

신용카드 사기 거래 탐지 AI 경진대회

EDA 보고서 (목차/표·그림 번호/캡션/해석 포함)

생성일: 2026-01-09

입력 데이터: train.csv(113,842; unlabeled), val.csv(28,462; labeled/학습불가), test.csv(142,503; unlabeled)

분석 목표: (1) 데이터 품질/무결성 확인, (2) train-test 분포 드리프트 점검, (3) val 라벨 기반 사기 시그널 추출, (4) 이상탐지 기준의 FP/FN 원인과 임계값 민감도 파악.

목차

Placeholder for table of contents	0
-----------------------------------	---

1. 데이터 개요

본 대회 데이터는 V1-V30의 비식별화 수치 피처로 구성되며, 학습 가능한 라벨이 존재하지 않는 train/test와 라벨이 포함되지만 학습에 사용할 수 없는 val로 나뉜다. 따라서 EDA는 (i) 데이터 품질과 누수 가능성 점검, (ii) 분포 기반 이상탐지의 안정성을 좌우하는 드리프트 및 꼬리 분포(heavy-tail) 특성 파악, (iii) val 라벨을 이용한 사기 시그널(feature signal) 추출을 중심으로 수행한다.

Split	Rows	Labeled
train	113842	No
val	28462	Yes (학습불가)
test	142503	No

표 1. 데이터셋 구성 요약

2. 데이터 품질 및 무결성(Integrity)

ID 중복, split 간 교집합(누수), 결측/무한대 값은 이상탐지 대회에서 기본적으로 제거해야 하는 실패 요인이다. 본 데이터는 해당 리스크가 매우 낮은 상태로 확인되었다.

dataset	rows	unique_ids	duplicate_ids
train	113842	113842	0
val	28462	28462	0
test	142503	142503	0

표 2. ID 유일성 및 중복 여부

pair	overlap
train \cap val	0
train \cap test	0
val \cap test	0

표 3. 데이터셋 간 ID 교집합(누수 가능성)

metric	value
NaN_total(train)	0
NaN_total(val)	0
NaN_total(test)	0
Inf_total(train)	0
Inf_total(val)	0
Inf_total(test)	0

표 4. 결측/무한대 값 총합(전체 피처 합산)

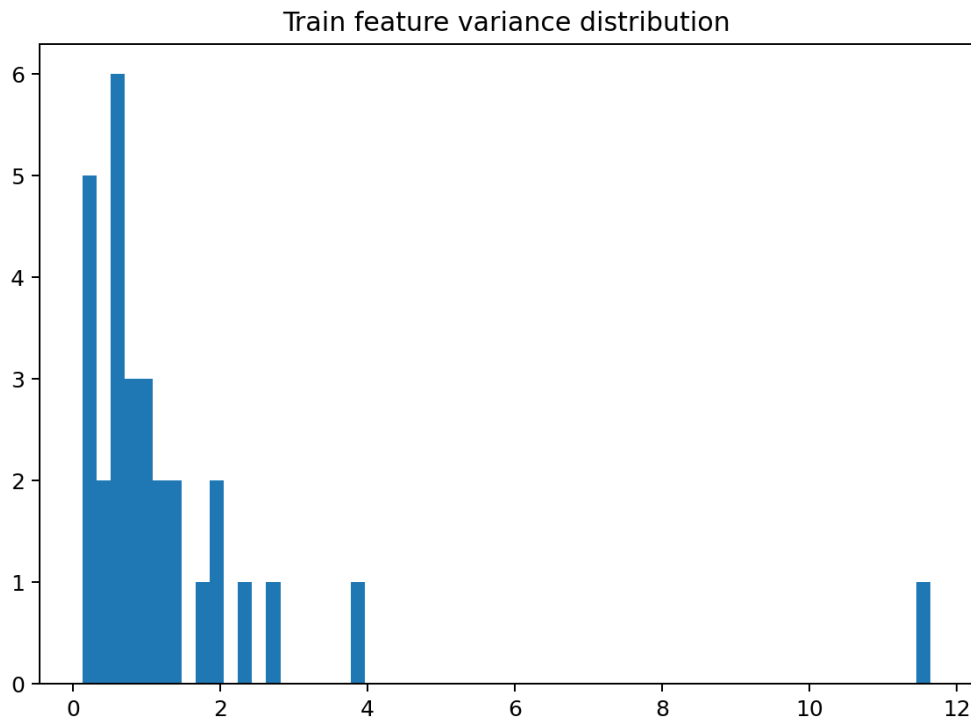


그림 1. Train 피쳐 분산 분포(분산이 매우 작은 피쳐는 준상수 후보).

3. Validation 라벨 분포 및 불균형성

Validation의 사기(Class=1) 비율은 0.001054 (약 0.1054%)로 극단적으로 작다. Macro F1에서는 정상(0)과 사기(1) F1을 동일 가중치로 평균하므로 FP(정상->사기)가 증가하면 정상 F1이 붕괴하여 점수가 급락할 수 있다. 따라서 후보 생성과 FP 억제를 분리하는 전략이 유리하다.

class	count	ratio
0	28432	0.998946
1	30	0.00105404

표 5. Validation 클래스 분포

4. 단변량 분포 특성 및 heavy-tail 분석

heavy-tail이 강한 피쳐는 정상에서도 극단값이 발생할 수 있어 이상탐지 모델이 이를 사기 신호로 오인(FP)할 위험이 있다. kurtosis와 분위수/극단값을 통해 위험 피쳐를 선별한다.

feature	train_kurtosis	train_skew	train_min	train_q99	train_max	abs_skew
V28	1381.59	19.5303	-9.61792	0.532718	33.8478	19.5303
V23	523.993	-7.58797	-44.8077	1.5189	22.5284	7.58797
V29	275.474	11.8209	-0.307413	13.9993	180.101	11.8209
V8	192.641	-8.2232	-50.9434	2.06009	20.0072	8.2232
V21	185.925	3.5988	-22.7575	1.94645	27.2028	3.5988
V20	167.781	-2.05612	-28.0096	2.39907	26.2374	2.05612
V2	109.487	-5.08187	-72.7157	3.80141	21.4672	5.08187
V7	100.876	0.580314	-41.5068	2.7436	44.0545	0.580314
V27	87.6568	-2.32878	-9.89524	0.946744	11.1357	2.32878

V17	62.3873	-2.19019	-21.2979	2.2615	9.25353	2.19019
-----	---------	----------	----------	--------	---------	---------

표 6. Train heavy-tail 상위 피쳐(상위 kurtosis 기준, 상위 10개)

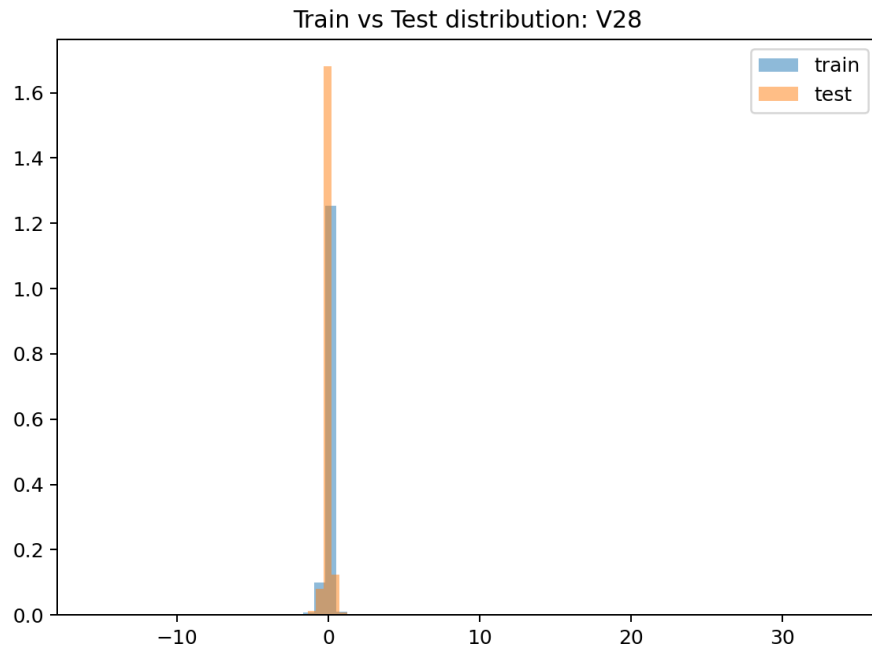


그림 2. V28의 Train vs Test 분포 오버레이(heavy-tail 예시).

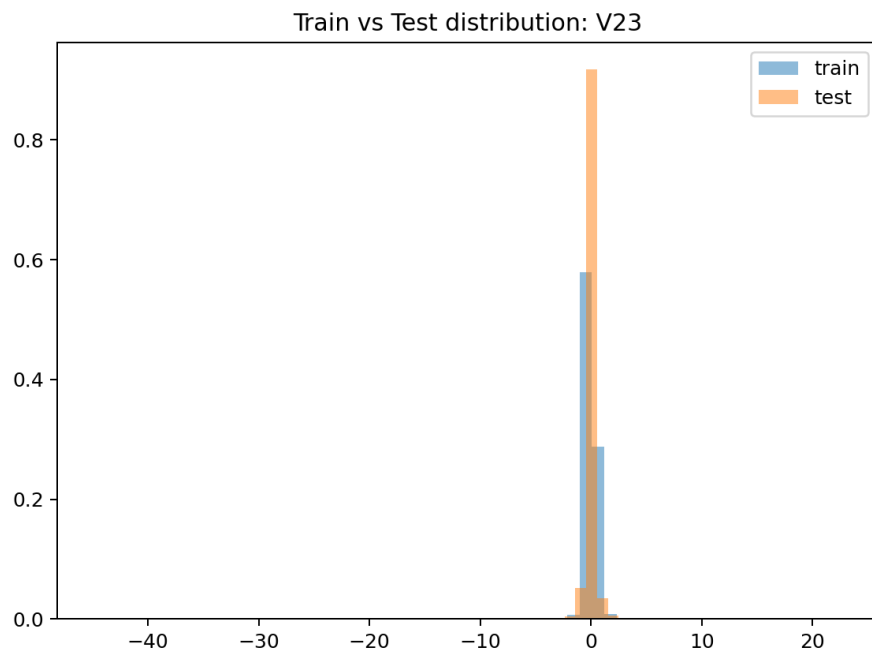


그림 3. V23의 Train vs Test 분포 오버레이(heavy-tail 예시).

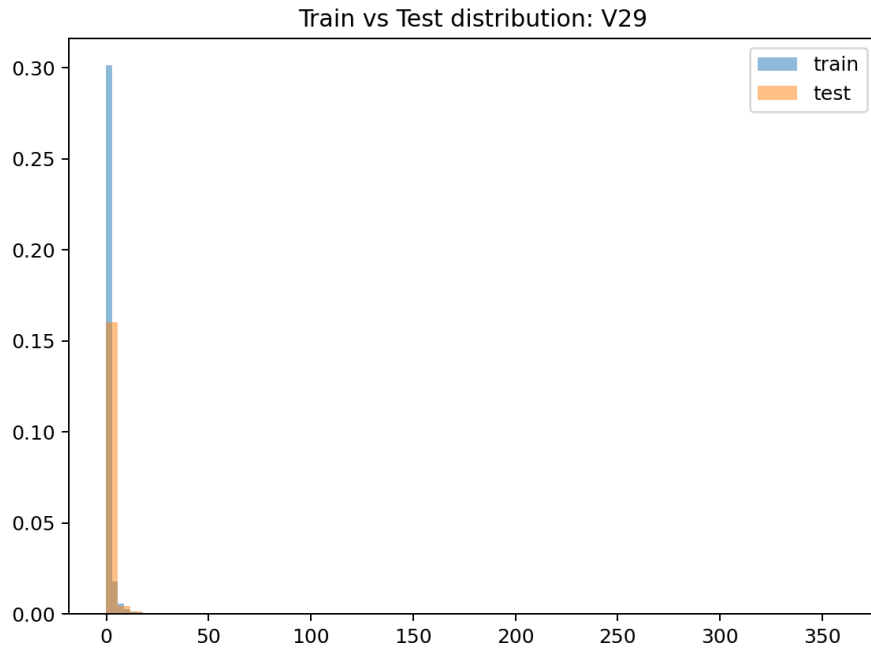


그림 4. V29의 Train vs Test 분포 오버레이(극단값 예시).

5. 분포 드리프트(Train-Val/Test) 분석

feature-wise로 KS 통계량, Wasserstein distance, PSI를 계산했다. PSI는 0.1 이상이면 유의미한 shift로 해석하는 것이 일반적이다.

feature	PSI_train_test	KS_train_test	Wass_train_test
V3	0.000544409	0.00476251	0.017969
V28	0.000520173	0.00332403	0.00386829
V14	0.000465076	0.00369801	0.011631
V6	0.000395643	0.00337825	0.00758424
V23	0.000365988	0.00405489	0.00336526
V17	0.000360315	0.00355874	0.0111663
V8	0.000354189	0.00453978	0.00853673
V7	0.000347613	0.00321969	0.0144874
V25	0.000325992	0.00521471	0.00410991
V24	0.000306557	0.00306729	0.00212194

표 7. Train vs Test 드리프트 상위 10개 피처(PSI 내림차순)

PSI 값이 전반적으로 매우 작아(train-test shift가 사실상 미미) train 기반 이상탐지 기준의 일반화 가능성이 높다.

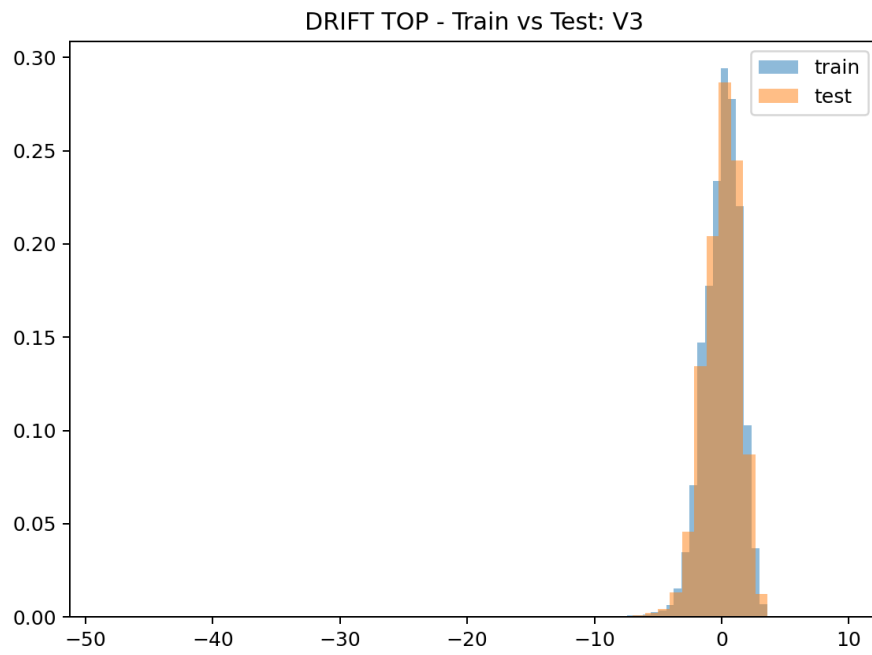


그림 5. 드리프트 상위 피처의 분포 비교: V3

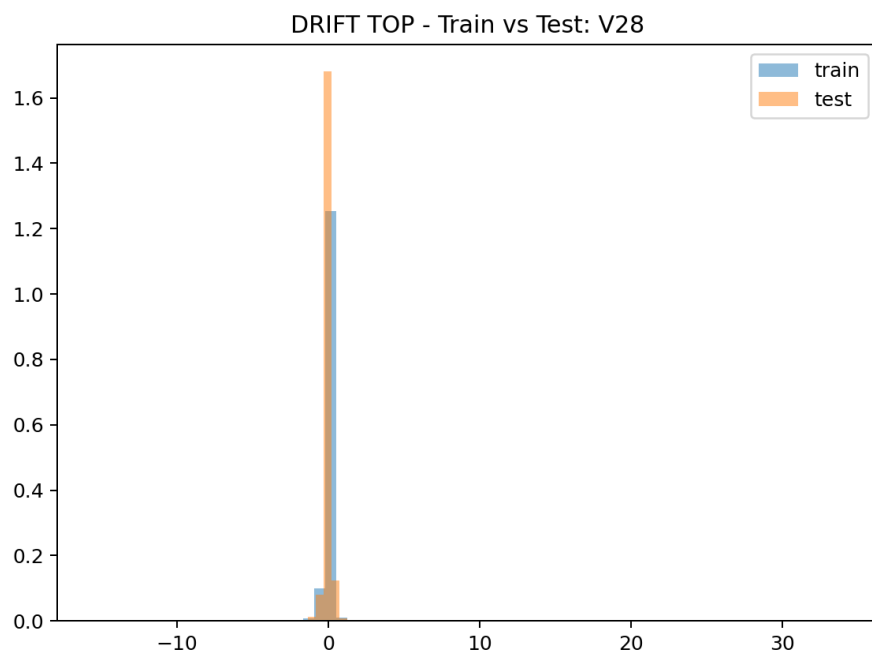


그림 6. 드리프트 상위 피처의 분포 비교: V28

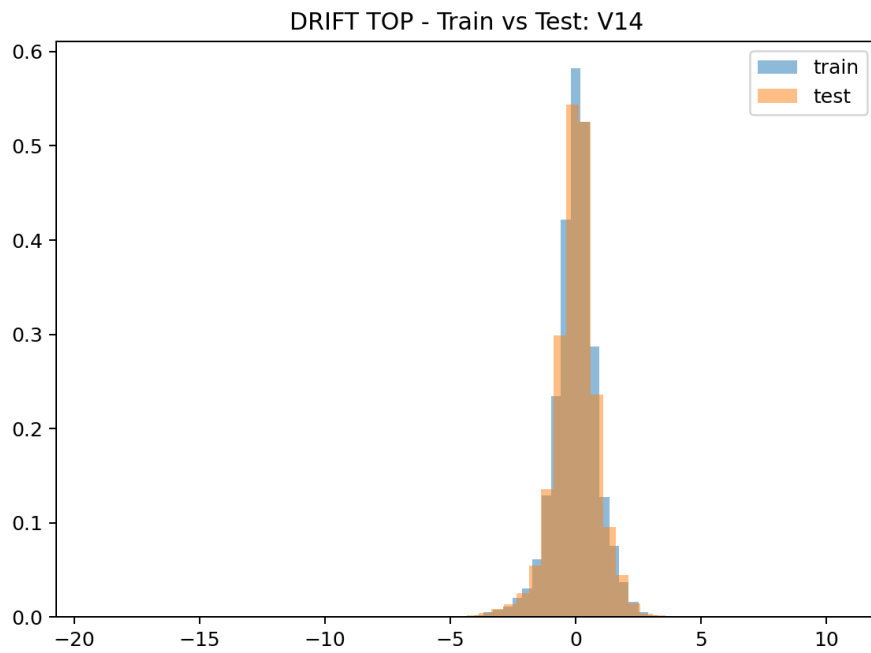


그림 7. 드리프트 상위 피처의 분포 비교: V14

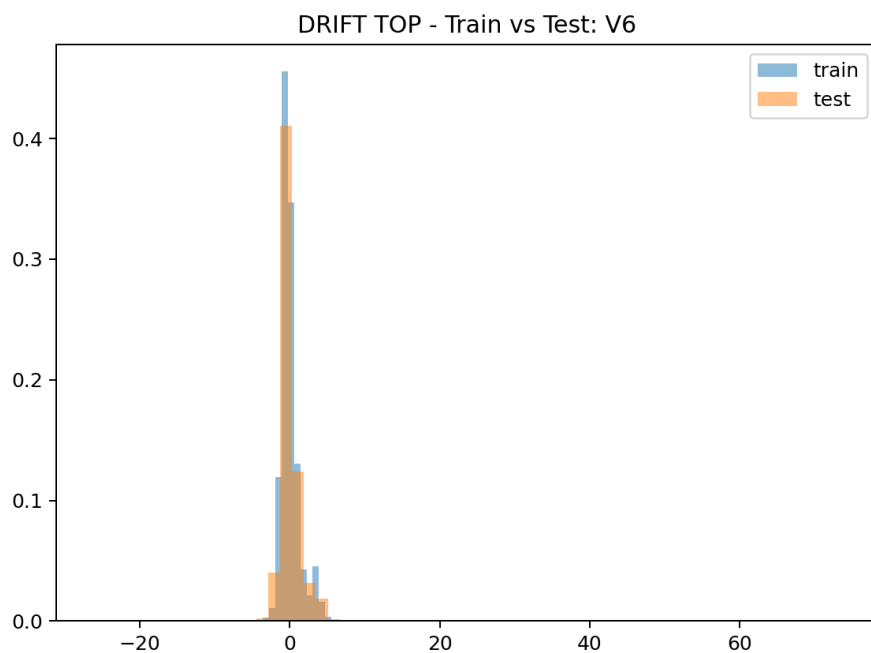


그림 8. 드리프트 상위 피처의 분포 비교: V6

6. Validation 라벨 기반 사기 시그널(피처 분리력) 분석

ranksum p-value, Cohen's d, 단일 피처 AUC, 사기 중앙값의 정상 IQR 이탈 여부를 사용해 분리력을 평가했다.

feature	ranksum_p	cohens_dauc	single_feature_auc	fraud_outside_IQR
V10	5.97452e-19	-7.17665	0.0310683	1
V14	7.4318e-19	-8.67085	0.0323485	1
V11	2.94609e-17	4.42336	0.945527	1
V4	8.61658e-17	3.9187	0.938871	1

V12	1.01255e-16	-7.59389	0.0621389	1
V3	1.34349e-15	-6.69648	0.0786321	1
V9	9.13257e-15	-3.05548	0.0912751	1
V2	1.69578e-14	2.99784	0.904561	1
V7	9.15177e-12	-6.77795	0.140398	1
V6	4.58275e-11	-1.40904	0.152809	0
V16	6.80996e-10	-5.30146	0.174612	1
V1	1.12602e-08	-3.38805	0.198863	0

표 8. Val 기준 피쳐 분리력 상위 12개

상위 피쳐에서 p-value가 매우 작고 $|d|$ 가 크며, 사기 중앙값이 정상 IQR 밖인 피쳐가 다수다. 이는 IQR 기반 피쳐 선택/후처리 규칙의 근거가 된다(단, 사기 표본 수가 작으므로 과적합 주의).

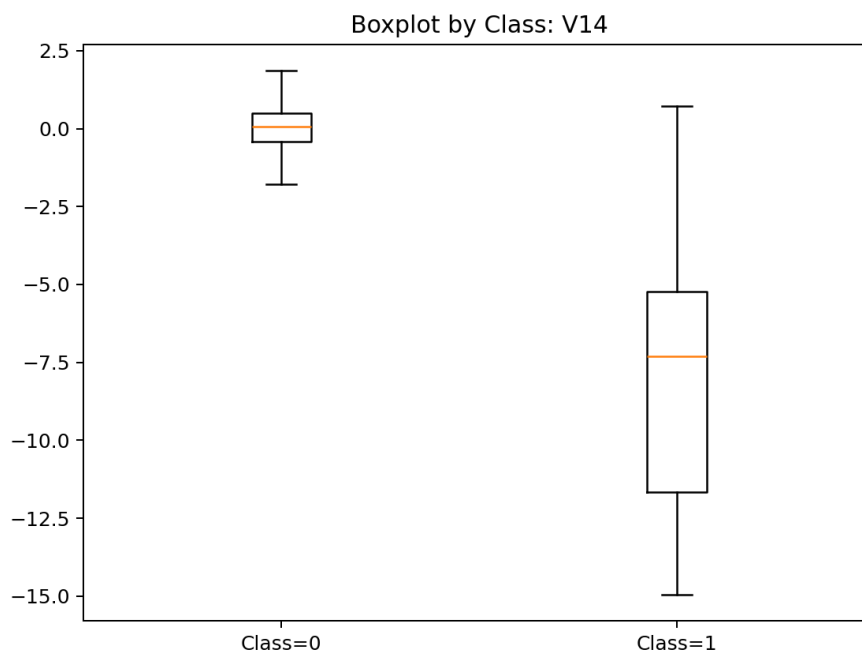


그림 9. Val 클래스별 boxplot: V14

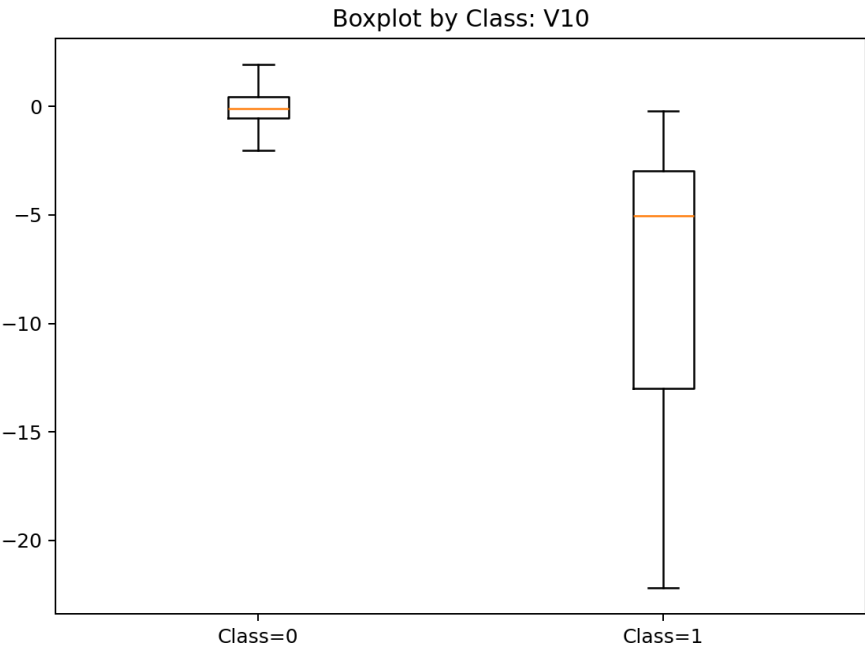


그림 10. Val 클래스별 boxplot: V10

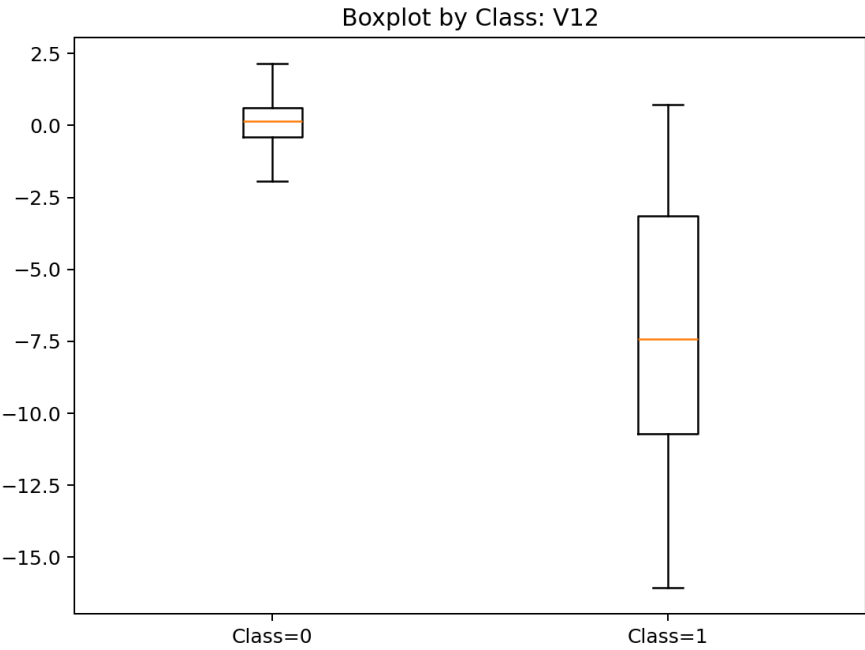


그림 11. Val 클래스별 boxplot: V12

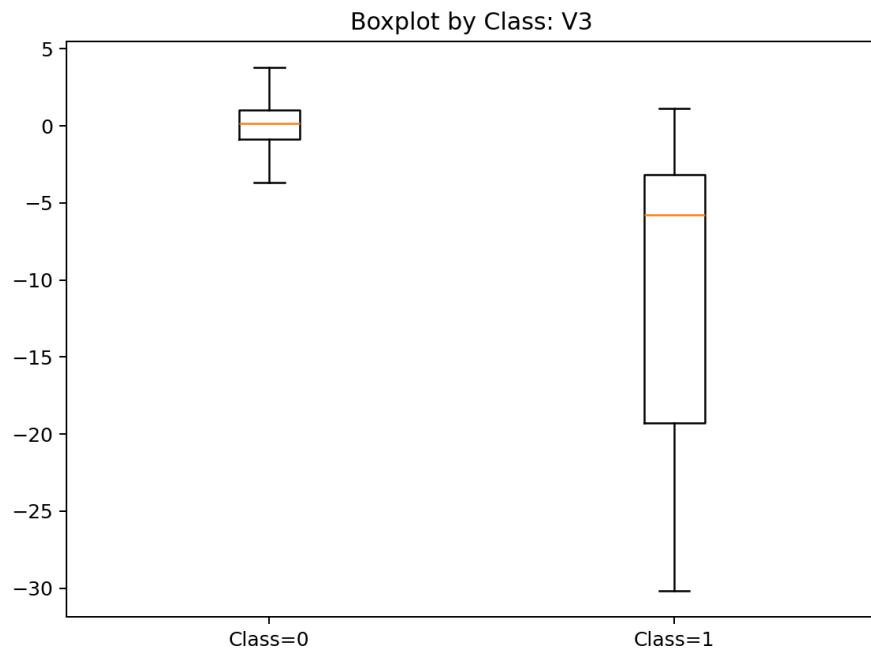


그림 12. Val 클래스별 boxplot: V3

7. 피처 상관 및 중복(Redundancy) 분석

Train 피처 상관쌍을 정리하여 중복 축 후보를 확인했다. 중복 축은 차원만 증가시키고 이상탐지 분할 효율을 떨어뜨릴 수 있다.

f1	f2	corr	abs_corr
V2	V29	-0.530737	0.530737
V3	V30	-0.429787	0.429787
V7	V29	0.391239	0.391239
V5	V29	-0.382747	0.382747
V20	V29	0.348219	0.348219
V11	V30	-0.247051	0.247051
V1	V29	-0.231744	0.231744
V25	V30	-0.231473	0.231473
V6	V29	0.210381	0.210381
V3	V29	-0.207185	0.207185
V15	V30	-0.181832	0.181832
V5	V30	0.175038	0.175038
V22	V30	0.14425	0.14425
V12	V30	0.126405	0.126405
V1	V30	0.118519	0.118519

표 9. Train 상관 상위 피처쌍 (상위 15개)

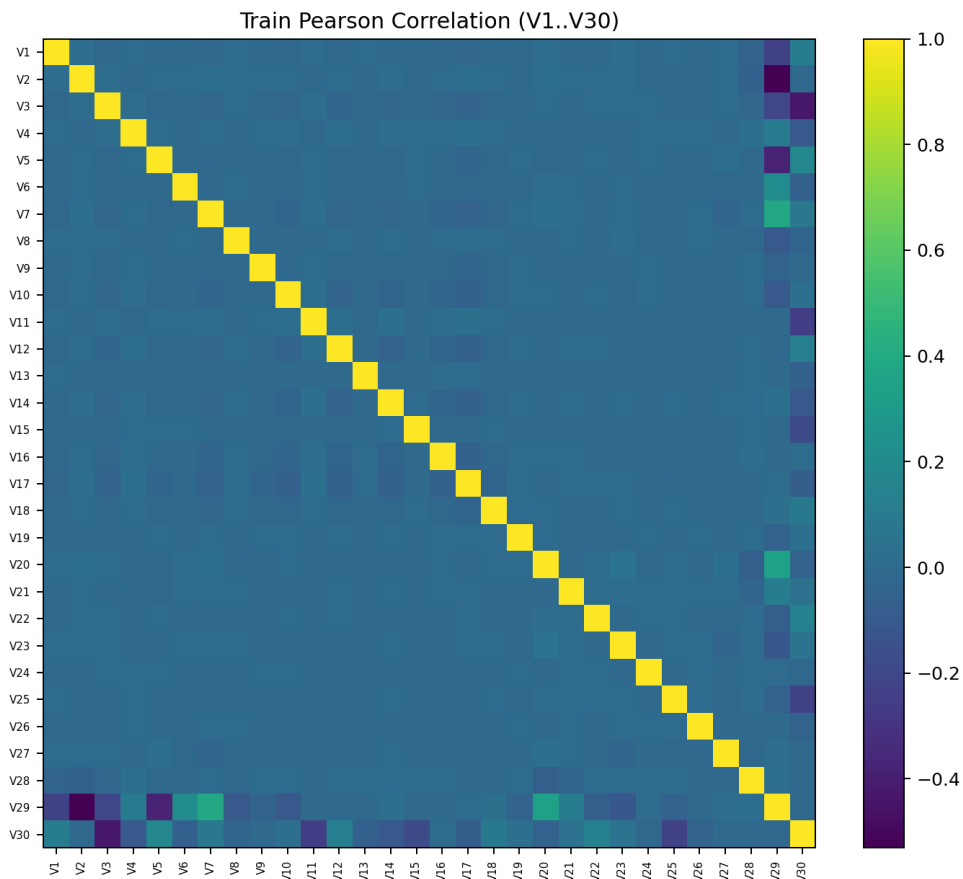


그림 13. Train 피쳐 상관 히트맵(Pearson).

8. PCA 기반 구조 확인(차원축소 EDA)

PCA 2차원 설명 분산 비율은 $PC1=0.320$, $PC2=0.088$ (합 0.408)이다. 저차원에서도 일부 구조가 보존되므로 비선형 임베딩 적용 시 분리 가능성이 있다.

pca_explained_var_ratio_dim1	pca_explained_var_ratio_dim2
0.320271	0.0875638

표 10. PCA 설명 분산 비율(2차원)

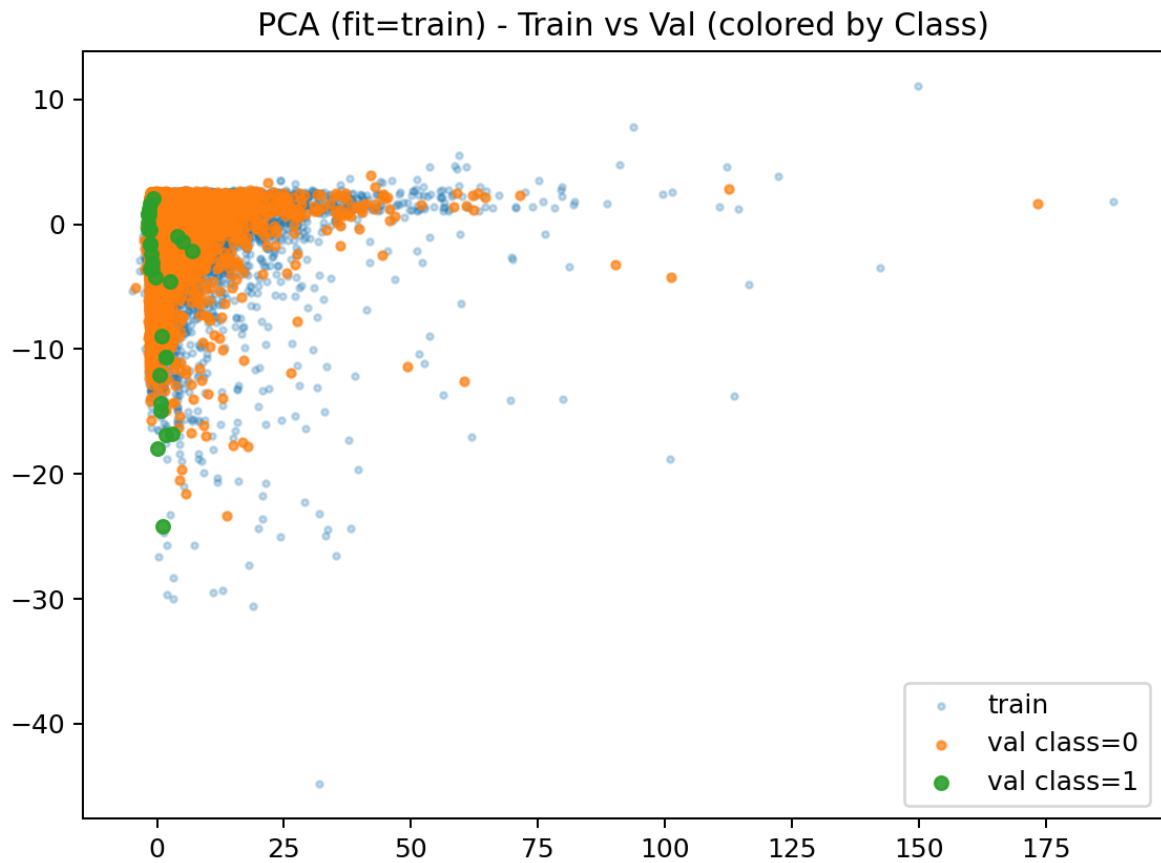


그림 14. PCA (fit=train)에서 Train vs Val(라벨 색칠) 시각화.

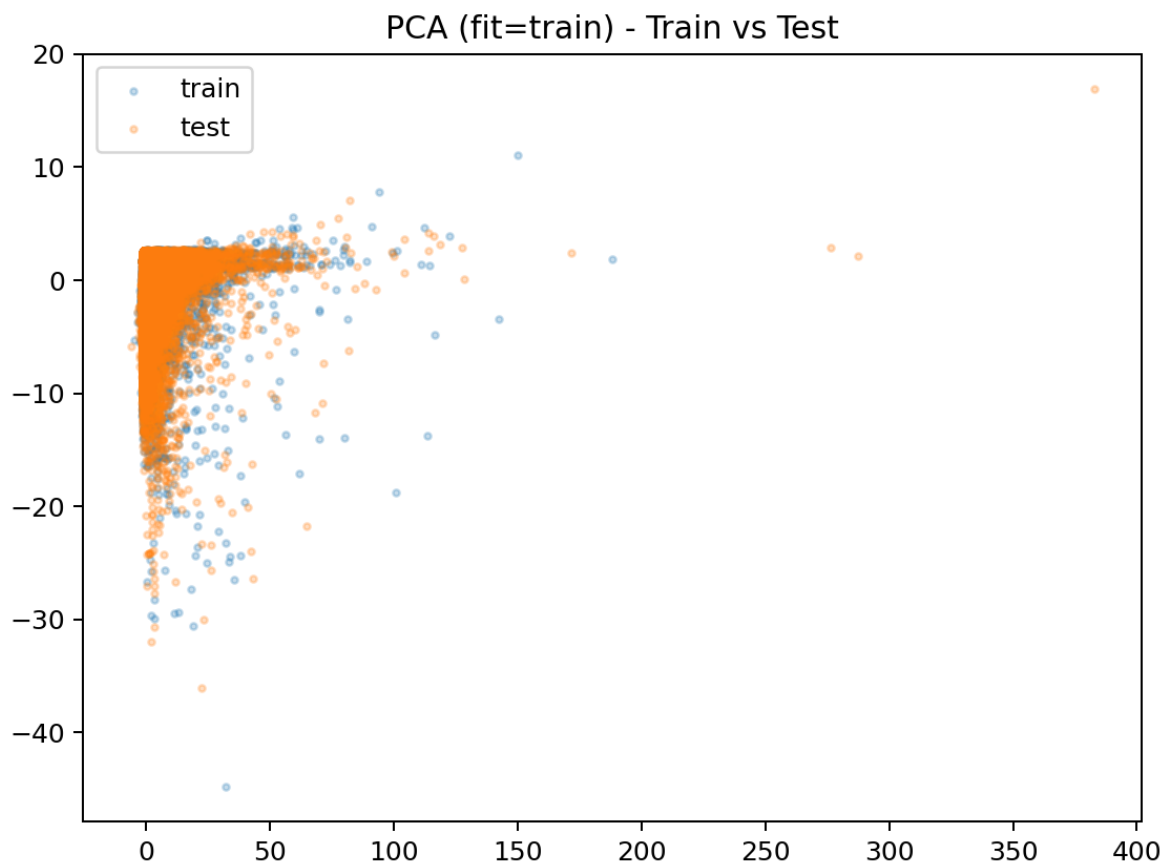


그림 15. PCA (fit=train)에서 Train vs Test 시각화.

9. IsolationForest 기반 이상점수/임계값 EDA

IsolationForest를 train에만 fit한 뒤 val에서 성능을 측정하고, contamination sweep 및 Recall@k로 threshold/랭킹 특성을 확인했다.

contamination	macro_f1	TN	FP	FN	TP	predicted_ones
0.00105404	0.69613	28417	15	19	11	26

표 11. IsolationForest baseline (fit=train) - val 혼동행렬 요약

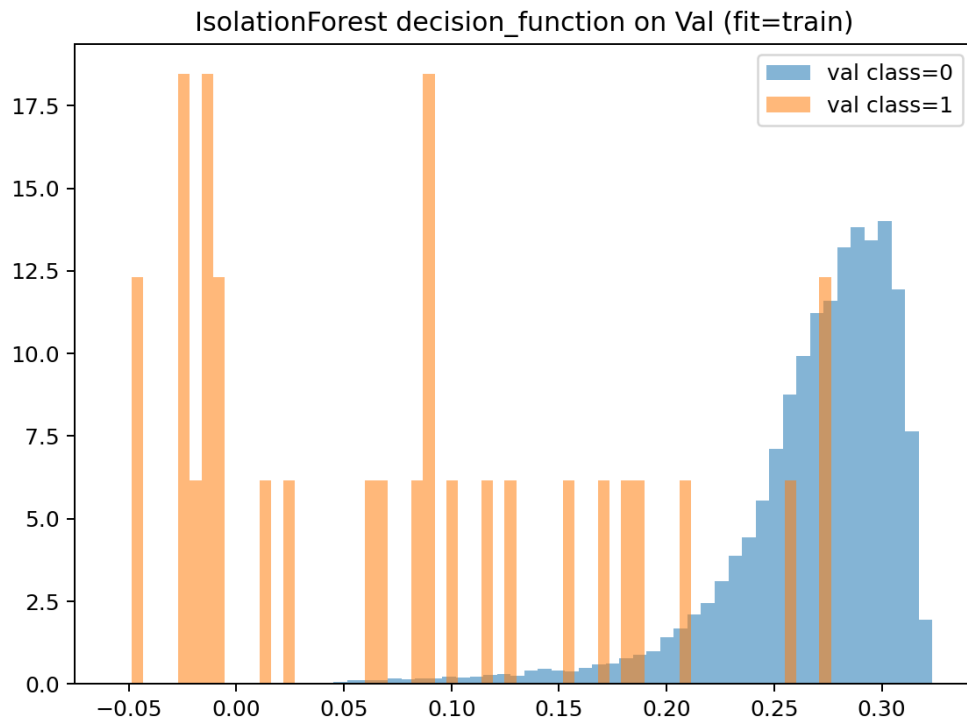


그림 16. IsolationForest decision_function 점수 분포(Val).

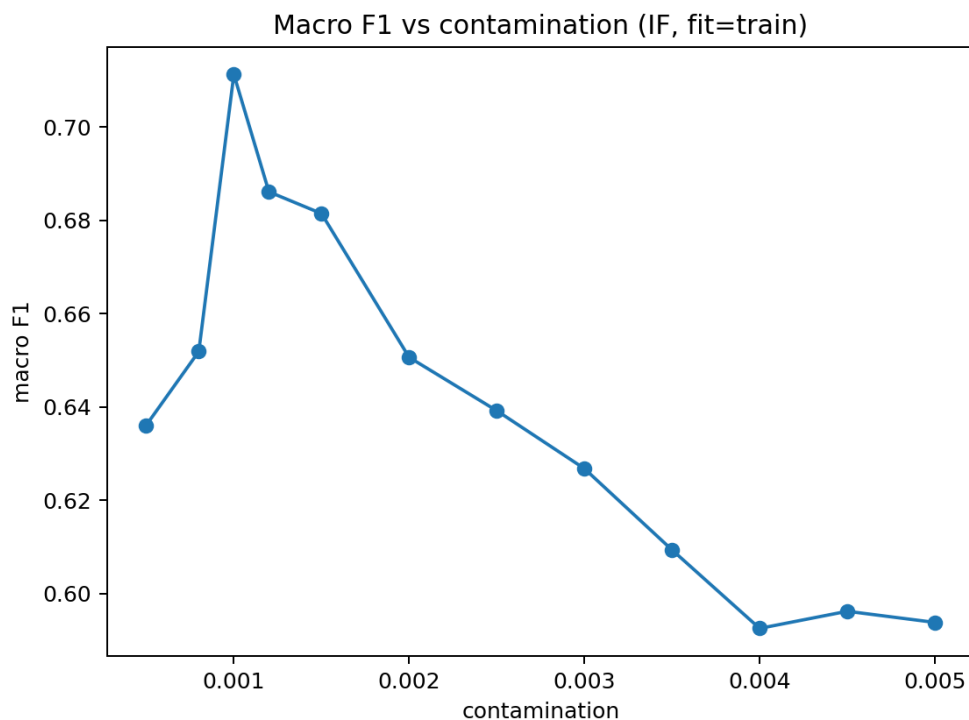


그림 17. contamination 변화에 따른 Val macro F1.

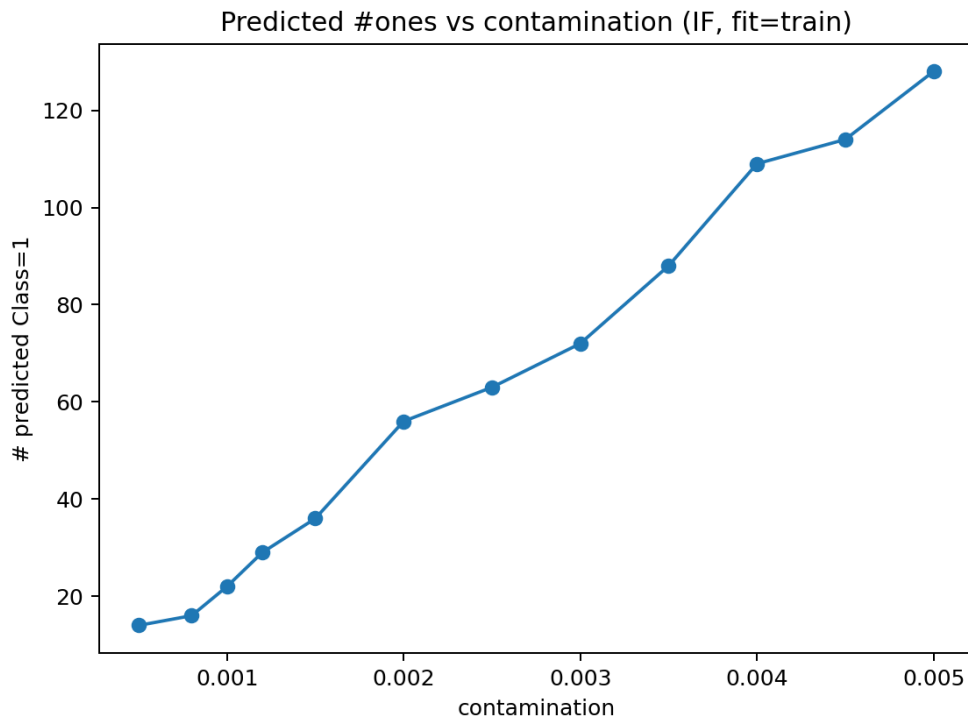


그림 18. contamination 변화에 따른 예측 Class=1 개수.

k	hits_fraud_in_topk	fraud_total	recall_at_k
10	4	30	0.133333
20	11	30	0.366667
30	11	30	0.366667
50	13	30	0.433333
100	13	30	0.433333
200	15	30	0.5
300	20	30	0.666667
500	22	30	0.733333
1000	23	30	0.766667

표 12. Val 이상치 상위 k개 내 사기 포함률(Recall@k).

Recall@k가 유의미하면, 모델이 사기를 상위 이상치로 올려두는 랭킹 능력은 있으나 최종 컷과 FP 제어가 성능을 좌우한다. outlier pool + FP 제거(투표/규칙) 구조가 유리하다.

10. FP/FN 케이스 스터디: 오탐/미탐 원인 피처

FP/FN 상위 사례를 추출하고 train 통계 기준 mean |z|로 극단 피처를 요약했다.

feature	FP_top50_mean_abs_z
V20	14.8985
V23	13.3995
V2	11.969
V29	10.5422

V28	9.9325
V1	9.83084
V21	8.79064
V27	8.71077
V7	7.65295
V5	6.32794
V8	5.37746
V3	4.92745
V4	4.47768
V22	4.21122
V25	4.12574

표 13. FP 상위 50개 샘플에서 mean |z| 상위 피처(15개).

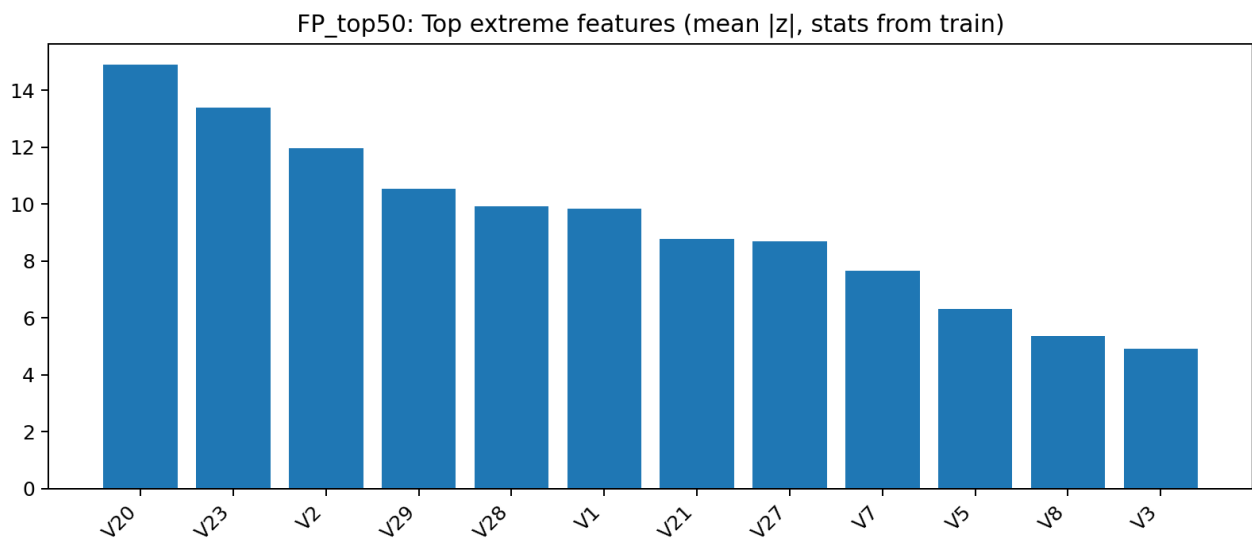


그림 19. FP 상위 50개에서 극단 피처(평균 |z|) 막대그래프.

feature	FN_top50_mean_abs_z
V14	6.51639
V17	4.85218
V12	4.18325
V10	3.05462
V11	2.88782
V3	2.64713
V4	2.63143
V16	2.4698
V18	1.94951
V7	1.79524
V9	1.75748
V2	1.53953

V27	1.36577
V19	1.13884
V1	1.12696

표 14. FN 상위 50개 샘플에서 mean |z| 상위 피쳐(15개).

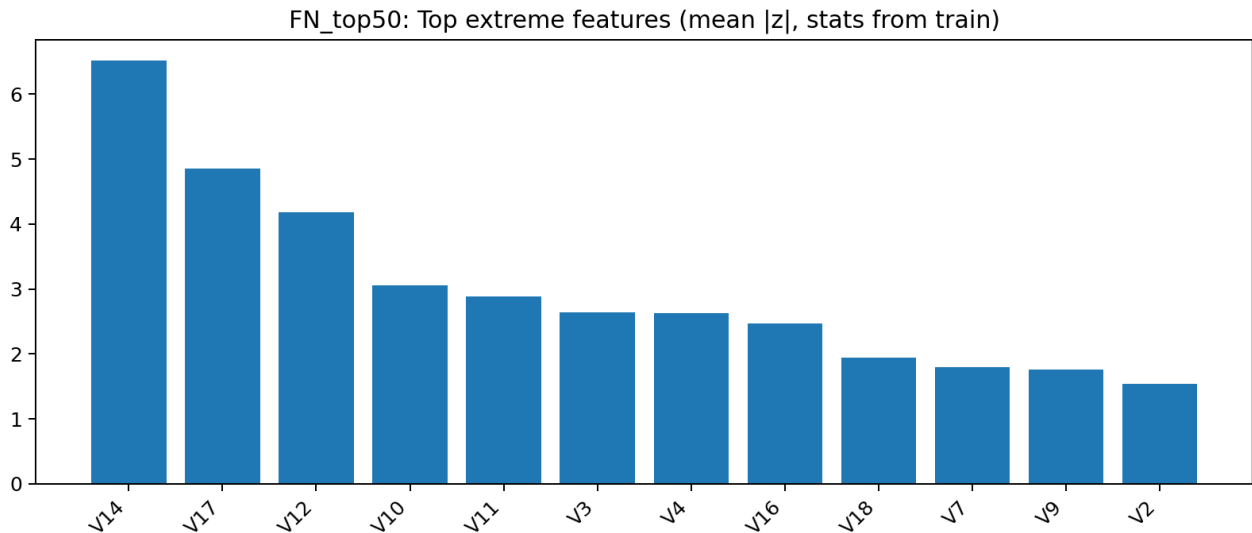


그림 20. FN 상위 50개에서 극단 피쳐(평균 |z|) 막대그래프.

11. 결론 및 다음 단계

데이터 품질은 매우 양호하고(train/val/test 중복·결측·누수 없음), train-test 드리프트가 작아 일반화 가능성이 높다. Val 라벨 기준 사기 시그널이 강한 피쳐가 존재하며, heavy-tail 피쳐는 FP 유발 위험이 있다. IsolationForest는 랭킹 능력이 일부 있으나 threshold와 FP 제어가 성능을 좌우하므로 outlier pool + FP 제거 후처리 전략이 유리하다.

다음 단계 제안

- Val 분리력 상위 피쳐를 힌트로 사용하되, train 기반 비지도 피쳐 선택과 결합하여 과적합을 완화한다.
- heavy-tail 피쳐는 제거/다운웨이트 또는 robust 변환(클리핑/분위수 변환)으로 FP를 완화한다.
- 단일 threshold 튜닝(contamination sweep) 외에 outlier pool을 만들고 voting/규칙 기반으로 FP를 제거한다.
- PaCMAP/KernelPCA 등 임베딩 기반 분리와 앙상블(IF seed/피쳐셋)을 결합해 private 안정성을 강화한다.