



Card

Machine Learning

team

project



천안형 글로벌 인재 육성 사업 초급 프로젝트

DACON 카드 연체 예측 AI 경진대회

#데이터 2조

#김동환 - 호서대학교

#김현서 - 단국대학교

#박현우 - 호서대학교

#한현민 - 상명대학교



Card

Machine Learning

team

project

Project outline

프로젝트 명

DACON 카드 연체 예측 AI 경진대회 (39회)

예측 대상

정상(0) 또는 연체(1) 여부

주제

고객의 카드 사용 데이터를 바탕으로 "연체 여부" 를 예측하는 이진 분류 모델 개발

[0
1



Card

Machine Learning

team

project

주제 선정 이유

삼성에 1위 내준 신한카드, 반년 만에 또 희망퇴직

송고 2025-06-04 17:21

신한카드의 1분기 말 연체율은 1.61%로 작년 동기(1.56%), 전 분기(1.51%) 대비 각각 0.05%p, 0.10%p 올라 2015년 3분기 말(1.68%) 이후 **최고치를 기록** 했다.

연체 : 고객이 약정된 날짜까지 신용카드 대금을 상환하지 못한 상태를 의미



Card

Machine Learning

team

project

주제 선정 이유

'서민 급전창구' 카드론, 높아진 문턱에 '그림의 떡'

1분기 말 기준 저신용자 금리 19%에도 "건전성·수익성 문제에 대출 어려워"

김정후 기자 | kjh@newsprime.co.kr | 2025.06.05 16:37:28

'서민 급전창구' 카드론의 대출 문턱이 낮아질 기미가 보이지 않는다. 연체율이 10년 만에 최고치를 경신하는 가운데, 금융당국도 관리를 요구하고 있어서다. 업계 안팎에서는 경기가 회복되기 전까지 이런 현상은 지속될 것이라고 전망하면서 서민들의 한숨은 더욱 깊어갈 전망이다.



web

programming

team

project

왜 연체 예측이 중요한가?

손실 예방 목적



연체 고객의 부실 채권



대손충당금 증가



수익 감소

리스크 관리



예측 모델을 통한 고위험 고객 선별



조기 대응 가능

✓ 1. 738개의 열을 어떻게 뽑아내서 붙일 것인가?

```
1 for i in train_df.columns:  
2     print(i)
```

	청구서 발송여 부_R3M	청구금 액_R3M	포인트_ 마일리 지_건별 _R3M	포인트_ 건별 _R3M	포인트_ 마일리지 _월적립 _R3M	포인트_ 월적립 _R3M	포인트_ 적립포 인트 _R3M	포인트_ 이용포 인트 _R3M	마일_ 적립포 인트 _R3M	마일_ 이용포 인트 _R3M	...
0	1	1678	0	0	0	0	0	0	0	0	...
1	1	7778	0	0	0	0	0	0	0	0	...
2	0	0	0	0	0	0	0	0	0	0	...
3	1	15957	0	3743	0	0	3743	2285	0	0	...
4	1	2661	0	0	0	0	0	0	0	0	...

5 rows × 88 columns

어떤 칼럼을 정할까?

연체율 3개월 데이터
연체율 6개월 데이터
연체율 금액

어떤 데이터를 정해야 할까?
→ 편차가 있어서
이거를 다 합해서 결과를 내야할까?

변수 선정

변수명	예상 기여 이유
상환개월수_결제일_R6M	상환 기간이 길면 연체 위험 증가 가능성
선결제건수_R6M	선결제 많이 하면 연체 가능성 낮음
카드이용한도금액	한도가 낮거나 과도한 사용은 연체 유발 가능성 있음
연체건수_R3M	과거 연체 이력은 미래 연체 위험과 강한 연관
포인트_이용포인트_R12M	혜택만 쓰고 상황에 관심 없는 패턴일 수도 있음
할인금액_R3M	지나친 소비 혜택 의존은 연체 리스크로 이어질 수 있음
이용금액_카드론_R3M	고금리 단기대출 사용자는 상황능력 불안정 가능성
RV최소결제비율	최소결제만 하고 있는 경우 연체 위험이 높을 수 있음
한도증액금액_R12M	자주 한도 상향한 고객은 소비 과다일 수 있음



Card

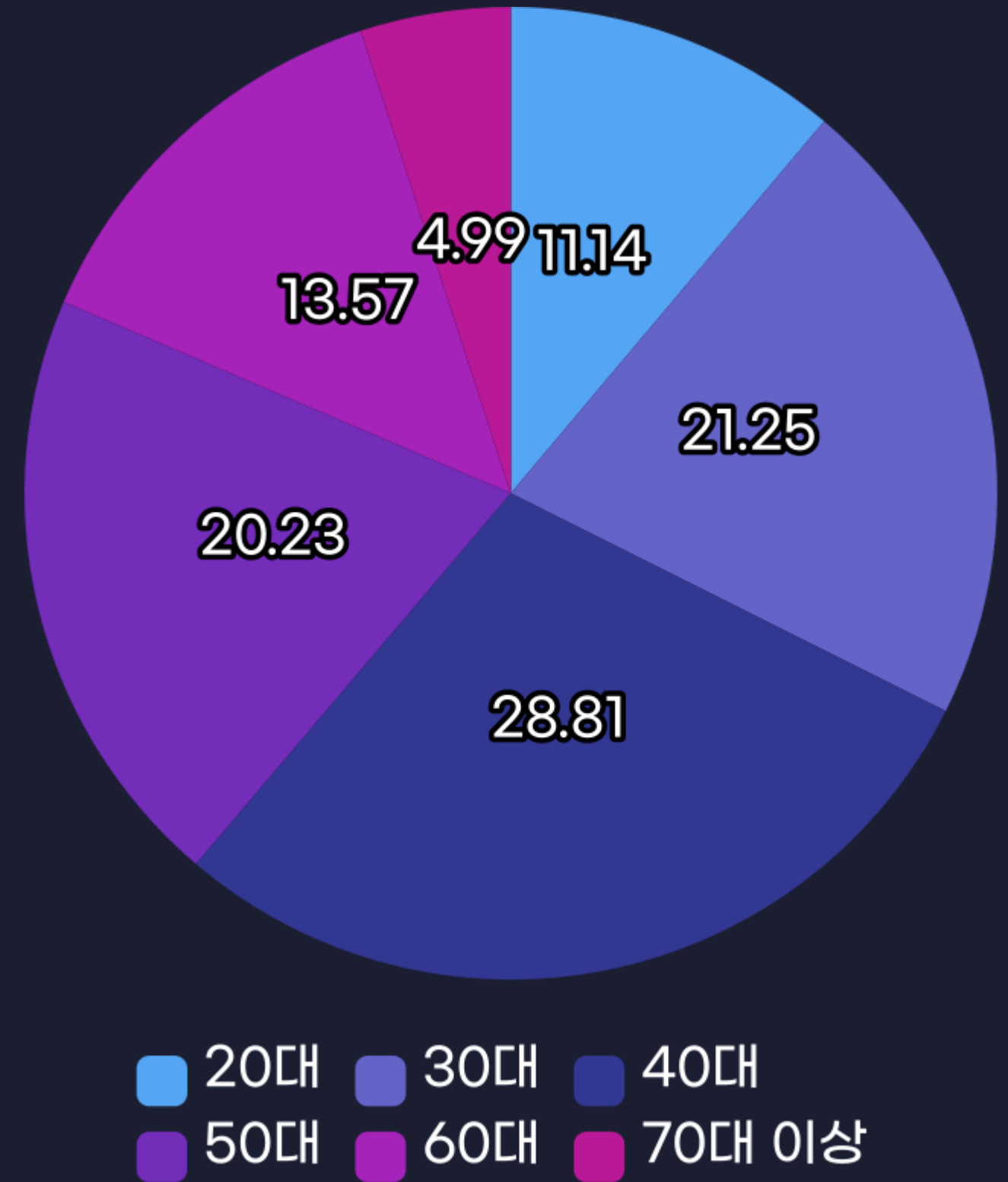
Machine Learning

team

project

연령대별 데이터 분포

총 데이터 개수는 781개로
40대가 28.81%로 가장 많고,
70대 이상이 4.99%로 가장 적음.





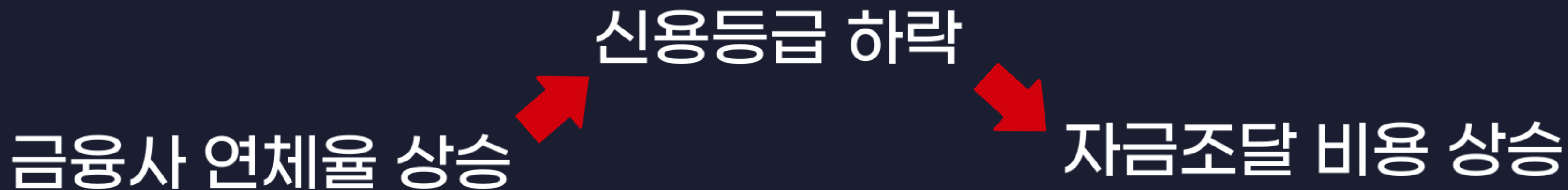
Card

Machine Learning

team

project

왜 연체 예측이 중요한가?



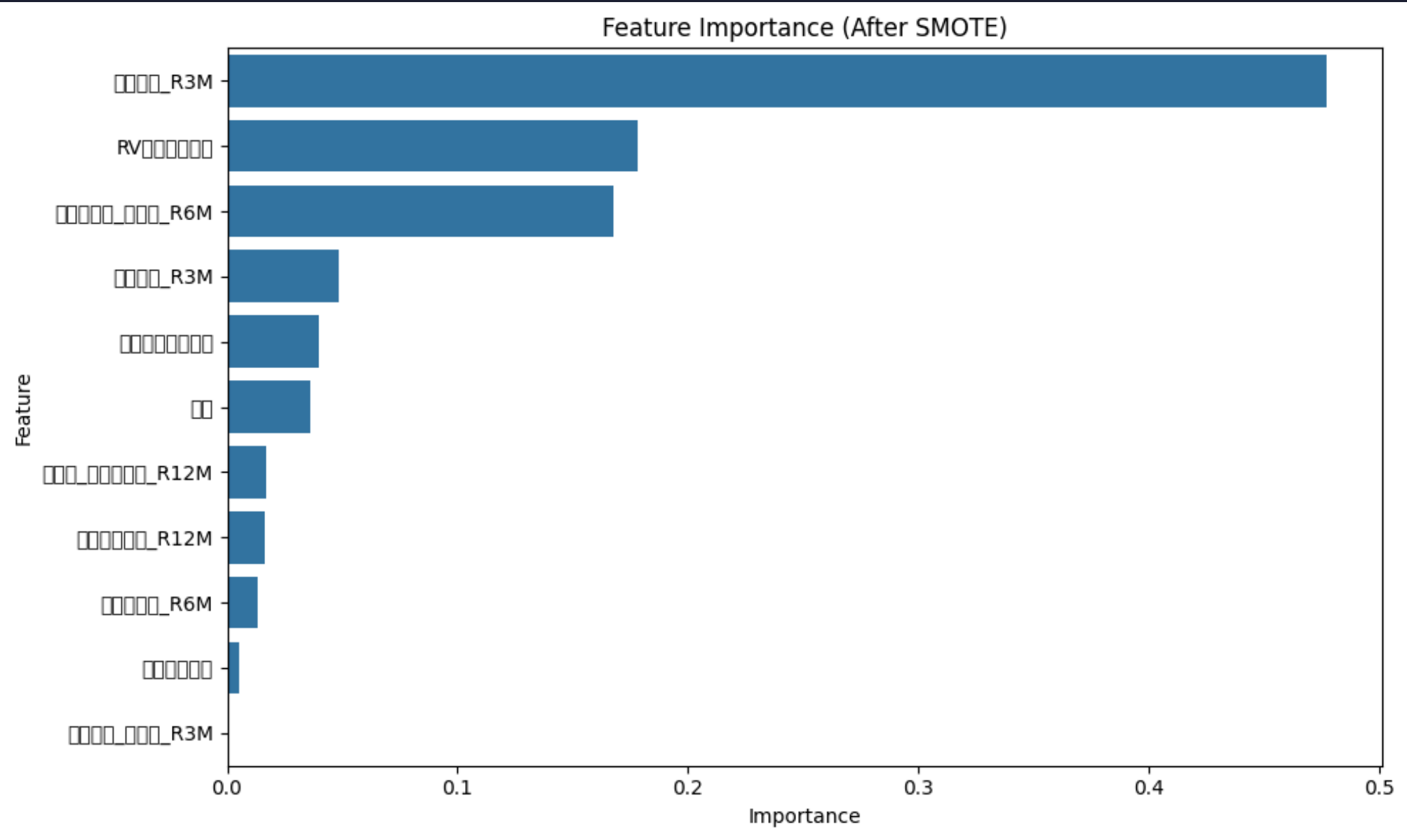
과정

1. 범주형 변수 인코딩 (One-Hot Encoding)
2. 로지스틱 회귀로 모델 학습 및 평가
3. 변수 중요도 확인 > 결과는 정밀도, 재현율, F1 매우 낮음
4. 개선된 불균형 처리 + XGBoost 모델 코드 > 이것도 좀 높아
지긴 했지만 아직 낮음 (24%정도)

20	113
30	341
40	444
50	335
60	160
70	47

문제: 연체율 데이터가 평균에 비해 부족

데이터 샘플링이 필요함





Card

Machine Learning

team

project

데이터 문제점 및 해결

문제 발견

연체(1) 클래스 데이터 부족
으로 불균형 발생



Card

Machine Learning

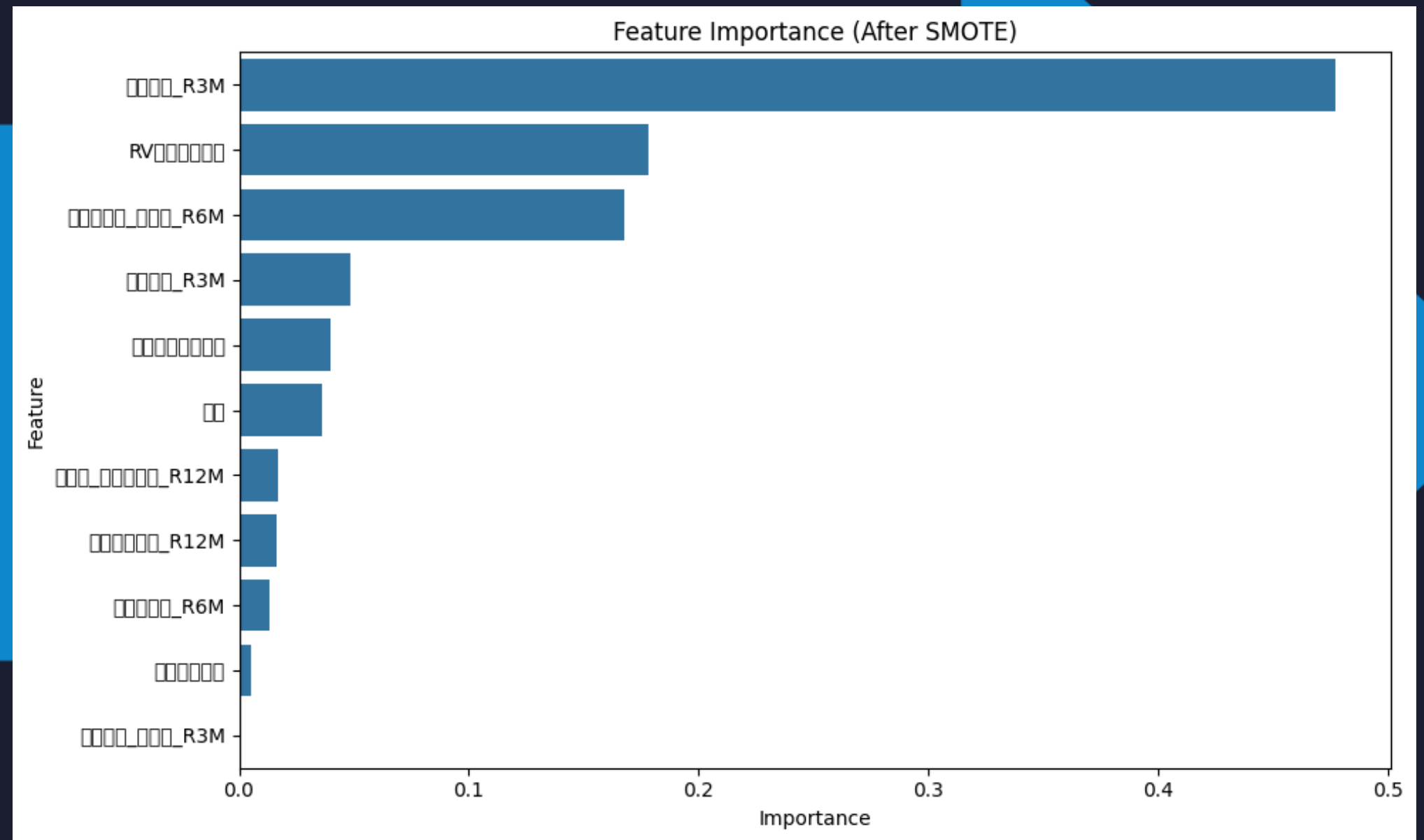
team

project

데이터 문제점 및 해결

문제 발견

연체(1) 클래스 데이터 부족
으로 불균형 발생





Card

Machine Learning

team

project

데이터 문제점 및 해결

문제 발견

연체(1) 클래스 데이터 부족
으로 불균형 발생

해결 방안

SMOTE 기법 적용

결과

소수 클래스 데이터 합성으
로 균형 있는 학습 가능



Card

Machine Learning

team

project

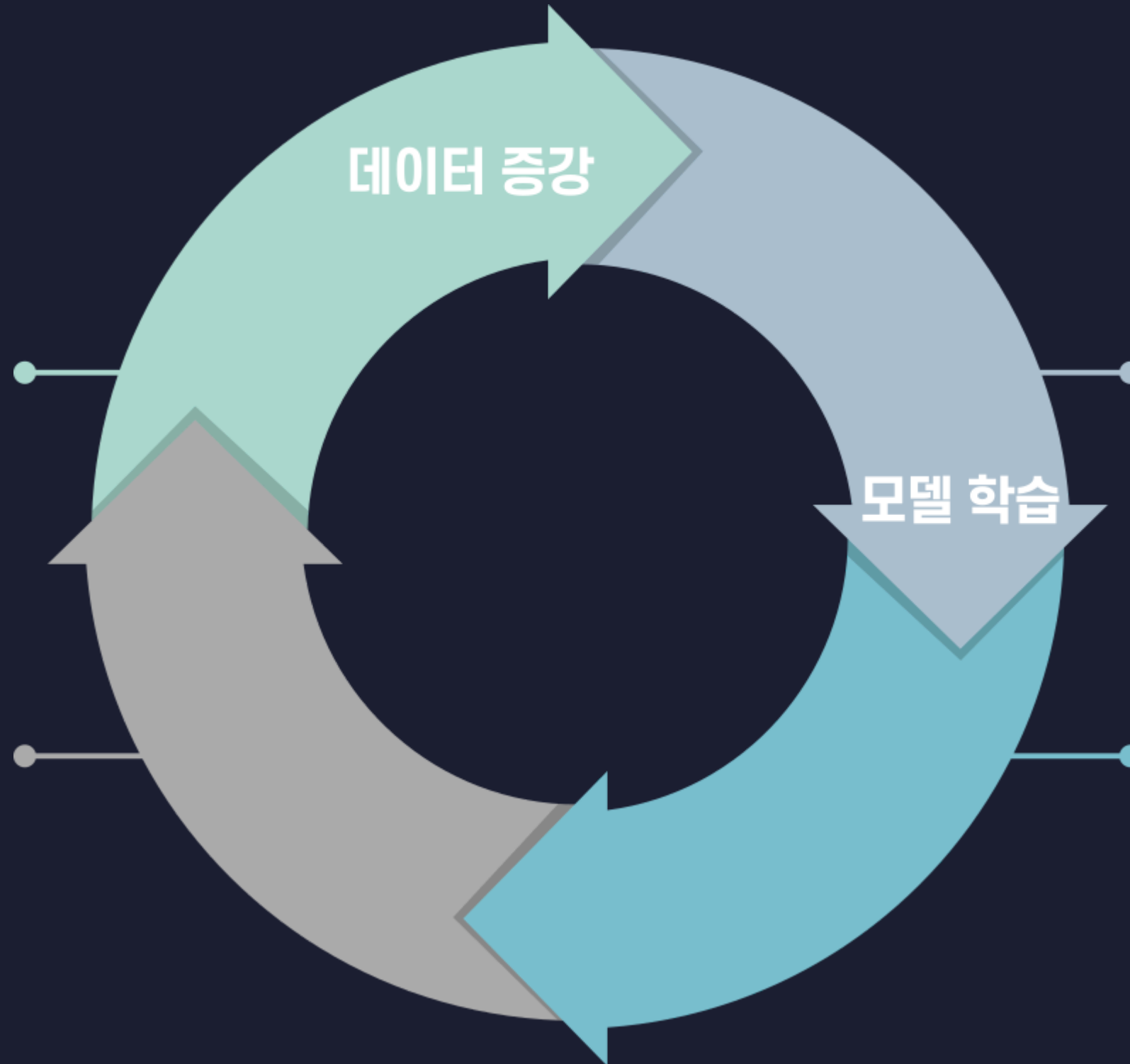
모델링 및 평가 지표

SMOTE로 연체
클래스 샘플 증가

데이터 증강

모델 학습

로지스틱 회귀, 랜덤
포레스트 등 적용





Card

Machine Learning

team

project

평가 지표 설명



정밀도(Precision)

예측이 맞다고 한 것 중 실제로 맞은 비율

F1-score

Precision과 Recall의 조화 평균

(불균형 데이터에 적합)

재현율(Recall)

실제 연체 중 모델이 맞춘 비율

Support

각 클래스별 실제 샘플 개수

9
10
11
12
13



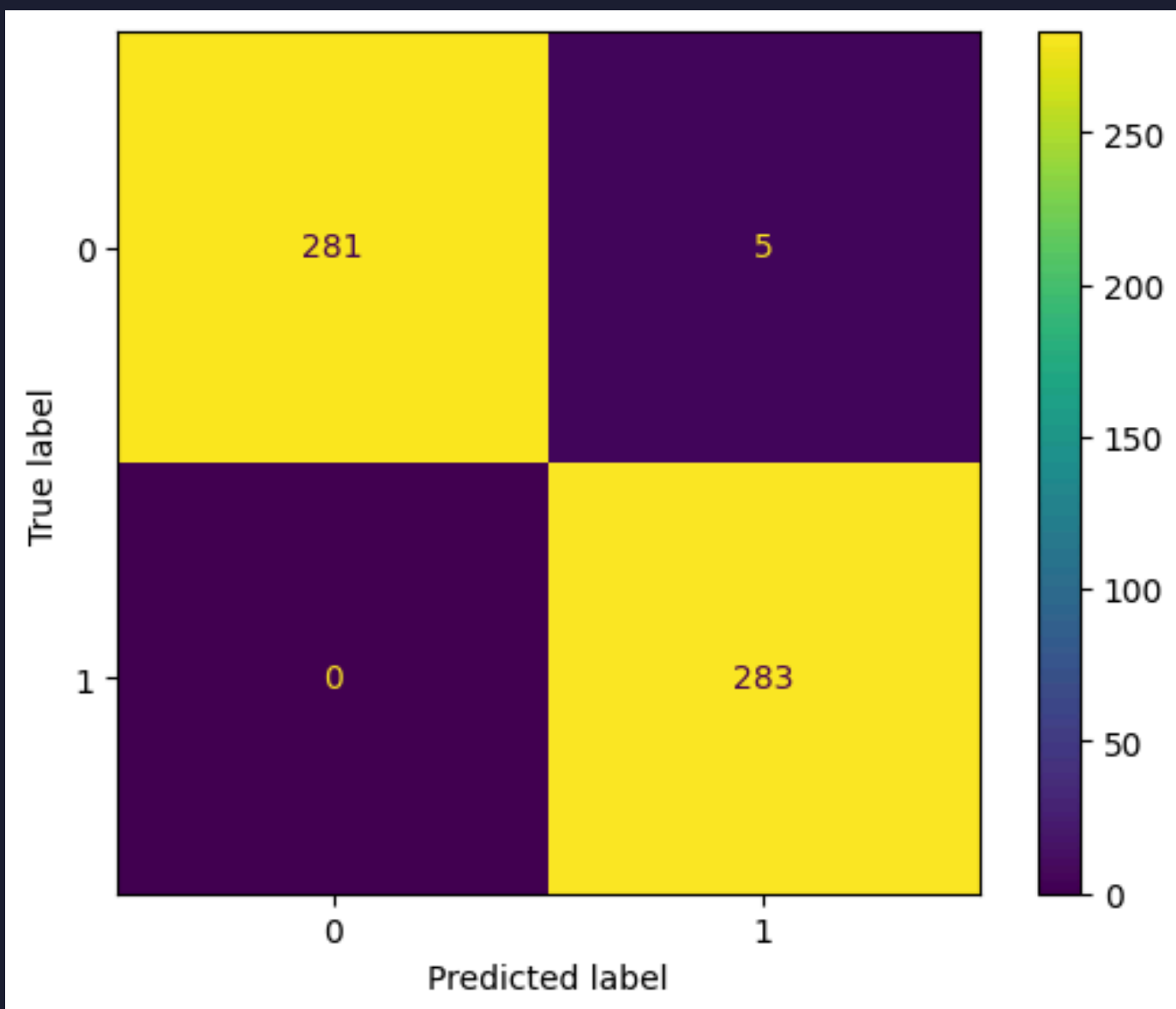
Card

Machine Learning

team

project

모델 성능 예시



Class	Precision	Recall	F1-score	Support
0 (정상)	1.00	0.98	0.99	286
1 (연체)	0.98	1.00	0.99	283

전체 예측 정확도(Accuracy)는 0.99입니다.
두 클래스 모두에서 높은 성능을 달성했습니다.

25
26



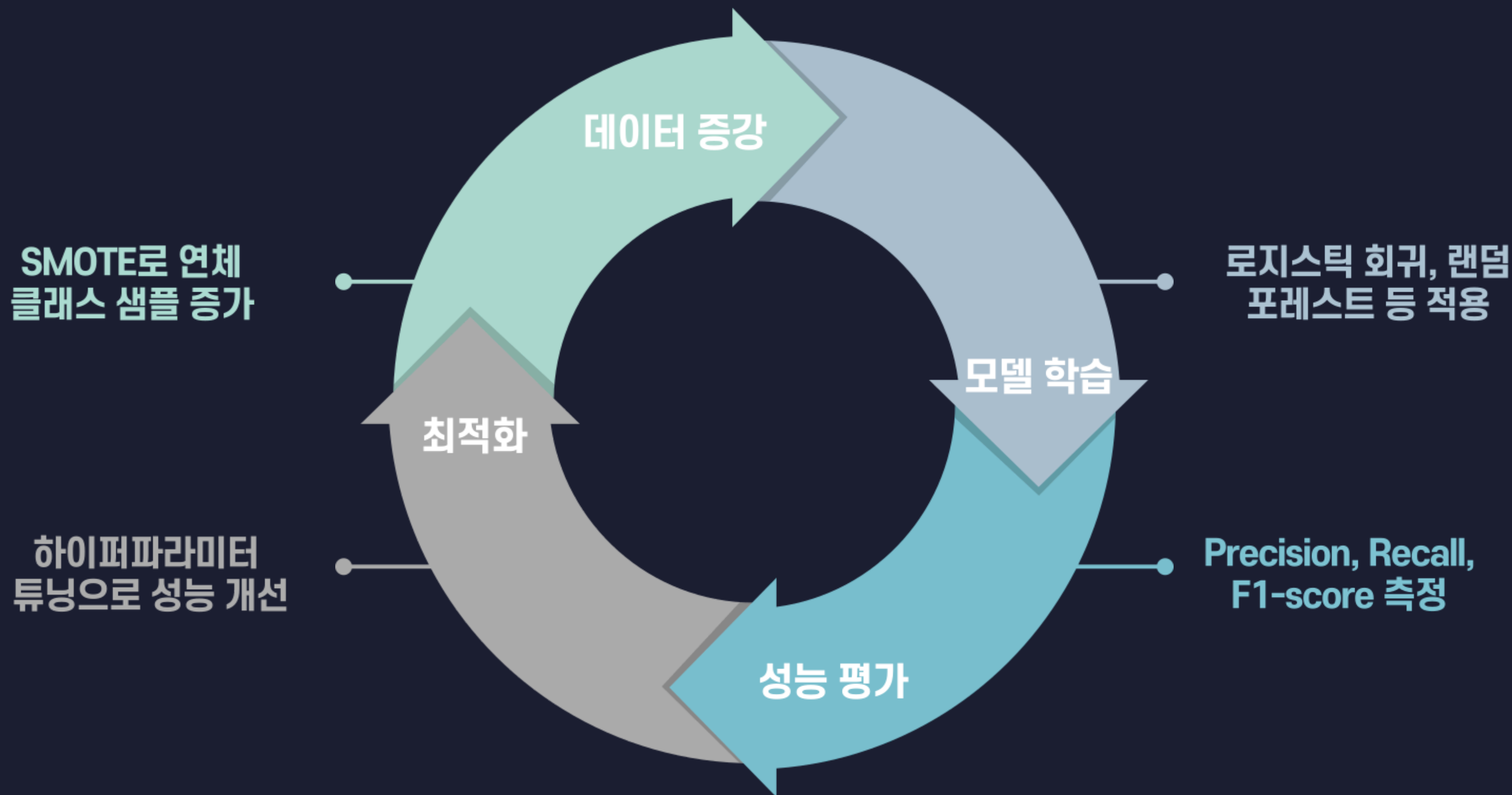
Card

Machine Learning

team

project

모델링 및 평가 지표





Card

Machine Learning

team

project

성별 연체율 분석

여성



평균 대출 : 6377만원

연체율 : 0.57%

남성



평균 대출: 3,717만원

연체율: 0.40%

여성의 연체율이 남성보다 0.17% 낮음.

국내외 연구에서도 여성 대출자의 연체율이 남성보다 낮음이 확인되었습니다.



Card

Machine Learning

team

project

성별 연체 특성 연구

연구 결과

동일 신용 점수에서 여성의 연체율이 남성
보다 약 15% 낮음(스페인 은행 데이터)

모델 시사점

성별은 연체 예측 모델에서 중요한 변수로 활용 가능

행동 특성

여성은 신용 정책 준수율이 높고
위험 관리에 더 보수적인 경향

