Part. 06

Class Imbalanced Problem

| Class Imbalanced Problem을 해결하기 위한 방법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

강사. 이경택

I Class Imbalanced Problem이란

- Class Imbalanced Problem이란
  - 클래스 불균형은 다수 클래스(majority class)의 수가 소수 클래스(minority class)의 수보다 월등히 많은 학습 상황을 의미하며 클래스 불균형 데이터를 이용해 분류 모델을 학습하면 분류 성능이 저하되는 문제가 발생함.
  - 모델이 소수의 데이터를 무시하는 경향이 생김
  - 클래스 불균형 데이터는 의료, 반도체, 보험, 텍스트 등 여러 분야에 걸쳐서 발생하고 있는 문제임
  - IR (class imbalanced ratio) $= \frac{\#\ of\ majority\ class}{\#\ of\ minority\ class}$

I Class Imbalanced Problem이란

- Class Imbalanced Problem이란
  - Binary문제에서 일반적으로 모델은 각 데이터에 대한 확률 값을 output으로 함
  - IR이 높은 경우 대부분의 데이터 셋에 대하여 0에 가까운 확률 예측 값을 냄
  - 예측 threshold 값(기본 0.5)이 달라 져야하는 문제가 생김

- Class Imbalanced Problem에서 사용하는 모델 성능 지표
  - G-mean, F1 measure, AUC

I Class Imbalanced Problem이란

- Class Imbalanced Problem을 해결하기 위한 방법
  - Resampling method

    - Over sampling : 소수의 데이터를 부풀리는 방법

    - Under sampling : 다수의 데이터를 줄이는 방법

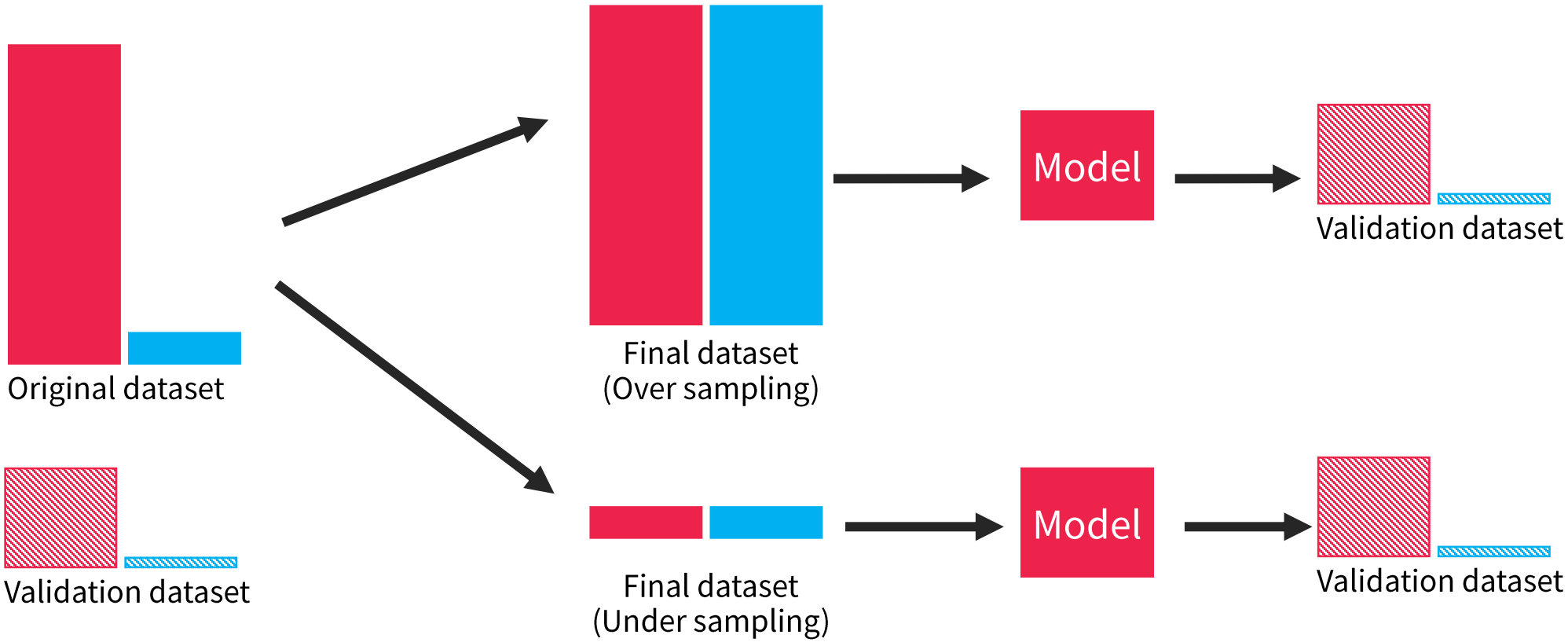
    - Hybrid resampling : Over & Under sampling을 결합해서 사용하는 방법

  - Cost-sensitive learning

    - Class의 오 분류에 대한 cost의 가중치를 조절하여 학습하는 방법

I Class Imbalanced Problem이란

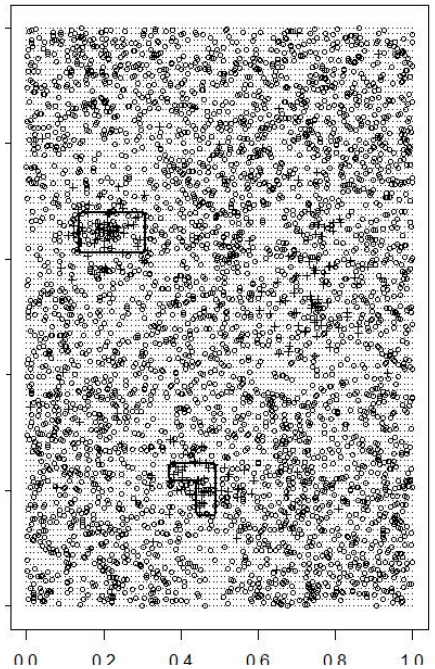
- Class Imbalanced Problem을 해결하기 위한 방법
  - Resampling method



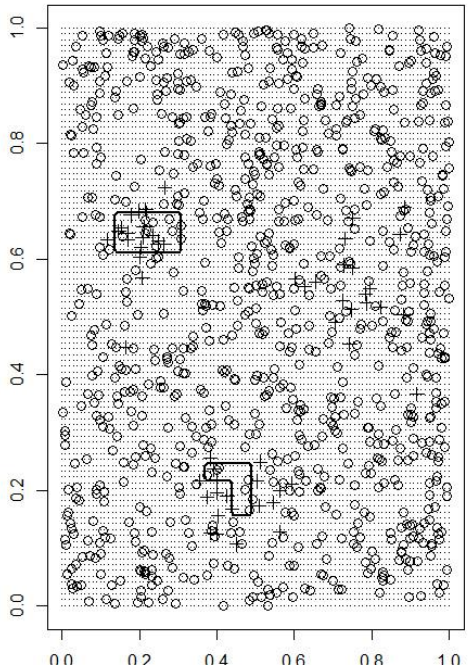
I Class Imbalanced Problem이란

- Class Imbalanced Problem을 해결하기 위한 방법
  - Resampling method

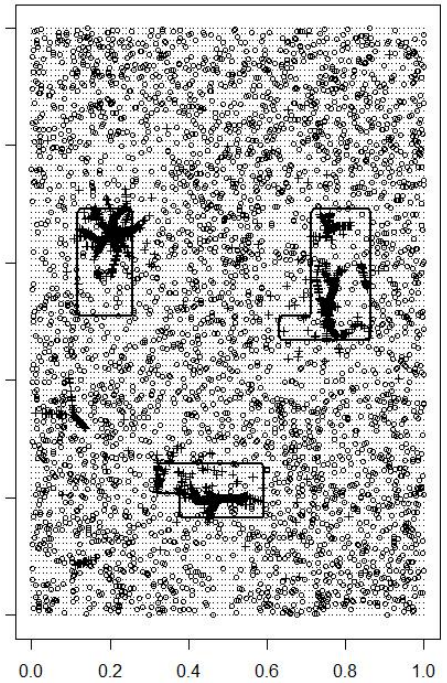
    Oversampling를 사용한 예시



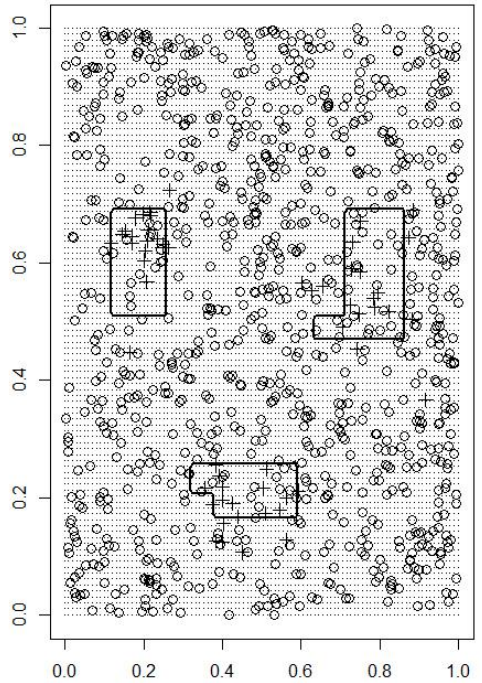
학습데이터



검증데이터
(F-measure : 0.25)



학습데이터



검증데이터
(F-measure : 0.37)

Part. 06

Class Imbalanced Problem

| Oversampling기법

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z|

강사. 이경택