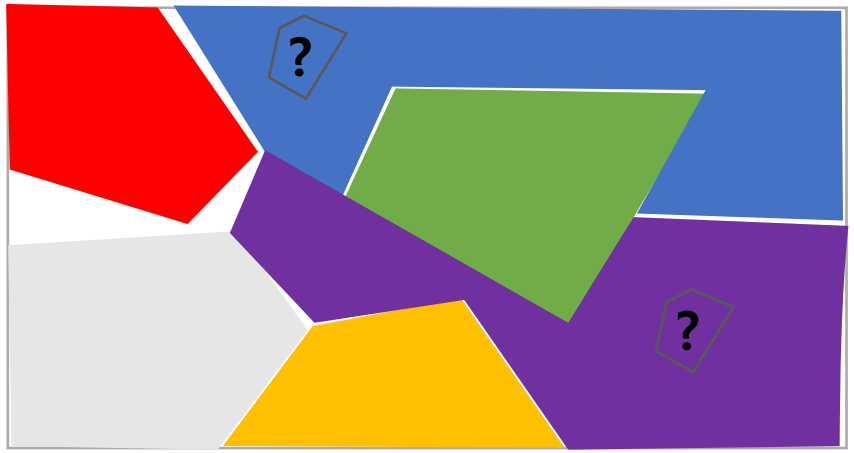
단순히 데이터를 복제한다고해서 얻을수있는 정보가 증가하지 않음
-> 판단범주의 영역은 변화없음

단순히 비율만 맞추면 안되는 이유

보다 정확한 예측을 위해선 다양한 정보를 습득하는게 유리하다

목표 : 초록색 데이터 판별하기

데이터간 비율은 이전과 동일
3:1



추가된 정보를 바탕으로
초록색 데이터가 아님을 분명히 예측할수 있음