Transaction and synthetic identity fraud detection for smaller institutions

OPEN METHODOLOGY · FREE TO RUN, VERIFY, AND ADOPT

FEATURE ENGINEERING SPECIFICATION · VERSION [VERSION]

The same construction, on two different payment systems, and where it goes next

Exact definitions for every detection signal, written so that a community bank, a credit union, or a regional lender computes the identical measurement on its own data and obtains a number comparable to one computed anywhere else. Comparability is the point: a scheme run across many institutions is invisible to each of them separately.

WHY THIS GENERALISES

ONE CONSTRUCTION PRINCIPLE, TWO STRUCTURALLY UNRELATED DATASETS

CONSTRUCTION PRINCIPLE	ENVIRONMENT A · CARD TRANSACTIONS	ENVIRONMENT B · ACCOUNT TRANSFERS	CARRIED INTO SYNTHETIC IDENTITY
Amount is the primary signal MAGNITUDE	`amt`. Strongest single correlate of the fraud label at r = 0.219. Extraction is maximised per attempt because the instrument will be blocked.	`amount`, and the balance deltas around it. Right skewed with extreme outliers that are genuine signal and must not be trimmed.	Utilisation and drawdown magnitude rather than a single transaction. The same logic on a longer clock.
Fraud concentrates in specific channels STRUCTURE	`category`. Card not present categories carry elevated risk; everyday in person categories carry the least.	`type`. Fraud occurs **only** in TRANSFER and CASH_OUT. No fraudulent PAYMENT, DEBIT, or CASH_IN appears at all.	Origination channel. Remote application paths carry the exposure that in branch paths do not.
Identifiers are excluded, not used LEAKAGE CONTROL	`cc_num`, `trans_num`, and nine further personal fields dropped before training.	`nameOrig` and `nameDest` are account identifiers and carry the same memorisation risk.	Social Security number, name, and address are inputs to verification, never to scoring.
Severe imbalance governs evaluation MEASUREMENT	0.58% fraud, a ratio of 172 to 1. Predicting legitimate everywhere scores 99.42% and detects nothing.	0.0536% fraud. Predicting legitimate everywhere scores 99.9% and detects nothing.	Rarer still, and the label is worse, because a synthetic identity that never defaults is never labelled.
Signal survives the model choice TRANSFERABILITY	Three algorithms exceeded 99% on a held out set of 555,719 records, Random Forest at 99.74%.	Four algorithms independently exceeded 99% on a structurally different dataset.	The reason to expect the approach to carry: the signal sits in the construction, not in one architecture.

Environment A Card transaction dataset, 23 raw columns, 1,296,675 training records.

Environment B Account transfer dataset, 11 raw columns, 6,362,620 records over 744 hourly steps.

FOR **Community banks, credit unions, regional lenders** And smaller origination platforms	ASSUMES **One or two technical staff** No analytics specialist	PARTS **Four** Two validated, two specified
ENTRIES **42 features defined** 22 validated, 20 specified	COMPANION **Reference implementation** [LINK]	VERSION **[VERSION]** Published [DATE]
LICENSE **CC BY 4.0** Code under [CODE LICENSE]	ARCHIVED AT **[DOI]** Permanent identifier	

0.1 Why precision in the definitions matters more here than usual

A loosely defined measurement produces numbers that cannot be compared between institutions, and a distributed detection capability built on numbers that cannot be compared is not a capability.

Card fraud operations run against institutions in many states at once. Synthetic identities are cultivated deliberately across accounts at several institutions, precisely because no single institution observes the complete pattern. Detection capability confined to the largest banks therefore leaves both schemes viable; the same capability distributed across the smaller institution tier degrades them everywhere at the same time.

That only works if a measurement computed at one credit union means the same thing as the measurement computed at another. So every definition here fixes the observation window in days rather than in words like recent, states the reference population explicitly, and names the data the measurement requires. Where an institution cannot obtain that data, the entry says so rather than leaving the reader to discover it.

STATUS LABELS, AND ONE HONEST STATEMENT

Parts 1 and 2 describe feature sets that were built and evaluated. Parts 3 and 4 are **specified and not implemented**. In particular, **no synthetic identity model has been trained**. Part 4 is a specification written in advance so that it can be checked against the eventual implementation, and so that an institution with the data can compute the measurements before that implementation exists. Any claim elsewhere that a synthetic identity detector is running today would be inaccurate.

VALIDATED	Computed and evaluated on the datasets named in the entry. Results reported in Part 5.
SPECIFIED	Defined here, not implemented. Fixed so that the eventual implementation can be checked against it.
EXCLUDED	Present in the source data and deliberately removed. The reason is recorded, because a reader who does not know why will put it back.
DATA DEPENDENT	Computable only if the institution holds a particular data source. The requirement is named in the entry.

Comparability rules, binding on every entry

- **Windows are stated in days**, counted backward from the observation timestamp, never as "recent" or "current period".
- **Timestamps are UTC.** Local time introduces a discontinuity twice a year that looks like behaviour.
- **Amounts are computed unscaled** in the institution's own currency, and scaled only at model fitting time. A measurement that has been scaled cannot be compared to one that has not.
- **Reference populations are defined by an explicit rule**, not by an analyst's judgement, and the rule is published alongside any score derived from it.
- **Feature identifiers are stable.** A changed definition receives a new identifier rather than a redefinition, so that results computed under an earlier version remain interpretable.
- Anything marked `[LIKE THIS]` must be set by the adopting institution and recorded.

Part 1

Environment A: card transaction features

Eleven retained features and eleven deliberate exclusions, from a card transaction dataset of 1,296,675 training records at a fraud rate of 0.58 percent.

A1 amt

VALIDATED

Transaction amount in dollars. The single most predictive retained feature.

TYPE	WINDOW	CORRELATION
Float, USD	Point in time	r = 0.219

SCALING

Model dependent

COMPUTATION

Used as recorded. Observed range in the training set was $1.00 to $952.49, mean $70.35, standard deviation $160.32.

RATIONALE

Extraction per attempt is maximised because the instrument will be blocked shortly after detection. That behavioural constraint is what puts amount ahead of every other single field.

CAUTION

Do not trim outliers. High value transactions are exactly the population of interest and removing them removes the signal.

Amount alone is a poor threshold rule. A correlation of 0.219 is a strong single field correlate and a weak standalone decision rule, and institutions that build a rule on it produce a large false positive burden and conclude that detection does not work.

A2 category

VALIDATED

Merchant category. Fourteen values in the source data.

TYPE	CARDINALITY	ENCODING
Categorical	14	Label, then one hot for linear models

CORRELATION

r = 0.020

RATIONALE	Card not present categories carry elevated risk and everyday in person categories carry the least, which is consistent with card not present being the dominant fraud vector. The weak direct correlation understates its value, because the field matters mainly in interaction with amount rather than on its own.
CAUTION	Fourteen categories is low cardinality and one hot encoding is cheap. Do not label encode this for a linear or distance based model, because the imposed ordering is not real.

A3 · A4 merchant, job

VALIDATED

Merchant name and cardholder occupation. High cardinality categorical fields.

TYPE	CARDINALITY	ENCODING
Categorical	About 693 and about 494	Label

CORRELATION

Near zero individually

COMPUTATION

```
enc = LabelEncoder().fit(train[col])     # fit on training data only
train[col] = enc.transform(train[col])
test[col]  = enc.transform(test[col])    # transform, never fit_transform
```

RATIONALE

Neither field predicts on its own. Both become useful when aggregated into a rate, which is what feature C4 in Part 3 specifies.

KNOWN DEFECT

The published implementation applied `fit_transform` independently to the training and test sets. Where the two sets contain different value orderings, the same merchant receives different integer codes on each side, and the evaluation is measured on a silently different encoding from the one the model learned.

The code block above is the correction. It is recorded here rather than quietly fixed because an institution reproducing the published result needs to know which version produced it. Correcting this is the first item in Part 6.

A5 gender

VALIDATED

Cardholder sex as recorded, mapped to a binary integer.

TYPE	ENCODING	CORRELATION
Binary integer	M to 0, F to 1	r = -0.008

RECOMMENDATION

Exclude

COMPUTATION

```
df['gender'] = df['gender'].map({'M': 0, 'F': 1})   # applied identically to both splits
```

RECOMMENDATION

Exclude this from any deployed scoring model. Its measured correlation with the label is indistinguishable from zero, so it contributes nothing, while a protected characteristic in a scoring feature matrix creates a fair lending exposure that no institution should take on for no gain. Retain it for fairness auditing only, per Part 6.

A6 to A9 lat, long, merch_lat, merch_long

VALIDATED

Cardholder and merchant geographic coordinates.

TYPE	WINDOW	CORRELATION
Float, degrees	Point in time	About 0.002 each

BETTER AS

Feature C1

RATIONALE

Raw coordinates carry essentially no direct signal, which the near zero correlations confirm. Their value is entirely as inputs to a derived distance, specified at C1 in Part 3.

CAUTION

Four raw coordinate columns in a feature matrix invite a tree ensemble to carve the map into regions that reflect the sampling of the dataset rather than any behaviour. Prefer the derived distance and drop the raw pairs once it exists.

A10 unix_time

VALIDATED

Transaction timestamp as seconds since the Unix epoch.

TYPE	WINDOW	CORRELATION
Integer seconds	Point in time	r = -0.005

BETTER AS

Features C2, C3

RATIONALE

As a raw integer this is close to a row index and carries almost nothing. It is retained because it is the input to every temporal and velocity measurement in Part 3, all of which are unexploited in the validated work.

CAUTION

A raw monotonic timestamp is a leakage risk under a random split, because the model can learn where in the file a record sits. Split temporally, and derive the cyclical features at C2 rather than passing the epoch value to the model.

A11 city_pop

VALIDATED

Population of the cardholder's city.

TYPE	RANGE	CORRELATION
Integer	23 to 2,906,700	r = 0.002

SCALING

Required for KNN and linear

RATIONALE

A population proxy for the density of the cardholder's environment. Negligible on its own; retained because unscaled it dominates distance computations and its exclusion should be a decision rather than an oversight.

A12 Eleven excluded columns

EXCLUDED

Fields present in the source data and removed before training, with the reason for each.

EXCLUDED

COLUMN	REASON FOR EXCLUSION
`cc_num`	**Leakage.** A card identifier lets the model memorise individual cards. Validation performance rises and deployment performance does not follow.
`trans_num`	Transaction UUID. Unique per row and therefore pure noise or pure memorisation.
`trans_date_trans_time`	Superseded by `unix_time`. Two representations of one fact invite inconsistent parsing.
`first, last`	Direct identifiers with no general fraud signal.
`street, city, state, zip`	Direct identifiers. Geography is already carried by the coordinate fields and by `city_pop`.
`dob`	Direct identifier. Age is a legitimate derived feature and is not computed in the validated work.
`Unnamed: 0`	CSV index artefact.

Eleven of twenty three columns removed, leaving eleven features and the target. The personal fields are excluded on principle as well as on utility: a detection model that requires a customer's street address to work is a model an institution should not deploy.

Part 2

Environment B: account transfer features

A structurally unrelated dataset of 6,362,620 records over 744 hourly steps, at a fraud rate of 0.0536 percent. Its purpose here is to show which construction principles survived a change of environment.

B1 type

VALIDATED

Transaction type: PAYMENT, TRANSFER, CASH_OUT, DEBIT, or CASH_IN.

TYPE	CARDINALITY	ENCODING
Categorical	5	One hot

DISCRIMINATIVE

Very high

OBSERVED

Every fraudulent record in the dataset is a TRANSFER or a CASH_OUT. No fraudulent PAYMENT, DEBIT, or CASH_IN occurs at all, and the two fraudulent types appear at roughly equal frequency.

RATIONALE

Consistent with the mechanics of exfiltration: funds leave through a transfer to a controlled account or through a cash withdrawal. A payment to a merchant does not move value anywhere the operator can retrieve it.

CAUTION

This is the single strongest field in the environment and the one most likely to be an artefact of how the dataset was generated. In production, restricting scoring to two transaction types on the strength of this observation would blind the institution to any scheme that adapts.

Use it as a stratification variable, so that models are compared within type, rather than as a filter applied before scoring.

B2 to B6 amount and the four balance fields

VALIDATED

Transfer amount, and originator and recipient balances before and after the transaction.

TYPE	DISTRIBUTION	SCALING
Float	Heavily right skewed	Applied to all models in the study

RANGE

Up to about 40 million

FIELDS	`amount`, `oldbalanceOrg`, `newbalanceOrig`, `oldbalanceDest`, `newbalanceDest`. Extreme outliers are characteristic of high value fraudulent transfers and were preserved as genuine signal.
RATIONALE	These fields carry the same information as `amt` in Environment A, with the addition of a conservation check: the change in each balance should equal the amount moved, and a discrepancy is mechanically implausible rather than merely unusual.
[illegible]D, [illegible]LD HAVE BEEN	The balance reconciliation residuals were available and were not constructed. They are specified at C5 in Part 3 and are, on the face of the data, likely to be among the strongest available features in this environment. Recording that omission is more useful than presenting the feature set as complete.

B7 step

VALIDATED

Integer time index. One step equals one hour, across 744 steps covering thirty simulated days.

TYPE	RESOLUTION	SPAN
Integer	1 hour	744 steps

BETTER AS

Features C2, C3

RATIONALE

Equivalent in role to `unix_time` in Environment A, and equally unexploited. Hour of day and day of week are directly recoverable from it, and velocity measurements are computable per account because the account identifiers are present.

B8 nameOrig, nameDest

EXCLUDED

Originating and destination account identifiers.

REASON

Identifiers, excluded for the same memorisation reason as `cc_num` in Environment A. They are nonetheless **required inputs** to the account level velocity and recipient concentration measurements specified at C3 and C6, where they are used as grouping keys rather than as model features.

Part 3 — Specified extensions to the transaction feature set

Six measurements the validated work identified as available and did not construct. Defined here so that an adopting institution gets them without repeating the omission.

ALL ENTRIES IN THIS PART ARE SPECIFIED, NOT IMPLEMENTED

ID	FEATURE	DEFINITION AND COMPUTATION	RATIONALE AND REQUIREMENT
C1	`txn_haversine_km`	Great circle distance in kilometres between cardholder coordinates and merchant coordinates, computed on the WGS 84 sphere with radius 6371.0 km. Replaces A6 to A9 in the feature matrix.	A purchase placed a thousand kilometres from the cardholder's home is anomalous in a way that four raw coordinates cannot express. Requires only fields already present.
C2	`txn_hour_sin`, `txn_hour_cos`, `txn_dow`	Hour of day encoded cyclically as $\sin(2\pi h/24)$ and $\cos(2\pi h/24)$, plus day of week as a categorical. Derived from the timestamp field in either environment, in UTC.	Fraud clusters in hours when the account holder is asleep and unlikely to notice. Cyclical encoding is required so that hour 23 and hour 0 are adjacent rather than maximally distant.
C3	`acct_txn_velocity_1h`, `_24h`	Count of transactions on the same instrument or account in the trailing 1 hour and 24 hours, computed with a strictly backward looking window that excludes the current record.	A burst of activity in a short window is the attack itself rather than a coincidence. **The exclusion of the current record is not optional**; including it leaks the present into its own features.
C4	`merchant_fraud_rate_90d`	Historical fraud rate for the merchant over the trailing 90 days, computed **only on data strictly preceding** the record being scored, with Laplace smoothing and a minimum volume floor of `[SET]` transactions.	Turns a near useless high cardinality identifier into a rate. The strict time ordering is what separates a legitimate aggregate from target leakage, and getting it wrong produces spectacular validation results that collapse in production.
C5	`bal_residual_orig`, `bal_residual_dest`	`oldbalanceOrg - newbalanceOrig - amount` and `newbalanceDest - oldbalanceDest - amount`. Environment B only.	A conservation check. A nonzero residual is mechanically implausible rather than statistically unusual, which puts it in the same class as a posthumous claim in healthcare billing. Requires no additional data.
C6	`dest_inflow_concentration_24h`	Number of distinct originating accounts sending to the same destination in the trailing 24 hours, and the share of that destination's inflow from the top originator.	Exfiltration converges on a small number of controlled accounts. This is the transaction level analogue of the beneficiary overlap measurement in the healthcare work, and it is invisible to any per transaction rule.

EVERY AGGREGATE IN THIS PART IS A LEAKAGE HAZARD

C3, C4, and C6 summarise history. If the window includes the record being scored, or any record that arrived after it, the feature contains information the model would not have at decision time and the evaluation is worthless while looking excellent. Compute all three with a strictly backward looking window, evaluate on a temporal split, and assert the ordering in code rather than trusting it.

Part 4

The synthetic identity feature set

Twenty measurements across six families. All specified, none implemented. This is the part the endeavour exists to produce, and it is written before the code so that the code can be checked against it.

A synthetic identity is a fabricated person assembled from real fragments: a genuine Social Security number belonging to somebody who does not actively use it, paired with a fabricated name, a plausible date of birth, and a real address. Identity verification checks whether each element is valid, and the identity passes every check, because every element is valid. Only the combination is false.

Detection therefore cannot operate on verification. It has to operate on behaviour over time: how the credit file was assembled, whether the borrowing pattern is internally consistent, and how balances and repayments move relative to stated capacity. That is the same class of analysis as transaction anomaly detection, applied to a longer clock and a different unit of observation.

The families below are ordered by how early in the scheme they become visible. The first three are observable at origination, which is where an institution can still decline. The last three are observable only after an account is open, which is where an institution can still limit exposure before the drawdown.

> **A LABEL PROBLEM THAT HAS NO CLEAN SOLUTION**
>
> A synthetic identity that is never drawn down is never labelled, and one that is drawn down is usually labelled as a charge off rather than as fraud. Any institution training on historical labels here is training on a subset defined by which schemes concluded and were correctly classified afterwards. The realistic first deployment is **unsupervised or rule assisted ranking for manual review**, not a supervised classifier, and an institution told otherwise should ask to see the labels.

ID	FEATURE	DEFINITION, WINDOW, COMPUTATION	WHY IT INDICATES A CONSTRUCTED IDENTITY
FAMILY 1 · FILE GENESIS · OBSERVABLE AT ORIGINATION			
S1	si_file_age_days	Days between the earliest record of any kind on the credit file and the application timestamp. BUREAU DATA	An organic file accumulates from a first student account or a first card and has depth. A constructed file begins abruptly, and abruptness is the most basic signature available.
S2	si_file_age_vs_age	Ratio of S1 to the applicant's claimed age in days, capped at 1.0. Undefined below a file age of [SET, suggested 90] days.	A forty year old applicant with an eleven month file is the central case. The ratio makes the implausibility comparable across ages in a way that raw file age does not.
S3	si_nonbureau_footprint	Count of distinct non credit records associating the name with the address before file open: utility, voter, property, employment. THIRD PARTY DATA	Real people leave traces before they borrow. A constructed identity typically has none, because it was assembled to borrow.
FAMILY 2 · ELEMENT COHERENCE · OBSERVABLE AT ORIGINATION			
S4	si_name_ssn_multiplicity	Count of distinct names observed against the same identity number across the institution's own application history, over the trailing [SET, suggested 1095] days.	One number under several names is the defining structure of the scheme. Computable on an institution's own records with no external data at all, which makes it the cheapest entry in this table.
S5	si_addr_multiplicity	Count of distinct applicants sharing a normalised address within the trailing window. Address normalisation rule must be fixed and published, since formatting differences otherwise fragment the count.	Constructed identities cluster on addresses the operator controls. Genuine multi occupancy exists, so this is a ranking input rather than a decision rule.
S6	si_contact_multiplicity	Count of distinct applicants sharing a telephone number, or an email local part, within the trailing window.	Contact elements are reused far more than addresses because they cost nothing to hold and are rarely checked across applications.
S7	si_element_age_min_days	Age of the youngest identity element among telephone, email, and address occupancy, in days. THIRD PARTY DATA	An identity assembled recently has recently created components. A single very young element among otherwise established ones is the specific pattern.
FAMILY 3 · TRADELINE CULTIVATION · OBSERVABLE AT ORIGINATION			
S8	si_au_ratio	Authorised user tradelines divided by total tradelines on the file. BUREAU DATA	Adding a constructed identity as an authorised user on a real, long standing account transplants that account's history onto the file. A high ratio means most of the apparent history was inherited rather than earned.
S9	si_au_seasoning_gap_days		A genuine authorised user is usually added near account opening, a spouse or a child. A large gap means an aged

ID	FEATURE	DEFINITION, WINDOW, COMPUTATION	WHY IT INDICATES A CONSTRUCTED IDENTITY
		Median days between an authorised user tradeline's own open date and the date it was added to this file.	account was attached to a new file, which is the paid tradeline market.
S10	si_tradeline_velocity_180	Count of tradelines opened in the trailing 180 days, excluding the application under assessment.	Cultivation is deliberate and fast once it starts. The exclusion prevents an application from counting itself.
S11	si_limit_growth_rate	Slope of total extended credit limit against time over the trailing 365 days, in currency per day.	The scheme's objective is limit accumulation, so limit growth outpacing what the stated profile would support is the trajectory itself rather than a proxy for it.
FAMILY 4 · APPLICATION BEHAVIOUR · OBSERVABLE AT ORIGINATION			
S12	si_inq_velocity_30, _90	Hard inquiries in the trailing 30 and 90 days. BUREAU DATA	Applications precede accounts, so inquiries are the earliest externally visible signal, arriving before any tradeline does.
S13	si_app_cluster_30	Count of other applications at this institution in the trailing 30 days sharing two or more identity elements with this one, where an element is address, telephone, email local part, or employer.	Operators submit in batches. Two shared elements across otherwise unrelated applicants is difficult to explain innocently and needs no external data.
S14	si_income_coherence	Ratio of stated income to the median stated income for the same occupation and geography in the institution's own application history, with a minimum group size of [SET].	Stated income is the field most often inflated and the one an institution can most cheaply benchmark against itself.
FAMILY 5 · UTILISATION TRAJECTORY · OBSERVABLE AFTER OPENING			
S15	si_util_slope_90	Ordinary least squares slope of monthly utilisation over the trailing 90 days, per line and aggregated across lines.	The bust out is a coordinated climb after a long flat period. The slope is what distinguishes it from ordinary borrowing, which drifts.
S16	si_sync_drawdown_14	Count of distinct lines crossing [SET, suggested 90] percent utilisation within any 14 day window.	Simultaneity across lines that have no reason to move together is the signature. A single maxed card is a life event; four in a fortnight is a plan.
S17	si_payment_regularity	Coefficient of variation of payment amounts over the trailing 365 days, computed per line.	Genuine repayment is irregular, because life is. A cultivation phase is often unnaturally regular, which makes **low** variance the anomaly here and inverts the usual direction.
FAMILY 6 · STRUCTURAL ABSENCE · OBSERVABLE THROUGHOUT			
S18	si_secured_absence	Binary flag: revolving credit present above [SET] total limit, with no secured	Constructed identities accumulate revolving credit because it is granted remotely. Secured lending involves

ID	FEATURE	DEFINITION, WINDOW, COMPUTATION	WHY IT INDICATES A CONSTRUCTED IDENTITY
		tradeline of any kind, at a claimed age above `[SET]`.	collateral, inspection, and title, which a fabricated person cannot survive.
S19	si_geographic_dispersion	Count of distinct states across the file's tradeline originators, and the distance between the application address and the modal originator location.	A real borrower's lenders cluster near the places they have lived. Remote origination platforms leave a scattered footprint with no residential logic.
S20	si_composite_rank	Percentile rank of the applicant within an explicitly defined reference population, computed from the standardised measurements above. The population rule must be published with any score derived from it.	The output an institution can act on. A rank against a stated peer group is defensible when challenged in a way that a raw score is not, and it is the form in which measurements from different institutions can be compared.

WHAT A SMALL INSTITUTION CAN COMPUTE TODAY, WITH NO PURCHASED DATA

S4, S5, S6, S13, and S14 require nothing but the institution's own application history. They are the cheapest entries in the table and they address the structure of the scheme directly. An institution with no bureau feed and no third party identity data can build a usable manual review queue from those five alone, and should, rather than waiting for the rest.

Part 5

Measured performance of the validated feature sets

Reported with the caveats that make the numbers readable, rather than as headline figures.

ENVIRONMENT A · CARD TRANSACTIONS · TEST SET OF 262,144 RECORDS

MODEL	ACCURACY	ROC AUC	FRAUD PRECISION	FRAUD RECALL	FRAUD F1
Random Forest	**99.95%**	**0.84**	**0.82**	**0.67**	**0.74**
Decision Tree	99.94%	0.84	0.74	0.68	0.71
K Nearest Neighbours	99.92%	0.72	0.71	0.44	0.54
Logistic Regression	99.92%	0.61	1.00	0.23	0.37

READ THE FOURTH COLUMN, NOT THE SECOND

All four models exceed 99.9 percent accuracy, and a model predicting legitimate for every record would score 99.42 percent while detecting nothing. The accuracy column separates these models by five hundredths of a percentage point and is close to meaningless.

The recall column is where they actually differ. Logistic Regression achieves perfect precision on the fraud class and finds under a quarter of it, which is a conservative boundary rejecting anything borderline. Random Forest finds two thirds at a precision of 0.82. **Roughly a third of fraud goes undetected by the best model here**, and in a production environment that is direct loss. An institution reading only the accuracy column would not know that.

In a separate run on a card dataset of approximately 1.9 million records, three algorithms each exceeded 99 percent accuracy on a held out set of 555,719 records, Random Forest at 99.74 percent, at a fraud rate of 0.58 percent and an imbalance ratio of 172 to 1. In Environment B, four algorithms independently exceeded 99 percent on a structurally different dataset at a fraud rate of 0.0536 percent.

The cross environment consistency is the finding that matters, and it is a claim about the feature construction rather than about any model. A result that holds across a linear model, a single decision boundary model, and two ensemble methods, on two datasets with different columns and different fraud mechanics, indicates the signal is being captured by how the features are built. That property is what makes the approach worth publishing for other institutions to run.

Part 6

Defects, corrections, and limitations

Recorded here rather than discovered later by the institution that adopts this.

Defects in the published implementation, with corrections

- **Encoders fitted on both splits.** `fit_transform` was applied independently to training and test data for the high cardinality categorical fields. Fit on training data only and use `transform` on everything else. Corrected form is at A3.
- **Class imbalance not addressed at training time in Environment A.** No cost sensitive weighting or resampling was applied, which is consistent with the low recall observed. Apply `class_weight='balanced'` as a first step, as was done in the healthcare work.
- **Precision, recall, and area under the precision recall curve were not reported in the original credit card run.** The confusion matrix in that write up is reconstructed from accuracy and class composition rather than measured. Measure them directly.

- **Random splits were used where temporal splits were available.** Both environments carry timestamps. A random split over time ordered data allows the model to learn from the future.
- **Six available features were not constructed**, including the balance reconciliation residuals in Environment B, which are mechanically implausible when nonzero. All six are specified in Part 3.

Fairness handling, binding on any deployment

- Exclude the sex field from any deployed scoring model. Its correlation with the label is indistinguishable from zero and its presence in a lending decision creates fair lending exposure for no gain.
- Do not use geography as a direct scoring feature in an origination model without testing for proxy effects. Distance measurements such as C1 and S19 are relational rather than locational and do not carry the same exposure.
- Compute score and decline rate distributions across protected classes after training, record the result, and publish it with the model.
- Where a synthetic identity score contributes to an adverse action, the institution must be able to state the reasons in specific terms. Prefer measurements that can be described in a sentence a customer would understand.

Limitations of this specification

- **No synthetic identity model exists.** Part 4 is a specification. It has not been fitted, tuned, or evaluated against any label.
- **Both validated datasets are synthetic.** They reproduce structure, not the missingness, coding drift, and adversarial adaptation of production data. Expect performance to fall.
- **Environment B's strongest field may be an artefact.** Fraud appearing exclusively in two transaction types is a property of how that dataset was generated and should not be assumed of a real payment system.
- **Several Part 4 measurements need data a small institution may not hold.** Each is marked. The five that need nothing but the institution's own records are named at the end of Part 4.
- **The synthetic identity label problem has no clean solution** and constrains what any supervised approach can achieve, as set out in Part 4.

Compute it, compare it, and tell me where it is underspecified

This specification exists so that a measurement computed at one institution means the same thing as the measurement computed at another, because the schemes it addresses are run across many institutions at once. If a definition here is loose enough that two analysts would compute it differently, that is a defect in this document and the most useful thing you could report.

COMPANION DOCUMENTS

- Reference implementation, transaction detection [LINK]
- Implementation guide for smaller institutions [LINK]
- Threat model note, synthetic identity [LINK]
- Healthcare feature dictionary and validation protocol [LINK]

VERSIONING RULE

Feature identifiers are stable. A changed definition receives a new identifier rather than a redefinition, so that measurements computed under an earlier version remain comparable. Changelog: [LINK]

CORRECTIONS

Report an ambiguous definition, a defect, or a measurement that could not be computed on your data: [CONTACT]

Ayeni, A. [YEAR]. Feature engineering specification: transaction and synthetic identity fraud detection for smaller financial institutions, version [VERSION]. [DOI] · Licensed CC BY 4.0.

This document describes measurements used to rank applications and transactions for human review. No measurement defined here establishes that any person committed fraud, and a score derived from them is not a determination about any applicant or account holder.