International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 16(2024) pp. 1-11
https://doi.org/10.70917/ijcisim-2026-6100
Article

Why Your Fraud Rules Are Rejecting Good Customers and How to Measure It

Olid Sarker[1*], Habibur Rahman[2]

[1]Feliciano School of Business, Montclair State University, olidsarker@gmail.com.
[2]Northern University of Bangladesh, nubhabib@gmail.com.

Abstract: Online merchants who screen card-not-present orders with fraud rules cannot see what happens to the orders they decline. This retrospective examines how merchants and their advisers estimated those false declines in the last half-decade, and proposes a way for a small merchant to measure them. We tested ten claims from the period on DECLINESIM-23, a seeded synthetic benchmark of era-styled order streams in which the true outcome of every declined order is known, across 300 store-level parameter draws. Under red-flag rules that declined a median of 1.5% of orders, genuine customers accounted for a median of 89% of declines, far more than the one-third industry sources reported. Counting customers who reordered after a decline captured about a fifth of declines as genuine, and performance measured on approved orders alone understated the rules' accuracy. Measured as lost profit rather than lost sales, false declines exceeded fraud losses in only 57% of stores. Approving a random 5% of would-be declines gave nearly unbiased estimates, and in 80% of stores it earned more from recovered sales than it lost to fraud. Holdouts were too small for precise reporting at small-merchant volumes, but large enough for decisions: a 2% holdout lifted a blanket ban on international orders in a median of one month and cut false declines on those orders from 135 to 21 per 1,000 orders.
Keywords: Card-not-present fraud; False declines; Selective labels; Randomized holdout; Off-policy evaluation; E-commerce

1. Introduction

Every fraud rule that declines an order makes two kinds of decisions at once. It stops some criminals, and it turns away some customers who meant to pay. Merchants see the first kind of decision clearly, because fraud that gets through comes back weeks later as a chargeback. They rarely see the second kind at all. A declined customer does not file a dispute; most simply leave, and the order disappears from the record. A merchant can therefore know its fraud losses to the dollar and still have no idea how many good customers its rules reject.

The industry's estimates of this hidden cost were large. A Javelin Strategy & Research study reported in 2015 that wrongly declined card transactions led to roughly $118 billion of lost sales for United States merchants in 2014, set against roughly $9 billion of card fraud (Stewart, 2015). Survey results on how shoppers reacted to a false decline circulated widely in the same period (Zohar, 2015). A 2018 white paper from Card Not Present and Radial cited a study in which one-third of declines were false positives, and it noted that merchants lacked practical ways to measure the problem (Card Not Present & Radial, 2018). Card networks, meanwhile, continued to advise merchants to treat a list of order features as red flags (Visa, 2016).

Academic work in neighbouring fields had already named the underlying difficulty. When a decision determines which outcomes are observed, evaluation on the observed outcomes alone is biased; Lakkaraju et al. (2017) called this the selective labels problem. Credit scoring had wrestled with the same issue for decades under the name of reject inference (Crook & Banasik, 2004; Banasik & Crook, 2007; Ehrhardt et al., 2021), and web experimentation and contextual bandits had developed randomized ways to evaluate a policy that was not the one in force (Kohavi et al., 2009; Li et al., 2011; Dudík et al., 2014). Fraud research recognised verification latency and the partial feedback that investigators provide (Dal Pozzolo et al., 2018), and one study used contextual bandits to adapt fraud screening

to drift (Soemers et al., 2018). These ideas rarely reached the small merchants who wrote their own rules on hosted platforms.

This paper studies that gap as a retrospective. This study was designed and written in 2024 as a retrospective; it makes no claim to have been produced in the period it studies. The analysed corpus is bounded to 2016–2023, with a few earlier methodological sources and two 2024 preprints cited only to position the contribution. We ask three questions. How many of the orders that era-style rules declined came from genuine customers? Which of the measurement shortcuts available to merchants at the time would have revealed that number? And what is the cheapest design that measures false declines well enough to act on?

Because no public data set records the true outcome of declined orders, we built DECLINESIM-23, a seeded synthetic benchmark in which every order has a known label and the merchant sees only what a real merchant would see: chargebacks on shipped orders, reviewers' decisions, customers' retries, and the outcomes of any orders approved at random. The study makes three contributions. It turns ten claims from the period into testable statements and reports a verdict for each. It compares six ways of estimating false declines against the truth, including a doubly robust estimator built on a randomized holdout. And it gives merchants a decline-audit protocol, with sample-size rules and evidence that a holdout usually pays for itself.

2. Background

2.1 What the era said about false declines

Practitioner sources of the period agreed that false declines were common and costly, and disagreed about how common. The Javelin estimate compared lost sales with fraud losses and found the former far larger (Stewart, 2015). The Card Not Present and Radial white paper reported that merchants lost 2.79% of revenue to false positives against 0.52% to chargebacks, and it described a measurement method based on customers who returned and bought a replacement for a cancelled order (Card Not Present & Radial, 2018). Consumer surveys reported that many shoppers who were wrongly declined reduced or stopped their purchases at that merchant (Zohar, 2015). These figures came from vendors with an interest in the problem, and none rested on a design that observed the outcome of a declined order.

Academic fraud research of the period paid more attention to fraud that was missed than to good orders that were stopped. Production studies documented class imbalance, concept drift, and the delay before investigators or chargebacks supply a label (Dal Pozzolo et al., 2018; Lebichot et al., 2021). At Microsoft, an e-commerce fraud system was optimised for long-term profit across approval decisions (Nanduri et al., 2020), and a study of an online retailer described a system that supported manual review (Carneiro et al., 2017). Work on false positives in fraud detection aimed to reduce them (Baader & Krcmar, 2018; Vorobyev & Krivitskaya, 2022), but it measured them on cases that had been investigated, not on cases that had been turned away. Reviews of e-commerce fraud catalogued the techniques in use (Rodrigues et al., 2022) and noted the distance between academic models and business practice (Mekterović et al., 2021).

2.2 Why declines are hard to measure

Missing-data theory gives the problem its general form. When whether an outcome is observed depends on the same information that predicts the outcome, estimates from the observed cases are biased unless the selection is modelled (Heckman, 1979; Little & Rubin, 2019). In fraud screening, an order is shipped, and its label is later observed, only if the rules approve it. The labels that a merchant collects are therefore drawn from the region the rules consider safe (Lakkaraju et al., 2017).

Three families of remedies were available. Reject inference, developed in credit scoring, infers the outcomes of rejected applicants from a model fitted on accepted ones, sometimes with augmentation, semi-supervised learning, or deep generative models (Li et al., 2017; Mancisidor et al., 2020; Ehrhardt et al., 2021). Earlier evaluations found that such methods offered little or no improvement in predictive performance (Crook & Banasik, 2004; Banasik &

Crook, 2007). Randomized exploration takes a different route: it approves a known, random share of the cases the policy would reject, and weights their outcomes by the inverse of the probability of being approved (Horvitz & Thompson, 1952). The method underlies online controlled experiments (Kohavi et al., 2009; Kohavi et al., 2020) and offline evaluation of contextual bandits (Li et al., 2011). Doubly robust estimators combine a model with the randomized cases, so that the estimate stays consistent if either the model or the exploration probabilities are correct (Robins et al., 1994; Dudík et al., 2014). Learning-to-rank research showed how propensity weighting corrects biased feedback in practice (Joachims et al., 2017).

The present study takes the merchant's view, tests the measurement methods that were available to merchants in the period, and evaluates them on simulated order streams.

3. Claims audited

Table 1 lists the claims. C1 is the working assumption behind any fraud dashboard that reports chargebacks but not declines: that declined orders were mostly fraudulent. C2 and C3 are the industry estimates of the share and cost of false declines (Stewart, 2015; Card Not Present & Radial, 2018). C4 is the measurement method described in the 2018 white paper (Card Not Present & Radial, 2018). C5 is a shortcut that we found no source recommending but that a merchant with a review team could easily take: treating the share of reviewed orders that reviewers clear as the share of genuine declines. C6 is the promise of reject inference (Li et al., 2017; Mancisidor et al., 2020). C7 is the selective labels warning (Lakkaraju et al., 2017). C8 is the claim from the experimentation literature that a small randomized holdout gives unbiased estimates (Kohavi et al., 2009; Dudík et al., 2014). C9 restates C8 for the precision that small merchants can reach. C10 concerns the advice to treat shipments abroad as a red flag (Visa, 2016), which some merchants turned into a blanket ban on international orders.

Table 1. Claims audited

Claim	Statement	Source	Test
C1	Declined orders are mostly fraudulent	Implicit in chargeback-only reporting	Share of genuine declines
C2	About one-third of declines are false positives	Card Not Present & Radial (2018)	Same
C3	False declines cost merchants more than fraud	Stewart (2015); Card Not Present & Radial (2018)	Ratio of lost profit to fraud losses
C4	Customers who reorder after a decline reveal false declines	Card Not Present & Radial (2018)	Bias of the retry count
C5	Reviewers' clearance rate shows how many declines are wrong	Untested shortcut	Bias of the review-band rate
C6	A model fitted on approved orders can infer the outcome of declines	Reject inference literature	Bias of the model estimate
C7	Performance measured on approved orders is misleading	Lakkaraju et al. (2017)	Observed versus true AUC
C8	A random holdout gives unbiased counterfactual estimates	Experimentation literature	Bias and interval coverage
C9	A small holdout is precise enough to report	Same	Months to a ±5-point estimate
C10	Blocking international orders is a safe way to limit risk	Visa (2016), as applied	False declines with and without a holdout

4. Method

4.1 The DECLINESIM-23 benchmark

DECLINESIM-23 simulates one year of orders for a single online store. Each order comes from one of seven segments. Genuine orders come from regular customers, gift buyers shipping to another address, international customers whose banks rarely support address verification, and customers browsing through a virtual private network or while travelling. Fraudulent orders come from stolen cards, card testing, and friendly fraud, in which a cardholder disputes a purchase he or she made. Each order carries nine binary features that a hosted store records: differing billing and shipping addresses, expedited shipping, an international address, a mismatch between IP and billing country, a failed address check, a failed card security code check, repeated payment attempts, a high order value and a new customer. Within a segment, the features are independent; across segments, they are correlated, as in practice.

Each store draws its own parameters. Monthly volume is log-uniform between 1,500 and 20,000 orders, the fraud rate is log-uniform between 0.3% and 2%, the international share of genuine orders is uniform between 2% and 25%, gross margin is uniform between 25% and 65%, and every feature probability is perturbed on the logit scale with a standard deviation of 0.4. All central values and ranges are the authors' assumptions, informed by the qualitative logic of the sources in Section 2. They are not estimates of any merchant's data, and the analysis reports which conclusions hold across the range.

The store screens orders with an era-style rule. It counts the red flags on each order, excluding the new-customer flag, declines orders with at least k flags (k is 4 or 5), sends orders with k − 1 flags to manual review up to a capacity of 0.5–2% of orders (orders beyond capacity are approved), and approves the rest. In 28% of stores, the rule also declines every international order.

4.2 What the merchant observes

Shipped third-party fraud is charged back with a probability between 75% and 95%, and friendly fraud is always disputed. Reviewers clear 95% of genuine orders, 90% of friendly fraud and 20% of third-party fraud. After a decline, a genuine customer tries again with a probability between 15% and 60%, and the retry is approved with a probability between 50% and 80%. A third-party fraudster also tries again, with a probability between 20% and 80%, and succeeds less often (15–45%); friendly fraudsters behave like genuine customers. When a holdout is running, a random share h of the orders that the rule would decline is approved instead and shipped, and its outcome is observed through chargebacks.

The target quantity is the genuine share of declines, G = P(genuine | declined), together with the number of false declines per 1,000 orders. Lost profit on a false decline is the order's margin net of shipping and processing, multiplied by (1 + λ), where λ is the share of future profit lost when a customer does not return; we report λ = 0, 0.5 and 1. The loss on approved fraud is the cost of the goods, shipping, a $20 chargeback fee, and processing.

4.3 Measurement methods

We compare six estimators of G. The first assumes that every decline was fraud (G = 0). The second counts declined orders that were followed by a successful retry. The third applies the reviewers' clearance rate in the review band to the decline band. The fourth fits a logistic regression on shipped orders, with chargebacks as labels, and averages one minus its predicted fraud probability over the declined orders. The fifth uses the random holdout:

$$\hat{G}\mathrm{H} = 1 - (1/n\mathrm{H}) \sum i \in H \; yi \quad (1)$$

where H is the set of held-out orders, nH its size, and yi an indicator of a chargeback. Its 95% interval uses the Wilson method (Wilson, 1927; Agresti & Coull, 1998). The sixth is doubly robust: it starts from the model's prediction $\hat{p}i$ and corrects it with the residuals in the holdout (Robins et al., 1994; Dudík et al., 2014),

$$\hat{G}\mathrm{DR} = 1 - [\, (1/n\mathrm{D}) \sum i \in D \; \hat{p}i + (1/n\mathrm{H}) \sum i \in H \; (yi - \hat{p}i) \,] \quad (2)$$

where D is the set of would-be declines and nD its size. Because every would-be decline has the same holdout probability h, inverse-probability weights cancel within the decline band. Orders that reviewers clear are approved, but in this benchmark, they are left out of the labelled data used by the model and by the observed AUC.

4.4 Experiments

E1 draws 300 stores, simulates a year for each, and computes the true G, the cost ratio, and every estimator, with holdouts of 1%, 2%, 5% and 10%. It also computes the AUC of the red-flag count (excluding the new-customer flag, with a small order-value tie-breaker) against the true labels and against the labels the merchant observes on shipped orders. E2 computes the number of held-out orders needed to estimate G within ±5 points with 95% confidence, $n = 1.962\ G(1 - G)/0.052$, and converts it into months for monthly decline volumes of 50 to 1,500. E3 takes 40 stores that block all international orders and runs the rule for twelve months with no holdout, a 2% holdout, and a 5% holdout. The ban is lifted when the upper Wilson bound on the chargeback rate of held-out international orders falls below the break-even fraud probability at which approving and declining an international order cost the same, with $\lambda = 0.5$.

4.5 Verdict rules

A claim about a quantity (C1, C2, C3) is held if the benchmark agrees with it in at least 80% of draws, overturned if it disagrees in at least 80%, and qualified otherwise. C1 disagrees when the genuine share exceeds one half, C2 when it exceeds one-third, and C3 when the profit-based cost ratio is below 1. A measurement method (C4, C5, C6, C8) is held if its estimate is within 5 points of the truth in at least 80% of draws and its mean bias is under 2 points, overturned if it is within 5 points in fewer than 20% of draws, and qualified otherwise. C7 is held if the AUC on approved orders differs from the true AUC by more than 0.02 in at least 80% of draws, and overturned if it does so in fewer than 20%. C9 is held if a 5% holdout reaches a ±5-point interval within twelve months at the median decline volume, and overturned otherwise. C10 is overturned if lifting the ban with a holdout cuts false declines and every lift is correct. The rules were set after a pilot run at reduced scale and before the full run reported here.

4.6 Transparency of the analysis

All data in this study are synthetic. They are era-styled, in that they follow the screening practices, scales, and feedback typical of the last 7 years, but they were not collected in that period or from any merchant. The generator is seeded, and every number in the text and tables was written to a results file by the shipped code and inserted from that file. The analysis used Python with NumPy and scikit-learn. Google Colab implemented and executed the simulation and analysis code and drafted the text under the authors' direction, as set out in the declaration before the references.

5. Results

5.1 How many declines came from genuine customers?

Across the 300 stores, the rules declined a median of 1.5% of orders, about 66 a month, and a median of 89.1% of those declines came from genuine customers (10th–90th percentile 57–99%). The genuine share exceeded one half in 93% of stores and one-third in 98%. The result follows from base rates. When fraud is below 1% of orders, even a rule that is several times more likely to flag a fraudulent order than a genuine one will draw most of its declines from the much larger genuine population. The share fell as fraud became more common, from a median of 93% in stores with fraud below 0.6% to 79% in stores with fraud of 1.2% or more. Stores that blocked international orders had the highest share (99% against 83%). Claim C1 is overturned. Claim C2 is overturned as well, but in the opposite direction from its critics' expectation: one-third understated the problem.

5.2 Did false declines cost more than fraud?

The answer depends on what is counted. Measured as lost sales, as in the industry estimates (Stewart, 2015), declined genuine revenue was a median of 3.0 times the fraud losses of the same store. Measured as lost profit, the

ratio fell to 0.99 when declined customers were assumed to come back and to 1.48 when they took half of their future profit elsewhere. False declines cost more than fraud in 50% of stores without lost future profit, 57% with half of it lost, and 62% when all of it was lost. Claim C3 is qualified. False declines were a cost of the same order as fraud, not an order of magnitude larger, and comparisons of lost revenue with fraud losses overstated them.

5.3 Which measurement methods worked?

Fig. 1 and Table 2 compare the estimators with the truth. Counting successful retries (C4) recovered a median genuine share of 20.5% against a truth of 89.1%, a mean bias of −61.0 points. The method can only count customers who tried again and got through, and fraudsters also retry, so it is a weak lower bound. The review-band shortcut (C5) was nearly unbiased on average, with a mean error of +2.7 points, but it missed the truth by more than 5 points in 71% of stores, because the orders that reviewers see are less risky than the orders that are declined. The model fitted on approved orders (C6) came within 5 points in 62% of stores and overstated the genuine share by 4.8 points on average; it extrapolates into a region it has never seen, and missing chargebacks make that region look safer than it is.

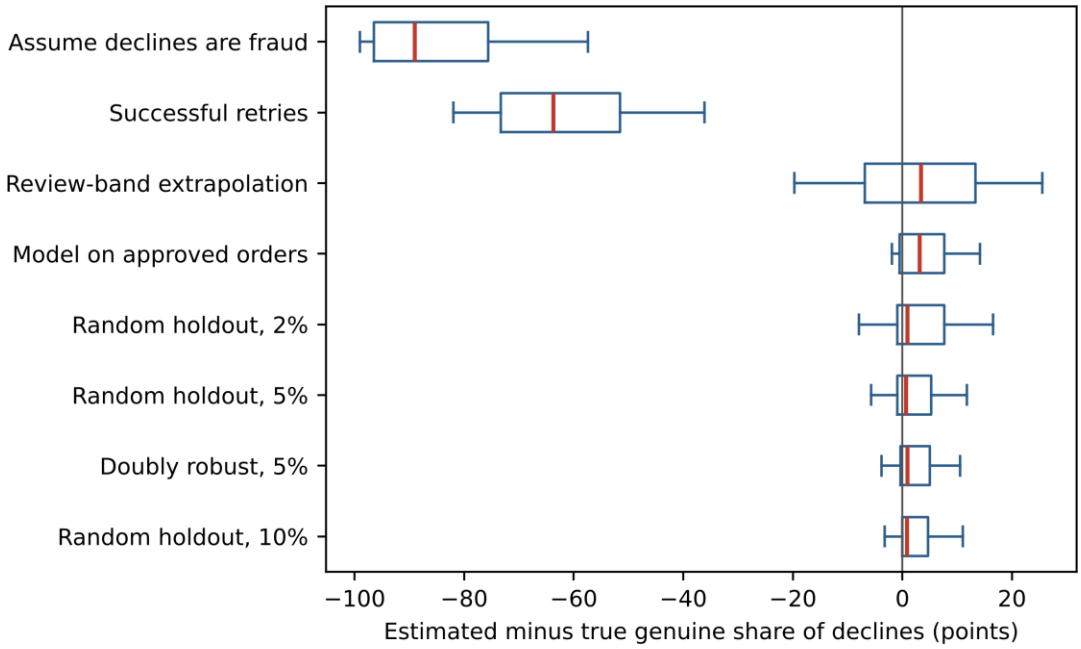


Fig. 1. Error of each estimator of the genuine share of declines across 300 stores. Boxes show the interquartile range, whiskers the 10th and 90th percentiles, and red lines the median.

The random holdout (C8) was close to unbiased. With 5% of would-be declines approved, the mean error was +2.3 points, and the 95% Wilson interval contained the truth in 93% of stores. The small positive bias has a known source: fraud that is never charged back looks genuine to the merchant. The doubly robust estimator reduced the error further, coming within 5 points in 66% of stores against 62% for the plain holdout, because the model absorbs much of the noise from the few held-out orders. Neither met the strict 80% criterion at 5%, and C8 is therefore qualified: unbiased, as claimed, but noisy at small-merchant volumes.

Table 2. Error of estimators of the genuine share of declines (300 stores; bias and mean absolute error in percentage points)

Estimator	Median estimate (%)	Mean bias	Mean absolute error	Within 5 points (%)
True value	89.1	–	–	–

Assume declines are fraud (C1)	0.0	−83.0	83.0	0
Successful retries (C4)	20.5	−61.0	61.0	0
Review-band extrapolation (C5)	89.5	+2.7	14.6	29
Model on approved orders (C6)	92.6	+4.8	5.9	62
Holdout 2%	95.8	+2.8	8.5	55
Doubly robust 2%	92.8	+3.2	7.5	60
Holdout 5%	91.5	+2.3	6.5	62
Doubly robust 5%	90.4	+2.4	5.7	66
Holdout 10%	91.1	+2.1	5.2	69
Doubly robust 10%	91.1	+1.9	4.5	70

5.4 How large a holdout, and at what cost?

Precision is the binding constraint. A 5% holdout in the median store produced 44 held-out orders in a year, and the median width of the 95% interval was 19 points. Fig. 2 shows why. Estimating a genuine share near one half within ±5 points requires about 384 held-out orders. A store that declines 50 orders a month would need 154 months at a 5% holdout and 77 months at 10%. Only a store declining around 1,500 orders a month reaches that precision within half a year at 5%. At the median store's 66 declines a month, a 5% holdout would take 116 months for a genuine share near one half, and about 45 months at the median genuine share of 89%. Claim C9 is overturned for small merchants.

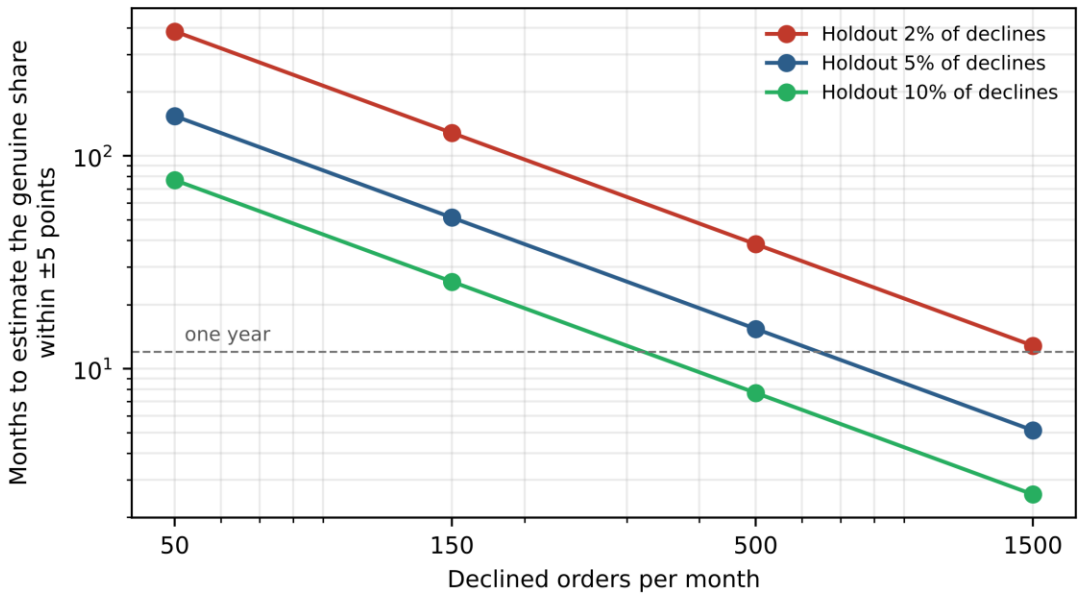


Fig. 2. Months needed for a random holdout to estimate a genuine share of 50% within ±5 points with 95% confidence, by monthly decline volume and holdout rate.

The cost of the holdout was usually negative. Because most declines were genuine, approving a random few recovered more profit than it lost to fraud. With a 5% holdout, the net effect was a gain in 80% of stores, with a median gain of 15.4 dollars per 1,000 orders, before any value from what the merchant learned. The common objection that exploration lets fraud through is true in each case and false on average for rules that decline mainly genuine customers.

5.5 Do metrics on approved orders mislead?

Yes, but not in the direction that is usually feared. The AUC of the red-flag count, computed on shipped orders with chargebacks as labels, had a median of 0.713, against 0.782 on the true labels of all orders. The observed value was lower in 97% of stores and differed from the truth by more than 0.02 in 93%. The rules looked worse than they were for two reasons: the most suspicious orders, which the rules separate best, never reach the labelled data, and some shipped fraud is never charged back. Claim C7 is held: a merchant comparing two rules on approved orders alone would compare them on the part of the data where they differ least.

5.6 Blanket bans and the evidence they suppress

Table 3 reports E3. Without a holdout, no store ever lifted its ban on international orders, because no international order was ever shipped and so no evidence could accumulate. The ban caused a median of 135 false declines per 1,000 orders. With a 2% holdout, 98% of stores lifted the ban, in a median of 1 month (90th percentile 3.2 months), and every lift was correct in the sense that the true fraud rate of international orders lay below the break-even point. False declines on international orders fell to 20.5 per 1,000 orders, and the holdout itself earned a median of 17.1 dollars per 1,000 orders. A decision needs far less evidence than a report: the ban could be lifted as soon as the first held-out international orders showed no sign of the fraud rate that would justify it. Claim C10 is overturned.

Table 3. Twelve months of a blanket ban on international orders (40 stores)

Holdout	Stores that lifted the ban (%)	Median month of lifting	International false declines per 1,000 orders	Net cost of holdout per 1,000 orders ($)
None	0	–	135.3	0
2%	98	1	20.5	-17.1
5%	100	1	16.6	-31.8

5.7 Verdict matrix

Table 4. Verdicts on the ten claims

Claim	Finding	Verdict
C1 Declines are mostly fraud	Genuine share above one half in 93% of stores	Overturned
C2 One-third of declines are false	Genuine share above one-third in 98% of stores	Overturned (understated)
C3 False declines cost more than fraud	True in 57% of stores at $\lambda = 0.5$; lost-sales comparisons inflate the ratio	Qualified
C4 Reorders reveal false declines	Mean bias −61.0 points	Overturned
C5 Review clearance rate	Within 5 points in 29% of stores	Qualified
C6 Model on approved orders	Within 5 points in 62% of stores; bias +4.8	Qualified
C7 Approved-only metrics mislead	AUC off by more than 0.02 in 93% of stores	Held

C8 Random holdout is unbiased	Bias +2.3 points at 5%; coverage 93%	Qualified
C9 A small holdout is precise	116 months at the median store's 66 declines a month (5% holdout)	Overturned
C10 Blanket country bans are safe	Lifted correctly with a 2% holdout; false declines cut sharply	Overturned

6. Discussion

6.1 What the era got right and wrong

The period's practitioners were right that false declines were a high and hidden cost, and they were right that merchants lacked a way to see them (Card Not Present & Radial, 2018). They underestimated how many declines were wrong and overestimated how much those declines cost, because they compared lost sales with fraud losses rather than profit with profit. The measurement method they described, counting customers who came back, saw only the most persistent genuine customers and could not separate them from returning fraudsters. The academic literature had the right tools (Lakkaraju et al., 2017; Dudík et al., 2014; Kohavi et al., 2020), but those tools were developed for large platforms that could afford to wait for thousands of randomized observations. For a small merchant, a randomized holdout cannot produce a precise monthly figure. It can still answer the question that matters most, which is whether a specific rule should stay.

6.2 A decline audit for small merchants

The results suggest a simple protocol; its rule-by-rule step was tested here only for a country ban, so the other steps are recommendations rather than tested results. First, log every would-be decline with the rule that triggered it, including orders declined by blanket bans. Second, approve a random 2–5% of would-be declines, below a value cap that the merchant can afford to lose, and record the random assignment. Third, review the evidence by rule, not in aggregate: a rule whose held-out orders show a chargeback rate whose upper bound lies below its break-even probability should be relaxed or removed, and the check can be repeated monthly. Fourth, when reporting the overall genuine share, use the doubly robust estimate, which combines the held-out orders with a model fitted on approved orders, and report its interval. Fifth, compare the cost of false declines with fraud losses in profit, not revenue. Where strong customer authentication is available, sending a held-out order through authentication instead of approving it outright lowers the cost of the holdout further, although it also changes which outcomes are observed.

6.3 Limitations

The data are synthetic, and the parameter ranges reflect the authors' reading of the period's sources rather than measured values. The share of genuine declines depends strongly on the fraud rate and the rule threshold, so a merchant with much higher fraud, or a much stricter rule, will see different numbers. The benchmark treats chargebacks as arriving within the year and omits the delay that real disputes involve (Dal Pozzolo et al., 2018), and it assumes that holdout orders are approved without any change in the customer's behaviour. Fraudsters who learn that some declined orders are approved at random could probe for them, which argues for a value cap and for keeping the holdout rate private. Reviewers were modelled with fixed accuracy, and lost future profit was set by assumption. Finally, the measurement methods were tested on a single family of rule-based policies; score-based policies with continuous thresholds would allow finer exploration designs.

7. Conclusion

A fraud rule that declines orders without any record of what those orders would have done is unaccountable by construction. In a benchmark built to reflect the screening practices of half a decade, most declined orders came from genuine customers; the measurement shortcuts that merchants had at hand missed most of them, and metrics computed on approved orders understated how well the rules discriminated. The fix that the experimentation literature offered

was sound but had to be adapted for small merchants: a small random holdout of would-be declines, combined with a model, measures enough to decide on a specific rule such as a country ban and, in most stores, costs nothing. Measuring false declines turns out to be cheaper than not measuring them.

Supplementary Information

"Not applicable"

Ethical Approval and Consent to participate

"Not applicable"

Consent for publication

"Not applicable"

Availability of supporting data

"Not applicable"

Competing interests

"Not applicable"

Funding

"Not applicable"

Authors' contributions

Two authors equally and actively contributed to the conception, design, implementation, data analysis, interpretation, manuscript writing, and submission of this study. No other contributors were involved.

8. Acknowledgements

During the preparation of this manuscript, the author used ChatGPT (OpenAI) and Grammarly to assist with editing and formatting. The author has reviewed and edited all generated content and takes full responsibility for the final version of this publication. The author thanks their colleagues at www.saachistyle.com for insightful discussions and constructive feedback that informed the development of the payment fraud and loss report.

REFERENCES

1. Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550
2. Baader, G., & Krcmar, H. (2018). Reducing false positives in fraud detection: Combining the red flag approach with process mining. *International Journal of Accounting Information Systems*, *31*, 1–16. https://doi.org/10.1016/j.accinf.2018.03.004
3. Banasik, J., & Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, *183*(3), 1582–1594. https://doi.org/10.1016/j.ejor.2006.06.072
4. Card Not Present, & Radial. (2018). *False positives… the monster that's really killing you and how to survive* [White paper]. https://www.radial.com/files/CNP-Radial-White-Paper-False-Positives%20-%20final.pdf
5. Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining-based system for credit-card fraud detection in e-tail. *Decision Support Systems*, *95*, 91–101. https://doi.org/10.1016/j.dss.2017.01.002
6. Crook, J., & Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, *28*(4), 857–874. https://doi.org/10.1016/S0378-4266(03)00203-6
7. Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643
8. Dudík, M., Erhan, D., Langford, J., & Li, L. (2014). Doubly robust policy evaluation and optimization. *Statistical Science*, *29*(4), 485–511. https://doi.org/10.1214/14-STS500
9. Ehrhardt, A., Biernacki, C., Vandewalle, V., Heinrich, P., & Beben, S. (2021). Reject inference methods in credit scoring. *Journal of Applied Statistics*, *48*(13-15), 2734–2754. https://doi.org/10.1080/02664763.2021.1929090

10. Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161. https://doi.org/10.2307/1912352
11. Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446
12. Joachims, T., Swaminathan, A., & Schnabel, T. (2017). Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 781–789). https://doi.org/10.1145/3018661.3018699
13. Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, *18*(1), 140–181. https://doi.org/10.1007/s10618-008-0114-1
14. Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press. https://doi.org/10.1017/9781108653985
15. Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 275–284). https://doi.org/10.1145/3097983.3098066
16. Lebichot, B., Paldino, G. M., Siblini, W., He-Guelton, L., Oblé, F., & Bontempi, G. (2021). Incremental learning strategies for credit card fraud detection. *International Journal of Data Science and Analytics*, *12*(2), 165–174. https://doi.org/10.1007/s41060-021-00258-0
17. Li, L., Chu, W., Langford, J., & Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (pp. 297–306). https://doi.org/10.1145/1935826.1935878
18. Li, Z., Tian, Y., Li, K., Zhou, F., & Yang, W. (2017). Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*, *74*, 105–114. https://doi.org/10.1016/j.eswa.2017.01.011
19. Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260
20. Mancisidor, R. A., Kampffmeyer, M., Aas, K., & Jenssen, R. (2020). Deep generative models for reject inference in credit scoring. *Knowledge-Based Systems*, *196*, 105758. https://doi.org/10.1016/j.knosys.2020.105758
21. Mekterović, I., Karan, M., Pintar, D., & Brkić, L. (2021). Credit card fraud detection in card-not-present transactions: Where to invest? *Applied Sciences*, *11*(15), 6766. https://doi.org/10.3390/app11156766
22. Nanduri, J., Jia, Y., Oka, A., Beaver, J., & Liu, Y. (2020). Microsoft uses machine learning and optimization to reduce e-commerce fraud. *INFORMS Journal on Applied Analytics*, *50*(1), 64–79. https://doi.org/10.1287/inte.2019.1017
23. Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.1080/01621459.1994.10476818
24. Rodrigues, V. F., Policarpo, L. M., da Silveira, D. E., da Rosa Righi, R., da Costa, C. A., Barbosa, J. L. V., Antunes, R. S., Scorsatto, R., & Arcot, T. (2022). Fraud detection and prevention in e-commerce: A systematic literature review. *Electronic Commerce Research and Applications*, *56*, 101207. https://doi.org/10.1016/j.elerap.2022.101207
25. Soemers, D., Brys, T., Driessens, K., Winands, M., & Nowé, A. (2018). Adapting to concept drift in credit card transaction data streams using contextual bandits and decision trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v32i1.11411
26. Stewart, J. (2015, August 27). Lost sales from unfounded rejections dwarfs actual fraud losses, report says. *Digital Transactions*. https://www.digitaltransactions.net/lost-sales-from-unfounded-rejections-dwarfs-actual-fraud-losses-report-says/
27. Visa. (2016). *Layer up for safer holiday sales* [Merchant bulletin]. https://by.visa.com/content/dam/VCOM/global/support-legal/documents/cyber-monday-holiday-bulletin-2016.pdf
28. Vorobyev, I., & Krivitskaya, A. (2022). Reducing false positives in bank anti-fraud systems based on rule induction in distributed tree-based models. *Computers & Security*, *120*, 102786. https://doi.org/10.1016/j.cose.2022.102786
29. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953
30. Zohar, S. (2015, September 10). How consumers respond to false declines. Riskified. https://riskified.com/?p=409