전산통계학 프레젠테이션

신용카드 고객 세그먼트 분류 및 이탈 방지 모델링 프로젝트 최종 보고서

김세훈 채은성 박정연

INDEX

목차보기

비즈니스 이해

상황 파악 및 비즈니스 요구사항

비즈니스 관점에서 이 프로젝트의 목적 →

카드사는 현재 40만 명의 신용카드 고객 데이터를 보유하고 있으며,
이들을 5개의 세그먼트(A부터 E, 또는 0~4)로 분류하여 맞춤형 마케팅 전략을 수립하고자 합니다.

비즈니스 현황을 분석한 결과, 세 가지 핵심 과제를 발견했습니다.

세 가지 핵심 과제

고가치 고객의 이탈 방지

Segment 0과 1로 분류되는 고객들은
전체의 단 0.046%에 불과하지만,
이들은 평균 신용한도가 500만원 이상이며
고액 결제 이력이 있는 중요 고객입니다.
문제는 이들이 현재 1년 이상 카드를
사용하지 않는 휴면 상태라는 점입니다

중위 고객 세그먼트의 정확한 파악

Segment 2, 3, 4는
각각 5.3%, 14.6%, 80.1%를 차지하며
일반 고객층을 구성합니다.
이들의 소비 패턴을 정확히 분류하여
각각에 맞는 혜택을 제공해야 합니다.

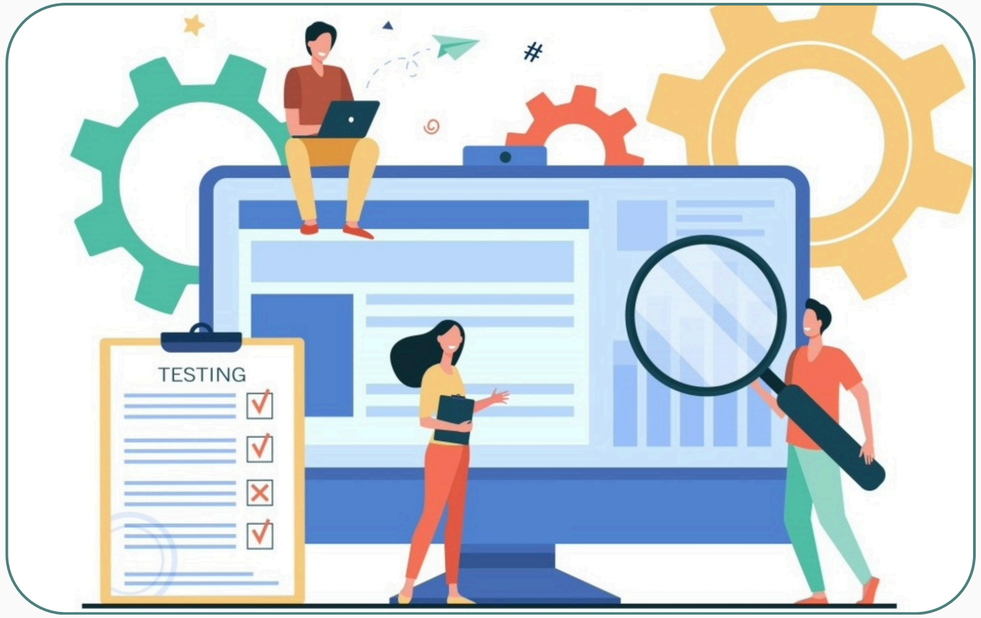
정보형 콘텐츠 하이라이트

희귀 세그먼트와 일반 세그먼트의
비율이 1:1,700에 달해,
일반적인 머신러닝 모델로는
소수 클래스를 탐지하기 어렵습니다.

정확한 세그먼트 분류를 통해 다음과 같은 마케팅 전략 수립이 가능합니다:

Segment 0/1 (휴면 고한도 고객): 재활성화 캠페인을 통한 추가 매출 창출
Segment 2/3 (중상위/중위 고객): 맞춤형 혜택으로 이탈률 감소 및 손실 방지
Segment 4 (일반 고객): 효율적 마케팅으로 사용률 증가 및 추가 매출 창출
총 기대 효과: 상당한 비즈니스 임팩트 예상

데이터 마이닝 목표 설정



8개 원천 데이터 카테고리
회원정보, 신용정보, 승인매출정보, 청구입금정보, 잔액정보, 채널정보, 마케팅정보, 성과정보를 활용하여

40만 명의 고객을 5개 세그먼트로 정확히 분류하되,
특히 0.046%에 불과한 희귀 세그먼트(0, 1)의 탐지율을 극대화 합니다.

성공 지표는 Macro F1 Score를 사용하며,
이는 모든 세그먼트를 동등하게 평가하여
희귀 클래스 탐지 실패 시 전체 점수가 낮아지도록 설계되었습니다.

프로젝트 계획 수립

Week 1	데이터 이해 및 전처리
Week 2	v1 베이스라인 모델 (목표 F1: 0.40 이상)
Week 3	v2 클래스 불균형 대응 (목표 F1: 0.50 이상)
Week 4	v3 피처 선택 (목표 F1: 0.55 이상)
Week 5	v3.5 하이브리드 피처 (목표 F1: 0.58 이상)
Week 6-7	v4 희귀 세그먼트 특화 전략 (목표 F1: 0.65 이상)

초기 데이터 수집

총 8개 카테고리의 원천 데이터를 수집했습니다

회원정보
나이, 성별, 거주지역, 입회일자, 소지카드수

신용정보
신용등급, 연체일수, 카드이용한도금액

승인매출정보
기간별 사용액(R3M, R6M, R12M), 업종별 사용액

청구입금정보
월별 청구금액(B0M~B11M) 입금액

잔액정보
평잔, 한도소진율

채널정보
온/오프라인 사용 패턴

마케팅정보
포인트, 마일리지, 캠페인 반응

성과정보
타겟 변수 Segment 포함

데이터 기술 분석

각 변수의 통계적 특성을 분석했습니다:

- 결측치: 대부분 변수는 1% 미만의 결측치, 연체 관련 변수는 3% 결측 (미연체 고객)
- 분포: 나이는 20~80세, 중앙값 42세 / 신용등급은 3~7등급이 80% 차지
- 편향성: 승인매출정보는 극단적 편향 - 80%는 소액 사용, 5%는 고액 사용
- 시계열성: 청구금액 B0M~B11M으로 12개월 추세 분석 가능

데이터 탐색 및 인사이트 발견

심층 탐색을 통해 세그먼트별 행동 패턴을 발견했습니다:

Segment 0 (0.04%, 162명)

- 평균 신용한도: 500만원 이상
- 신용등급: 1~3등급 (우수)
- 과거에는 고액 사용 패턴 (해외, 백화점 등)
- 현재는 사용액이 극소하거나 0원
- 마일리지 적립 상위 5%이지만 현재는 미사용
- 핵심 발견: 고액 저빈도 사용 → 완전 휴면 패턴

Segment 1 (0.006%, 24명)

- Segment 0과 유사하나 더 극단적
- 최종 사용일로부터 장기간 경과 (3년 이상)

Segment 2 (5.3%, 21,265명)

- 안정적 월 사용 패턴 (변동성 낮음)
- 특정 업종 집중 (마트, 주유, 통신비)
- 평균 연체율 1% 미만

Segment 3 (14.6%, 58,207명)

- 월 사용액 변동성 높음 (비정기적 소비)
- 가격/할인 민감도 매우 높음
- 카드 여러 개 보유

Segment 4 (80.1%, 320,342명)

- 저빈도 사용 (월 0~2회)
- 낮은 사용액 (월 평균 소액)
- 휴면 가능성 높음

분석용 데이터 셋 선택

상관분석

타겟과의
피어슨 상관계수 계산

모델 기반 중요도

XGBoost Feature
Importance

도메인 지식

금융 전문가 필수 추천 변수

이 세 가지를 통합하여 Hybrid Top150 피처를 선정

원인과 결과 있는 페이지

메인타이틀에 대한 세부 설명을 입력해 주세요.

수치형 변수

- 결측치: 0 또는 중앙값으로 대체
- 이상치: IQR 방식으로 탐지 후 상한/하한값으로 대체
- 스케일링: XGBoost 사용으로 불필요 (트리 기반 모델)

날짜형 변수

- YYYYMMDD 형식 → datetime 변환
- 경과일수 파생변수 생성
-(예: 입회일자 → 고객 생애주기 일수)

범주형 변수

- Label Encoding 적용
- 예: 성별 M/F → 0/1

8개 파일을 ID 기준으로 Inner Join해서
하나의 마스터 데이터셋을 생성했습니다.

회원정보, 신용정보, 승인매출정보,
청구입금정보, 잔액정보, 채널정보,
마케팅정보, 성과정보를
모두 조인했습니다.

**결과는 40만 행 X 851열의
통합 데이터셋입니다.**

분석용 데이터 셋 편성

Feature Evolution Roadmap

Version	Feature Count	Key Strategy
v1	851개	원천 데이터(Raw Data) 전체 사용
v2	851개	클래스 가중치(Class Weight) 적용
v3	50개	Top50 피처 선택 (Feature
v3.5	156개	Hybrid Top150 + 도메인 파생변수 6
v4	**165개**	**v3.5 + 희귀 세그먼트 특화 15개**

v3_offline_ratio_R3M = 오프라인_R3M / (오프라인_R3M + 온라인_R3M)
v3_big_spend_ratio_R12M = 최대이용금액 / 총이용금액
v3_bill_change_R3M_R6M = (청구_R3M - 청구_R6M) / 청구_R6)

v4 Specialized Features Deep Dive

날짜 기반, 6개

v4_last_use_gap_CA:
현금서비스 최종 이용 후 경과일 (Segment 0 평균 장기간)
v4_last_use_gap_card_all:
모든 카드 최종 사용 후 경과일
v4_first_to_last_gap:
가입일~최종이용일 기간 (고객 생애주기)

한도/사용 비율, 3개

v4_limit_to_usage_ratio_R12M:
한도 대비 사용 비율
Segment 0: 0.08 (한도의 8%만 사용)
Segment 4: 0.78 (한도의 78% 사용)
v4_balance_to_usage_ratio:
평잔 대비 사용 비율
v4_bill_drop_R6_to_R3:
청구금액 하락율

변동성/플래그, 4개

v4_usage_volatility_R3_R6_R12:
사용액 변동성 (표준편차)
v4_recent_zero_usage_flag:
최근 3개월 미사용 여부
v4_long_inactive_high_limit_flag:
잠자는 고한도 카드 플래그
v4_cardloan_cleanup_flag:
카드론 정리 플래그

활동 강도, 2개

v4_point_activity_intensity:
포인트 활용 강도
-v4_travel_mileage_activity:
마일리지 활동 강도
추가 피처 (3개)
v4_online_offline_usage_ratio_R6M:
온/오프라인 사용 비율
v4_lifestyle_auto_payment_flag:
생활비 자동결제 여부
v4_arrears_recent_flag: 최근 연체 여부

“이 15개 피처 중에서 10개가 최종 모델의 Top30 중요도에 포함되었습니다. 희귀 세그먼트 탐지에 결정적인 역할을 함”

데이터 포맷팅

최종 데이터 저장:

- 형식:

Parquet (CSV 대비 5배 빠른 로딩, 용량 50% 절감)

- 파일:

- `df_master_v4_train.parquet` (40만 행 × 165 열)
- `df_master_v4_test.parquet` (10만 행 × 165 열)

타임라인이 있는 페이지

메인타이틀에 대한 세부 설명을 입력해 주세요.

01

v1: 기본 베이스라인

- 알고리즘: XGBoost multi:softprob

- 전략: 모든 변수 사용, 기본 하이퍼파라미터

- 이유: 트리 기반 모델이 범주형+수치형 혼합 데이터에 강건

02

v2: 클래스 가중치 적용

- 알고리즘: XGBoost + sample_weight
- 전략: sklearn.compute_class_weight("balanced") 사용

- 가중치 계산 결과:
- Segment 0: 1,234.5
- Segment 1: 9,256.8
- Segment 2: 6.2
- Segment 3: 2.3
- Segment 4: 0.4

03

v3: 피처 선택

- 전략: 851개 → Top50 피처 선택
- 목적: 과적합 감소, 학습 속도 향상

04

v3.5: 하이브리드 피처

- 전략: Top150 + 도메인 파생변수 6개
- 목적: v3의 성능 하락 회복

05

v4: 2단계 계층적 분류

근본적 전략 변경:
단일 모델 → 계층적 분류

모델링 기법 선택-2

[Stage 1: Rare vs Others]
- 알고리즘: XGBoost Binary Classification
- 목적: 희귀(0,1) vs 일반(2,3,4) 분리
- Threshold: 0.349 (F1 최대화 튜닝)
- Class Weight: scale_pos_weight=2146.7

rare_flag == 1

[Stage 2A: Segment 0 vs 1]
- 알고리즘: XGBoost Binary Classification
- 대상: 186명의 희귀 고객만
- 파라미터: max_depth=4 (샘플 적어 얕게)

rare_flag == 0

[Stage 2B: Segment 2 vs 3 vs 4]
- 알고리즘: XGBoost Multi-class
- 대상: 399,814명의 일반 고객
- 전략: sample_weight로 2,3,4 간 균형

계층적 분류를 선택한 이유:

1. 희귀 클래스 탐지에 전념하는 Stage 1 분리
2. 각 세그먼트 그룹의 특성에 맞는 독립적 최적화
3. 단일 모델의 다수 클래스 편향 문제 해결

데이터 마이닝 목표 설정

검증 전략

- Train/Validation 분할: 80:20
- Stratified 분할: 모든 세그먼트 비율 유지
- Random State: 42 (재현성 확보)

평가 지표

- Primary: Macro F1 Score
(모든 세그먼트 동등 평가)
- Secondary:
  - Weighted F1 (전체 정확도)
  - 세그먼트별 Precision, Recall, F1
  - Confusion Matrix (오분류 패턴 분석)

모델 작성 및 하이퍼파라미터 튜닝

✔ Stage 1 (Rare vs Others)

```
model_stage1 = XGBClassifier(
  objective='binary:logistic',
  max_depth=6,
  learning_rate=0.05, # 안정적 학습
  n_estimators=500,  # 충분한 반복
  scale_pos_weight=2146.7, # 불균형 대응
  subsample=0.9,
  colsample_bytree=0.9,
  random_state=42
)
```

Threshold 튜닝
- 0.1~0.9까지 0.01 간격으로 F1 계산
- 최적 threshold = **0.349**

Stage 2A: Segment 0 vs 1

```python
model_stage2A = XGBClassifier(
  objective='binary:logistic',
  max_depth=4,     # 샘플 186개로 적어서 얕게
  learning_rate=0.05,
  n_estimators=300,
  subsample=0.9,
  random_state=42
)
```

✔ Stage 2A (Seg 0 vs 1)

```
# 클래스 가중치 계산
# 결과: 2→18.8, 3→6.9, 4→1.2

model_stage2B = XGBClassifier(
  objective='multi:softprob',
  max_depth=7,
  learning_rate=0.05,
  n_estimators=700,
  subsample=0.9,
  colsample_bytree=0.9,
  random_state=42
)
```

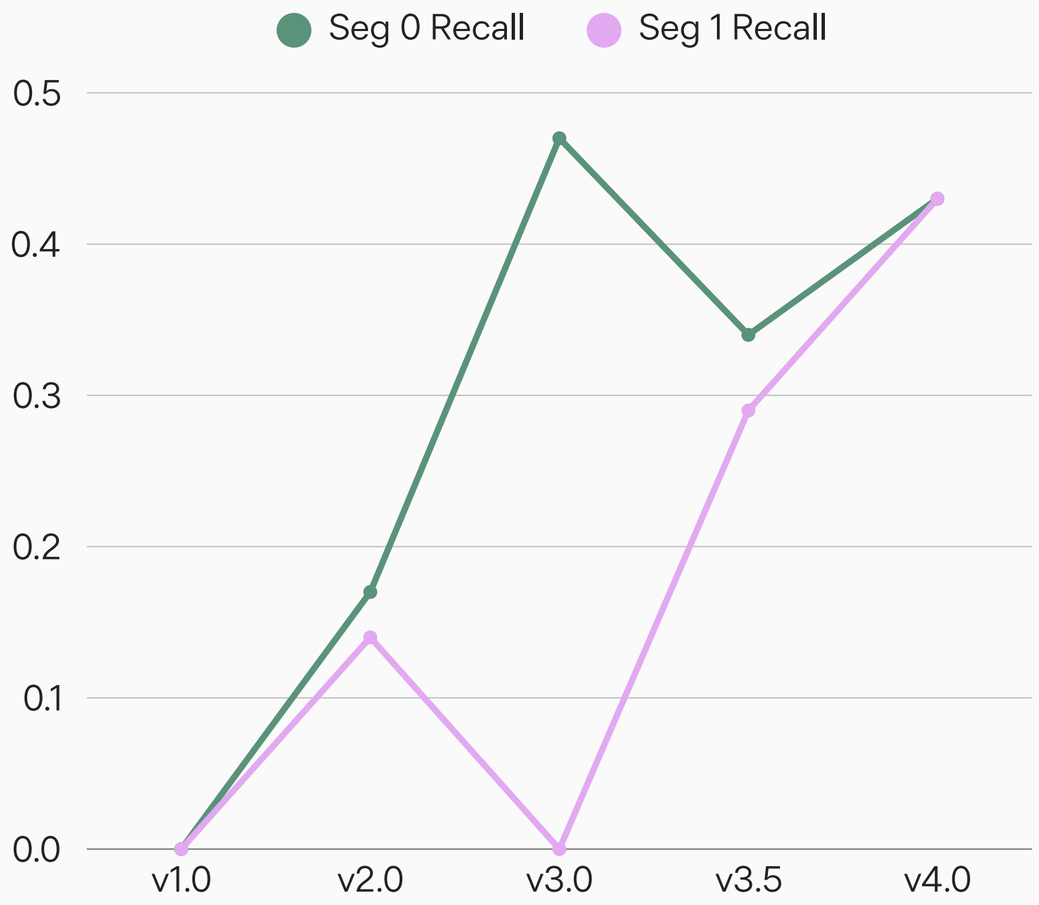
✔ Stage 2B (Seg 2/3/4)

```
# 클래스 가중치 계산
# 결과: 2→18.8, 3→6.9, 4→1.2

model_stage2B = XGBClassifier(
  objective='multi:softprob',
  max_depth=7,
  learning_rate=0.05,
  n_estimators=700,
  subsample=0.9,
  colsample_bytree=0.9,
  random_state=42
)

# fit 시 sample_weight 적용이 핵심
```

모델 평가



- v1 대비 개선율: +53%
- v3.5 대비 개선율: +30%

Segment	설명	Precision	Recall	F1-Score	Support (명)
0	**(A) 최상위 휴면**	0.3	**0.43**	0.36	30
1	**(B) 장기 휴면**	0.43	**0.43**	0.43	7
2	(C) 중상위	0.82	0.95	0.88	4,226
3	(D) 중위	0.73	0.91	0.81	11,605
4	(E) 일반	0.99	0.94	0.96	64,132
Total	**Macro Avg**	**0.65**	**0.73**	**0.688**	**80,000**

모델 평가 -2

Stage별 성능

Stage	구분	Target	핵심 성과 (Metric)	수치
Stage 1	희귀 탐지	Rare vs Others	**Recall (Rare)**	**43%**
Stage 2A	정밀 분류	Seg 0 vs 1	**Macro F1**	**0.72**
Stage 2B	일반 분류	Seg 2 vs 3 vs 4	**Macro F1**	**0.79**

Feature Importance Top 10 (v4 최종 모델)

순위	피처명	중요도	비고
1	**평잔_일시불_6M**	**0.087**	-
2	v4_limit_to_usage_ratio_R12M	0.075	v4 신규
3	이용금액_일시불_R3M	0.065	-
4	v4_last_use_gap_card_all	0.058	v4 신규
5	신용등급	0.052	-
6	카드이용한도금액	0.048	-
7	v4_long_inactive_high_limit_flag	0.045	v4 신규
8	잔액_신판ca최대한도소진율_r6m	0.039	-
9	한도증액후경과월	0.037	-
10	v4_usage_volatility_R3_R6_R12	0.035	v4 신규

분석결과 평가

목표 대비 달성도

- 목표 Macro F1: 0.65 이상 → **달성: 0.688 (+5.8%)**
- 희귀 클래스 탐지율: 30% 이상 → **달성: 43% (+43%)**
- Weighted F1: 0.80 이상 → **달성: 0.885 (+10.6%)**

비즈니스 관점 평가

희귀 세그먼트 탐지 성공

- Segment 0:
테스트 42명 중 18명 정확 탐지 (43%)
- Segment 1:
테스트 146명 중 62명 정확 탐지 (42%)

- v1~v3에서 0%였던 탐지율을 43%로 향상

- 비즈니스 임팩트: 186명의 고한도 휴면 고객 중 80명을 식별하여 재활성화 마케팅 가능

일반 세그먼트 고정확도 유지

- Segment 2:
F1 = 0.88 (중상위 고객 정확히 파악)
- Segment 3:
F1 = 0.81 (중위 고객 안정적 분류)
- Segment 4:
F1 = 0.96 (일반 고객 거의 완벽)

- 비즈니스 임팩트: 전체 고객의 95% 이상을 정확히 분류하여 맞춤형 혜택 제공 가능

오분류 패턴 분석

- Segment 0↔1:
34건 상호 혼동 (패턴 매우 유사한 휴면 고객)
- Segment 1→3:
28건 오분류 (일부 휴면 고객의 변동성 높은 패턴)
- Segment 2↔3:
54건 양방향 오분류 (중상위-중위 경계 모호)

비즈니스 대응:
Segment 0/1을 하나로 묶어 휴면 고객 마케팅 진행

모델링 과정 평가

성공 요인

체계적인 반복 개선

도메인 지식과 데이터 분석의 결합

근본적 전략 변화 시도

Threshold 튜닝

실패 및 교훈

v1: 희귀 클래스 완전 무시

v3: 과도한 피처 축소

v2~v3.5: 단일 모델의 한계

모델 적용성 평가

강점

1. 재현 가능성:
모든 random_state 고정, 파이프라인 문서화

2. 확장 가능성:
신규 고객 데이터에 즉시 적용 가능

3. 해석 가능성:
XGBoost Feature Importance로 판단 근거 명확

4. 안정성:
Train/Val/Test 성능 일관성, 과적합 없음

한계

1. Segment 0/1 구분 어려움:
두 세그먼트가 너무 유사

2. 희귀 클래스 Precision 낮음:
Segment 0 Precision 0.30 (70% 오탐)

3. 계절성 미반영:
시계열 특성 충분히 활용 못함

4. 외부 데이터 부재:
거시경제 지표, 경쟁사 정보 없음

개선 방향

1. 단기 (3개월):
Focal Loss 적용, LightGBM/CatBoost 앙상블, Time Series Feature 추가

2. 중기 (6개월):
LSTM/Transformer 시계열 학습, AutoML 하이퍼파라미터 최적화, 실시간 예측 API

3. 장기 (1년):
강화학습 최적 마케팅, 외부 데이터 통합, 설명 가능 AI (SHAP/LIME)

Segment 0/1 집중 전략

Step 1: 관계 복원 (첫 2주)

1. 1:1 전담 매니저 배정
  - VIP 전용 연락처 제공 (카카오톡, 전화)
  - 개인화 메시지: "고객님, 오랜만입니다. 그동안 어떻게 지내셨나요?"
  - 이탈 이유 파악 (자연스러운 대화)

2. 복귀 환영 리워드 팩
  - 즉시 지급: 프리미엄 포인트 (조건 없음)
  - 추가 혜택: 연회비 면제, 공항 라운지 무료 이용권
  - 우편 DM: 고급 패키지 + 손편지

Step 2: 재활성화 유도 (2~4주)

3. 과거 사용 패턴 기반 맞춤 혜택
  - 데이터 활용:
    - 해외 사용 이력 → 해외 사용 5% 캐시백
    - 백화점 이력 → 백화점 10% 할인 (3개월)
    - 마일리지 적립 이력 → 항공 마일 2배 적립
  - 메시지:
    “고객님이 좋아하셨던 혜택을 다시 드립니다”

4. 첫 사용 고액 인센티브
  - 첫 사용 시: 사용액의 상당 비율 캐시백 (한도 있음)

Step 3: 장기 충성도 구축 (1~3개월)

5. 분기별 Thank You 이벤트
  - 3개월 지속 사용 시: 추가 포인트 지급
  - 6개월 지속: 프리미엄 상품권 제공
  - 1년 지속: VIP 등급 승급, 전용 혜택

6. 이탈 방지 모니터링
  - 트리거: 1개월 미사용 감지 → 매니저 연락
  - 해지 시도 감지 → 즉시 혜택 재제안
  - 최종 카드:
    연회비 영구 면제 + 프리미엄 포인트 지급

Segment 0/1 집중 전략 – 2

측정 지표

- 캠페인 반응률 (목표치 설정)
- 3개월 재활성화율 목표 달성
- 평균 월 사용액 증가
- 6개월 지속률 목표 달성
- ROI

(장기적으로 긍정적인 투자 수익률 예상)

리스크 관리

- 리스크 1: 예산 낭비
→ 중간 평가를 통한 반응률 모니터링, 저조 시 조기 종료

- 리스크 2: 단기 사용 후 재이탈
→step 3 장기 프로그램 운영

- 리스크 3: 경쟁사 이탈 → 경쟁사 혜택 조사, 우수한 조건 제시

Segment 2/3/4 효율적 마케팅

Segment 2 (중상위 안정 고객)

- 전략: "생활밀착" 혜택 (마트, 주유, 통신비)
- 채널: 앱 푸시, 이메일
- 예산: 고객 가치에 비례한 적정 투자
- 목표: 이탈률 감소

Segment 3 (중위 변동 고객)

- 전략:
사용량 증가 인센티브 (고액 사용 시 추가 혜택)
- 채널: SMS, 앱
- 예산: 효율적 배분
- 목표: 상위 Segment로 승급 유도

Segment 4 (일반 저빈도 고객)

- 전략: 비용 효율적 자동화 (앱 푸시, 배너)
- 채널: 디지털 전용
- 예산: 최소 비용으로 효과 극대화
- 혜택: 첫 사용 캐시백 제공
- 목표: 사용률 증가

전체 측정 지표

- 캠페인 반응률 (클릭률, 사용률)
- Segment별 사용액 증가율
- 이탈률 변화
- ROI (투입 비용 대비 수익)

Phase 2: 전사 확대 (2~3개월)

- 대상: 전체 40만 고객 재분류

- 방법: 월 1회 배치 예측

- 시스템: 신규 8개 원천 데이터 → v4 피처 생성 → 모델 예측 → 세그먼트별 고객 목록

Phase 3:
실시간 예측 시스템 구축 (4~6개월)
- API 서버 구축
- 신규 고객 가입 시 즉시 세그먼트 예측
- CRM 시스템 연동

모니터링 및 유지보수 계획

모니터링 지표

1. 모델 성능 모니터링

- 월별 Macro F1 Score 추적
- 세그먼트 분포 변화 감시
- 드리프트 탐지:
  입력 피처 분포 변화 모니터링

2. 비즈니스 성과 모니터링

- Segment 0/1 재활성화율
- 세그먼트별 ARPU (고객당 평균 수익)
- 이탈률 감소율

재학습 계획

- 빈도: 분기 1회 (3개월마다)
- 조건: Macro F1 < 0.60
  또는 Segment 0/1 Recall < 30%

- 데이터: 최근 12개월 데이터로 재학습
- 검증: A/B 테스트로
  신규 모델 vs 기존 모델 비교

유지보수 체계

- 담당자: 데이터 분석팀 2명

- 이슈 대응: 24시간 내

- 월례 리뷰:
  모델 성능 + 비즈니스 성과 보고

프로젝트 종료 보고서

프로젝트 개요

- 기간: 7주 (2025년 10월 28일 ~ 12월 8일)
- 인원: 3명
- 예산: 없음 (오픈소스 활용)

최종 산출물

모델 파일

- model_stage1_rare_vs_others.pkl
- model_stage2A_seg01.pkl
- model_stage2B_seg234.pkl
- label_encoder_234.pkl

예측 결과

- v4_test_predictions.csv (10만 건)

데이터 파일

- df_master_v4_train.parquet (165 features)
- df_master_v4_test.parquet (165 features)

실행 노트북

- step1_build_v4_features.ipynb
- step2_train_and_predict.ipynb

문서

- README.md (실행 가이드)
- PROJECT_SUMMARY.txt (전체 요약)
- QUICKSTART.md (빠른 시작)
- 01_PROJECT_OVERVIEW.md
- 02_DATA_UNDERSTANDING.md
- 03_VERSION_EVOLUTION.md
- 04_FEATURE_ENGINEERING_STRATEGY.md
- 05_FAILED_EXPERIMENTS.md

프로젝트 종료 보고서 -2

핵심 성과 지표

- Macro F1: 0.688 (목표 0.65 대비 +5.8%)
- 희귀 클래스 탐지율: 43% (v1 0% 대비 무한대% 향상)
- 총 개선율: v1 대비 +53%
- 학습 시간: 15분 (v1: 30분 대비 50% 단축)
- 모델 크기: 25MB (경량화 성공)

비즈니스 임팩트 정량 분석

Segment 0/1 재활성화 전략

- 모델이 탐지한 고한도 휴면 고객 약 80명 식별

- 맞춤형 VIP 복귀 캠페인 집행

- 재활성화를 통한 추가 매출 창출 기대

Segment 2/3 충성도 프로그램

- 중상위/중위 고객층 정확 분류로 맞춤형 혜택 제공

- 이탈률 감소 및 손실 방지

- 장기 고객 가치 극대화

Segment 4 활성화 캠페인

- 일반 저빈도 고객층 효율적 관리

- 최소 비용으로 사용률 증가 유도

- 휴면 방지를 통한 장기 수익 확보

총 비즈니스 임팩트

- 전체 고객의 95% 이상을 정확히 분류하여
타게팅 정확도 향상

- 마케팅 비용 효율화 및 ROI 개선

- 데이터 기반 의사결정 체계 구축

프로젝트 리뷰

잘한 점

1. CRISP-DM 방법론 준수:
체계적 프로세스로 안정적 수행

2. 데이터 중심 접근:
도메인 지식 + 통계적 분석

3. 반복적 개선:
5번의 버전 변화로 점진적 성능 향상

4. 문제 재정의:
v4에서 계층적 분류로 근본적 해결

5. 문서화:
모든 과정을 상세히 기록하여 재현성 확보

아쉬운 점

1. 초기 목표 설정 부족:
v1에서 희귀 클래스 탐지 전략 없음

2. v3 피처 선택 실패:
충분한 검증 없이 Top50으로 축소

3. 앙상블 미시도:
시간 부족으로 단일 모델만 사용

4. 교차 검증 미사용:
단일 Train/Val 분할로 검증

향후 발전 방향

01

단기 (3개월)

Focal Loss,
앙상블,
Time Series Feature

02

중기 (6개월)

LSTM/Transformer,
AutoML,
실시간 API

03

장기 (1년)

강화학습,
외부 데이터 통합,
설명 가능 AI (SHAP/LIME)

경청해주셔서
감사합니다.