EDA Report

Generated: 2026-02-04 15:41:28

Dataset Overview

- Data path: ./data/synthetic_aml_mixed_50k_20260204_153242.csv
- Rows used: 50000
- Target column: sar_actual
- Time column: txn_ts

Data Quality

- Dataset has 50000 rows and 50 columns.
- Duplicate row ratio: 0.00%.
- Highest missing rate is 60.44% in 'app_version'.

Missingness

Column	Missing Count	Missing Rate
app_version	30,218	60.44%
case_status	1,318	2.64%
risk_segment	1,288	2.58%
device_fingerprint	1,274	2.55%
case_notes	1,265	2.53%
merchant_category	1,255	2.51%
payment_reference	1,236	2.47%
merchant_name	1,235	2.47%
counterparty_id	1,234	2.47%
kyc_level	1,233	2.47%
payment_memo	1,223	2.45%
merchant_description	1,222	2.44%
device_id	1,217	2.43%
dest_country	1,188	2.38%
origin_country	1,170	2.34%
ip_risk_score	808	1.62%
customer_tenure_days	787	1.57%
geo_lon	773	1.55%
geo_lat	750	1.50%
account_age_days	749	1.50%
sum_amount_24h	749	1.50%
velocity_score	744	1.49%
fee_amount	742	1.48%
txn_amount	736	1.47%
num_txn_24h	729	1.46%
num_txn_1h	727	1.45%

Column	Missing Count	Missing Rate
account_balance	718	1.44%
is_crypto_related	557	1.11%
is_business_account	543	1.09%
sanctions_match	539	1.08%
is_international	524	1.05%
is_new_device	507	1.01%
is_pep	504	1.01%
is_high_risk_country	479	0.96%

The following columns have no missing values: txn_id, txn_ts, settlement_time, platform, channel, payment_rail, txn_type, txn_direction, account_id, customer_id (+6 more) (16 columns total).

Null-like / Placeholder Values

These values are not nulls but may represent missing or invalid data placeholders.

Column	Null-like Count	Null-like Rate	Example Values
counterparty_id	269	0.54%	unknown
merchant_name	241	0.48%	unknown
merchant_category	241	0.48%	unknown
merchant_description	238	0.48%	unknown
payment_memo	241	0.48%	unknown
payment_reference	247	0.49%	unknown
origin_country	225	0.45%	unknown
dest_country	230	0.46%	unknown
device_id	247	0.49%	unknown
device_fingerprint	257	0.51%	unknown
app_version	213	0.43%	unknown
kyc_level	265	0.53%	unknown
risk_segment	268	0.54%	unknown
case_status	261	0.52%	unknown
case_notes	272	0.54%	unknown

Column Type Classification

Type	Columns
numeric	account_age_days
numeric	customer_tenure_days
numeric	txn_amount
numeric	fee_amount
numeric	account_balance
numeric	num_txn_1h
numeric	num_txn_24h
numeric	sum_amount_24h
numeric	velocity_score
numeric	ip_risk_score
numeric	geo_lat

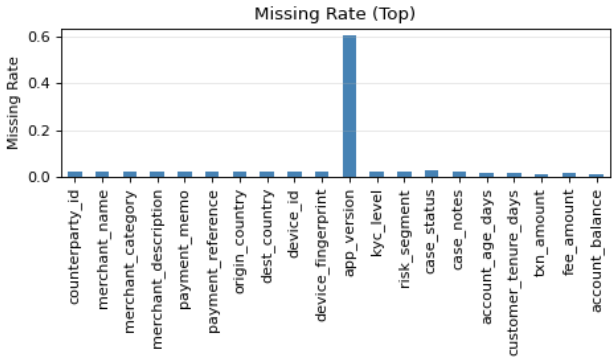
Type	Columns
numeric	geo_lon
numeric	sar_actual
categorical	platform
categorical	channel
categorical	payment_rail
categorical	txn_type
categorical	txn_direction
categorical	merchant_name
categorical	merchant_category
categorical	payment_memo
categorical	origin_country
categorical	dest_country
categorical	currency
categorical	device_type
categorical	app_version
categorical	account_type
categorical	kyc_level
categorical	risk_segment
categorical	case_status
categorical	is_international
categorical	is_new_device
categorical	is_high_risk_country
categorical	is_crypto_related
categorical	is_pep
categorical	sanctions_match
categorical	is_business_account
datetime	txn_ts
datetime	settlement_time
datetime	customer_birth_date
text	txn_id
text	account_id
text	customer_id
text	counterparty_id
text	merchant_description
text	payment_reference
text	device_id
text	device_fingerprint
text	ip_address
text	case_notes

Outlier Ratio (IQR)

Column	Outlier Ratio
sum_amount_24h	0.086

Column	Outlier Ratio
txn_amount	0.079
num_txn_1h	0.075
account_balance	0.063
num_txn_24h	0.011
velocity_score	0.008
customer_tenure_days	0
account_age_days	0
geo_lon	0
geo_lat	0

Charts



Target

- Target 'sar_actual' has 2 classes.
- Target rate over time plotted.

Target Distribution

Value	Count	Rate
0	46,849	93.70%
1	3,151	6.30%

Target Rate by dest_country

dest_country	Target Rate
PK	8.22%
RU	7.88%
NG	7.02%
UNKNOWN	6.52%
SG	5.90%
HK	5.89%
BR	5.83%
GB	5.81%
DE	5.73%
AE	5.68%

Target Rate by origin_country

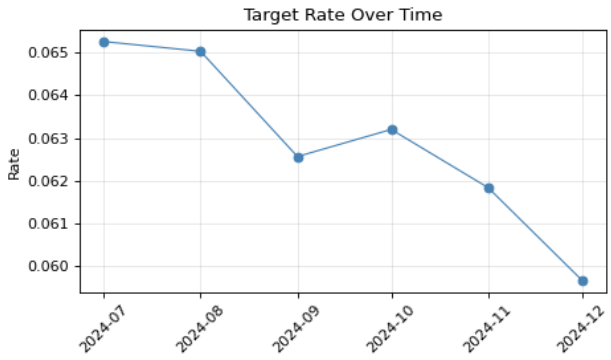
origin_country	Target Rate
NG	8.43%
RU	7.36%
DE	6.68%
AE	6.30%
SG	6.23%
UNKNOWN	6.22%
PK	6.20%
US	6.20%
CA	6.18%
GB	6.15%

Target Rate by txn_type

txn_type	Target Rate
crypto_trade	9.66%
p2p_transfer	6.85%
cash_out	6.72%
refund	6.62%
merchant_payment	6.45%
card_purchase	6.12%
deposit	6.10%

txn_type	Target Rate
ach	5.89%
cash_withdrawal	5.62%
bill_pay	5.58%

Charts



Univariate - Summary

Univariate

- Numeric columns analyzed: account_balance, customer_tenure_days, account_age_days, sum_amount_24h, geo_lon, txn_amount, geo_lat, num_txn_24h, num_txn_1h, velocity_score
- Categorical columns analyzed: dest_country, origin_country, txn_type, payment_memo, currency, merchant_category, merchant_name, channel, payment_rail, platform

Numeric Summary Statistics

Column	count	mean	std	min	25%	50%	75%	max
account_balance	49,282	4,122.414	3,911.714	111.920	1,741.565	2,984.245	5,098.545	83,883.760
customer_tenure_days	49,213	2,494.199	1,436.424	10	1,255	2,492	3,732	4,999
account_age_days	49,251	1,836.632	1,043.450	30	933	1,834	2,738	3,649
sum_amount_24h	49,251	226.802	355.855	0.590	54.165	119.140	258.305	12,788.150
geo_lon	49,227	0.525	97.862	-169.999	-84.104	0.733	85.599	169.999
txn_amount	49,264	44.503	58.425	0.530	13.690	26.980	52.550	1,726.020
geo_lat	49,250	-0.172	49.074	-84.994	-42.691	-0.290	42.299	85.000
num_txn_24h	49,271	4.832	2.220	0	3	5	6	16
num_txn_1h	49,273	1.563	1.266	0	1	1	2	9
velocity_score	49,256	1.731	0.867	0	1.109	1.676	2.291	5.905

Categorical Frequency (Top K)

Top K: dest_country

Category	Count	Rate
HK	4,496	8.99%
SG	4,478	8.96%
AE	4,436	8.87%
DE	4,433	8.87%
PK	4,427	8.85%
BR	4,424	8.85%
CA	4,422	8.84%
NG	4,402	8.80%
RU	4,393	8.79%
US	4,336	8.67%

Top K: origin_country

Category	Count	Rate
US	21,936	43.87%
DE	3,939	7.88%
GB	3,903	7.81%
CA	3,349	6.70%
SG	2,907	5.81%
AE	2,478	4.96%
BR	2,377	4.75%

Category	Count	Rate
HK	2,355	4.71%
NG	1,958	3.92%
RU	1,903	3.81%

Top K: txn_type

Category	Count	Rate
ach	10,349	20.70%
p2p_transfer	10,209	20.42%
card_purchase	6,012	12.02%
bill_pay	4,460	8.92%
merchant_payment	4,065	8.13%
deposit	3,032	6.06%
wire	3,021	6.04%
cash_withdrawal	2,973	5.95%
cash_out	2,874	5.75%
crypto_trade	1,947	3.89%

Top K: payment_memo

Category	Count	Rate
invoice	4,939	9.88%
loan	4,908	9.82%
rent	4,887	9.77%
refund	4,882	9.76%
salary	4,865	9.73%
subscription	4,840	9.68%
crypto	4,820	9.64%
gift	4,817	9.63%
tips	4,790	9.58%
family	4,788	9.58%

Top K: currency

Category	Count	Rate
USD	22,581	45.16%
EUR	4,051	8.10%
GBP	4,026	8.05%
CAD	3,435	6.87%
SGD	2,987	5.97%
AED	2,542	5.08%
BRL	2,455	4.91%
HKD	2,419	4.84%
NGN	2,004	4.01%
RUB	1,959	3.92%

Top K: merchant_category

Category	Count	Rate
travel	6,163	12.33%
gaming	6,138	12.28%
crypto	6,115	12.23%
utilities	6,087	12.17%
grocery	6,054	12.11%
services	6,046	12.09%
marketplace	5,984	11.97%
electronics	5,917	11.83%
nan	1,255	2.51%
UNKNOWN	241	0.48%

Top K: merchant_name

Category	Count	Rate
Foxtrot Groceries	7,064	14.13%
Echo Gaming	6,964	13.93%
Delta Travel	6,946	13.89%
Cobalt Services	6,905	13.81%
Bright Electronics	6,894	13.79%
Alpha Market	6,884	13.77%
P2P Contact	6,867	13.73%
nan	1,235	2.47%
UNKNOWN	241	0.48%

Top K: channel

Category	Count	Rate
mobile	23,123	46.25%
online	13,469	26.94%
web	4,932	9.86%
branch	4,372	8.74%
atm	3,062	6.12%
api	1,042	2.08%

Top K: payment_rail

Category	Count	Rate
p2p	14,125	28.25%
ach	13,377	26.75%
card	11,410	22.82%
wire	6,073	12.15%
cash	3,014	6.03%
crypto	2,001	4.00%

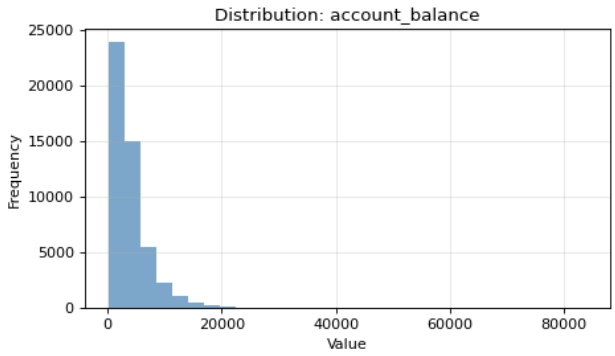
Top K: platform

Category	Count	Rate
bank	29,847	59.69%

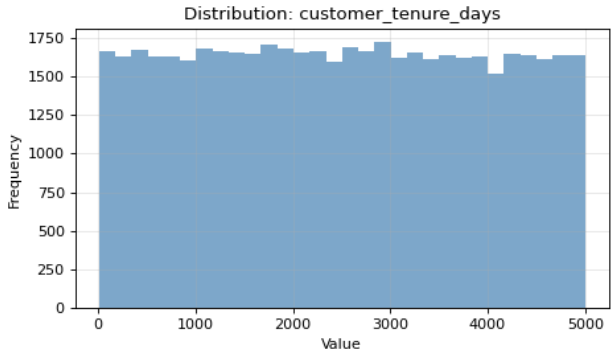
Category	Count	Rate
venmo	7,629	15.26%
paypal	6,071	12.14%
cashapp	3,949	7.90%
zelle	2,504	5.01%

Univariate Plots

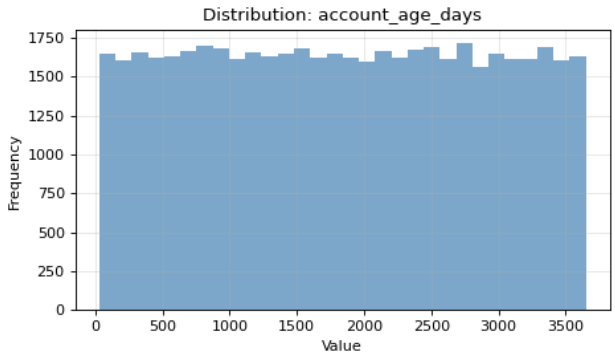
account_balance distribution



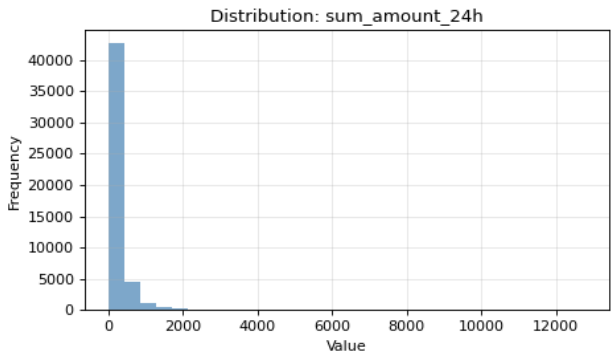
customer_tenure_days distribution



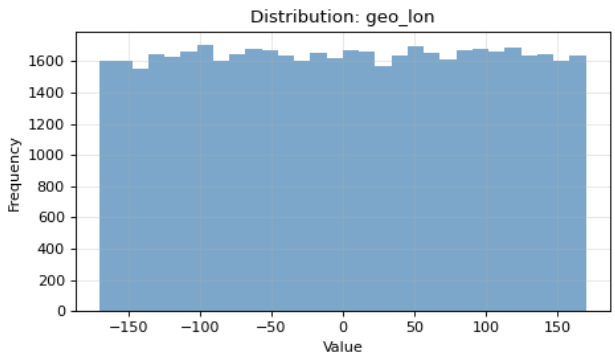
account_age_days distribution



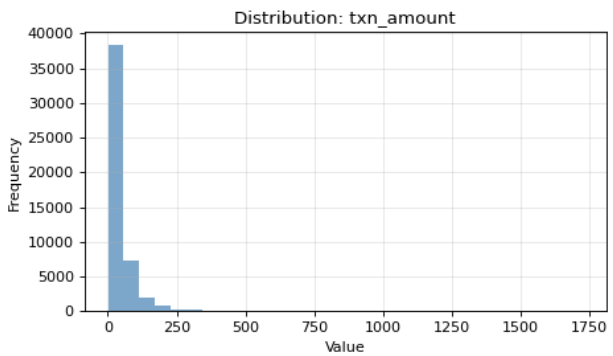
sum_amount_24h distribution



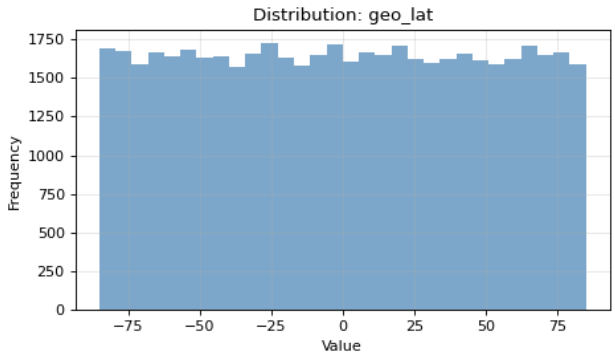
geo_lon distribution



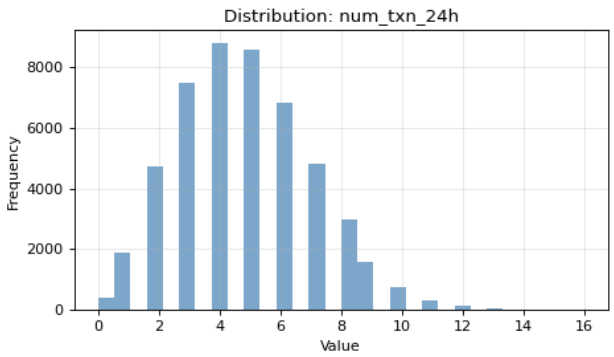
txn_amount distribution



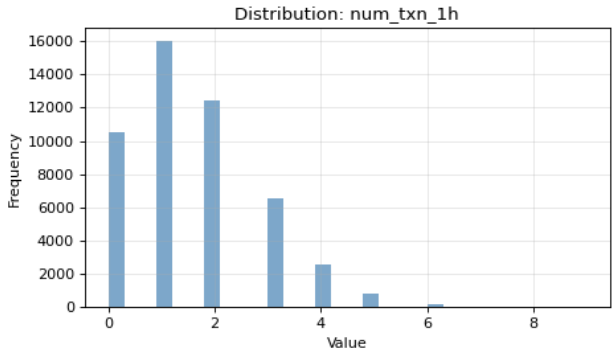
geo_lat distribution



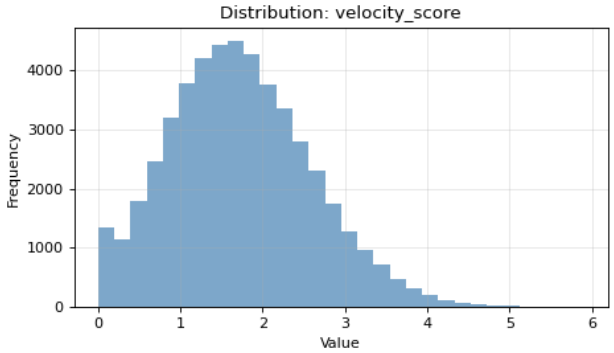
num_txn_24h distribution



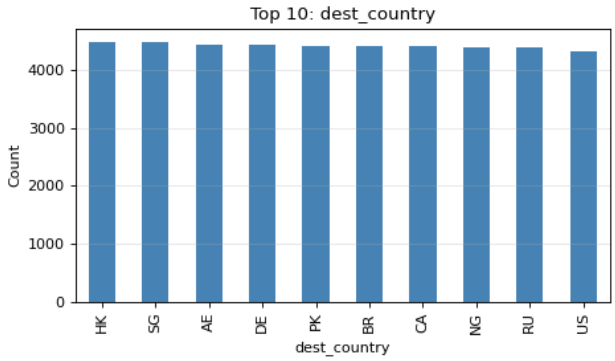
num_txn_1h distribution



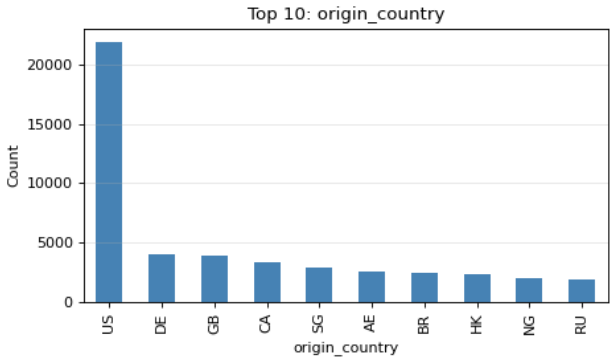
velocity_score distribution



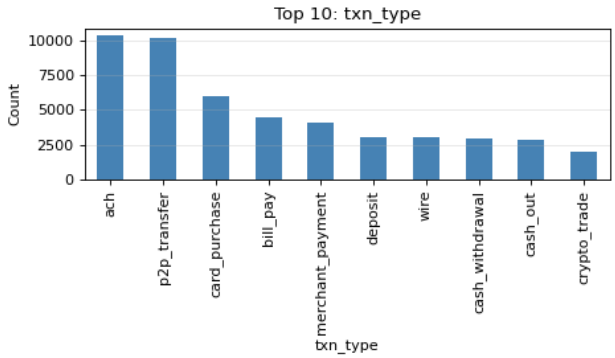
dest_country top 10



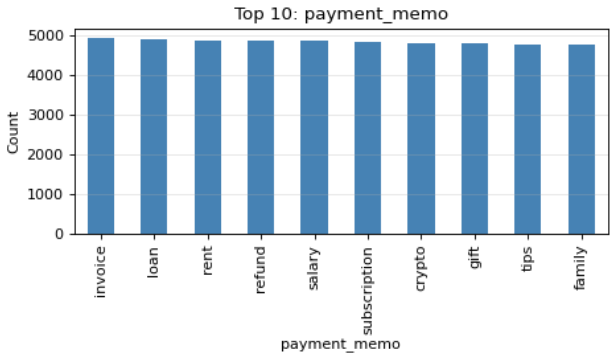
origin_country top 10



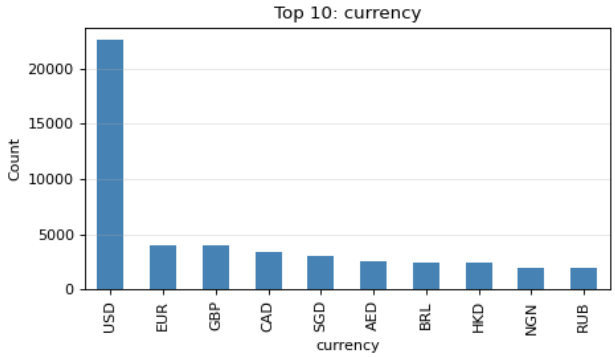
txn_type top 10



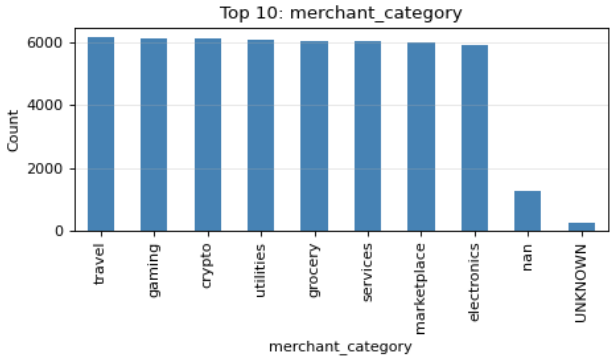
payment_memo top 10



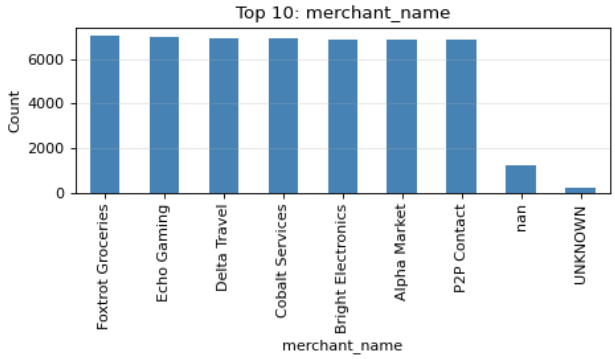
currency top 10



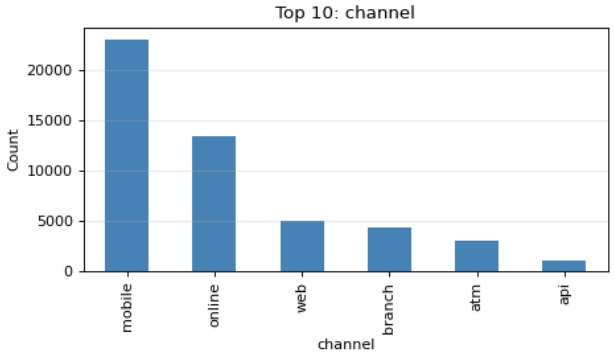
merchant_category top 10



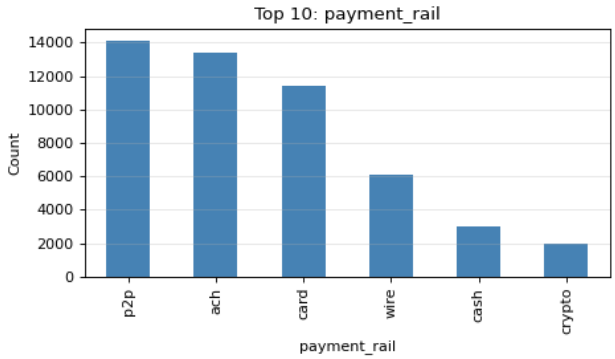
merchant_name top 10



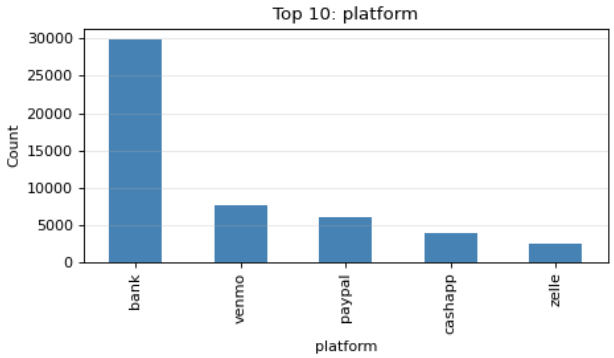
channel top 10



payment_rail top 10



platform top 10



Bivariate Target

- Bivariate target analysis completed.

Numeric vs Target (Binned)

Column	Bin	Target Rate
account_balance	(111.919, 1526.878]	6.73%
account_balance	(1526.878, 2436.358]	6.28%
account_balance	(2436.358, 3659.088]	5.96%
account_balance	(3659.088, 5837.766]	6.21%
account_balance	(5837.766, 83883.76]	6.31%
customer_tenure_days	(9.999, 1011.0]	6.21%
customer_tenure_days	(1011.0, 1991.8]	5.84%
customer_tenure_days	(1991.8, 2976.2]	6.72%
customer_tenure_days	(2976.2, 3981.0]	6.56%
customer_tenure_days	(3981.0, 4999.0]	6.24%
account_age_days	(29.999, 755.0]	5.99%
account_age_days	(755.0, 1473.0]	6.11%
account_age_days	(1473.0, 2199.0]	6.40%
account_age_days	(2199.0, 2921.0]	6.46%
account_age_days	(2921.0, 3649.0]	6.54%
sum_amount_24h	(0.589, 44.39]	6.13%
sum_amount_24h	(44.39, 88.64]	6.01%
sum_amount_24h	(88.64, 160.55]	5.98%
sum_amount_24h	(160.55, 312.56]	6.37%
sum_amount_24h	(312.56, 12788.15]	7.02%
geo_lon	(-170.0, -100.853]	6.49%
geo_lon	(-100.853, -33.445]	5.87%
geo_lon	(-33.445, 34.87]	6.09%
geo_lon	(34.87, 102.172]	6.46%
geo_lon	(102.172, 169.999]	6.65%
txn_amount	(0.529, 11.58]	5.95%
txn_amount	(11.58, 20.912]	6.28%
txn_amount	(20.912, 34.8]	6.22%
txn_amount	(34.8, 62.44]	6.59%
txn_amount	(62.44, 1726.02]	6.44%
geo_lat	(-84.995, -51.229]	6.08%
geo_lat	(-51.229, -17.276]	5.96%
geo_lat	(-17.276, 16.647]	6.66%
geo_lat	(16.647, 50.816]	6.15%
geo_lat	(50.816, 85.0]	6.73%
num_txn_24h	(-0.001, 3.0]	6.02%
num_txn_24h	(3.0, 4.0]	5.86%
num_txn_24h	(4.0, 5.0]	5.85%
num_txn_24h	(5.0, 7.0]	6.17%

Column	Bin	Target Rate
num_txn_24h	(7.0, 16.0]	8.48%
num_txn_1h	(-0.001, 1.0]	6.20%
num_txn_1h	(1.0, 2.0]	6.48%
num_txn_1h	(2.0, 3.0]	6.39%
num_txn_1h	(3.0, 9.0]	6.42%
velocity_score	(-0.001, 0.98]	6.03%
velocity_score	(0.98, 1.463]	5.95%
velocity_score	(1.463, 1.899]	5.92%
velocity_score	(1.899, 2.447]	6.20%
velocity_score	(2.447, 5.905]	7.38%

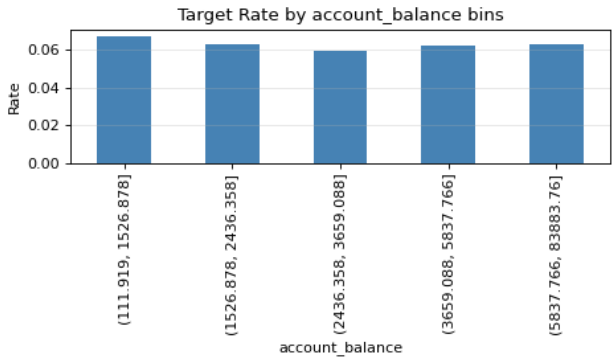
Categorical vs Target

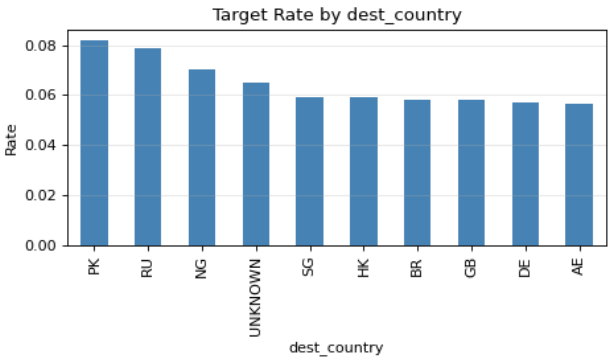
Column	Category	Target Rate
dest_country	PK	8.22%
dest_country	RU	7.88%
dest_country	NG	7.02%
dest_country	UNKNOWN	6.52%
dest_country	SG	5.90%
dest_country	HK	5.89%
dest_country	BR	5.83%
dest_country	GB	5.81%
dest_country	DE	5.73%
dest_country	AE	5.68%
origin_country	NG	8.43%
origin_country	RU	7.36%
origin_country	DE	6.68%
origin_country	AE	6.30%
origin_country	SG	6.23%
origin_country	UNKNOWN	6.22%
origin_country	PK	6.20%
origin_country	US	6.20%
origin_country	CA	6.18%
origin_country	GB	6.15%
txn_type	crypto_trade	9.66%
txn_type	p2p_transfer	6.85%
txn_type	cash_out	6.72%
txn_type	refund	6.62%
txn_type	merchant_payment	6.45%
txn_type	card_purchase	6.12%
txn_type	deposit	6.10%
txn_type	ach	5.89%
txn_type	cash_withdrawal	5.62%
txn_type	bill_pay	5.58%

Column	Category	Target Rate
payment_memo	UNKNOWN	7.88%
payment_memo	rent	7.14%
payment_memo	crypto	6.93%
payment_memo	loan	6.40%
payment_memo	family	6.39%
payment_memo	gift	6.35%
payment_memo	refund	6.29%
payment_memo	invoice	6.22%
payment_memo	tips	5.82%
payment_memo	subscription	5.81%
currency	NGN	8.38%
currency	RUB	7.40%
currency	EUR	6.64%
currency	AED	6.29%
currency	SGD	6.26%
currency	USD	6.19%
currency	PKR	6.16%
currency	GBP	6.14%
currency	CAD	6.11%
currency	HKD	5.79%
merchant_category	grocery	6.87%
merchant_category	electronics	6.61%
merchant_category	services	6.57%
merchant_category	gaming	6.35%
merchant_category	travel	6.34%
merchant_category	crypto	6.07%
merchant_category	marketplace	5.85%
merchant_category	utilities	5.77%
merchant_category	UNKNOWN	4.98%
merchant_name	UNKNOWN	9.54%
merchant_name	P2P Contact	6.45%
merchant_name	Cobalt Services	6.44%
merchant_name	Bright Electronics	6.44%
merchant_name	Delta Travel	6.44%
merchant_name	Alpha Market	6.36%
merchant_name	Foxtrot Groceries	6.02%
merchant_name	Echo Gaming	5.92%
channel	web	6.87%
channel	api	6.62%
channel	mobile	6.61%
channel	atm	5.91%
channel	branch	5.86%
channel	online	5.77%

Column	Category	Target Rate
payment_rail	crypto	7.90%
payment_rail	p2p	6.95%
payment_rail	wire	6.14%
payment_rail	cash	6.10%
payment_rail	card	6.01%
payment_rail	ach	5.74%
platform	zelle	7.23%
platform	cashapp	7.22%
platform	venmo	7.10%
platform	paypal	6.65%
platform	bank	5.83%

Charts

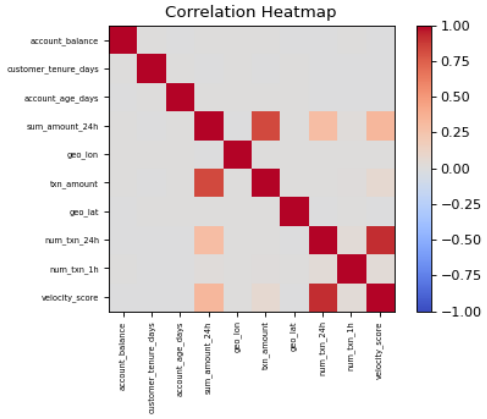




Feature Vs Feature

- No highly correlated feature pairs detected.

Charts



Time Drift

- Time series and drift analysis completed.

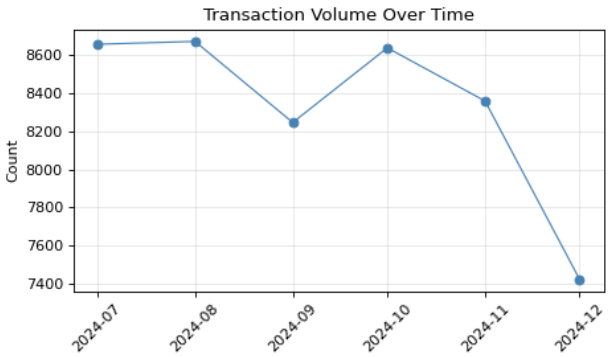
PSI Drift (Numeric)

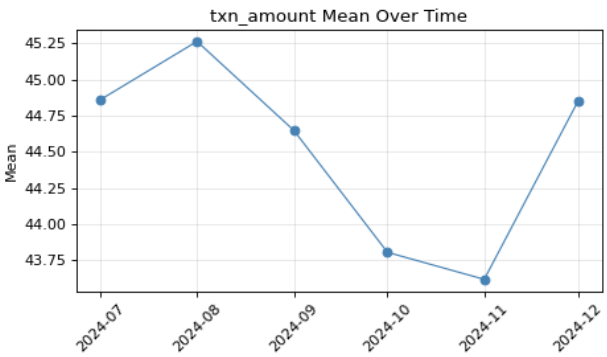
Column	PSI
account_age_days	0.002
geo_lat	0.001
txn_amount	0.001
account_balance	0.001
sum_amount_24h	0.001
velocity_score	0.000
customer_tenure_days	0.000
geo_lon	0.000
num_txn_1h	0.000
num_txn_24h	0.000

Categorical Drift (Total Variation)

Column	Drift Score
txn_type	0.012
currency	0.011
origin_country	0.010
payment_memo	0.010
merchant_category	0.008
payment_rail	0.007
dest_country	0.006
platform	0.006
merchant_name	0.005
channel	0.004

Charts





Summary

- High missingness columns: ['app_version'].
- Heavy-tailed numeric features: ['sum_amount_24h', 'fee_amount', 'txn_amount', 'sar_actual', 'account_balance'].
- Next steps: consider feature engineering, handling missing values, and monitoring drift.

Skipped EDA Sections

No sections were skipped.

Run Configuration

Key	Value
data_path	./data/synthetic_aml_mixed_50k_20260204_153242.csv
rows_original	50,000
rows_used	50,000
target_col	sar_actual
time_col	txn_ts
time_parse_ratio	1