

3 장

종합 실습 문제 답안 샘플

※ 다음 답안 샘플은 2026년 6월 25일부터 7월 7일 기간 동안 챗지피티 5.5 Pro로 테스트한 것입니다. 사용하는 모델과 환경에 따라 답은 달라질 수 있음을 참고하세요.

■ 단계 1. 이탈 정의와 관측/예측 창 정리

[문제 1의 답안 샘플]

Dplus30_churn_label은 스냅샷 기준일(2025-11-01) 이후 30일 예측 창 안에서 이탈이 발생했는지를 나타내는 이진 라벨이다. 값이 1이면 예측 창 내 이탈 고객, 0이면 유지 고객이다. 이 데이터의 예측 창은 2025-11-01부터 2025-11-30까지로 해석한다.

PlayNow 맥락에서 하드 이탈은 구독 해지, 계정 탈퇴, 결제 실패 후 최종 정지처럼 명시적인 이탈 이벤트가 발생한 경우다. 소프트 이탈은 해지 버튼을 누르지는 않았지만 장기간 미접속, 핵심 콘텐츠 미재생, 세션 급감 등으로 이용 의지가 사실상 약화된 상태다. 본 데이터의 라벨은 하드/소프트 이탈 유형을 별도로 구분하지 않으므로 모델링에서는 “향후 30일 내 이탈 발생 여부”라는 단일 이진 목표변수로 사용하였다.

[문제 2의 답안 샘플]

변수	집계 의미	날짜 범위
recent30_active_days	최근 30일 동안 서비스에 접속한 날짜 수	2025-10-02~2025-10-31
recent14_core_plays	최근 14일 동안 핵심 콘텐츠 재생 횟수	2025-10-18~2025-10-31

변수	집계 의미	날짜 범위
recent14_session_change	최근 14일 세션 수 - 그 이전 14일 세션 수	최근 14일: 2025-10-18~2025-10-31 / 직전 14일: 2025-10-04~2025-10-17
recent60_complaint_type	최근 60일 VOC 유형	2025-09-02~2025-10-31
recent90_avg_revenue	최근 90일 월평균 매출	2025-08-03~2025-10-31

■ 단계 2. 정보 누수 점검과 피처 설명

[문제 1의 답안 샘플]

스냅샷 이후 30일 쿠폰 사용 여부, 예측 기간 내 VOC 해결 여부, 예측 창 내 결제 성공 여부는 예측 시점에는 알 수 없는 미래 정보다. 이러한 변수를 모델에 넣으면 모델이 실제 운영 시점에는 사용할 수 없는 정보를 학습하게 되므로 성능이 과대평가된다.

누수 위험 변수	사용 불가 이유	관측 창 대체 변수
스냅샷 이후 30일 쿠폰 사용 여부	예측 창 안에서 발생한 미래 행동	recent30_coupon_used: 기준일 이전 최근 30일 쿠폰 사용 여부
예측 기간 내 VOC 해결 여부	예측 기간 중 CS 처리 결과	recent60_complaint_type, sla_delay_flag: 기준일 이전 최근 60일 VOC 및 SLA 지연 여부
예측 창 내 결제 성공/실패 여부	이탈 라벨과 직접 연결될 수 있는 미래 결제 결과	recent_pay_fail_flag, last_success_pay_gap: 기준일 이전 결제 실패 및 마지막 성공 결제 경과일
예측 창 내 접속 일수·재생 횟수	이탈 여부를 사실상 사후적으로 반영	recent30_active_days, recent14_core_plays: 기준일 이전 활동 정보

[문제 2의 답안 샘플]

전처리 기준은 “독립변수는 2025-10-31까지 이미 관측된 정보만 사용하고, 종속변수만 2025-11-01~2025-11-30의 결과로 생성한다”로 설정한다. 따라서 모든 활동·결제·VOC·마케팅 반응 피처는 snapshot_date 이전 관측 창에서 집계하고, 예측 창 이후에 생성된 이벤트는 피처에서 제외한다.

■ 단계 3. 이탈률과 기본 탐색

[문제 1의 답안 샘플]

전체 10,000명 중 Dplus30_churn_label=1인 고객은 269명이며, 전체 이탈률은 269/10,000=2.69%이다. 유지 고객이 97.31%로 대다수이므로 이 문제는 명확한 불균형 이진 분류 문제다. 따라서 정확도만 보면 안 되며, 정밀도·재현율·리프트를 함께 평가해야 한다.

[문제 2의 답안 샘플]

변수	유지 평균	유지 중앙값	이탈 평균	이탈 중앙값	이탈-유지 평균 차이
최근 30일 접속일수	13.15	13	8.89	8	-4.26
최근 14일 핵심 콘텐츠 재생	22.64	20	15.13	12	-7.51
마지막 성공 결제 이후 경과일	15.80	16	23.43	21	7.63
최근 90일 평균 매출	10,250.89	9,000	10,015.99	9,000	-234.90

이탈 고객은 유지 고객보다 최근 30일 접속일 수가 약 4.26일 낮고, 최근 14일 핵심 콘텐츠 재생 횟수도 약 7.51회 낮다. 반대로 마지막 성공 결제 이후 경과일은 약 7.63일 더 길다. 최근 90일 평균 매출은 큰 차이가 없으므로, PlayNow의 이탈 신호는 “고객 가치 자체”보다 “최근 이용 약화와 결제 문제”에서 더 강하게 나타난다.

조건	고객 수	이탈률
최근 30일 결제 실패 있음	209	33.97%
결제 관련 VOC 있음	347	16.14%
SLA 지연 경험 있음	161	14.29%
최근 14일 앱 비정상 종료 경험 있음	700	5.43%
OS/앱 버전 불일치 있음	484	4.75%
VOC 없음	8,405	1.90%

■ 단계 4. 로지스틱 회귀 모델

[문제 1의 답안 샘플]

피처 축	입력 변수	전처리
활동·참여	recent30_active_days, recent14_core_plays, recent14_session_change	연속형 표준화
결제·거래	recent_pay_fail_flag, last_success_pay_gap, recent30_coupon_used	연속형 표준화, 이진형 0/1 유지
VOC·품질	recent60_complaint_type, sla_delay_flag, crash14_flag, os_app_mismatch_flag	VOC는 None 포함 원-핫 인코딩, 이진형 0/1 유지

피처 축	입력 변수	전처리
가치·마케팅 반응	recent90_avg_revenue, recent30_open_rate, plan_type	연속형 표준화, 요금제 원-핫 인코딩
추가 통제	days_since_signup	연속형 표준화

모델은 8:2 층화 분할(random_state=42)로 학습/검증 데이터를 나누고, 로지스틱 회귀에 class_weight="balanced"를 적용했다. recent60_complaint_type의 결측은 실제 결측이라기보다 “최근 60일 VOC 없음”으로 해석하여 None 범주로 대체했다.

[문제 2의 답안 샘플]

이탈률이 2.69%로 낮기 때문에 무작위로 10%를 고르면 정밀도는 약 2.7%에 불과하다. 따라서 실무적으로는 Top 10%에서 최소 3배 이상의 리프트, 즉 정밀도 8~10% 이상을 보이면 의미 있는 기준선으로 볼 수 있다. 예를 들어 Top 10% 정밀도 0.10, 리프트 3.7배, 재현율 0.35라면 콜·메시지 대상을 무작위보다 훨씬 효율적으로 좁힌 것이다.

지표	검증 결과	해석
ROC-AUC	0.849	무작위 0.5보다 충분히 높아 순위화 성능이 있음
Average Precision	0.316	낮은 이탈률을 고려하면 무작위 기준 대비 개선
Top 10% 정밀도	16.50%	전체 이탈률 2.69% 대비 고위험군 농축
Top 10% 재현율	61.11%	검증셋 실제 이탈자 중 61.11% 포착
Top 10% 리프트	6.11배	무작위 선정 대비 약 6.1배 효율

검증셋 Top 10% 혼동행렬	예측 유지	예측 이탈
실제 유지	1,779	167
실제 이탈	21	33

위험군 컷	대상 수	정밀도	재현율	리프트	포착 이탈자 수
Top 5%	100	27.00%	50.00%	10.00배	27
Top 10%	200	16.50%	61.11%	6.11배	33
Top 15%	300	12.00%	66.67%	4.44배	36
Top 20%	400	10.00%	74.07%	3.70배	40

주요 계수 해석은 다음과 같다. 결제 실패, SLA 지연, 앱 비정상 종료, 품질 VOC는 이탈 위험을 높이고, 최근 30일 접속일 수와 세션 증가 세는 이탈 위험을 낮추는 방향으로 나타났다. 연속형 변수의 오즈비는 표준화된 1표준편차 증가 기준이다.

주요 변수	계수	오즈비	방향
최근 30일 결제 실패	3.325	27.80	위험 증가
SLA 지연 처리 경험	1.137	3.12	위험 증가
최근 14일 앱 비정상 종료	1.073	2.92	위험 증가
VOC=품질(없음 대비)	0.977	2.66	위험 증가
최근 30일 접속일 수(표준화)	-0.967	0.38	위험 감소
VOC=콘텐츠(없음 대비)	0.660	1.93	위험 증가
최근 14일 세션 증감(표준화)	-0.423	0.65	위험 감소
VOC=결제(없음 대비)	0.321	1.38	위험 증가

■ 단계 5. 의사결정나무 규칙 해석

[문제 1의 답안 샘플]

해석용 의사결정나무는 최대 깊이 3, class_weight="balanced", min_samples_leaf=50으로 학습했다. 검증셋 ROC-AUC는 0.807이며, 아래 표는 학습된 나무를 전체 데이터에 적용했을 때 이탈률이 높은 리프트 규칙이다.

구분	조건	표본 수	이탈자 수	이탈률	리프트
규칙 1	최근 30일 결제 실패 있음 AND 최근 14일 핵심 콘텐츠 재생 $\leq$ 20회	112	48	42.86%	15.93배
규칙 2	최근 30일 결제 실패 있음 AND 최근 14일 핵심 콘텐츠 재생 $\geq$ 21회	97	23	23.71%	8.81배
규칙 3	최근 30일 결제 실패 없음 AND 최근 30일 접속일 수 $\leq$ 8일 AND 최근 14일 핵심 콘텐츠 재생 $\leq$ 5회	656	61	9.30%	3.46배
규칙 4	최근 30일 결제 실패 없음 AND 최근 30일 접속일 수 $\leq$ 8일 AND 최근 14일 핵심 콘텐츠 재생 $\geq$ 6회	1,346	54	4.01%	1.49배
규칙 5	최근 30일 결제 실패 없음 AND 최근 30일 접속일 수 $\geq$ 9일 AND 최근 14일 세션 변화 $\leq$ -7회	564	20	3.55%	1.32배

[문제 2의 답안 샘플]

가장 위험한 리프는 최근 30일 결제 실패가 있고 최근 14일 핵심 콘텐츠 재생이 20회 이하인 고객군이다. 이 집단의 이탈률은 42.86%로 전체 평균의 약 15.94배다. 결제 실패가 있지만 핵심 콘텐츠 재생이 21회 이상인 집단도 이탈률이 23.71%로 높아, 결제 실패 자체가 강력한 위험 신호임을 보여준다.

결제 실패가 없는 고객 중에서도 최근 30일 접속일 수가 8일 이하이고 핵심 콘텐츠 재생이 5회 이하인 고객은 이탈률이 9.30%로 전체 평균보다 크게 높다. 즉, 조용히 이용량이 줄어드는 고객도 결제 실패 고객만큼은 아니지만 중요한 관리 대상이다. 로지스틱 회귀에서도 최근 결제 실패의 오즈비가 27.80으로 가장 컸고, 최근 30일 접속일 수의 오즈비는 0.38로 보호 요인으로 나타났다. 두 모델은 공통적으로 “결제 문제, 핵심 이용 감소, 활동 급감, 품질/VOC 문제”를 반복적인 위험 축으로 제시한다.

■ 단계 6. A/B/C 전략과 간단 ROI 계산

[문제 1의 답안 샘플]

우선순위 점수는 “이탈 확률 × 가치 가중치 × 타이밍 가중치”로 계산했다. 가치 가중치는 recent90_avg_revenue가 중앙값 이상이면 1.25, 아니면 1.00으로 두었다. 타이밍 가중치는 월 결제주기 30일을 가정하여 last_success_pay_gap이 25~27일이면 D-5~D-3으로 보고 1.20, 그 외는 1.00으로 두었다.

우선순위 점수 상위 1,000명의 실제 사후 이탈률은 15.10%로 전체 이탈률 2.69% 대비 5.61배 높았다. A/B/C 구역별 위험 특성은 다음과 같다.

구역	대상 수	평균 예측 이탈확률	실제 이탈률	실제 이탈자 수	결제 실패 비중	저활동 비중
A구역(콜)	200	90.81%	28.00%	56	57.00%	52.00%
B구역(메시지+ 라이트 쿠폰)	400	78.44%	15.00%	60	17.75%	68.50%
C구역(라이트 터 치)	400	68.80%	8.75%	35	4.25%	67.00%

구역	대상 수	성공률	예상 추가 유지 고객	총비용	추가 공헌이익	순기여
A구역(콜)	200	22.00%	44	1,200,000원	748,000원	-452,000원
B구역(메시지+ 라이트 쿠폰)	400	6.00%	24	96,000원	408,000원	312,000원
C구역(라이트 터 치)	400	1.00%	4	0원	68,000원	68,000원
합계	1,000	-	72	1,296,000원	1,224,000원	-72,000원

현재 가정에서는 전체 순기여가 -72,000원으로 소폭 음수다. 핵심 원인은 A구역 콜의 비용이다. 콜 1건당 비용 6,000원, 유지 고객 1인당 추가 마진 17,000원이므로 콜이 손익분기점을 넘으려면 성공률이 $6,000/17,000=35.3\%$ 이상이어야 한다. 주어진 성공률 22%는 이 기준보다 낮다.

[문제 2의 답안 샘플]

콜은 단가가 높기 때문에 Top 10% 전체에 넓게 쓰기보다 Top 5% 또는 결제 실패·고가치·결제 임박 고객으로 좁히는 것이 적절하다. 검증셋 기준 Top 5%는 정밀도 27.00%, 리프트 10.00배로 Top 10%보다 더 농축된 위험군이다. 반면 메시지와 라이트 터치는 비용이 낮으므로 Top 10~15%까지 넓혀 재현율을 높이는 운영이 가능하다.

운영 목표	권장 컷	권장 채널	이유
정밀도 우선	Top 5%	상담 콜, 결제 복구 콜	고비용 채널이므로 실제 이탈 가능성이 가장 높은 고객에게 집중
균형	Top 10%	콜+메시지 조합	정밀도와 재현율의 균형이 가장 무난

운영 목표	권장 컷	권장 채널	이유
재현율 우선	Top 15%	메시지, 앱 푸시, 라이트 쿠폰	저비용 채널로 더 많은 이탈자를 포착

실제 고객 커뮤니케이션에서는 “곧 이탈하실 것 같아서 연락드렸습니다”, “해지 위험 고객으로 분류되었습니다”, “탈퇴 예정으로 보여서” 처럼 예측 결과를 직접 드러내는 표현을 피해야 한다. 대신 “최근 이용 중 불편한 점이 없으셨는지 확인드리고 싶습니다”, “결제/재생 경험을 더 원활하게 도와드리겠습니다”, “선호하실 만한 콘텐츠와 혜택을 정리해 드렸습니다”처럼 고객 경험 개선 관점의 언어를 사용해야 한다.

최종 정리:

PlayNow 이탈 예측의 핵심 신호는 최근 결제 실패, 최근 활동 감소, 핵심 콘텐츠 재생 감소, 세션 급감, VOC·품질 문제다. 로지스틱 회귀는 검증셋 기준 ROC-AUC 0.849와 Top 10% 리프트 6.11배를 보여 기준선 모델로 충분히 의미가 있다. 의사결정나무는 결제 실패와 저 활동 조합을 현업 규칙으로 설명하는 데 유용하다. 실행 전략은 고비용 쿨을 Top 5% 또는 결제 실패·고가치·결제 임박 고객으로 좁히고, 나머지 위험군은 메시지와 라이트 터치 중심으로 운영하는 방향이 바람직하다.