

사무소 법인카드 이상거래 탐지 모델 개선 방안

디지털전략부 R&D센터 AI파트

인턴 김범수

Contents

I. 추진배경

- i) 현행 프로세스
- ii) 문제점 및 추진방향

II. 초도모델 분석

- i) 데이터 레이아웃
- ii) 분석 결과
- iii) 데이터 품질 향상

III. 성능 고도화

- i) EDA(탐색적 자료 분석)
- ii) 파생변수 생성 예시
- iii) 파생변수 현황
- iv) AI 기반 모델링 결과

IV. 결 론

- i) 개인 기여도 및 소감

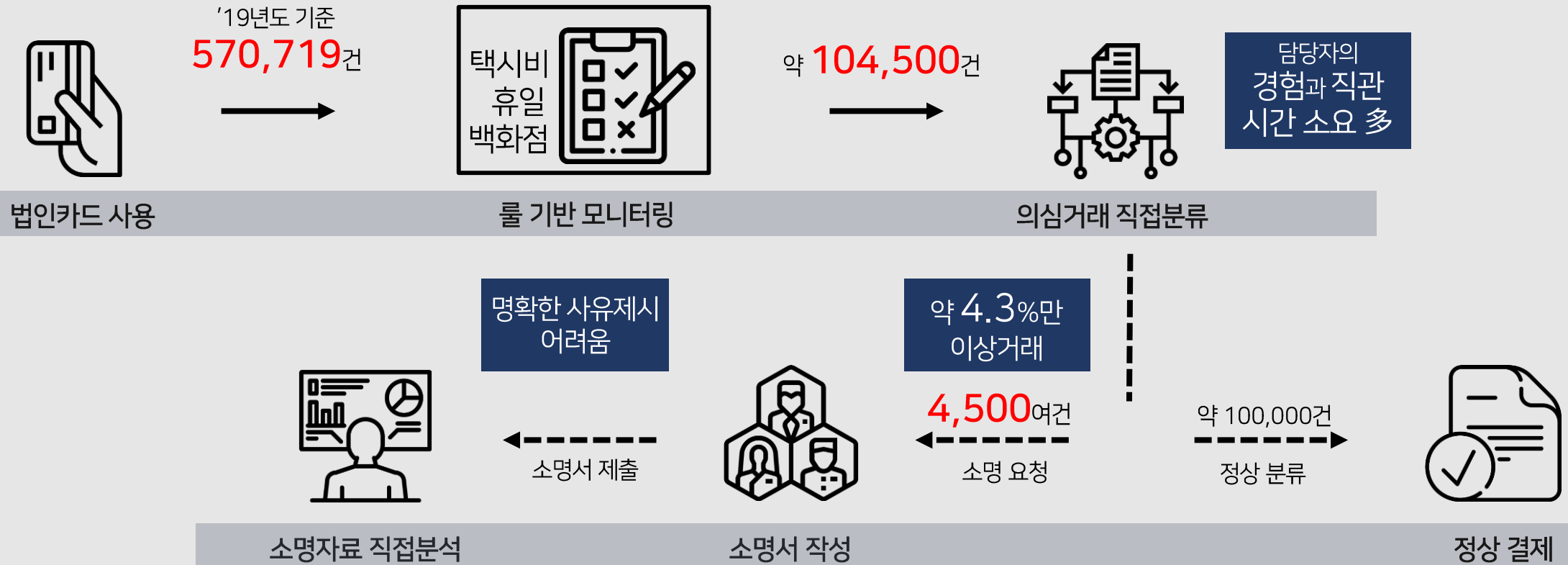
I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

i) 현행 프로세스



I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

ii) 문제점 및 추진방향

PoC 추진방향 (현재 진행중)

現 시스템

- ✓ 현 프로세스는 담당자에게 업무 과다(부담)
- ✓ 담당자의 경험과 직관에 의존하여 소명요청
→ 명확한 사유를 제시하기 어려움
- ✓ 새로운 유형의 이상거래 기존 시스템 적용 불가

PoC

- ✓ 기존 거래 데이터를 Random Forest, Cat Boost, LightGBM 등 AI 기반 학습모델을 통해 분석하여 이상거래 탐지 시스템 고도화
- ✓ 다양한 분석 지표를 활용하여 객관적인 성능 평가
- ✓ 향후 은행 내 FDS(이상금융거래 탐지 시스템)의 확대 가능성을 점검



개인 추진방향

- ✓ 데이터 품질 고도화를 통해 향후 확장성을 증가
- ✓ 데이터 분석을 통한 이상거래 탐지 시스템 고도화에 기여

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

i) 데이터 레이어아웃

| 변수명 | 데이터 예시 | 종 류 | 특이사항 |
|------------|---------------------|------------|-----------------|
| Label_소명여부 | 여(이상 거래) / 부(정상 거래) | 2 개 | . |
| 개인id | kkikkk9b | 5,057 개 | 데이터 비식별화 |
| 사무소코드 | 2066 | 1,202 개 | . |
| 사무소명 | 000지점 | 1,200 지점 | . |
| 소명대상_구분 | 마트, 가전, 가구 | 15 개 | . |
| 사용일 | 20180605 | 1,324 일 | . |
| 결제요일 | 월 ~ 일 | 7 개 | . |
| 시간 | 000000 ~ 235959 | . | . |
| 가맹점명 | (주)스타벅스커피코리아 | 219,364 지점 | . |
| 업종코드 | 2104 | 330 개 | . |
| 업종명 | 외식유흥관련가맹점 | 198 개 | . |
| 가맹점세부업종명 | 양식 | 278 개 | . |
| 금액 | 0 ~ 30,450,624,560 | . | 0원, 백억단위의 데이터 有 |
| 가맹점 주소 | 서울특별시 강남구 봉은사로... | 266,976개 | 주소 이상 |
| 사무소 주소 | 서울 서대문구 미군동21-1 ... | 1,176개 | . |

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

ii) 데이터 품질 이상

결제 가맹점 주소 이상

데이터 품질 확인 결과 **주소를 입력할 때 에러가 발생**하는 경우가 전체의 **41%**

→ 약 270,000개 데이터 중 110,000개

| 띄어쓰기 오기입 | 일부 주소만 입력 | 숫자만 입력 |
|--|-------------------|--------------|
| 서울특별시 강서구 공항대로 269-15120호 (마곡동,힐스테이트에코마곡) | 603-26 동원원룸a동 | 278-14 G-101 |
| 세종특별자치시 만남로 1461층 103,104,105호 (고운동,삼성프라자) | 4-2백양상가 | 108 |
| 서울특별시 서대문구 통일로9안길 36-41층 | 1193번지 대주 상가 102호 | 2992-1 13/5 |
| 서울특별시 종로구 새문안로5길 111층 (당주동) | 546-23 A동1층 | 427-10 |

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

ii) 데이터 품질 향상

[외부 API를 활용한 데이터 품질을 개선]

As-Is

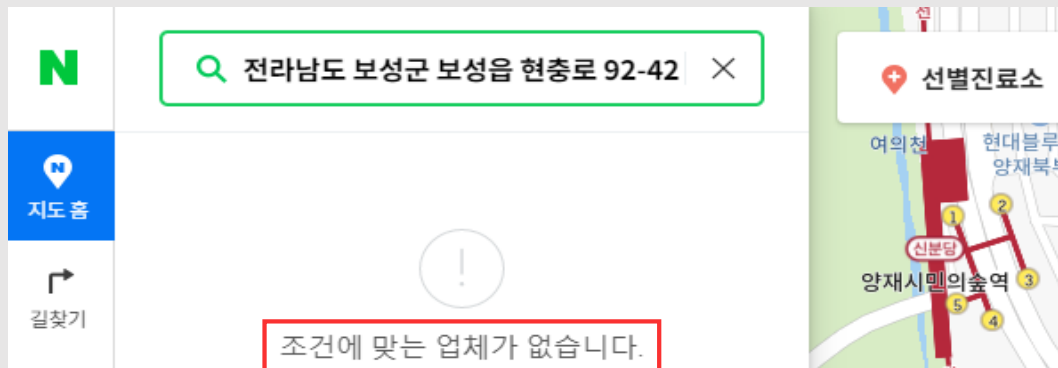
비정확한 데이터 입력시 에러 발생

전라남도 보성군 보성읍 현충로 92-42

경기 안산시 단원구 원시동동산로 76 지하110호

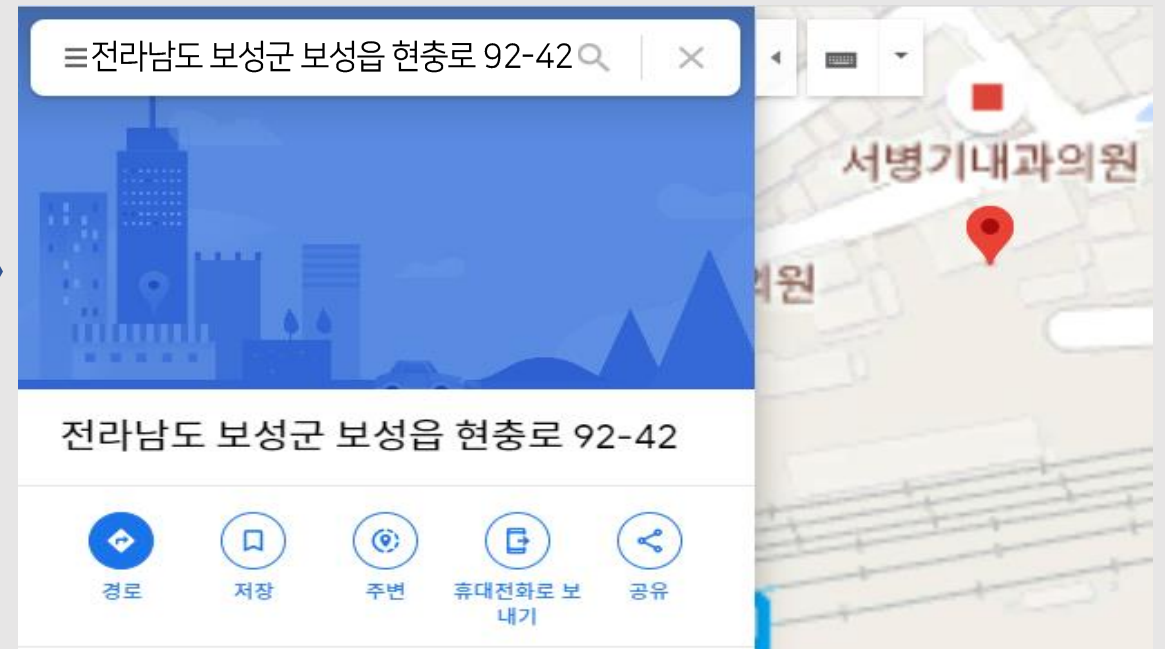
경기 안산시 단원구 목내동광양프론티어1동121호

서울특별시 강서구 마곡중앙6로 66108호, 109호



To-Be

Google Geocoding의 경우 가장 일치하는 위치를 반환



I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

ii) 데이터 품질 향상

[개인 아이디어를 통한 데이터 품질 개선]

As-Is

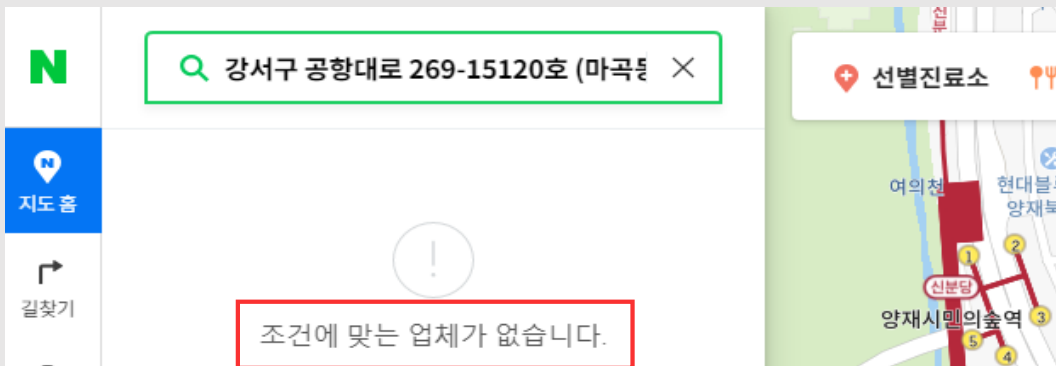
띄어쓰기 오기입시 에러 발생

서울특별시 강서구 공항대로 269-15

서울특별시 서대문구 통일로9안길 36-41층

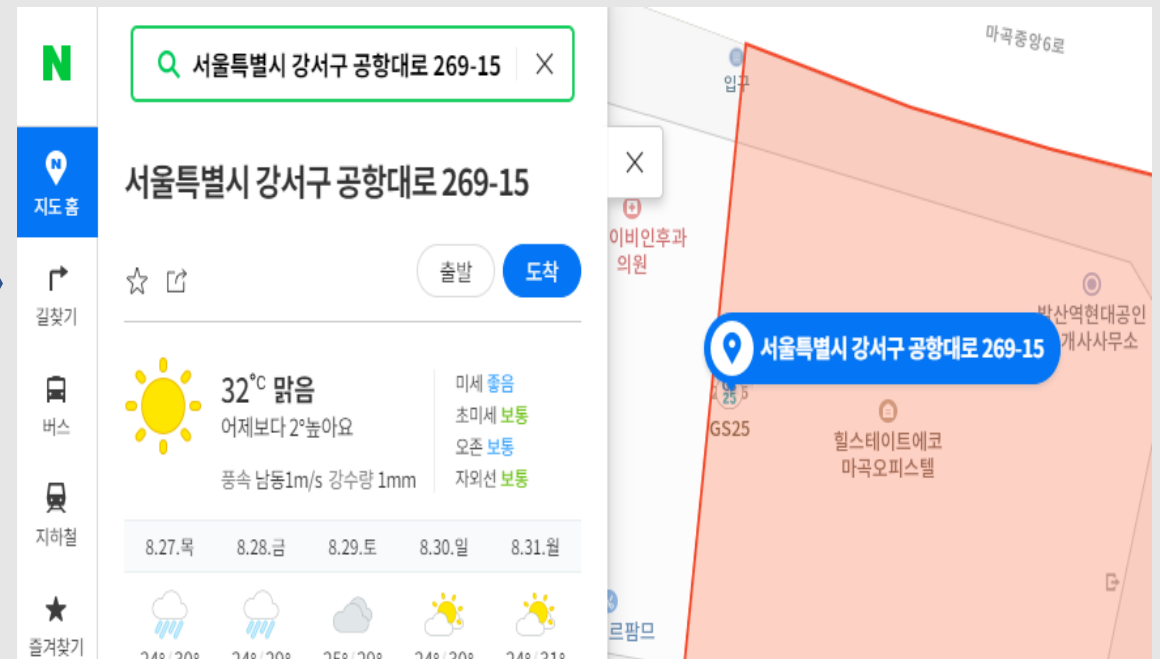
서울특별시 종로구 새문안로5길 111층 (당주동)

세종특별자치시 만남로 1461층 103,104,105호 (고운동,삼성프라자)



To-Be

주소의 일정부분을 제거하여 정확한 주소를 반환



I. 추진 배경

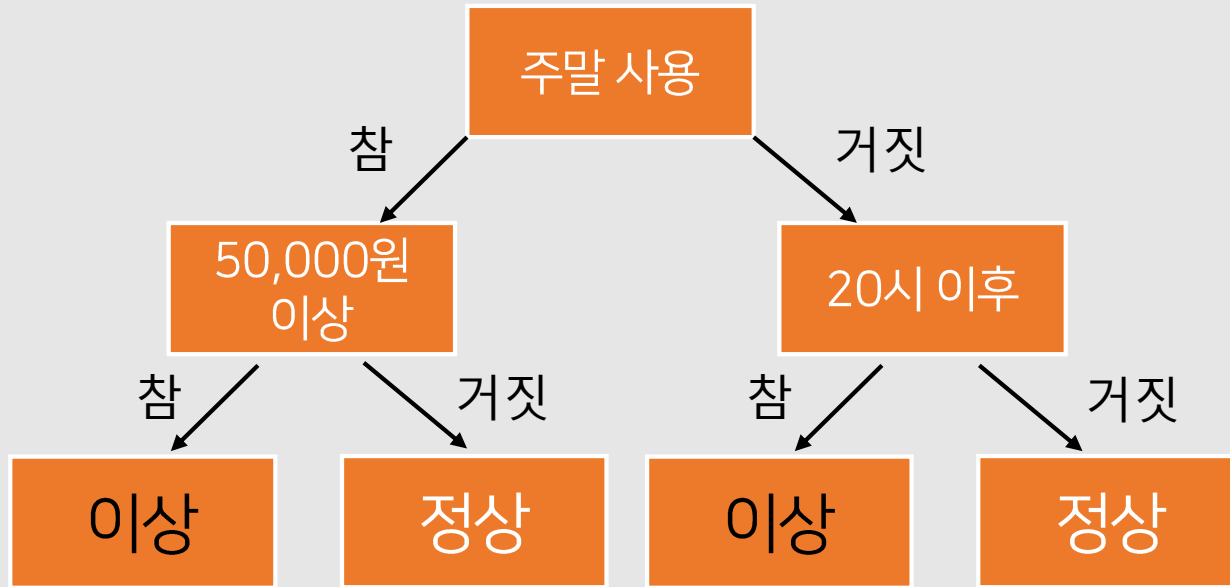
II. 초도모델 분석

III. 성능 고도화

IV. 결 론

iii) 분석 결과

Decision Tree (의사결정 나무)



사용모델 : Decision Tree
총 데이터 : 1,924,521개

초도모델 분석 결과(Raw Data)

| | | |
|-----|--------------------|-------|
| 주지표 | 재현율 (Recall) | 8.5% |
| 부지표 | 정밀도 (Precision) | 58.2% |
| | F1-Score | 14.9% |
| | 정확도 (Accuracy) | 99.3% |

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

iii) 분석 결과

정확도(Accuracy) : $\frac{\text{맞춘(정답) 결과}}{\text{전체 데이터}}$

테스트 데이터(전체) :

577,357개

맞춘(정답) 결과 :

573,070개

| | | 시스템 예측 결과 | |
|------------|------|-------------|---------|
| | | 정상거래 | 이상거래 |
| 실 제 결 과 | 정상거래 | 572,696(TN) | 269(FP) |
| | 이상거래 | 4,018(FN) | 374(TP) |

초도모델 분석 결과(Raw Data)

| 주지표 | 재현율 (Recall) | 8.5% |
|-----|--------------------|-------|
| 부지표 | 정밀도 (Precision) | 58.2% |
| | F1-Score | 14.9% |
| | 정확도 (Accuracy) | 99.3% |

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

iii) 분석 결과

대출 정상거래(1억)의 수익(1%) = 100만원
부도의 경우(이상거래) 발생하는 손해 = 1억

$$\text{정확도} : \frac{98(\text{TN}) + 1(\text{TP})}{100(\text{All Data})} = 99\%$$

| | | 시스템 예측 결과 | |
|------------|--------|-----------|--------|
| | | 정상 | 부도(이상) |
| 실 제 결 과 | 정상 | 98(TN) | 0(FP) |
| | 부도(이상) | 1(FN) | 1(TP) |

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

iii) 분석 결과

$$\text{재현율(Recall)} : \frac{TP}{TP + FN}$$

| 총 데이터 : 1,924,521개 학습용 : 1,347,164개 테스트용 : 577,357개 | | 시스템 예측 결과 | |
|---|------|-------------|---------|
| | | 정상거래 | 이상거래 |
| | | 572,696(TN) | 269(FP) |
| 실 제 결 과 | 정상거래 | 572,696(TN) | 269(FP) |
| | 이상거래 | 4,018(FN) | 374(TP) |

초도모델 분석 결과(Raw Data)

| 주지표 | 재현율 (Recall) | 8.5% |
|-----|--------------------|-------|
| 부지표 | 정밀도 (Precision) | 58.2% |
| | F1-Score | 14.9% |
| | 정확도 (Accuracy) | 99.3% |

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

iii) 분석 결과

$$\text{정밀도(Precision)} : \frac{TP}{TP + FP}$$

| 총 데이터 : 1,924,521개 학습용 : 1,347,164개 테스트용 : 577,357개 | | 시스템 예측 결과 | |
|---|------|-------------|---------|
| | | 정상거래 | 이상거래 |
| | | 572,696(TN) | 269(FP) |
| 실 제 결 과 | 정상거래 | 572,696(TN) | 269(FP) |
| | 이상거래 | 4,018(FN) | 374(TP) |

초도모델 분석 결과(Raw Data)

| 주지표 | 재현율 (Recall) | 8.5% |
|-----|--------------------|-------|
| 부지표 | 정밀도 (Precision) | 58.2% |
| | F1-Score | 14.9% |
| | 정확도 (Accuracy) | 99.3% |

I. 추진 배경

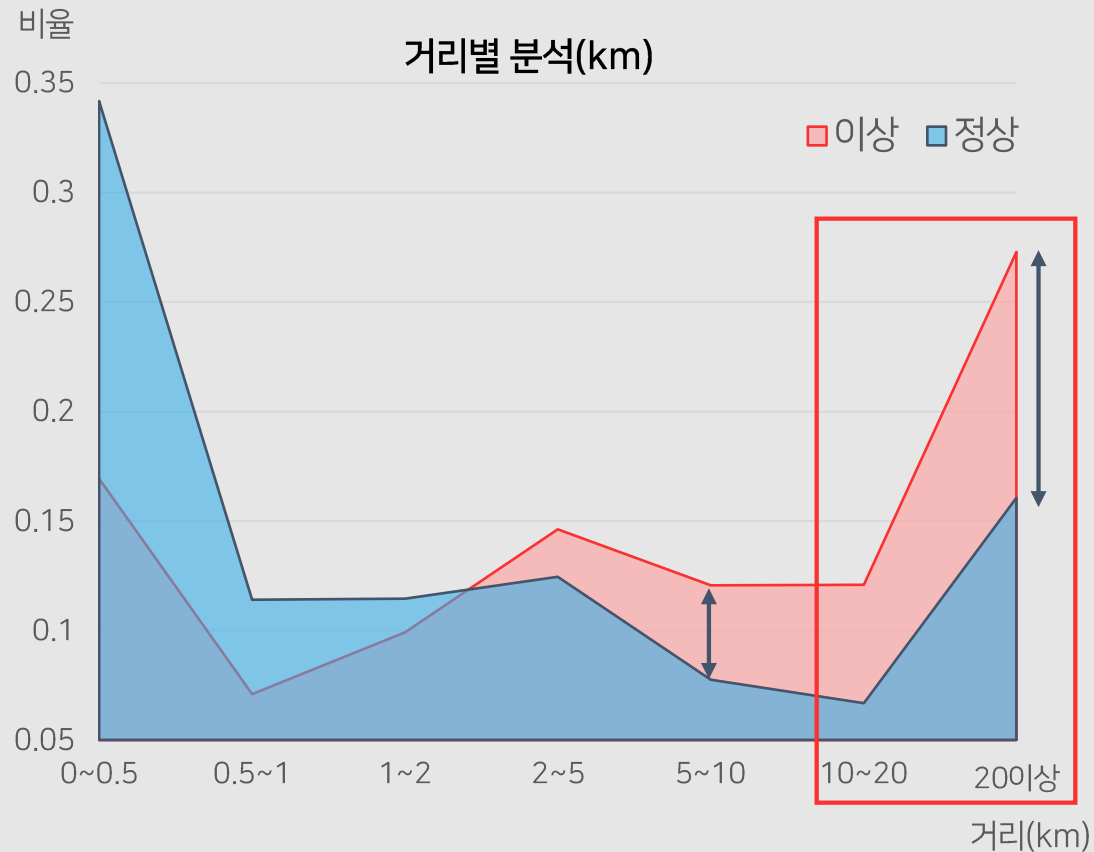
II. 초도모델 분석

III. 성능 고도화

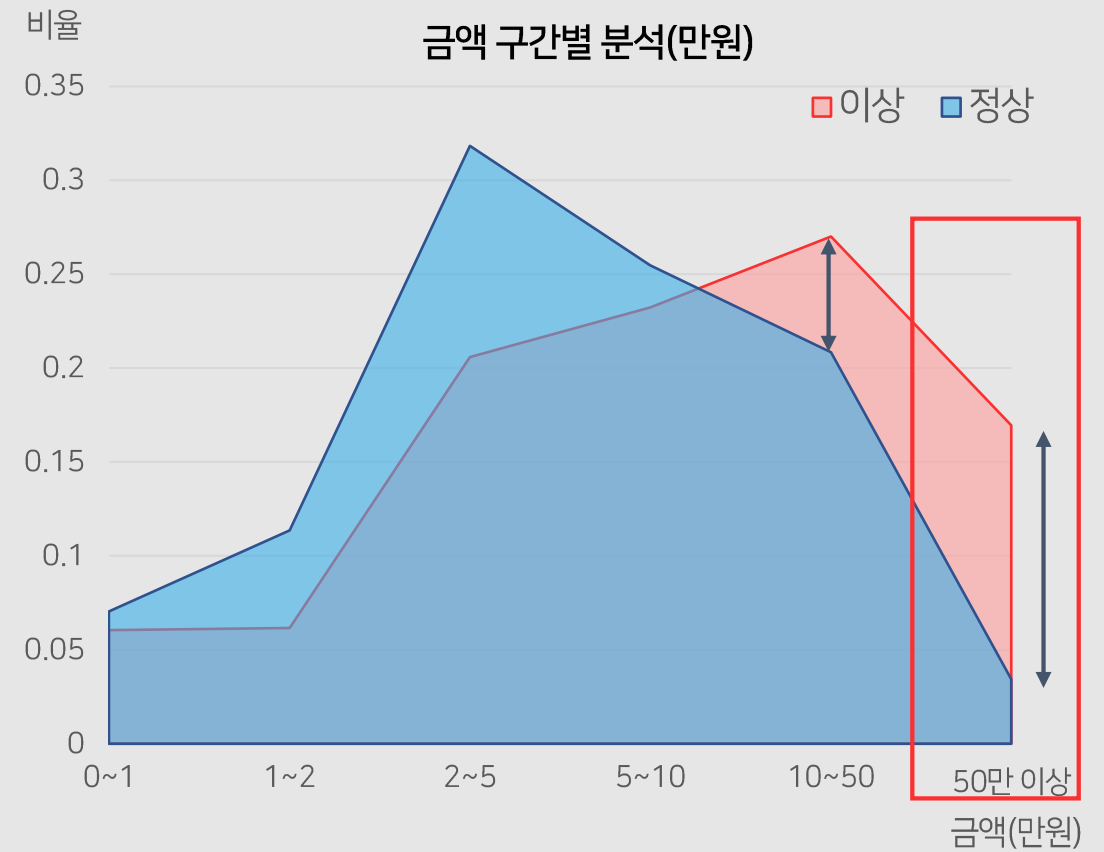
IV. 결 론

i) EDA(탐색적 자료분석)

사무소 - 가맹점 거리 10km이상부터 이상거래 ↑
→ 10km 이상 결제 여부 파생변수 생성



50만원 이상부터 이상거래 ↑
→ 50만원 이상 결제 여부 파생변수 생성



I. 추진 배경

II. 초도모델 분석

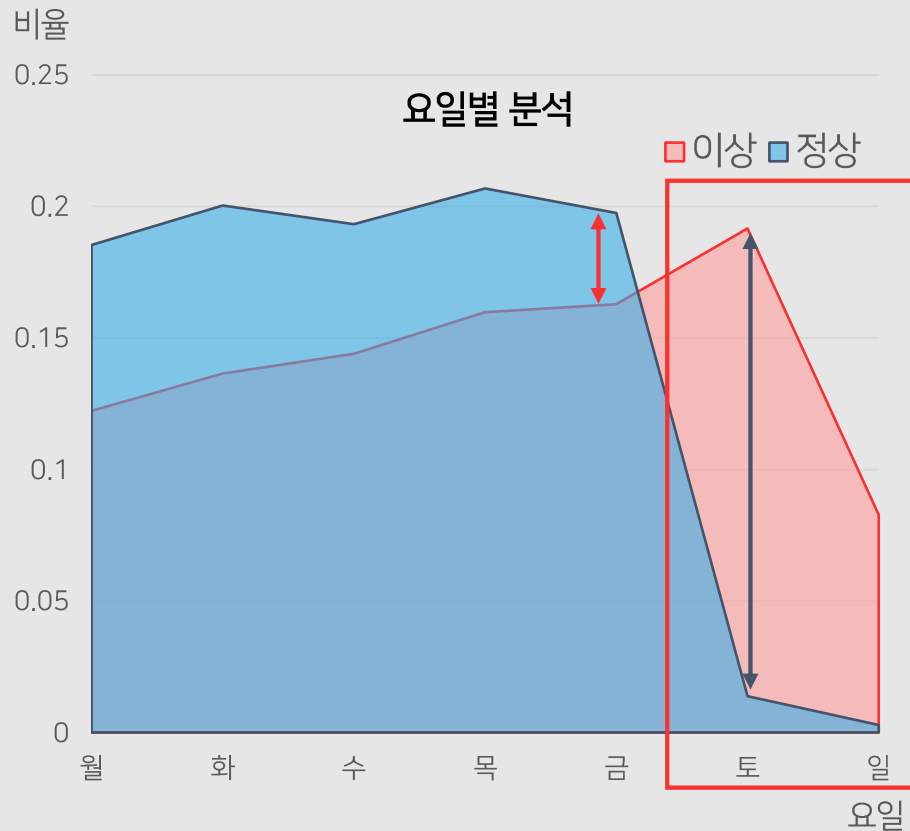
III. 성능 고도화

IV. 결 론

i) EDA(탐색적 자료분석)

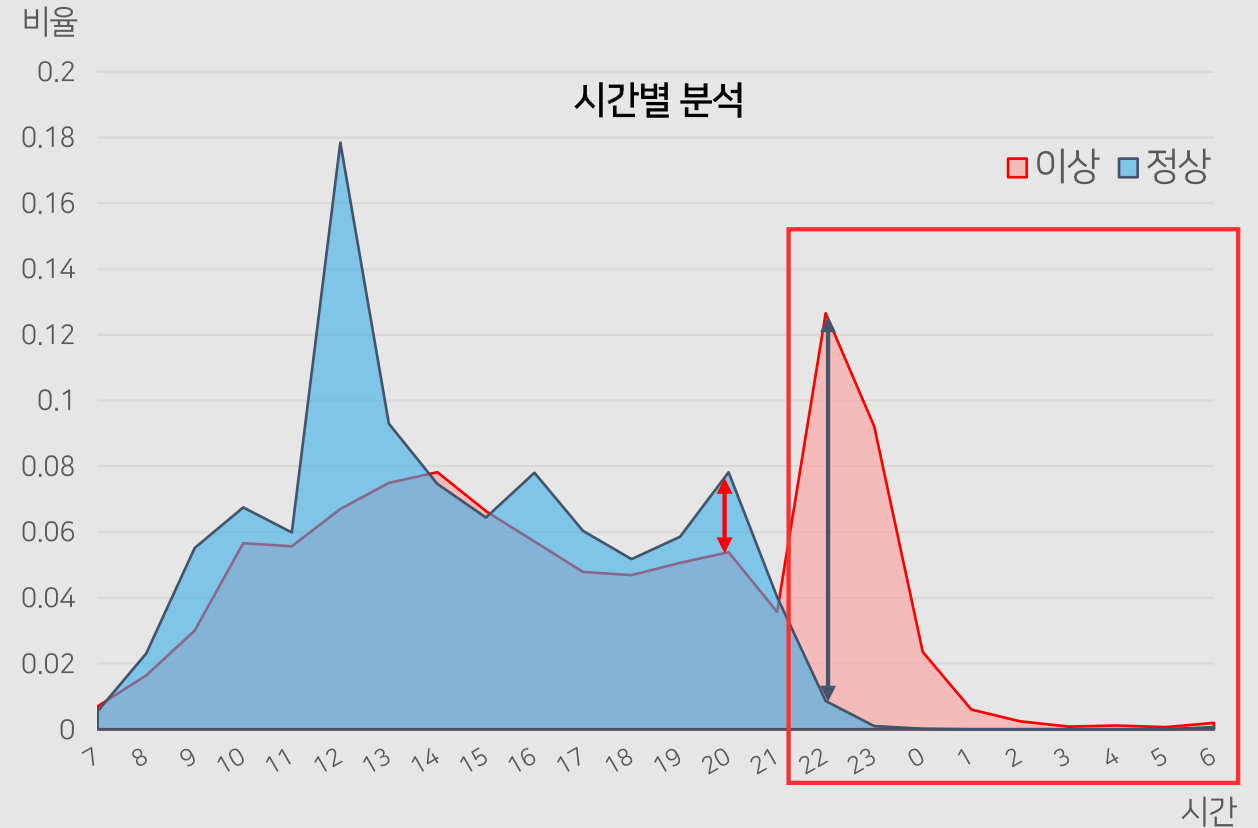
주말사용 결제 이상거래 ↑

→ 주말 사용, 공휴일 사용 여부 파생변수 생성



22시 이후부터 이상거래 ↑

→ 22 ~ 07시 사이 결제 여부 파생변수 생성



I. 추진 배경

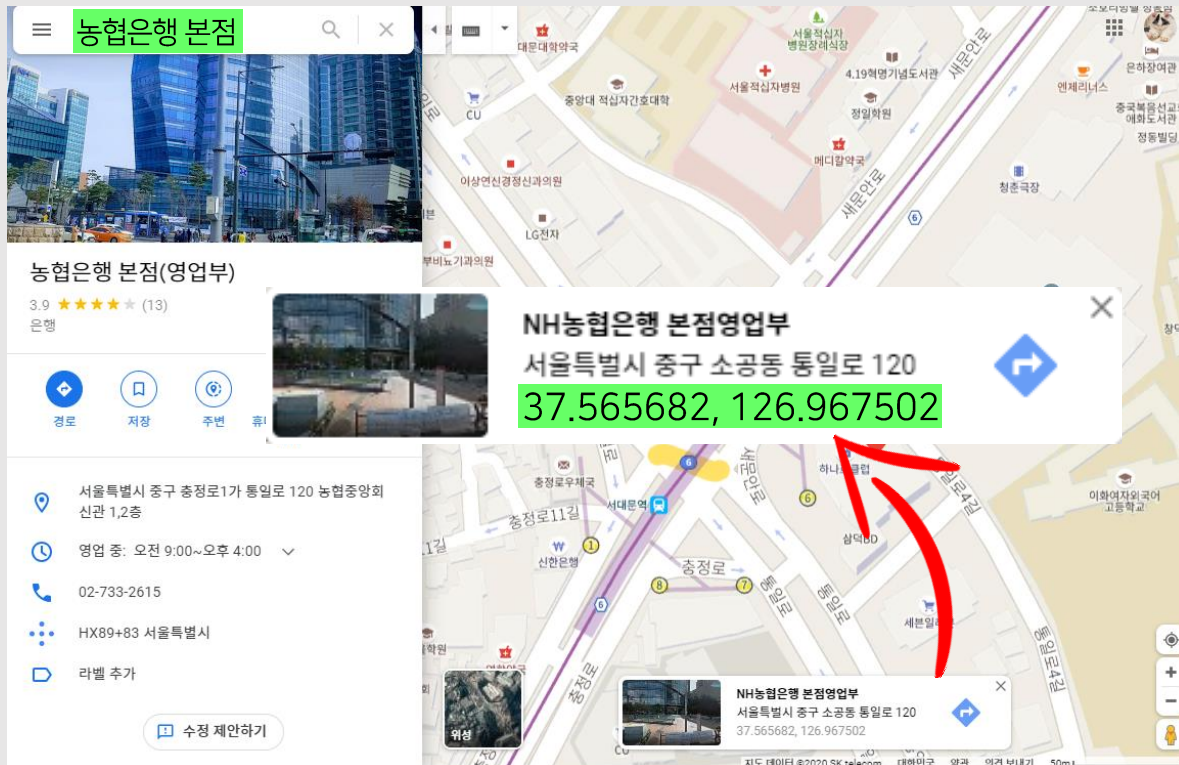
II. 초도모델 분석

III. 성능 고도화

IV. 결 론

ii) 파생변수 생성 예시(공통)

고도화 시킨 데이터를 활용한 위도, 경도 추출



Python 오픈소스를 활용하여 위·경도 기반의 거리 계산



I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

ii) 파생변수 생성 예시(개인)

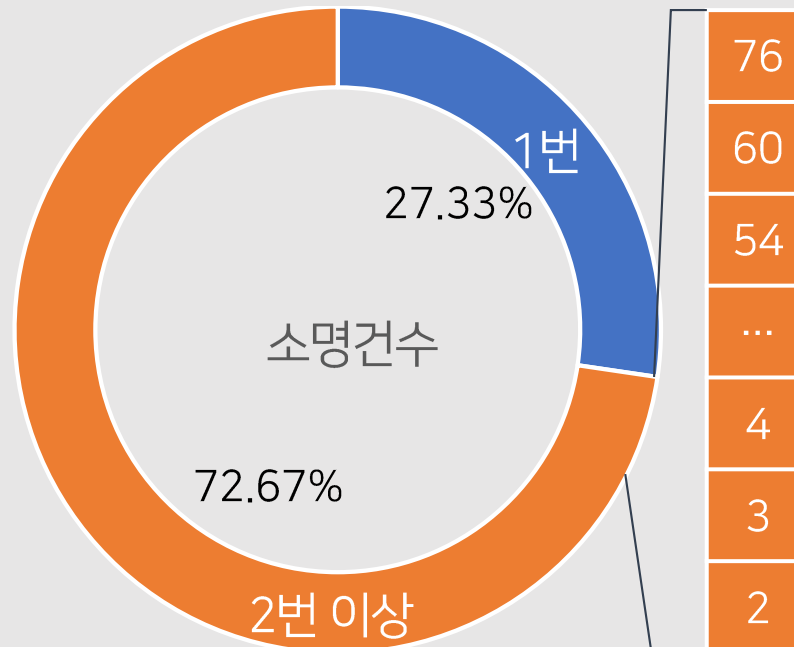
초도 모델

Only 기본 데이터

| | | |
|-----|-----------------|-------|
| 주지표 | 재현율 (Recall) | 8.5% |
| 부지표 | 정밀도 (Precision) | 58.2% |
| | F1-Score | 14.9% |
| | 정확도 (Accuracy) | 99.3% |

법인카드 id별 이상거래 가중치 생성

법인카드 id별 소명건수 EDA 진행
2번 이상 소명한 부서 多 (재소명률 ↑)



파생변수 포함 모델

기본 데이터 + 법인카드 id별 가중치

| | | |
|-----|-----------------|-------|
| 주지표 | 재현율 (Recall) | 35.3% |
| 부지표 | 정밀도 (Precision) | 58.2% |
| | F1-Score | 43.8% |
| | 정확도 (Accuracy) | 99.3% |

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

iii) 파생변수 현황

| 순번 | 파생변수 설명 | 순번 | 파생변수 설명 |
|----|-------------------|----|---------------------|
| 1 | 사무소-가맹점 거리 (공통) | 13 | 소명대상_구분별 이상 거래 비율 |
| 2 | 5km 이상 여부 | 14 | 결제가맹점 - 사무소 주소 일치여부 |
| 3 | 10km 이상 여부 | 15 | 금액이 1,000단위 여부 |
| 4 | id별 소명건수 2회 이상 여부 | 16 | 금액이 10,000단위 여부 |
| 5 | 공휴일 사용 여부 | 17 | 금액이 50,000단위 여부 |
| 6 | 공휴일 전후 사용 여부 | 18 | 거리 구간별 이상 거래 비율 |
| 7 | 주말 사용 여부 | 19 | 금액 구간별 이상 거래 비율 |
| 8 | 22시~07시 사용 여부 | 20 | id별 평균 결제 금액 |
| 9 | 50만원 이상 사용 여부 | 21 | 평균 결제 금액 초과 여부 |
| 10 | 10만원 이상 사용 여부 | 22 | id별 평균 결제 거리 |
| 11 | 업종별 이상 거래 비율 | 23 | 평균 결제 거리 초과 여부 |
| 12 | 세부업종별 이상 거래 비율 | 24 | id별 전월대비 평균 결제 금액 |

기설명한 파생변수(10개)



추가적인 파생변수(14개)

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

iv) AI 기반 모델링 결과

| | | 초도모델 (Decision Tree) | Decision Tree | Random Forest | Gradient Boosting | Light GBM |
|---|--------------------|-------------------------|---------------|------------------|----------------------|-----------|
| 주 | 재현율 (Recall) | 8.5% | 86.1% | 86.3% | 89.8% | 88.0% |
| 부 | 정밀도 (Precision) | 58.2% | 63.0% | 73.4% | 17.3% | 80.9% |
| | F1-Score | 14.9% | 72.8% | 79.3% | 29.1% | 84.3% |
| | 정확도 (Accuracy) | 99.3% | 99.5% | 99.6% | 96.5% | 99.7% |

I. 추진 배경

II. 초도모델 분석

III. 성능 고도화

IV. 결 론

i) 개인 기여도 및 소감

개인 기여도



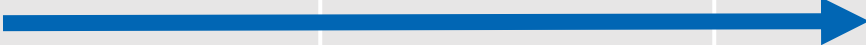


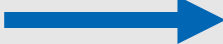
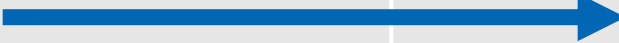

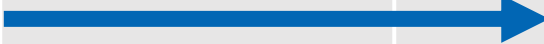
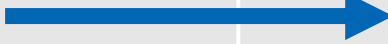
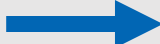
- ✓ 데이터 품질 개선을 통한 확장성 증가
→ 새로운 서비스 개발시 도움
- ✓ 실제 PoC에 일정부분 기여

개인 소감



Q & A

[참고 자료]

| | 8/3 ~ 8/7 | 8/10 ~ 8/14 | 8/17 ~ 8/21 | 8/24 ~ 8/28 | 8/31 ~ 9/4 |
|-------------------------|--|--|---|---|---|
| 문서 분류 Topic Modeling |  | | | | |
| 정보수집 |  | | | | |
| 잠재원인 도출 |  | | | | |
| 데이터 품질확인 |  | | | | |
| 외부 데이터 활용 | |  | | | |
| 데이터 전처리 | | |  | | |
| 파생변수 생성 | | | |  | |
| 데이터 분석 | | | |  | |
| 결론 도출 | | | | |  |

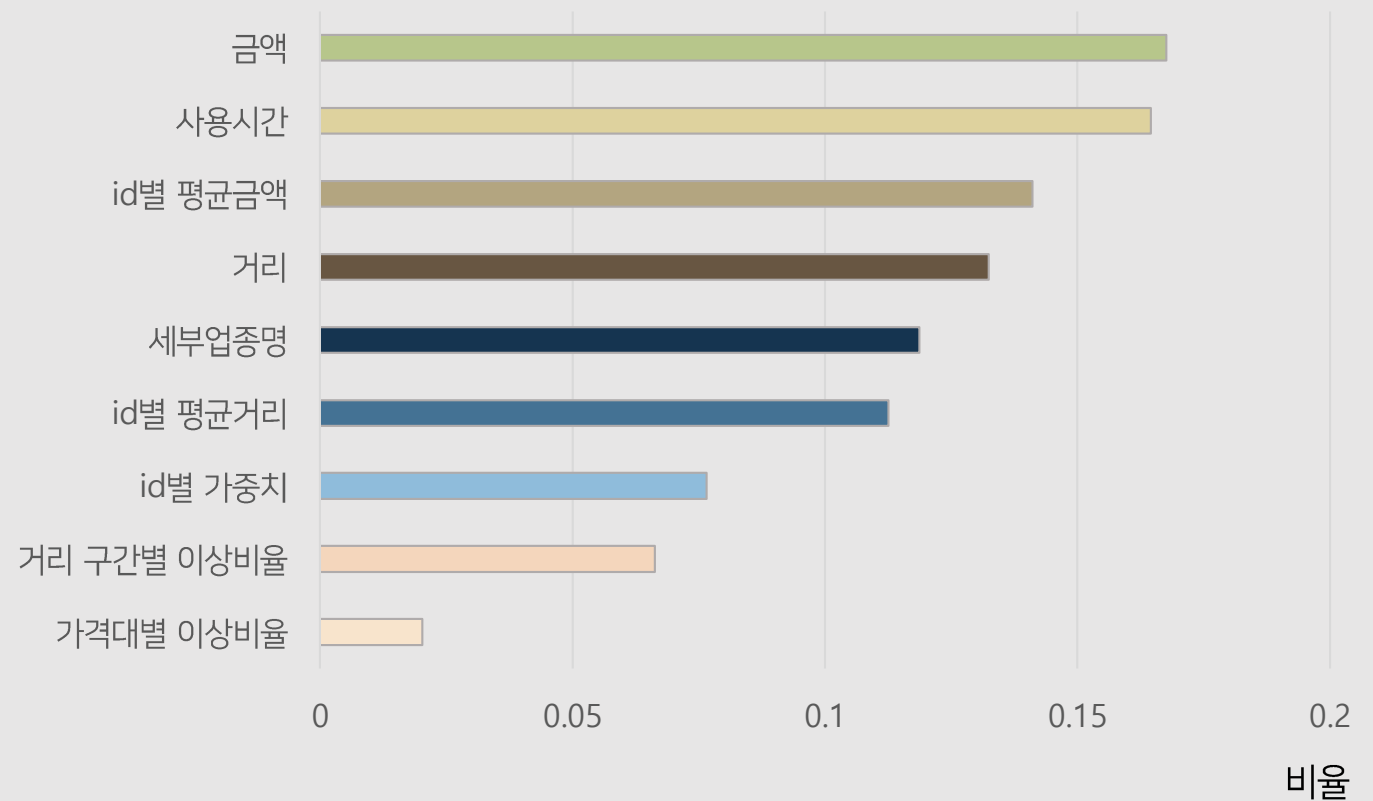
[참고 자료]

기대효과



- ✓ 1시간 80개 → 1,312시간 활용 가능
- ✓ 향후 FDS 시스템 도입을 위해 필요한 데이터 품질을 사전에 고도화

최종 모델 주요 파생변수 중요도



【 참고 자료 】

| | | 시스템 예측 결과 | |
|------------|--------|-----------|--------|
| | | 정상 | 부도(이상) |
| 실 제 결 과 | 정상 | 90(TN) | 4(FP) |
| | 부도(이상) | 1(FN) | 5(TP) |

$$Accuracy = \frac{TP + TN}{All\ Data}$$

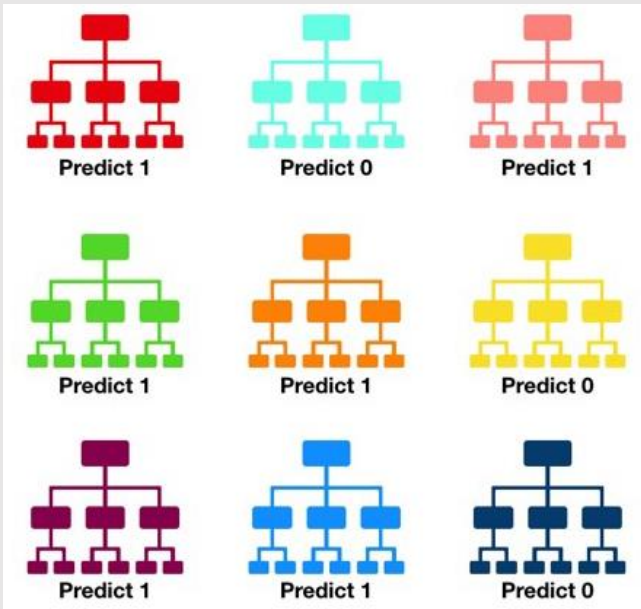
$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

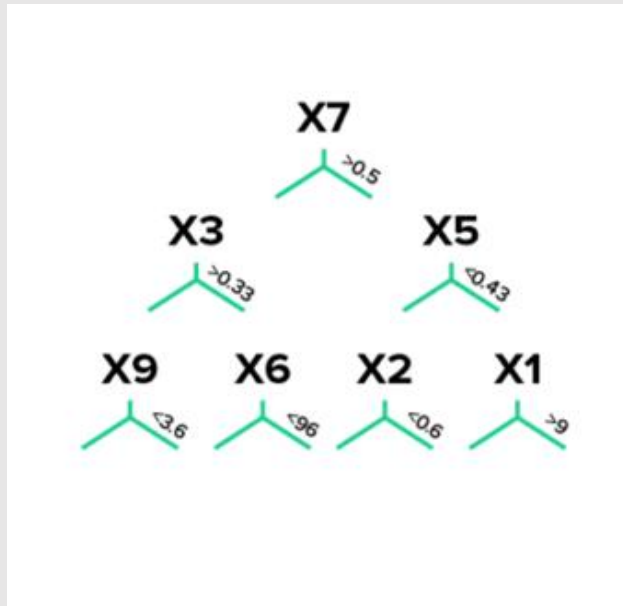
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

[참고 자료]

Random Forest



Gradient Boosting



lightGBM

