



부산대학교

PUSAN NATIONAL UNIVERSITY

Imbalanced Data

When Accuracy Lies

Rare Classes and the Cost of Being Wrong

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

데이터사이언스전공

yongkyung.oh@pusan.ac.kr

When Accuracy Lies

정확도가 높아도 쓸모없는 모델일 수 있다

신용카드 거래 자료의 범주 비율



한 칸은 약 199.4건, 사기 거래는 약 1.7칸에 해당

학습 자료 199,364건 중 정상 199,025건, 사기 339건으로 사기 비율은 약 0.17퍼센트

Source: UBC CPSC 330 (2026), Lecture 9, credit-card fraud example.

모두 정상으로 예측해도 정확도는 99.83퍼센트

모든 거래를 정상이라고 답하는 기준모델(baseline)의 정확도를 먼저 계산한다

예측 / 실제	실제 사기	실제 정상	합
사기라 답함	0	0	0
정상이라 답함	339	199,025	199,364

Caution 정확도는 99.83%지만 사기 거래 재현율은 0이다

혼동행렬은 숨은 오류를 드러낸다

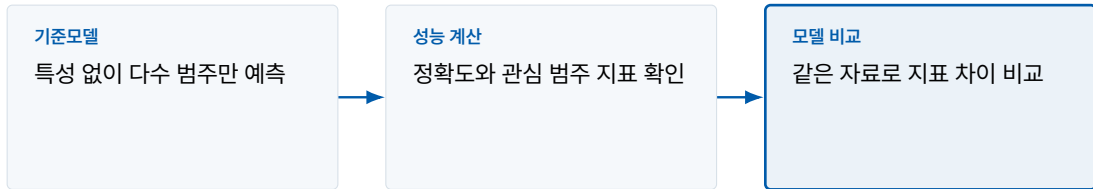
앞의 199,364건 중 30%를 떼어 낸 검증 자료 59,810건에서 얻은 로지스틱 회귀 결과

예측 / 실제	실제 사기	실제 정상	합
사기라 답함	63	8	71
정상이라 답함	39	59,700	59,739

정확도 99.92%만 보면 농친 사기 거래 39건이 보이지 않는다

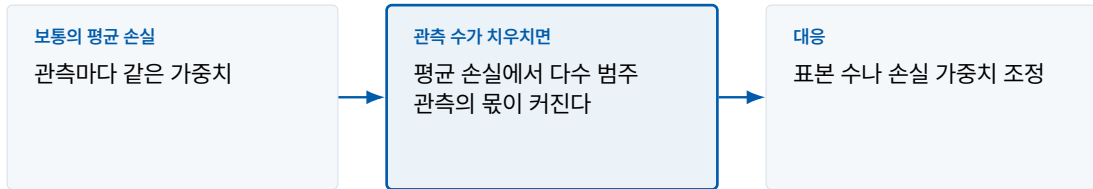
Source: UBC CPSC 330 (2026), Lecture 9, logistic-regression validation result.

같은 평가 자료에서 기준모델과 비교한다



다수 범주의 비율이 높으면 기준모델의 정확도도 높으므로 관심 오류를 별도로 확인한다

관측 수가 치우치면 다수 범주가 평균 손실을 좌우한다



Checkpoint

When Accuracy Lies

- 다수 범주로만 답하는 기준모델은 어떻게 만드는가
- 정확도가 99.83%인데 쓸모없을 수 있는 까닭은 무엇인가
- 네 칸 중 불균형에서 특히 봐야 하는 칸은 어느 것인가

Section 2 • Imbalanced Data

Metrics That Survive

치우친 자료에서도 비교 가능한 지표

균형 정확도는 범주별 재현율의 평균이다

$$\text{균형 정확도} = \frac{1}{2} \left(\underbrace{\frac{TP}{TP + FN}}_{\text{양성 재현율}} + \underbrace{\frac{TN}{TN + FP}}_{\text{음성 재현율}} \right)$$

이진분류에서 두 범주가 모두 존재할 때 사용하며, 양성을 반드시 소수 범주로 정할 필요는 없다. 사기 예에서 기준모델은 0.5, 로지스틱 회귀는 $(63/102 + 59,700/59,708)/2 \approx 0.81$ 이다

한 범주만 예측하는 모델의 균형 정확도는 0.5다

DEFINITION

정밀도(precision)

사기로 예측한 거래 중 실제 사기 거래의 비율

DEFINITION

재현율(recall)

실제 사기 거래 중 사기로 예측한 거래의 비율

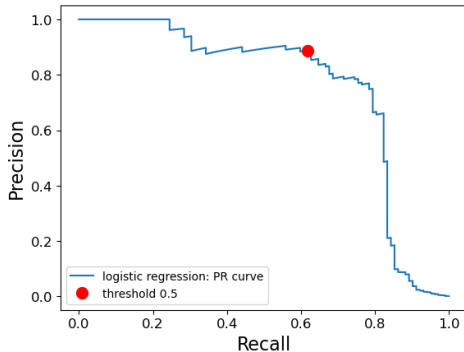
정밀도의 분모는 양성 예측 수, 재현율의 분모는 실제 양성 수이며 각각 0보다 커야 한다

PR 곡선을 계산하는 코드

PR 곡선(precision-recall curve)은 임계값을 바꾸며 (재현율, 정밀도) 점을 이은 곡선이다. 사기 범주를 1로 두고, `model.classes_[1] == 1`인 모델의 확률을 사용한다

```
1 from sklearn.metrics import precision_recall_curve
2
3 score = model.predict_proba(X_valid)[:, 1]
4 precision, recall, threshold = precision_recall_curve(
5     y_valid, score
6 )
```

PR 곡선은 드문 양성을 찾는 대가를 보여 준다



앞의 사기 검증 자료에서 로지스틱 회귀의 PR 곡선이다. 빨간 점은 임계값 0.5(재현율 0.62, 정밀도 0.89)이며, 재현율을 0.8 넘게 올리면 정밀도가 급격히 떨어진다. 무작위 순위라면 곡선 높이는 양성 비율(약 0.002)에 머문다

ROC와 PR 곡선의 차이

구분	ROC 곡선	PR 곡선
가로축	FPR	재현율
세로축	재현율	정밀도
중요한 분모	각 실제 범주의 관측 수	양성 예측 수와 실제 양성 수
함께 확인할 것	FP의 실제 건수	평가 자료의 양성 비율

앞의 로지스틱 회귀에서 FPR은 $8/59,708 \approx 0.0001$ 로 작지만 정밀도는 $63/71 \approx 0.89$ 다. 무작위 순위의 PR 곡선 높이는 양성 비율(약 0.002)이다

관심 오류와 범주 비율에 맞춰 곡선을 읽고 혼동행렬도 함께 확인한다

평가 목적에 따라 우선 볼 지표

평가 목적	우선 볼 지표	함께 확인할 것
놓친 양성 줄이기	재현율	정밀도 또는 FP 건수
잘못된 양성 예측의 비율 줄이기	정밀도	재현율 또는 놓친 수
두 범주를 같은 비중으로 평가	균형 정확도	혼동행렬

정밀도와 재현율을 F_1 로 요약한다

모델	정밀도	재현율	F_1
A	0.90	0.10	0.18
B	0.50	0.50	0.50
C	0.60	0.45	0.51

$F_1 = 2PR/(P + R)$, 정밀도 P 와 재현율 R 의 조화평균

F_1 은 두 값을 하나로 요약하지만 서로 다른 오류 비용까지 정해 주지는 않는다

Checkpoint

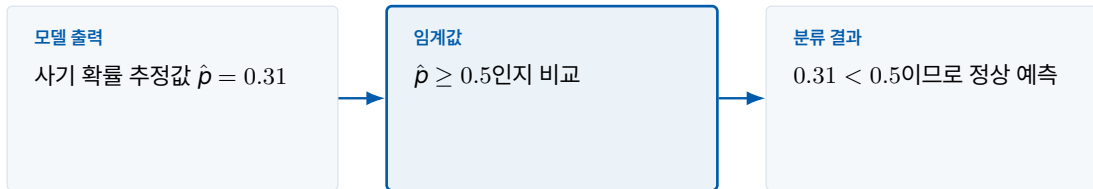
Metrics That Survive

- 균형 정확도는 무엇의 평균인가
- 임계값을 바꾸면 정밀도와 재현율을 함께 확인해야 하는 까닭은 무엇인가
- 드문 범주가 관심일 때 PR 곡선이 직접 드러내는 값은 무엇인가

Moving the Threshold

같은 모델의 분류 임계값을 바꾼다

확률 추정값에 임계값을 적용한다



임계값 0.5가 모든 평가 목적에 가장 좋은 것은 아니다

임계값에 따른 검출과 오류

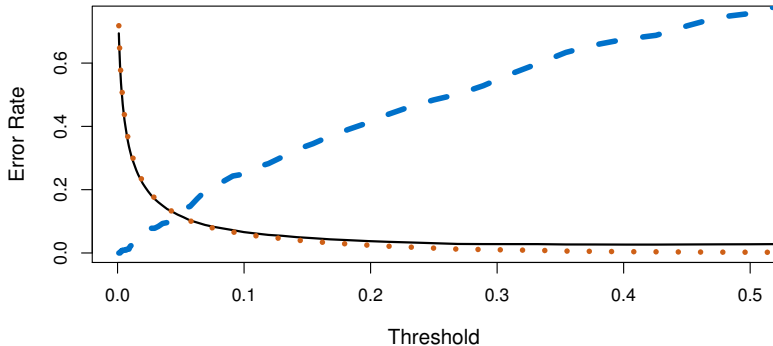
가상 거래 100건 중 사기 6건인 별도 예시

임계값	<i>TP</i>	<i>FP</i>	재현율
0.50	1	2	0.17
0.20	3	12	0.50
0.05	5	40	0.83

같은 점수에서 임계값을 낮추면 TP와 FP는 늘거나 같으며, 이 예에서는 둘 다 증가한다. 앞의 사기 검증 자료에서는 임계값 0.5에서 재현율 0.62·정밀도 0.89, 0.1에서 0.75·0.77이다

Source: UBC CPSC 330 (2026), Lecture 9, validation classification report.

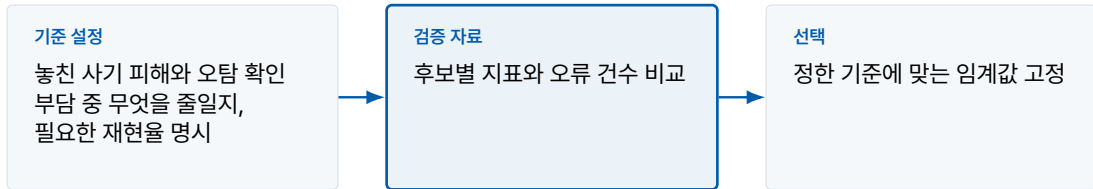
두 오류율이 엇갈린다



카드 채무 불이행(양성) 자료인 Default에서 파랑 점선은 FN 비율, 주황 점선은 FP 비율, 검정 실선은 전체 오류율

Source: James et al. (2023), Figure 4.7.

평가 기준을 정한 뒤 임계값을 선택한다



Caution FP 비율과 FN 비율이 같아지는 지점이 반드시 최적의 임계값은 아니다

확률 추정과 임계값 선택

단계	정하는 것
모델 학습	입력에 대해 사기 확률을 추정하는 규칙
임계값 선택	어느 확률부터 사기로 분류할지 정하는 기준

임계값은 검증 자료에서 선택하고, 최종 테스트 평가 전에는 고정한다

학습·검증·테스트 자료의 역할

자료	수행하는 작업	주의할 점
학습	모델의 매개변수 학습	학습 결과만으로 최종 성능을 판단하지 않음
검증	임계값 선택	선택에 사용했다는 점을 명시
테스트	고정한 모델과 임계값 평가	점수를 보고 임계값을 바꾸지 않음

Caution 임계값을 고른 자료의 점수를 독립적인 최종 성능처럼 보고하지 않는다

Checkpoint

Moving the Threshold

- 임계값을 내리면 재현율과 FP 건수는 각각 어떻게 되는가
- 임계값을 고르기 전에 어떤 평가 기준을 정해야 하는가
- 임계값은 어느 자료에서 고르는가

Section 4 • Imbalanced Data

Changing Data and Training

학습 자료와 손실을 바꾸는 방법

정확도는 범주별 재현율을 관측 수로 가중한다

$$\text{정확도} = \frac{N_+}{N} \frac{TP}{N_+} + \frac{N_-}{N} \frac{TN}{N_-}$$

N_+ 는 실제 양성 수, N_- 는 실제 음성 수이며 $N = N_+ + N_-$. 두 가중치를 모두 1/2로 두면 균형 정확도다

소수 범주의 관측 비중이 작으면 그 범주의 오류가 전체 정확도에 작게 반영된다

같은 검증 자료에서 비교한다

모델	FP	FN	정확도
모두 정상 예측	0	102	99.83%
로지스틱 회귀	8	39	99.92%

앞의 검증 자료: 실제 사기 102건, 정상 59,708건. 기준모델 수치는 이 구성에서 계산

Source: UBC CPSC 330 (2026), Lecture 9, validation confusion matrix.

두 오류가 초래할 수 있는 비용

오류	가능한 비용의 예
FN: 사기를 정상으로 예측	사기 거래를 발견하지 못해 발생하는 피해
FP: 정상을 사기로 예측	추가 확인이나 거래 제한에 따른 불편

실제 비용과 대응은 업무 절차에 따라 다르므로 문제에 맞게 기준을 정한다

지표·임계값·학습을 바꾸면 달라지는 것

변경 대상	바뀌는 것
평가지표	예측 결과를 요약하고 비교하는 기준
임계값	고정된 점수를 양성·음성으로 나누는 기준
학습 자료 또는 손실	모델의 매개변수를 학습하는 과정

Caution 표본 수 조정이나 범주 가중치로 학습한 모델의 예측 확률은 실제 사기 비율의 추정값으로 읽지 않는다

표본 수 조정의 두 방법

가상 학습 관측 1,000개 중 소수 범주 60개인 별도 예시

방법	소수 범주	다수 범주	합
원래 자료	60	940	1,000
다수 표본 줄이기	60	60	120
소수 표본 복제하기	940	940	1,880

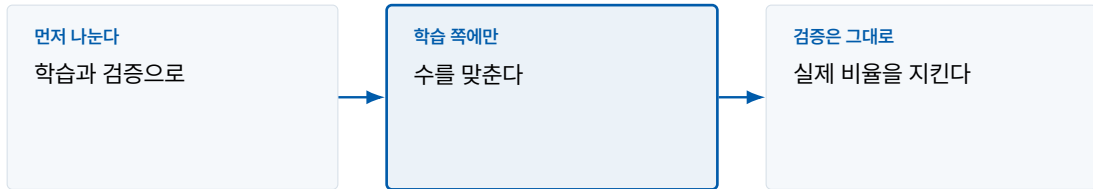
표본 수 조정(resampling)의 예로, 다수 표본 줄이기는 undersampling, 소수 표본 복제는 oversampling에 해당

표본 수를 바꿀 때의 대가

방법	주의할 점
다수 표본 줄이기	제외한 관측에 있던 유용한 정보도 잃을 수 있음
소수 표본 복제하기	같은 관측을 반복하므로 새로운 정보를 만들지 못하고, 복제된 점에 과적합되기 쉬움

학습 자료에만 적용하고 원래 범주 비율을 유지한 검증 자료에서 비교한다

표본 수 조정은 학습 자료에만 한다



Caution 검증 자료의 비율까지 바꾸면 실제로 마주할 상황과 다른 것을 재게 된다

분할할 때 범주 비율을 지키는 코드

먼저 학습·검증 자료를 나누고, 학습 자료에만 표본 수 조정을 적용한다

```
1 from sklearn.model_selection import train_test_split
2
3 X_train, X_valid, y_train, y_valid = train_test_split(
4     X, y, test_size=0.2, stratify=y, random_state=42
5 )
```

stratify=y는 두 자료의 사기 비율을 원자료와 같게 나눈다(층화 분할, stratified split)

막대 길이로 관측 수를 비교한다

원자료



$60 + 940$

다수를 줄임



$60 + 60$

소수를 복제



$940 + 940$

초록 칸은 소수 범주, 옅은 칸은 다수 범주이며 세 막대의 가로 길이에 동일한 관측 수 척도를 적용

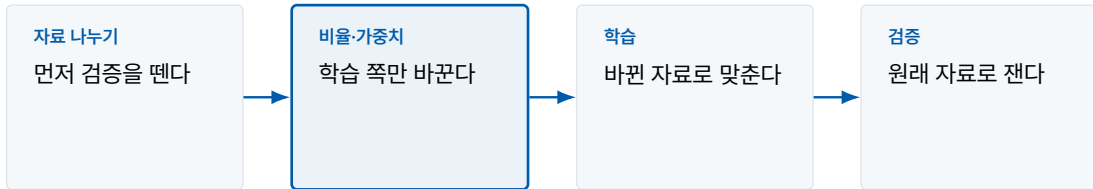
범주 가중치를 각 관측의 손실에 곱한다

$$J_w = -\frac{1}{n} \sum_{i=1}^n w_{y_i} [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)]$$

이진 로지스틱 손실의 가중 합 예시. $y_i \in \{0, 1\}$ 이며 $\hat{p}_i$ 는 양성 확률 추정값

틀린 예측에만 가중치를 주는 것이 아니라 해당 범주의 모든 관측 손실에 적용한다

분할부터 검증까지의 순서



Caution 분할 전에 표본 수를 조정하면 검증 분포가 왜곡되거나 복제 관측이 양쪽에 들어갈 수 있다

균형 가중치의 계산

$$w_c = \frac{n}{Cn_c}$$

범주	관측 수	관측당 가중치	가중치 합
소수	60	$1,000/120 \approx 8.33$	500
다수	940	$1,000/1,880 \approx 0.532$	500

C는 범주 수, n_c 는 범주 c 의 학습 관측 수. 이 가중치가 실제 업무의 오류 비용을 뜻하지는 않음

Source: scikit-learn (n.d.), LogisticRegression, class_weight documentation.

범주 가중치를 적용하는 코드

scikit-learn은 범주 빈도의 역수에 비례하는 가중치를 자동으로 계산할 수 있다

```
1 from sklearn.linear_model import LogisticRegression
2
3 model = LogisticRegression(
4     max_iter=500,
5     class_weight="balanced"
6 )
7 model.fit(X_train, y_train)
```

Adapted from UBC CPSC 330 (2026), Lecture 9, `class_weight`.

가중치 변경의 효과는 지표마다 다르다

교차검증 지표	기본 설정	균형 가중치
정확도	0.999	0.973
재현율	0.602	0.891
정밀도	0.889	0.054
F_1	0.717	0.102
ROC AUC	0.968	0.971
평균 정밀도(AP)	0.743	0.730

재현율은 오르고 정밀도· F_1 은 낮아졌지만, 임계값과 무관한 ROC AUC·AP는 거의 같다

Source: UBC CPSC 330 (2026), Lecture 9, five-fold cross-validation.

변경 전후의 지표를 함께 확인한다

확인할 것	이유
혼동행렬	FP와 FN의 실제 증감을 확인
관심 범주의 재현율	실제 양성을 얼마나 찾아냈는지 확인
정밀도	양성 예측 중 실제 양성의 비율 확인
기준모델과의 차이	학습한 모델의 개선 폭을 같은 자료에서 비교

불균형 대응만으로 해결하기 어려운 경우

상황	추가로 검토할 것
소수 범주 관측이 매우 적음	대표성 있는 관측을 더 확보할 수 있는가
서로 다른 범주가 같은 특성값을 가짐	추가 특성으로 구분할 수 있는가
소수 범주의 양상이 다양함	범주 정의와 하위 집단별 성능이 적절한가

표본 복제나 가중치 변경은 새로운 관측 정보를 만들어 내지 않는다

Checkpoint

Changing Data and Training

- 수를 맞추는 두 방법은 각각 무엇을 잃는가
- 표본 수 조정은 어느 자료에만 해야 하는가
- 가중치를 적용한 뒤 어떤 지표가 오르고 어떤 지표가 낮아졌는가

What If the Boundary Is Not a Line?

Next Lecture: Decision Trees

- 규칙의 연쇄가 특성 공간을 구역으로 나누는 것
- 위에서부터 분할을 고르는 절차
- 좋은 분할을 수로 재는 두 식