

Article

Value, Risk, and Recoverability: An Interpretable Order-Level Prioritization Framework for Service Recovery in E-Commerce

Youness Madane ¹ and Mohamed Azeroual ^{1,2,*}

¹ Laboratory of Economy & Management, Polydisciplinary Faculty of Khouribga, Sultan Moulay Slimane University, Beni Mellal 25000, Morocco; youness.madane@usms.ac.ma

² Laboratoire de Management, Finances, Marketing & Développement International, École Supérieure de Management, Télécommunications et Informatique, 22 Avenue Omar Ibn Al Khattab, Agdal, Rabat 10090, Morocco

* Correspondence: m.azeroual@ensaadmin.ma

Abstract

Customer prioritization in e-commerce remains dominated by value-based logics that allocate retention effort to the most profitable customers, even though risk-based targeting can be ineffective when intervention responsiveness is ignored. This study aims to develop and empirically test a Value–Risk–Recoverability (VRR) framework that prioritizes service-recovery effort under a fixed intervention budget. The framework draws its three axes from the synergy of three theoretical streams: customer-equity theory motivates the value axis, the churn and defection-management literature motivates calibrated dissatisfaction risk, and service-recovery theory—through the distinction between operational and structural causes of failure—motivates the recoverability axis, which operationalizes the intervention-responsiveness critique of risk-based targeting. The framework is instantiated on the public Brazilian marketplace dataset by Olist (91,954 customers; 93,663 delivered orders, 2016–2018) using unsupervised clustering for behavioral segmentation, calibrated gradient-boosting models to predict order-level dissatisfaction under a strictly temporal hold-out, and SHAP attribution to decompose predicted risk into operational and structural components. Results show that dissatisfaction becomes predictable mainly as fulfillment unfolds (out-of-sample AUC of 0.72 with in-fulfillment signals versus 0.61 at order time); that roughly 76% of predicted risk loads on operational, addressable factors; and that, at a 10% intervention budget, value-based targeting captures only about 30% of realized recoverable value against roughly 96% for risk-aware policies. The study contributes a theoretically grounded, interpretable, and reproducible prioritization logic for service recovery, together with an explicit account of the boundary conditions under which each axis carries decision-relevant information.

Keywords: customer prioritization; service recovery; dissatisfaction prediction; explainable machine learning; SHAP; e-commerce marketplace



Academic Editor: Jianwen Su

Received: 8 June 2026

Revised: 6 July 2026

Accepted: 10 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Few decisions in customer management are as consequential, or as routinely simplified, as the decision of whom to prioritize. Faced with finite budgets for retention, loyalty, and service recovery, firms must rank customers and act on a subset. The prevailing logic for doing so is value-based: customer-equity and customer-lifetime-value frameworks direct disproportionate attention to the most profitable customers, and a substantial body of evidence supports the profitability of such prioritization when it is implemented well [1,2].

However, value is only one of the quantities that should govern the allocation of retention effort and treating it as the sole criterion can be costly.

Two complementary studies suggest why. First, retention is fundamentally about risk: investing in a customer who was never going to defect yields little, so the risk of attrition or dissatisfaction is a natural second axis [3,4]. Second, and more subtly, Ascarza [5] shows that targeting high-risk customers can be ineffective—even counterproductive—when the intervention does not actually change their behavior. What matters is not risk per se but the responsiveness of risk to intervention. In a service context, this maps directly onto the recoverability of a dissatisfied customer: failures that stem from addressable operational causes are recoverable, whereas failures rooted in structural mismatch are not [6]. A prioritization logic that ignores recoverability will misallocate recovery budgets toward customers who cannot be saved.

The idea that customers differ in value, and that marketing resources should be allocated accordingly, is central to modern customer management. Customer-equity frameworks formalize the firm's customer base as a portfolio of discounted future cash flows and recommend differential investment across that portfolio [2]. Homburg, Droll, and Totzek [1] provide direct evidence that systematic customer prioritization improves average customer profitability and firm performance, while cautioning that prioritization must be implemented with care to avoid alienating lower-tier customers. In transactional and marketplace contexts where repeat purchasing is sparse, full lifetime-value estimation is difficult, and probability models of the customer base offer one principled route [7]; in their absence, observed monetary value is a defensible, if imperfect, proxy for the value axis.

A parallel tradition treats retention as a prediction problem: identify customers at risk of defection and intervene. Defection-detection models have a long methodological history [3], and customer-relationship-management investments have measurable effects on retention [4]. The decisive complication is Ascarza's [5] demonstration that ranking customers by churn risk can be ineffective because the highest-risk customers are not necessarily those whose behavior an intervention will change. Optimal targeting depends on treatment responsiveness, not risk alone. A parallel methodological literature on uplift modeling operationalizes this responsiveness directly by estimating individual treatment effects and, more recently, by integrating them with expected customer value to align targeting with business profit rather than response probability alone [8,9]. Our recoverability axis pursues a similar intuition with an interpretability-first, attribution-based proxy suited to settings where experimental treatment data are unavailable. This insight motivates a third axis beyond value and risk: a measure of how amenable a customer's risk is to the intervention on offer.

In e-commerce, perceived service quality is shaped heavily by fulfillment, and electronic service-quality scales explicitly foreground efficiency, reliability, and fulfillment as primary dimensions [10]. Satisfaction mediates the link from service quality to retention [11], and the service-recovery literature establishes that the recoverability of a failure depends on its cause and on the firm's response [6]. Failures arising from controllable operational breakdowns—late or unreliable delivery—are, in principle, both preventable and recoverable; failures arising from product–need mismatch are less so. This distinction is the conceptual basis for our recoverability axis. The empirical setting is the Brazilian marketplace, which over the mid-2010s was among the fastest-growing e-commerce markets worldwide, undergoing rapid expansion in order volume, marketplace penetration, and last-mile logistics build-out—conditions under which fulfillment reliability is both strained and decisive for customer experience, and therefore a particularly informative context for studying operationally driven dissatisfaction.

Behavioral segmentation has moved from pre-specified schemes such as RFM [12] toward data-driven clustering [13–16], trading interpretability for flexibility. Explainable-AI techniques—feature importance [17], SHAP [18], and related methods cataloged by Arrieta et al. [19]—restore a measure of transparency to flexible models, and recent work couples clustering with interpretable classifiers for customer profiling [20,21], including SHAP-based churn analyses and systematic evidence on the interpretability–performance trade-off [22,23]. Canhoto and Clear [24] argue that such transparency is a precondition for managerial adoption. We build on this strand but reorient it from description toward decision: the interpretable models here serve to construct and justify a prioritization signal, not merely to label segments.

Table 1 synthesizes these four research streams, their prioritization criteria, and what each leaves unaddressed. Read together, they reveal a consistent pattern: value-based, risk-based, service-recovery, and interpretable-segmentation approaches have largely developed in isolation, each supplying one ingredient of a prioritization decision while omitting the others. Value-based logics identify who is worth saving but not who is at risk or reachable; risk-based logics locate likely failures but, as Ascarza [5] shows, not whether those failures respond to intervention; service-recovery theory explains when a failure is recoverable yet has not been cast as a quantitative, order-level targeting rule under a fixed budget; and interpretable-segmentation methods describe customers without turning that description into an auditable decision signal [25]. No widely adopted framework integrates value, risk, and recoverability into a single, transparent prioritization logic, tests it on real marketplace data, and is explicit about the conditions under which each axis carries decision-relevant information. The present article addresses this gap: the distinctive contribution of the VRR framework is to unify these streams into one interpretable priority signal in which value, risk, and recoverability are jointly measured at the order level, and in which the responsiveness critique of risk-based targeting is operationalized—rather than merely acknowledged—through a measurable recoverability axis.

Table 1. Four research streams bearing on customer prioritization, their prioritization criteria, and what each leaves unaddressed, situating the contribution of the VRR framework.

Research Stream	Prioritization Criterion/Core Logic	Representative Works	What It Leaves Unaddressed	How VRR Extends It
Value-based/customer equity	Allocate effort to the most economically valuable customers (CLV, customer equity)	Homburg et al. [1]; Rust et al. [2]; Fader & Hardie [7]	Value is weakly informative about who is at risk or reachable; can concentrate budget on satisfied, loyal customers	Retains value as one axis (bounded monetary proxy), demoted from sole criterion to sequencing/threshold role
Risk-based/churn & defection	Identify and target customers at risk of defection or dissatisfaction	Neslin et al. [3]; Verhoef [4]; Ascarza [5]	High risk $\neq$ responsiveness to intervention [5]; risk alone ignores whether a failure is addressable	Makes calibrated risk the second axis and operationalizes Ascarza’s responsiveness critique via recoverability
Service recovery	Restore satisfaction post-failure; recoverability depends on failure cause and firm response	Smith et al. [6]; Gubela et al. [8]; Haupt et al. [9]	Largely descriptive; not cast as a quantitative, order-level targeting rule under a fixed budget	Converts cause-of-failure logic into a measurable recoverability axis (operational vs. structural SHAP share)
Interpretable segmentation/XAI	Data-driven clustering plus explainability to profile customers	RFM [10]; clustering [11–14]; SHAP/XAI [15–20]	Used to describe/label customers, not to build and justify a decision signal	Uses interpretable models to construct and audit the priority signal, supporting governance and adoption

Concretely, the study addresses a decision problem faced by any e-commerce firm operating under a fixed service-recovery budget: given a continuous stream of order–customer cases, which cases should receive scarce recovery effort, and on what basis, so that the budget captures the largest possible share of value that can actually be saved? Answering this question requires knowing not only which customers are economically valuable and which are at risk of dissatisfaction, but also whether their risk is of a kind that a recovery action can plausibly address. We therefore pursue four research objectives. Objective 1 (O1): To formalize a transparent prioritization logic that integrates economic value, predicted dissatisfaction risk, and the recoverable component of that risk into a single, interpretable decision signal. Objective 2 (O2): To establish at what point along the order-to-review timeline dissatisfaction becomes predictable, and with what accuracy, distinguishing information available at order time from information that resolves as fulfillment unfolds. Objective 3 (O3): To quantify the share of predicted dissatisfaction risk that is operationally attributable—and therefore potentially recoverable—rather than structural. Objective 4 (O4): To evaluate, out of sample and at realistic intervention budgets, how much recoverable value competing targeting policies capture, and thereby to identify the conditions under which each axis of the framework carries decision-relevant information.

This article develops and tests a prioritization framework that integrates these three axes—Value, Risk, and Recoverability (VRR)—within a transparent, interpretable workflow suited to e-commerce. Interpretability is treated here not as an academic nicety but as a practical condition for adoption and governance: as automated profiling falls under closer regulatory and organizational scrutiny [19,26,27], firms increasingly require segmentation and targeting logic that can be explained, documented, and contested [28]. We therefore build the framework from components whose contributions can be decomposed at the variable level using SHAP-based feature attribution for calibrated tree-based models [17,18,29]. Empirically, the contribution is an order-level service-recovery prioritization in a marketplace setting, with customer monetary value entering as a bounded proxy for the value axis.

We instantiate the framework on a real, large-scale, anonymized dataset from the Brazilian marketplace Olist [30], deliberately avoiding the constructed datasets whose artificial separability has limited the credibility of prior demonstrations. Three contributions follow. First, conceptually, we articulate VRR as a prioritization logic that subsumes value-only and risk-only targeting as special cases and that operationalizes Ascarza’s responsiveness argument through a measurable recoverability axis. Second, methodologically, we distinguish a proactive at-order-time risk model from a deployable in-fulfillment model and show, through a temporally held-out out-of-sample decision-policy evaluation, when and why each axis matters. Third, empirically, we document that dissatisfaction in this marketplace is overwhelmingly operational and recoverable, that value-based targeting badly misallocates recovery budgets, and that the three axes empirically coincide in fulfillment-driven settings—an honest, falsifiable boundary condition rather than a universal claim of dominance.

The remainder of the paper proceeds as follows. Section 2 describes the data, methods, and the VRR framework. Section 3 reports results. Section 4 discusses implications, governance, and limitations, and Section 5 concludes.

2. Materials and Methods

2.1. The Value–Risk–Recoverability Framework

We consider a firm that must allocate a limited service-recovery budget across its customer base. For each order–customer case i , the framework defines three axes.

Value (V). The economic worth of the customer, operationalized as observed order monetary value, is used as a bounded proxy for the value axis in a sparse-repeat marketplace where formal lifetime-value models are not identifiable [7]. V answers: how much economic value is at stake in this order–customer case?

Risk (R). The probability that the customer experiences a dissatisfaction event, estimated by a calibrated supervised model and expressed as a well-behaved probability. R answers: how likely is loss?

Recoverability (Re). The share of the predicted risk attributable to operationally addressable causes, estimated from the risk model’s SHAP attribution. It re-operationalizes Ascarza’s [5] responsiveness argument as an attribution-based proxy rather than a causal treatment effect: it measures the share of predicted risk that is operationally attributable and therefore plausibly addressable, distinguishing failures that may respond to recovery actions from those rooted in structural causes [6]. It does not establish that customers would, in fact, respond to a given intervention; estimating such responsiveness requires experimental evidence beyond the scope of this study. Recoverability is an attribution-based proxy, not a causal treatment effect: it estimates how much of the predicted risk is operationally attributable, not the response to any specific intervention.

These axes combine into a priority signal, in its simplest multiplicative form $\text{Priority}_i = V_n(i) \times R_i \times \text{Re}_i$, where V_n is the value expressed as a percentile to bound its scale. The signal is deliberately transparent: each case’s priority decomposes into interpretable contributions, and the special cases V alone (value-based targeting) and R alone (risk-based targeting) are nested within it. Cases are then mapped to actions by their position in the value–risk plane, modulated by recoverability: high-value, high-risk, recoverable customers warrant targeted service recovery; high-value, low-risk customers warrant value-tiered loyalty investment; low-value, high-risk customers warrant monitoring; and the remainder are de-prioritized. Figure 1 summarizes the logic.

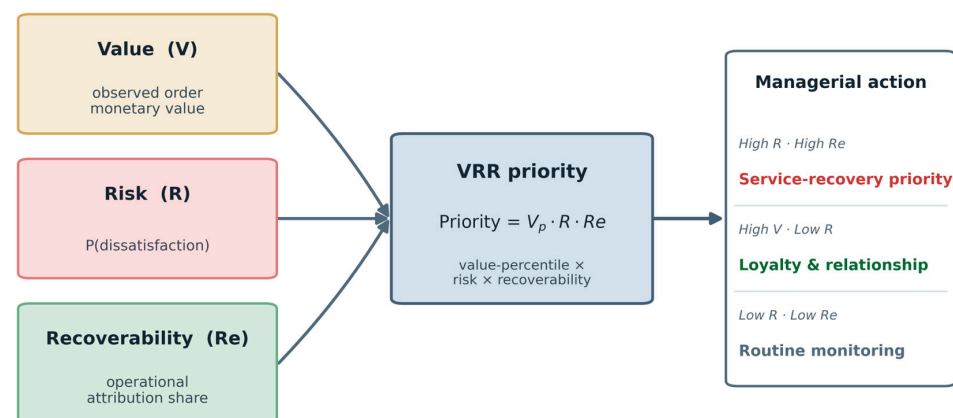


Figure 1. The Value–Risk–Recoverability prioritization logic. Three interpretable axes—value, predicted dissatisfaction risk, and the operational recoverability of that risk—combine into a priority signal that maps onto distinct managerial actions.

Consistent with the study’s objectives and its empirical, partly exploratory design, we address one research question and test three research hypotheses. Because the point at which dissatisfaction becomes predictable is an open empirical matter rather than a directional claim, we frame it as a research question (RQ), whereas the framework’s decision-relevant implications are stated as directional, falsifiable hypotheses (H1–H3) evaluated with the statistical procedures described below.

RQ (O2). At what point along the order-to-review timeline does order-level dissatisfaction become predictable, and how much does predictive accuracy improve once in-fulfillment information becomes available relative to information available at order time?

H1 (O4). *Ranking order–customer cases by economic value alone captures substantially less realized recoverable value at a fixed intervention budget than risk-aware ranking.*

H2 (O3). *In this e-commerce setting, dissatisfaction risk is predominantly operationally attributable—and therefore potentially recoverable—with the majority of risk attribution loading on fulfillment-related factors.*

H3 (O4). *The marginal contribution of the value and recoverability axes over risk alone is contingent on whether risk and recoverability coincide: where dissatisfaction is operationally driven, the three axes converge, and risk becomes a near-sufficient targeting statistic, whereas they diverge where dissatisfaction is structural.*

2.2. Data

The choice of the Olist dataset is deliberate and central to the study's aims rather than one of convenience. Three properties make it uniquely suited to the VRR framework. First, it is one of the very few large-scale, real, fully anonymized, and publicly available e-commerce datasets that link order-level fulfillment records—estimated and realized delivery timestamps, freight, and seller and customer geography—to the corresponding customer review, which is precisely the join required to construct and validate a recoverability axis grounded in operational versus structural causes of dissatisfaction. Constructed or survey-based datasets, whose artificial separability has limited the credibility of prior demonstrations, lack this order-to-outcome linkage. Second, its public availability makes every step of the analysis fully reproducible and independently auditable, in keeping with the interpretability and governance objectives of the framework. Third, with roughly 100,000 orders spanning a continuous marketplace ramp, it provides sufficient volume and temporal depth to support a strictly time-ordered train–test split, which is essential for the out-of-sample policy evaluation. The analysis period (September 2016–August 2018) corresponds to the full temporal coverage of the public release; it is not a selected sub-window but the entire span of the data, and it captures multiple seasonal cycles and a complete growth phase of the marketplace. Crucially, the mechanisms the framework isolates—the dependence of dissatisfaction on delivery performance, and the operational-versus-structural decomposition of risk—are structural features of marketplace logistics rather than period-specific artifacts, so the framework's logic is method-dependent rather than epoch-dependent: it can be re-estimated on any dataset offering the same order-to-outcome linkage.

We considered but did not adopt more recent data (e.g., 2022–2024): to our knowledge, no comparably granular, review-linked, order-level marketplace dataset with fulfillment timestamps is openly available for that period, and substituting proprietary data would defeat the reproducibility objective that motivates the framework. We therefore retain Olist as an established, reproducible benchmark and treat the temporal external validity of the findings—particularly in light of post-pandemic shifts in e-commerce logistics and consumer behavior—as an explicit boundary condition examined in the limitations.

The empirical basis is the Brazilian E-Commerce Public Dataset by Olist [30], a real, anonymized record of roughly 100,000 orders placed on a multi-seller marketplace between September 2016 and August 2018, distributed across relational tables for orders, items, payments, reviews, products, sellers, and customer geography. We restricted the analysis to delivered orders and joined the relevant tables. For the segmentation stage, transactions were aggregated to the unique customer identifier, yielding 91,954 customers after removing records with missing payment, review, or delivery information, and trimming the upper 0.5% of monetary and freight-ratio values; repeat purchasing is rare (about 3% of customers), so customer-level recency behaves as a tenure proxy, a boundary condition

we revisit in the limitations. For the risk and prioritization stages, the unit of analysis is the reviewed delivered order, of which 93,663 remain after analogous cleaning, with a global dissatisfaction base rate (a review score of 1 or 2) of 12.7%; the strictly later temporal test set used for risk evaluation has a lower base rate of 10.0%, reflecting period-specific fulfillment conditions.

2.3. Segmentation

Seven standardized, behaviorally interpretable features—recency, monetary value (log), item count (log), average review score, average payment installments, average delivery days, and freight-to-price ratio—were submitted to K-means clustering. The number of clusters was selected by jointly considering the elbow criterion, the average silhouette [31], the Davies–Bouldin index, the Calinski–Harabasz index, and managerial interpretability. Stability was assessed by the adjusted Rand index (ARI) across seven random seeds and by agreement with Ward agglomerative clustering on a sample.

2.4. Risk Models

Order-level dissatisfaction was modeled as a binary outcome (a review score of 2 or below). We deliberately separated two information regimes. The at-order-time model uses only features available when the order is placed: the estimated delivery window, freight value and ratio, item count, price, product category, payment type and installments, product weight and volume, and customer–seller geographic proximity (same state, same region). The in-fulfillment model adds realized fulfillment signals—actual delivery time, delay relative to the estimate, and a late-delivery indicator—that become available as fulfillment unfolds. It is useful to distinguish three windows: before delivery, when the estimated delay and carrier status support pre-emptive action; at the delivery event, when the realized delivery time and late-delivery flag are resolved; and after delivery but before the review is submitted, the window in which most predictive signals are fully observed and recovery outreach remains feasible. Models were trained on orders placed before 1 May 2018 and evaluated on later orders, a strictly temporal hold-out (68,892 training, 24,771 test orders). We compared a delivery-only logistic baseline, a random forest, and gradient boosting, each in uncalibrated and isotonicly calibrated form, reporting the area under the ROC curve (AUC), the area under the precision–recall curve (PR-AUC), and the Brier score. The deployed in-fulfillment model is a gradient-boosting classifier calibrated by isotonic regression using cross-validation during the training period. To probe robustness against circularity, we additionally re-estimated the in-fulfillment model under successive removal of the late-delivery flag, the realized delivery days, and all delay-related variables. Table 2 lists the predictor blocks and their operational or structural classification.

Table 2. Risk-model predictor blocks and their operational/structural classification. At-order features are available when the order is placed; in-fulfillment features become available during delivery, before the review.

Block	Features	Regime	Type
Fulfillment estimate	Estimated delivery days	At order	Operational
Logistics cost	Freight value, freight ratio	At order	Operational
Geographic proximity	Same state, same region	At order	Operational
Product physical	Weight, volume	At order	Operational
Basket/price	Items, price	At order	Structural
Payment	Installments, payment type	At order	Structural
Product category	Category (top-12 + other)	At order	Structural
Realized fulfillment	Delivery days, delay vs. estimate, late flag	In fulfillment	Operational

2.5. Recoverability and the VRR Out-of-Sample Decision-Policy Evaluation

Recoverability was computed from the calibrated risk model's SHAP attribution by partitioning predictors into operational (fulfillment and logistics) and structural (product and customer-side) blocks; an order's recoverability, Re , is the operational share of its absolute SHAP mass for the dissatisfaction class.

The theoretical rationale for this operationalization rests on two premises. First, service-recovery theory holds that the recoverability of a failure is governed by its cause: failures arising from controllable, firm-attributable operational breakdowns are, in principle, both preventable and recoverable, whereas failures rooted in structural product–need mismatch are not [6]. Recoverability is, therefore, at the conceptual level, a statement about the causal locus of a failure. Second, SHAP provides an additive, locally accurate, and consistent decomposition of a model's predicted risk into per-feature contributions [18], which allows us to measure, for each case, the share of predicted risk that loads on firm-controllable operational features rather than on structural ones. Combining the two premises, the operational SHAP share is an interpretable, case-level proxy for the theoretically defined recoverable fraction of risk: it estimates how much of a case's predicted dissatisfaction is attributable to causes the firm can act upon. This attribution-based construction is deliberately chosen over a causal treatment-effect estimate because the observational, non-experimental data contain no intervention and therefore cannot identify a response to one; the attribution proxy is what the data can support, and its interpretability aligns with the framework's governance objectives. It operationalizes the responsiveness intuition of Ascarza [5] and the value-driven targeting logic of the uplift literature [8,9] under an explicit interpretability-first constraint.

The out-of-sample decision-policy evaluation was conducted entirely out of sample, on the temporally held-out period. For a cohort of 8000 held-out orders, each customer's value V (order payment value), calibrated risk R , and recoverability Re were combined into the VRR priority. We defined realized recoverable value as the value accrued to late-delivered failed orders (a review score of 2 or below together with late delivery), used as a conservative proxy for realized, operationally attributable recoverable value, and compared five targeting policies—VRR, risk-only, value-adjusted risk, value-only, and random—by the cumulative share of realized recoverable value captured as the targeted fraction increases. Crucially, policies rank on model-derived quantities while evaluation uses realized outcomes, so no policy can trivially capture its own objective. Uncertainty was quantified by a paired bootstrap with 1000 resamples on the held-out decision cohort.

All analyses were conducted in Python 3.12 (Python Software Foundation, Wilmington, DE, USA), using the scikit-learn 1.6.1 and SHAP 0.46.0 libraries.

The authors designed the framework, specified and supervised all analyses, interpreted the results, and verified every reported figure; large-language-model assistance was used as a computational and drafting aid under author supervision, as disclosed in the Acknowledgments.

3. Results

3.1. Behavioral Segmentation

Exploratory analysis (Figure 2) confirms the heavy-tailed monetary distribution, the characteristic bimodality of online ratings, and a marked decline in review scores as delivery time lengthens, with dissatisfied customers experiencing delivery times roughly twice as long as satisfied customers. Pairwise correlations among the clustering features are modest, the largest being a negative association between monetary value and freight ratio; no variable mechanically determines the others. Segmentation was performed on the customer-level variables listed in Table 3.

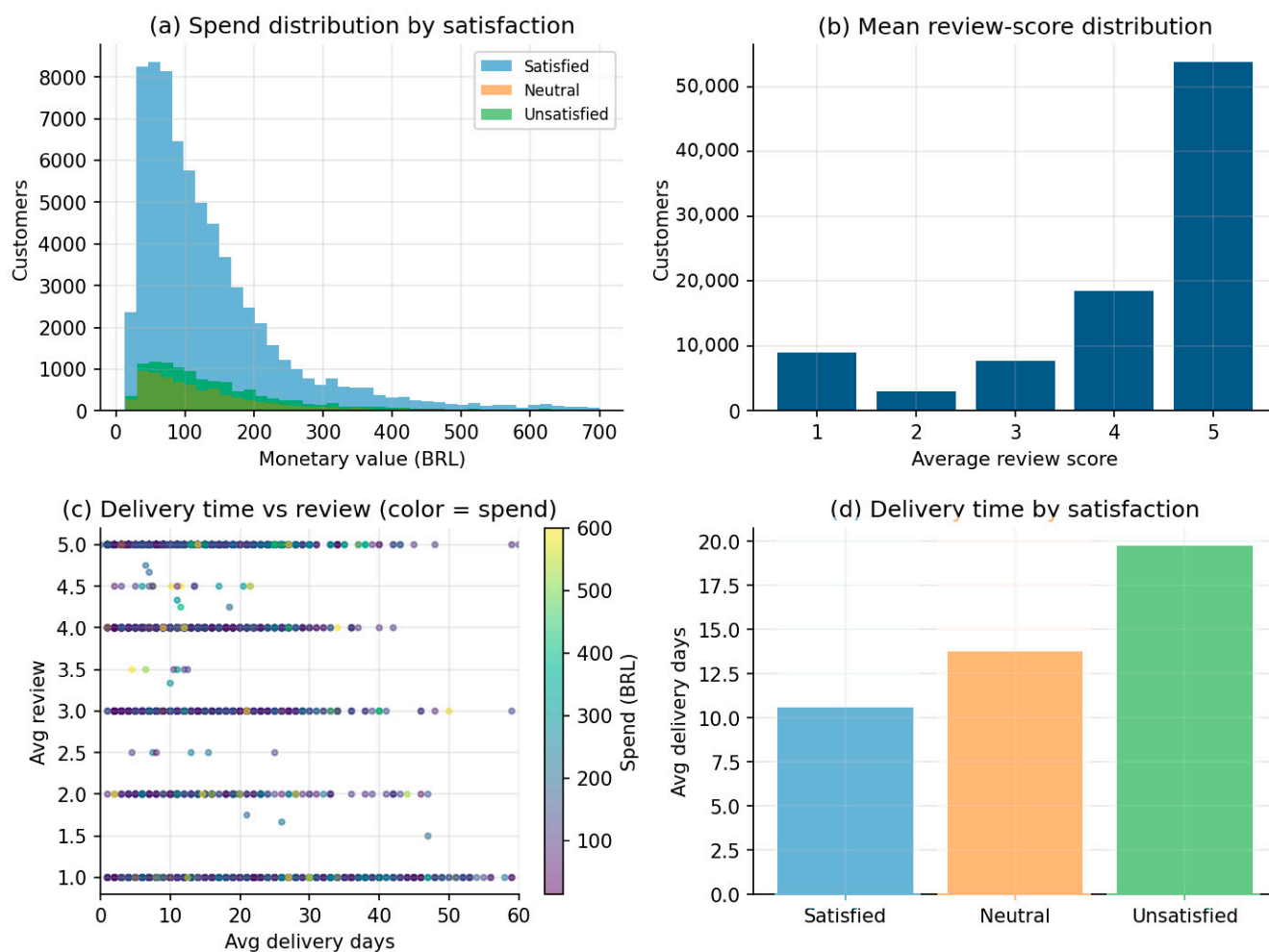


Figure 2. Exploratory overview. (a) Spend distribution by satisfaction band; (b) review-score distribution; (c) delivery time versus review, colored by spend; (d) mean delivery time by satisfaction band. Dissatisfaction tracks delivery time, not spend.

Table 3. Customer-level segmentation variables ($n = 91,954$). The customer identifier and state are retained for description but excluded from clustering.

Variable	Type	Description
Recency (days)	Integer	Days from last purchase to reference date (tenure proxy)
Monetary (BRL)	Continuous	Total payment value across delivered orders
Items	Integer	Total items purchased
Average review	Continuous	Mean post-purchase review score (1–5)
Avg. installments	Continuous	Mean payment installments (financing proxy)
Avg. delivery days	Continuous	Mean days from purchase to delivery
Freight ratio	Continuous	Mean freight value divided by item price

The cluster-selection diagnostics (Figure 3) flatten after four clusters, where the Davies–Bouldin index improves materially relative to three (1.55 versus 1.76). Silhouette values are modest throughout (about 0.19 at $k = 4$), as expected for real, overlapping behavioral data; higher- k solutions improved separation only by fragmenting small groups. We retained $k = 4$ on the joint basis of the diagnostics and interpretability. The solution is stable to initialization (pairwise ARI ≥ 0.97 across seeds) and in moderate agreement with Ward

clustering (ARI = 0.45), and the first two principal components account for 46.4% of feature variance (Figure 4).

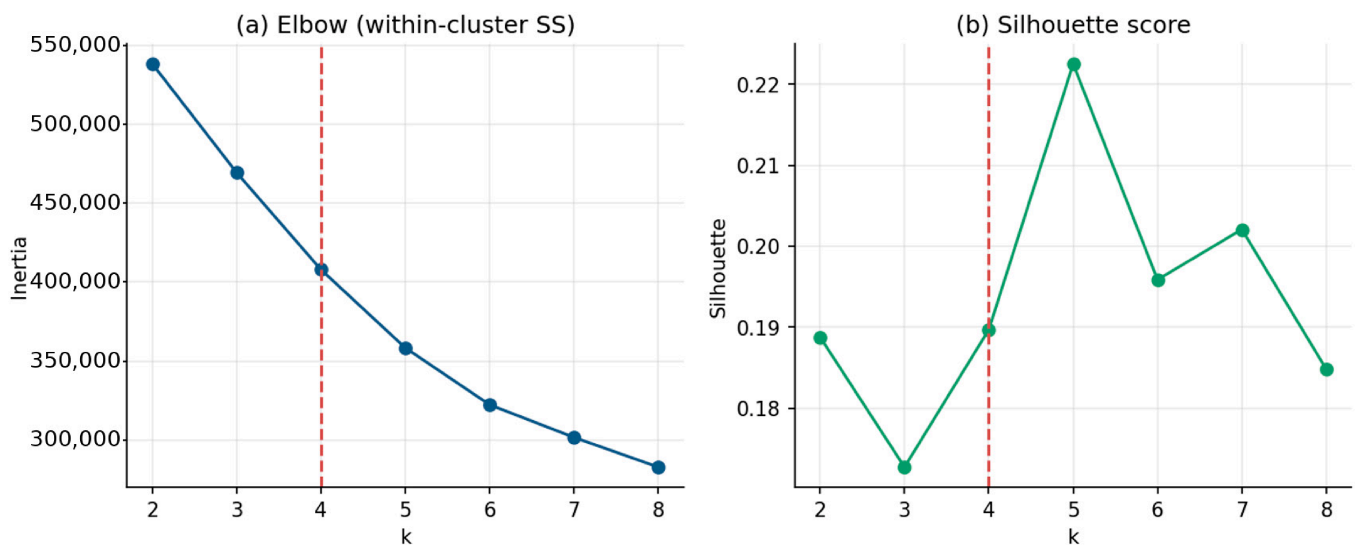


Figure 3. Cluster-selection diagnostics for $k \in \{2, \dots, 8\}$: (a) elbow curve, with a slope change at $k = 4$; (b) silhouette profile. The red dashed line marks the selected number of clusters ($k = 4$). Selection reflects a trade-off between compactness and interpretability.

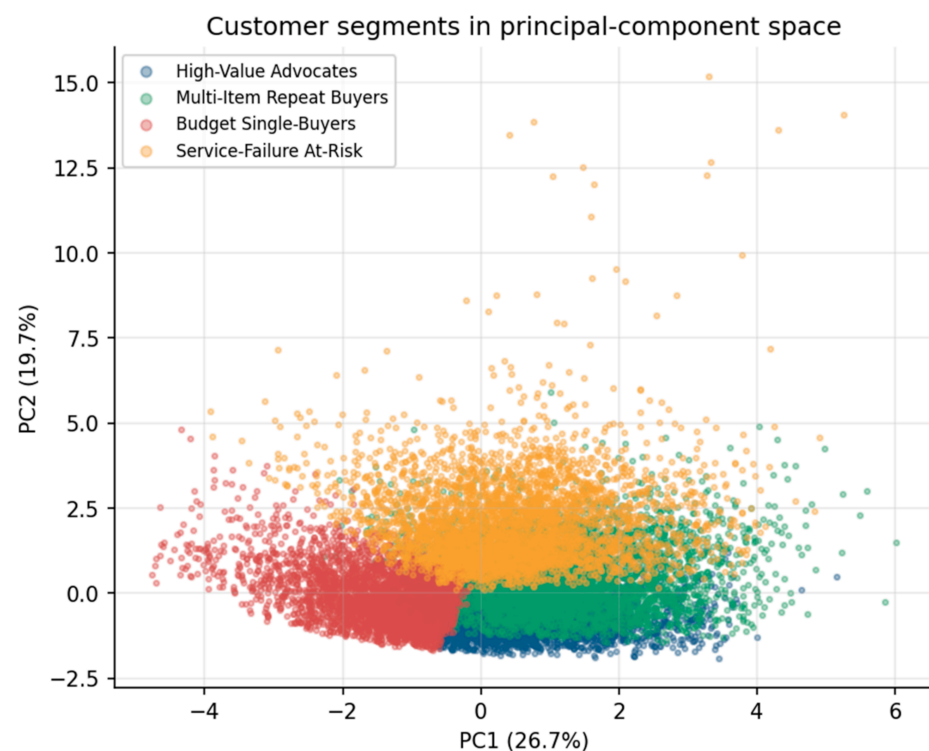


Figure 4. Principal-component projection of the four segments. PC1 and PC2 jointly explain 46.4% of the variance, so the projection is illustrative of overlap rather than a proof of compactness.

Four archetypes emerge (Table 4). High-Value Advocates (43.2%) combine solid spend, the highest review scores (4.63), fast delivery, and low freight burden. Multi-Item Repeat Buyers (11.2%) carry the largest baskets and the only above-unity purchase frequency. Budget Single-Buyers (33.4%) make small purchases bearing a high freight-to-price ratio (0.49) yet remain satisfied (4.54), indicating that relative shipping cost does not by itself depress experience. Service-Failure At-Risk customers (12.2%) are defined by experience

rather than spend: their mean review score is 1.59 and their mean delivery time is 24.3 days, more than double the rest, at mid-range spend (Figure 5).

Table 4. Segment behavioral profiles. Monetary values in BRL (mean $\pm$ 1 SD for the first column). “Review” is the mean score (1–5); “Delivery” is the mean days; “Freight” is the mean freight-to-price ratio; “Freq.” is the mean orders.

Segment	<i>n</i>	%	Monetary	Review	Delivery	Items	Freight	Freq.
High-Value Advocates	39,685	43.2	196.8 $\pm$ 148	4.63	10.6	1.00	0.16	1.00
Multi-Item Repeat Buyers	10,294	11.2	249.5 $\pm$ 185	3.85	10.6	2.46	0.31	1.29
Budget Single-Buyers	30,733	33.4	55.8 $\pm$ 24	4.54	10.0	1.02	0.49	1.00
Service-Failure At-Risk	11,242	12.2	152.6 $\pm$ 135	1.59	24.3	1.05	0.28	1.01

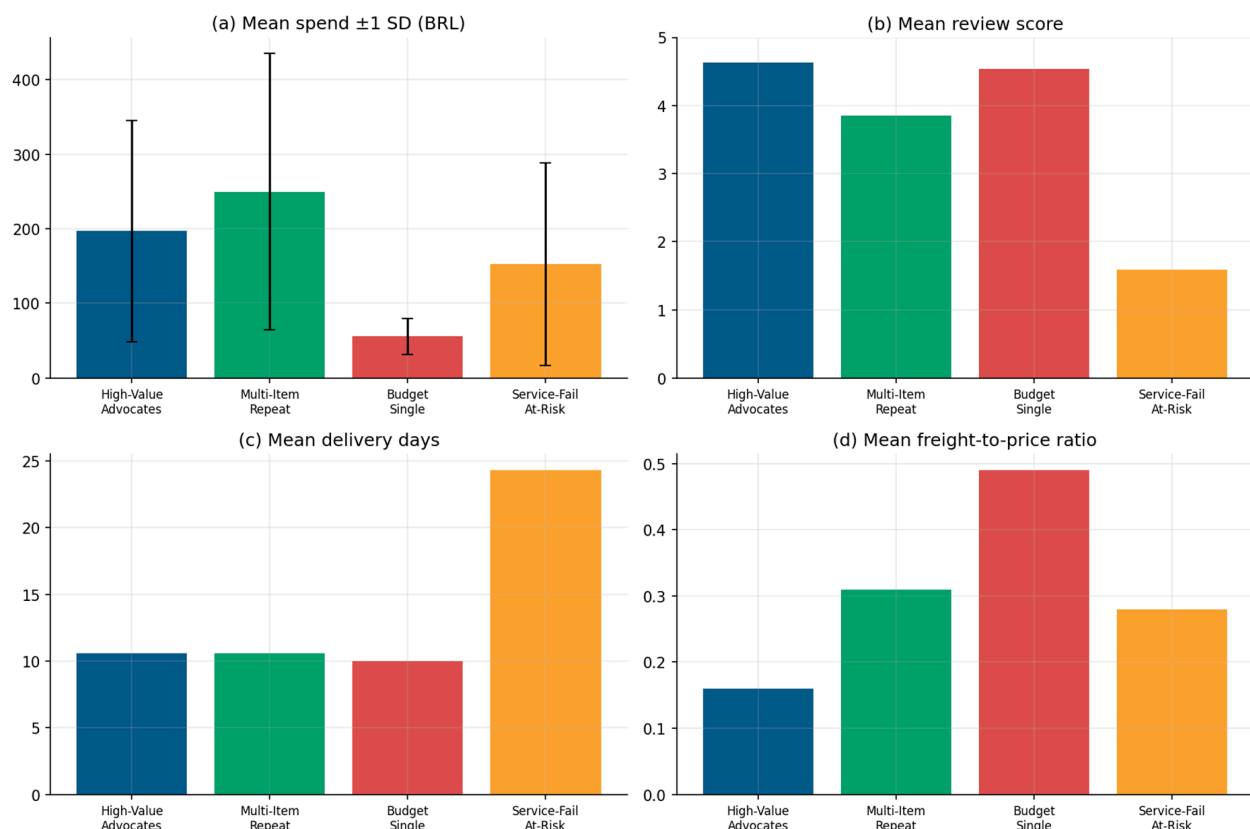


Figure 5. Segment profiles: (a) mean spend $\pm$ 1 SD; (b) mean review; (c) mean delivery time; (d) mean freight ratio. The Service-Failure At-Risk segment is distinguished by delivery time and review, not spend.

As an internal check on separability, a random forest reconstructs the segment labels from the full feature space with 0.982 cross-validated accuracy, while a parsimonious model using only four observable features achieves a realistic 0.901; the experience-defined at-risk segment is the hardest to recover from value features alone, confirming that it is constituted by service experience rather than spend. The segmentation is not itself the prioritization instrument—the decision-relevant analysis is the order-level prioritization reported in Section 3.4—but it contributes to the framework through three practical channels. First, it validates the framework’s central premise before any model is trained: the existence of an archetype defined by service experience rather than spend (low review, doubled delivery time, mid-range value) establishes empirically that dissatisfaction in this market is experience-driven rather than price-driven, which is precisely the pattern the operational/structural decomposition of risk presupposes. Second, it supplies the

managerial reading grid for the order-level scores: individual VRR scores are aggregated to segment level, which converts a stream of case-level numbers into a small set of named populations to which budgets, playbooks, and monitoring routines can be attached. Third, the archetypes inform how a prioritized case should be handled once selected: the same priority score warrants different treatment for a Multi-Item Repeat Buyer, where the relationship at stake extends beyond the focal order, than for a Budget Single-Buyer, where the freight-sensitive profile suggests compensation levers of a different kind. Figure 6 decomposes each archetype's normalized signature, with the at-risk segment dominated by low review and long delivery.

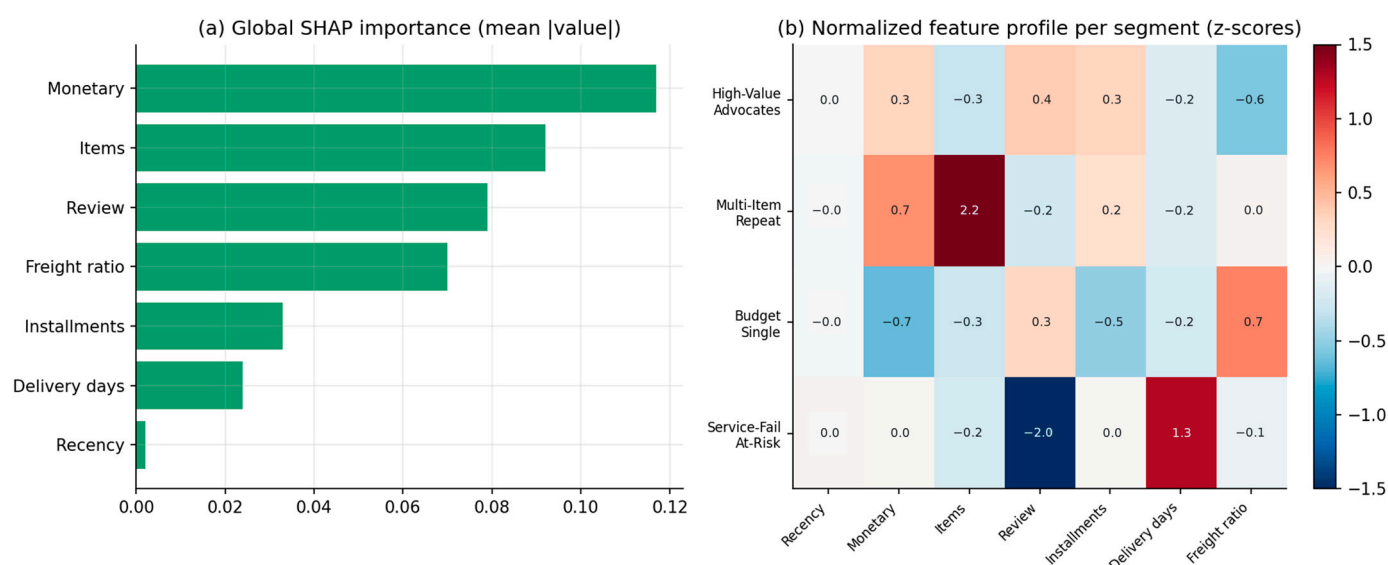


Figure 6. (a) Global SHAP importance for the segment classifier; (b) normalized feature profile per segment (z-scores). The Service-Failure At-Risk row loads on low review and long delivery; Multi-Item Repeat Buyers on basket size; Budget Single-Buyers on low value and high freight.

In terms of the study's design, this segmentation supports Objective 1 by supplying the behavioral context in which the VRR axes are read: it identifies for whom the framework's signals will later be interpreted, before any prediction is made.

3.2. When Does Dissatisfaction Become Predictable?

Table 5 reports the risk-model comparison on the temporal hold-out. At order time, dissatisfaction is barely predictable: a delivery-only logistic baseline is near chance (AUC 0.525), and richer at-order models—random forest and gradient boosting—plateau around 0.61. Adding in-fulfillment signals lifts performance markedly, raising the in-fulfillment models to an out-of-sample AUC of about 0.72; the isotonicly calibrated gradient-boosting model attains AUC 0.724, PR-AUC 0.348, and Brier 0.075 (Figure 7a,b). Because the test base rate is 0.10, the PR-AUC of 0.348 corresponds to roughly a three-and-a-half-fold improvement over the no-skill precision baseline, and Figure 7b displays the full precision–recall trade-off across all operating thresholds. We deliberately do not report single-threshold classification metrics such as the F1-score: the framework never converts scores into a fixed accept/reject classification, but ranks cases and acts on the top fraction that the intervention budget allows, so its operating characteristics are those of a ranking rule evaluated at budget levels, which Section 3.4 reports directly. The contrast is the result: the information that predicts dissatisfaction arrives with the shipment, not with the order. Isotonic calibration preserves discrimination while delivering the well-behaved probabilities required for prioritization (Figure 7c).

Table 5. Risk-model performance on the temporal hold-out (24,771 orders; test base rate 0.10). AUC = area under ROC; PR-AUC = area under precision–recall curve; lower Brier is better. Bold indicates the selected model (the best-performing configuration: highest AUC and PR-AUC, lowest Brier).

Model	Features	AUC	PR-AUC	Brier
Delivery-only logistic	At order	0.525	0.112	0.089
Gradient boosting	At order	0.607	0.176	0.086
Random forest	At order	0.611	0.181	0.087
Gradient boosting (calibrated)	At order	0.609	0.178	0.084
Random forest	In fulfillment	0.716	0.335	0.083
Random forest (calibrated)	In fulfillment	0.711	0.326	0.077
Gradient boosting (calibrated)	In fulfillment	0.724	0.348	0.075

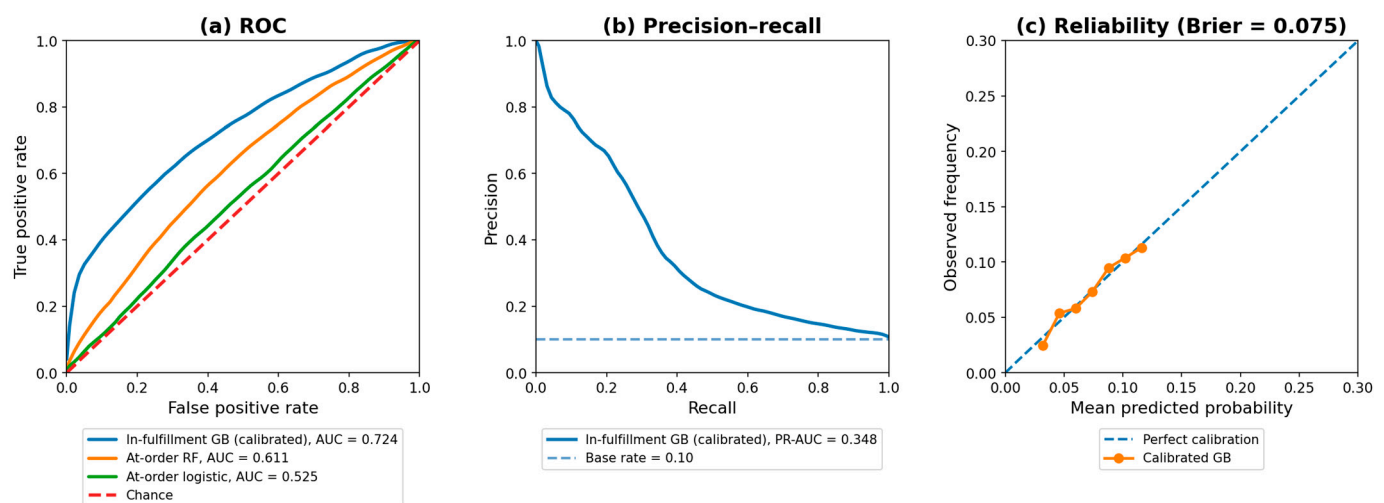


Figure 7. Risk-model curves on the temporal hold-out. (a) ROC: the in-fulfillment model (AUC 0.72) clearly exceeds at-order models (≈ 0.61); (b) precision–recall against a 0.10 base rate; (c) reliability after isotonic calibration (Brier 0.075).

A leakage-control ablation reduces, but does not eliminate, the concern that the in-fulfillment advantage is mechanically driven by delay-related information (Table 6). Removing the late-delivery flag lowers AUC only modestly, from 0.724 to 0.704; further removing the realized delivery days reduces it to 0.692, and removing all delay-related variables reduces it to 0.658—still well above the at-order ceiling of 0.609. The predictive signal is thus distributed across fulfillment dynamics rather than carried by one binary flag; this attenuates, without fully removing, the circularity concern, since delay-related information remains correlated with the outcome by construction.

Table 6. Leakage-control ablation on the in-fulfillment model (temporal hold-out). Each row removes the indicated fulfillment information; AUC and PR-AUC decline gradually rather than collapsing, and the at-order model bounds the effect from below. Bold indicates the full reference model against which each ablated specification is compared.

Specification	AUC	PR-AUC	Brier
Full in-fulfillment model	0.724	0.348	0.075
Without late-delivery flag	0.704	0.317	0.078
Without actual delivery days	0.692	0.301	0.080
Without delay-related variables	0.658	0.238	0.083
At-order only	0.609	0.178	0.084

These results answer the research question (RQ) posed under Objective 2: order-level dissatisfaction is only weakly predictable from information available at order time (AUC 0.61) and becomes materially predictable once fulfillment signals resolve (AUC 0.72). Managerially, this locates the actionable monitoring window inside the fulfillment process itself: recovery systems should score orders as shipping events unfold, rather than attempting to pre-identify dissatisfied customers at checkout.

3.3. Recoverability: Dissatisfaction Is Operational

SHAP attribution of the calibrated risk model (Figure 8) shows that the leading drivers of dissatisfaction are delivery delay relative to estimate, realized delivery time, and the late-delivery flag, with basket size the principal structural contributor. Partitioning attribution into operational and structural blocks, operational factors account for 76% of the total risk mass, indicating that a large share of predicted dissatisfaction risk is operationally attributable and therefore potentially recoverable. Dissatisfaction in this marketplace is thus plausibly addressable through service-recovery actions, in that it is associated with fulfillment performance the firm can act upon rather than with immutable product or customer characteristics. This supports Hypothesis H2 and identifies operations—not price or assortment—as the binding lever for service recovery.

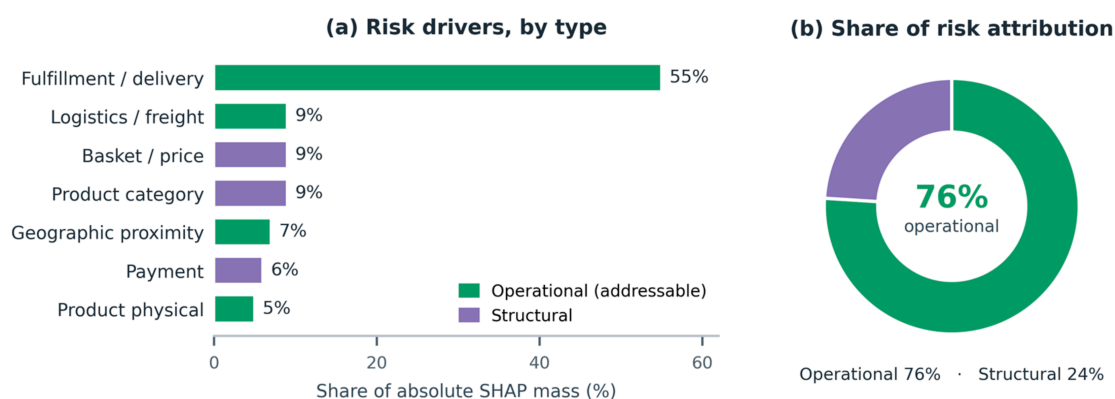


Figure 8. Recoverability via SHAP. (a) Risk drivers colored by type; delivery-related operational factors dominate. (b) Operational factors account for 76% of risk attribution, indicating a high share of operationally attributable, potentially recoverable risk.

This addresses Objective 3 and supports Hypothesis H2: the dominant share of predicted dissatisfaction risk loads on operational, firm-controllable factors rather than structural ones. For managers, the implication is that most of the risk the model surfaces is, in principle, addressable—through delivery monitoring, proactive communication, or expedited handling—rather than a fixed property of certain customers or products.

3.4. Prioritization and the Out-of-Sample Decision-Policy Evaluation

Table 7 reports the segment-level VRR profile on the held-out cohort. The Service-Failure At-Risk segment carries the highest predicted risk (0.229 versus about 0.09 elsewhere) and high recoverability (0.79), giving it the highest priority score; High-Value Advocates rank second on value; Budget Single-Buyers rank lowest, combining low value and low risk. The ordering is managerially coherent and operational: priority concentrates where valuable, recoverable risk resides. This table is the operational junction between the segmentation and the prioritization framework: it is where order-level VRR scores become segment-level budget and routing decisions, in the manner detailed in Section 4.1. At the individual order level, the same priority logic is visible in the VRR priority space (Figure 9), which plots customer value against predicted dissatisfaction risk, colored by recoverabil-

ity: high-risk customers are overwhelmingly recoverable, and the four quadrants map to distinct recovery actions.

Table 7. Segment-level VRR profile on the held-out cohort. V = mean order value (BRL); R = mean calibrated risk; Re = mean recoverability; “Fail” and “Late” are realized rates; Priority = value-percentile $\times$ R $\times$ Re.

Segment	V	R	Re	Fail	Late	Priority
Service-Failure At-Risk	148.6	0.229	0.79	0.878	0.308	0.091
High-Value Advocates	198.4	0.087	0.71	0.007	0.041	0.062
Multi-Item Repeat Buyers	194.1	0.121	0.57	0.180	0.037	0.054
Budget Single-Buyers	58.4	0.085	0.70	0.018	0.044	0.015

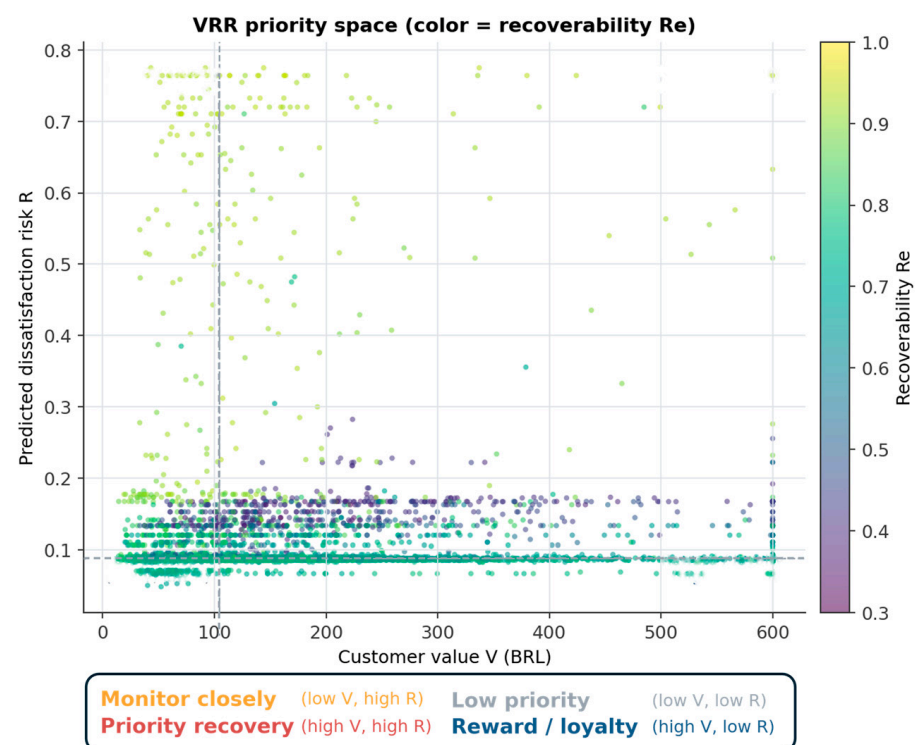


Figure 9. VRR priority space on the held-out cohort: customer value (x) against predicted dissatisfaction risk (y), colored by recoverability. High-risk customers are overwhelmingly recoverable (light), and quadrants map to distinct actions.

The temporally held-out policy evaluation (Figure 10, Table 8) is the framework’s central test. At a 10% intervention budget, value-only targeting—the common CRM default—captures only 30.4% of realized recoverable value, barely above random, whereas risk-aware policies capture roughly 96% (risk-only 96.1%, VRR 95.7%). In the vocabulary of campaign analytics, Figure 10 is a cumulative gain chart weighted by realized recoverable value—the share of total recoverable value captured as a function of the fraction of the cohort targeted—and the corresponding lift is the ratio of captured share to targeted fraction: at the 10% budget, risk-aware policies achieve a lift of approximately 9.6 against 3.0 for value-only targeting and 1.0 for random allocation by construction. This three-fold gap strongly supports Hypothesis H1: prioritizing by value alone badly misallocates recovery budgets. Risk-only marginally exceeds VRR here, and the explanation supports Hypothesis H3: because the deployable risk signal is itself operational, risk and recoverability coincide, so the recoverability axis adds little, and the value axis, by pulling some high-value non-failing customers forward, slightly dilutes recoverable-value capture. A

paired bootstrap (1000 resamples) confirms these readings: VRR and risk-only are statistically indistinguishable (paired $\Delta = -0.004$, 95% CI $[-0.017, 0.009]$), whereas both risk-aware policies dominate value-only by a wide, significant margin (VRR–value-only $\Delta = 0.653$, 95% CI $[0.621, 0.684]$; Tables 9 and 10). The three axes converge in this fulfillment-driven setting, making operational risk a near-sufficient targeting statistic—a contingency the framework renders explicit rather than a universal claim of composite dominance.

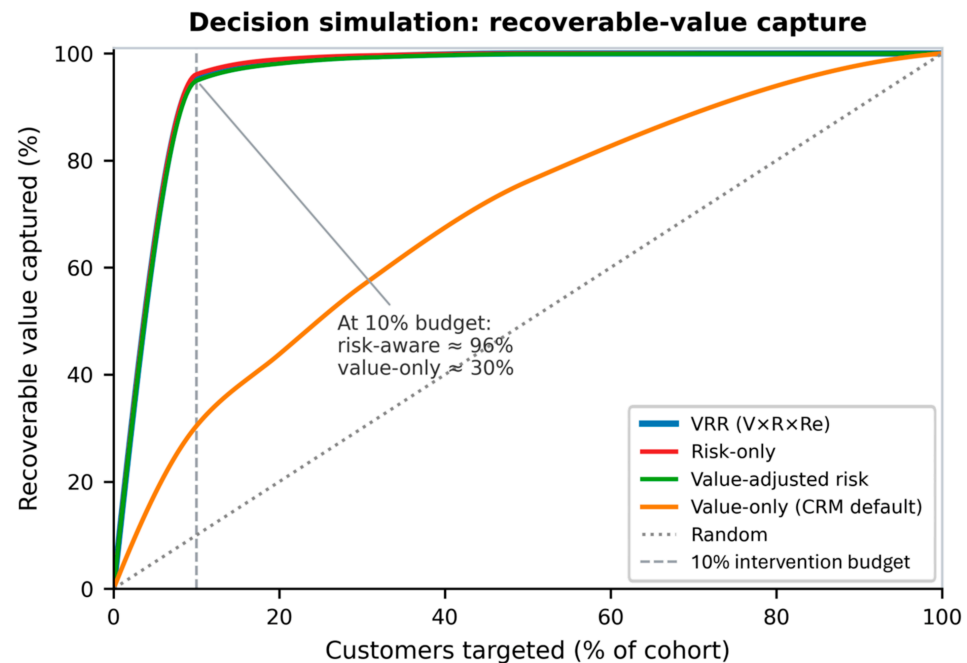


Figure 10. Recoverable-value capture by targeting policy. Risk-aware policies (VRR, risk-only) capture about 96% of recoverable value at a 10% budget; value-only targeting captures under one-third, near random. The dashed vertical line marks the 10% intervention budget.

Table 8. Out-of-sample decision-policy evaluation: share of realized recoverable value captured by each targeting policy at increasing budgets (held-out cohort, 8000 orders). Policies rank on model outputs; evaluation uses realized outcomes. The last column reports the lift at the 10% budget (captured share divided by the budget fraction); random allocation has a lift of 1 by construction.

Policy	10%	20%	30%	50%	Lift@10%
Risk-only	0.961	0.989	0.996	1.000	9.61
VRR ($V \times R \times Re$)	0.957	0.985	0.994	1.000	9.57
Value-adjusted risk	0.949	0.981	0.992	0.999	9.49
Value-only (CRM default)	0.304	0.438	0.566	0.761	3.04
Random	0.100	0.200	0.300	0.500	1.00

Table 9. Bootstrap confidence intervals for recoverable-value capture at the 10% intervention budget (1000 resamples, held-out cohort). Bold indicates the best-performing policy (highest recoverable-value capture at the 10% budget).

Policy	Capture at 10%	95% CI
Risk-only	0.961	[0.948; 0.972]
VRR	0.957	[0.943; 0.970]
Value-adjusted risk	0.949	[0.934; 0.964]
Value-only	0.304	[0.278; 0.331]
Random	0.100	[0.087; 0.115]

Table 10. Paired bootstrap differences between targeting policies at the 10% intervention budget (1000 resamples). Intervals excluding zero indicate a statistically meaningful difference. Bold indicates differences whose 95% confidence interval excludes zero (statistically meaningful).

Difference	Estimate	95% CI
VRR—Risk-only	−0.004	[−0.017; 0.009]
VRR—Value-adjusted risk	0.008	[−0.006; 0.021]
VRR—Value-only	0.653	[0.621; 0.684]
Risk-only—Value-only	0.657	[0.626; 0.689]

Table 9 reports bootstrap confidence intervals for recoverable-value capture at the 10% budget, and Table 10 reports the paired differences between policies. The intervals show that the risk-aware policies are tightly estimated and mutually indistinguishable, while their advantage over value-only targeting is large and unambiguous.

Taken together, these results address Objective 4 and its two hypotheses. The three-fold capture gap between value-based and risk-aware targeting confirms H1: under a realistic 10% budget, prioritizing by customer value alone forfeits roughly two-thirds of the value that could have been saved. The near-equivalence of the composite VRR policy and risk-only targeting confirms the contingency stated in H3: in a setting where dissatisfaction is operationally driven, risk is a near-sufficient targeting statistic, and the value and recoverability axes serve as tie-breakers and governance safeguards rather than as sources of additional capture. The managerial reading is direct: a firm in a fulfillment-driven setting can obtain most of the attainable benefit by ranking on calibrated risk, while retaining the value axis for budget sequencing and the recoverability axis to route each case toward the type of action that can plausibly address it. One scope statement bears emphasis. What this evaluation demonstrates is targeting efficiency: how much of the value attached to realized, operationally attributable failures each policy would have directed recovery effort toward. It does not estimate the causal impact of any service-recovery intervention—no intervention is observed in the data—so the capture figures quantify how well a budget is aimed, not how much of the aimed-at value a given recovery action would in fact save. The two questions are complementary: targeting efficiency bounds what any intervention can reach, while the intervention’s causal effect determines how much of what is reached is converted.

4. Discussion

4.1. Theoretical and Managerial Implications

The results carry a clear and somewhat uncomfortable message for value-centric customer management. The dominant practice of prioritizing service and retention effort by customer value is, in this marketplace, close to the least effective use of a recovery budget: it captures barely more recoverable value than random allocation at realistic budget levels. The reason is structural. Most value resides in satisfied customers who will not churn, so spending recovery resources on them is largely wasted, exactly the inefficiency that customer-equity scholars warn against when prioritization is implemented crudely [1]. Risk, by contrast, locates the failures, and because those failures are operational, addressing them is feasible.

The central theoretical contribution of the VRR framework is that it converts what the literature treats as a caveat into a measurable, decision-ready quantity. Prior work established that risk-based targeting can fail when risk does not respond to intervention [5]; the framework operationalizes this responsiveness concern as an order-level recoverability axis, estimated from the attribution structure of the risk model itself, and—critically—makes visible the boundary condition governing when the distinction matters. In our

fulfillment-driven setting, risk and recoverability largely coincide: the same operational failures that generate risk are the ones an operations-focused recovery can address, which is why the composite policy performs on par with risk-only targeting rather than dominating it. In categories where dissatisfaction is structural—product–need mismatch, taste, fit—the axes should diverge, and the recoverability axis is precisely what would prevent a recovery budget from being spent on unrecoverable cases. The framework thus does not merely restate the responsiveness critique; it supplies the instrument for detecting, in any given setting, whether that critique binds.

Beyond this, the framework advances customer relationship management along three lines that existing approaches do not jointly deliver. First, it shifts the unit of prioritization from the customer to the order–customer case, aligning the decision grain with the grain at which service failures actually occur and at which recovery actions are taken. Second, it integrates the three prioritization logics that CRM research has developed separately—value, risk, and cause-of-failure—into a single signal whose components remain individually inspectable, so that the trade-offs between them are explicit rather than buried in a composite score. Third, because the signal is built from calibrated probabilities and additive attributions, every prioritization decision can be traced, contested, and audited, which converts interpretability from a modeling preference into an operational governance capability. Together, these elements move CRM prioritization from segment-level heuristics toward transparent, case-level decision support.

For managers, three implications follow. First, recovery budgets should be allocated by operationally attributable risk—with recoverability treated as a prioritization proxy rather than a demonstrated treatment response—not by value; the value axis is best used to sequence interventions within the high-risk pool and to frame cost–benefit thresholds. Second, because the predictive signal resolves as fulfillment unfolds rather than at order time, the practical locus of intervention is the post-shipment, pre-review window: a late or stalled delivery is an actionable signal, and although the strongest features are only fully observed at or near the delivery event, recovery outreach before the customer submits a review remains feasible and well-targeted, while earlier in-transit signals such as predicted delay support pre-emptive action where available. Third, the dominance of operational drivers means that the highest-leverage investment is in fulfillment reliability itself, reducing the failures rather than only recovering from them.

Concretely, the prioritization signal maps onto a graded menu of service-recovery actions matched to the available window and the predicted risk: alerting the seller or carrier and expediting the shipment while the order is in transit; issuing a proactive apology and status update once a delay or failed delivery is detected; offering a compensation voucher or goodwill credit for high-risk, high-value orders; and prioritizing these orders in customer-support queues. The framework indicates where such actions should be concentrated; it does not, on its own, establish which action is most cost-effective—an estimation task we leave to future experimental work.

Implementing the matrix (Table 11) requires three operating choices. First, thresholds: the high/low cut-offs on each axis are set by the budget, not by fixed constants—rank cases by the priority signal and treat the top fraction the budget allows, using the value axis to sequence within it, so the matrix inherits the budget discipline of Section 3.4 rather than introducing new parameters. Second, cadence: because the predictive signal resolves during fulfillment, scoring must run continuously on shipping events (carrier scans, delay-versus-promise updates), not as a periodic batch over the customer base. Third, routing: the recoverability axis determines the type of action (operational remediation versus structural accommodation), while risk determines whether to act and value determines in which order—a division of labor that keeps each axis doing the one job for which Section 3.4 shows

it carries decision-relevant information. Firms can operate the matrix as a rule layer on top of the scores, which preserves the auditability rationale of Section 4.2: every intervention decision remains traceable to three inspectable quantities and one declared threshold.

Table 11. VRR decision matrix: recommended recovery posture by profile. V/R/Re are read as high (H) or low (L) relative to operating thresholds set by the budget; actions assume the post-shipment, pre-review window identified in Section 3.2.

VRR Profile	Diagnostic Reading	Recommended Recovery Actions	Budget Posture
H-V/H-R/H-Re	Valuable customer, likely failure, operationally addressable	Expedite or re-route shipment; proactive apology with status update; compensation voucher or goodwill credit; front of support queue	Intervene first; highest cost-per-case justified
L-V/H-R/H-Re	Low-stakes order, likely operational failure	Automated proactive status update and apology; carrier/seller alert; low-cost remediation	Intervene at scale via automation; minimal marginal cost
H-V/H-R/L-Re	Valuable customer, risk is structural (fit, taste, expectations)	Do not spend logistics effort; ease returns/exchange, provide product guidance, goodwill gesture	Route to non-operational remedies; protect relationship, not the order
L-V/H-R/L-Re	Low-stakes structural risk	No individual intervention; aggregate signal fed back to assortment, product content, and expectation-setting	Learn, do not spend
H-V/L-R	Valuable and satisfied	No recovery action; standard loyalty track	Preserve; recovery budget not applicable
L-V/L-R	Low value, low risk	No action; routine monitoring	None

These implications are, if anything, more salient in the contemporary e-commerce landscape than in the study period. Post-pandemic, fulfillment expectations have tightened, same- and next-day delivery has become a competitive default, and marketplaces increasingly compete on logistics reliability rather than assortment alone—conditions that enlarge the share of customer experience governed by exactly the operational signals the framework exploits. At the same time, richer real-time telemetry (carrier scan events, live tracking, delivery-promise engines) means the in-fulfillment information set available to a modern platform is strictly larger than the one used here, so the reported predictive performance is plausibly a lower bound. The framework is designed to travel: because its inputs are generic (order values, promised and realized fulfillment timestamps, and an outcome signal such as reviews, support contacts, or returns), it can be re-instantiated on more recent proprietary datasets, in other geographical markets with different logistics infrastructures, and across alternative digital business models—quick-commerce, subscription boxes, or food delivery—where the operational/structural decomposition of dissatisfaction may differ and where the boundary condition identified above can be tested directly.

4.2. Interpretability and Governance

Because every element of the priority signal decomposes into interpretable contributions, the framework is auditable in a way that opaque scoring is not. Operational, compliance, and audit functions can trace why a customer was prioritized, which supports internal documentation and contestability as automated profiling comes under closer scrutiny [19,26,28]. We frame this as a mechanism for organizational adoption and defen-

sibility rather than a guarantee of regulatory compliance, which depends on jurisdiction and use.

4.3. Limitations

Several limitations bound the findings. The data are cross-sectional and overwhelmingly single-purchase, so recency is a tenure proxy, the value axis is an observed monetary proxy rather than estimated lifetime value, and no forward retention outcome is observed; the risk models therefore predict an order-level dissatisfaction event rather than long-run defection. The recovery effect underlying the out-of-sample decision-policy evaluation is not estimated but is sidestepped by evaluating against realized recoverable value, so the simulation establishes targeting efficiency, not the causal return of any specific intervention; the reported capture shares are therefore upper bounds on what recovery actions could address, to be discounted by intervention effectiveness. Establishing that causal return requires designs the present data cannot support—randomized recovery experiments or, failing that, uplift-modeling approaches on data where interventions and their outcomes are logged—and constitutes the natural experimental sequel to this study. The recoverability axis inherits three distinct limitations from its attribution-based construction. First, it is a model-based proxy: it measures attributed risk, not demonstrated responsiveness, and therefore does not establish that any customer would in fact respond to a given recovery action, only that their risk loads on firm-controllable causes. Second, it depends on the operational/structural partition of predictors, which involves judgment at the boundary; we adopt a transparent, pre-declared classification (Table 2) so that the partition is auditable and can be contested or revised. Third, SHAP attributions are sensitive to dependence among correlated predictors, which can spread or concentrate attributed importance across collinear features; our leakage-control ablation (Table 6) partially probes this by showing that predictive signal is distributed across fulfillment features rather than carried by a single flag, but it does not fully eliminate the concern. We therefore present recoverability as a defensible, interpretable prioritization proxy whose validation against actual intervention response is left to future experimental work. Crucially, recoverability captures operationally attributable risk, not demonstrated responsiveness: it does not establish that customers would in fact respond to a recovery action. Estimating treatment effects or uplift, ideally through field experiments, is the natural next step. Finally, the evidence comes from one marketplace and period and is entirely correlational; external validity to other categories—especially those with structural dissatisfaction—should be tested rather than assumed, a test the framework is explicitly designed to invite. The data are also historical: they predate the accelerated digitalization, logistics reconfiguration, and shifts in consumer expectations that followed the COVID-19 pandemic. While the operational mechanism at the core of our findings—dissatisfaction driven by delivery performance—is likely to have persisted or intensified in that period, the magnitudes reported here (base rates, attribution shares, and capture curves) are period-specific and should be re-estimated on contemporary data before being generalized. The framework is explicitly designed to invite such re-estimation rather than to assert time-invariant point estimates.

More broadly, the single-dataset basis of the study calls for a differentiated statement of what should and should not generalize. Three layers must be distinguished. The framework itself—the three-axis decomposition, the calibrated-risk construction, and the attribution-based recoverability proxy—is method-level and portable to any setting offering order-level operational records linked to an outcome signal; nothing in its logic is specific to Brazil, to marketplaces, or to reviews as the outcome. The qualitative mechanism—dissatisfaction concentrating on fulfillment performance where logistics are strained—should travel to settings with comparable operational friction, such as emerging-market

marketplaces, cross-border retail, or quick-commerce, but is expected to weaken where fulfillment is commoditized and reliable, as in mature markets dominated by integrated logistics networks. The quantitative results—base rates, the 76% operational attribution share, and the near-sufficiency of risk-only targeting—are setting-specific and should be expected to shift: in industries where dissatisfaction is predominantly structural (fashion fit, taste-driven categories, digital services), the operational share should fall, risk and recoverability should decouple, and the recoverability axis should become decision-relevant rather than redundant. The framework is thus best understood as an instrument whose axes are general but whose readings are local: each new deployment re-estimates the decomposition and thereby tests, rather than assumes, the boundary condition identified in Section 4.1.

5. Conclusions

This article proposed and tested a Value–Risk–Recoverability framework for service-recovery prioritization at the order–customer interface, integrating customer-equity, churn-risk, and service-recovery logics into a single interpretable decision signal. Applied to 91,954 customers and 93,663 reviewed orders from a real marketplace, the framework shows that dissatisfaction is predominantly operationally attributable and therefore plausibly recoverable through service-recovery actions, that it becomes predictable only once fulfillment is underway, and that value-based targeting—still the default in practice—captures a small fraction of the recoverable value that risk-aware targeting reaches. It also shows, candidly, that the value and recoverability axes add little where risk and recoverability coincide, and explains why. The framework’s contribution is to make prioritization transparent, theoretically grounded, and falsifiable: it specifies the conditions under which each axis carries decision-relevant information and invites tests in settings where those conditions differ. Future work should estimate recovery effects experimentally, extend the value axis with formal lifetime-value models suited to sparse-repeat data [7], and replicate the framework in categories where dissatisfaction is structural rather than operational, where the recoverability axis should come into its own.

Author Contributions: Conceptualization, Y.M. and M.A.; methodology, Y.M. and M.A.; software, Y.M. and M.A.; formal analysis, Y.M. and M.A.; writing—original draft, Y.M. and M.A.; writing—review and editing, Y.M. and M.A.; visualization, Y.M. and M.A.; supervision, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable; the study uses publicly available, fully anonymized data.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is the publicly available Brazilian E-Commerce Public Dataset by Olist. All analysis code—preprocessing, model training, isotonic calibration, SHAP attribution, the leakage-control ablation, the paired bootstrap, and the decision-simulation and capture-curve scripts—together with the fixed random seed and the held-out predictions and outputs, will be made available in an anonymized public repository for peer review and released publicly upon acceptance/publication, ensuring full reproducibility.

Acknowledgments: The authors used large-language-model tools for language editing, code assistance, and figure formatting. All analyses, results, interpretations, and manuscript claims were independently verified by the authors, who take full responsibility for the content.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the ROC Curve
ARI	Adjusted Rand Index
BRL	Brazilian Real
CLV	Customer Lifetime Value
CRM	Customer Relationship Management
PR-AUC	Area Under the Precision–Recall Curve
RFM	Recency, Frequency, Monetary
SHAP	SHapley Additive exPlanations
VRR	Value–Risk–Recoverability
XAI	Explainable Artificial Intelligence

References

1. Homburg, C.; Droll, M.; Totzek, D. Customer Prioritization: Does It Pay off, and How Should It Be Implemented? *J. Mark.* **2008**, *72*, 110–130. [\[CrossRef\]](#)
2. Rust, R.T.; Lemon, K.N.; Zeithaml, V.A. Return on Marketing: Using Customer Equity to Focus Marketing Strategy. *J. Mark.* **2004**, *68*, 109–127. [\[CrossRef\]](#)
3. Neslin, S.A.; Gupta, S.; Kamakura, W.; Lu, J.; Mason, C.H. Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. *J. Mark. Res.* **2006**, *43*, 204–211. [\[CrossRef\]](#)
4. Verhoef, P.C. Understanding the Effect of Customer Relationship Management Efforts on Customer Retention and Customer Share Development. *J. Mark.* **2003**, *67*, 30–45. [\[CrossRef\]](#)
5. Ascarza, E. Retention Futility: Targeting High-Risk Customers Might Be Ineffective. *J. Mark. Res.* **2018**, *55*, 80–98. [\[CrossRef\]](#)
6. Smith, A.K.; Bolton, R.N.; Wagner, J. A Model of Customer Satisfaction with Service Encounters Involving Failure and Recovery. *J. Mark. Res.* **1999**, *36*, 356–372. [\[CrossRef\]](#)
7. Fader, P.S.; Hardie, B.G.S. Probability Models for Customer-Base Analysis. *J. Interact. Mark.* **2009**, *23*, 61–69. [\[CrossRef\]](#)
8. Gubela, R.M.; Lessmann, S. Uplift Modeling with Value-Driven Evaluation Metrics. *Decis. Support Syst.* **2021**, *150*, 113648. [\[CrossRef\]](#)
9. Haupt, J.; Lessmann, S. Targeting Customers Under Response-Dependent Costs: Response Transformation and Profit Decomposition for Revenue Uplift Modeling. *Eur. J. Oper. Res.* **2022**, *297*, 369–379. [\[CrossRef\]](#)
10. Parasuraman, A.; Zeithaml, V.A.; Malhotra, A. E-S-QUAL: A Multiple-Item Scale for Assessing Electronic Service Quality. *J. Serv. Res.* **2005**, *7*, 213–233. [\[CrossRef\]](#)
11. Cronin, J.J., Jr.; Brady, M.K.; Hult, G.T.M. Assessing the Effects of Quality, Value, and Customer Satisfaction on Consumer Behavioral Intentions in Service Environments. *J. Retail.* **2000**, *76*, 193–218. [\[CrossRef\]](#)
12. Hughes, A.M. *Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-Based Marketing Program*; Probus Pub. Co.: Chicago, IL, USA, 1994.
13. MacQueen, J. Multivariate Observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Oakland, CA, USA, 1967; Volume 1, pp. 281–297.
14. McLachlan, G.J.; Peel, D. *Finite Mixture Models*; John Wiley & Sons: Hoboken, NJ, USA, 2000.
15. Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the KDD*; Morgan Kaufmann Publishers: Burlington, MA, USA, 1996; Volume 96, pp. 226–231.
16. Wong, C.-G.; Tong, G.-K.; Haw, S.-C. Exploring Customer Segmentation in E-Commerce Using RFM Analysis with Clustering Techniques. *J. Telecommun. Digit. Econ.* **2024**, *12*, 97–125. [\[CrossRef\]](#)
17. Louppe, G.; Wehenkel, L.; Suter, A.; Geurts, P. Understanding Variable Importances in Forests of Randomized Trees. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 431–439.
18. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4768–4777.
19. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [\[CrossRef\]](#)
20. Joung, J.; Kim, H. Interpretable Machine Learning-Based Approach for Customer Segmentation for New Product Development from Online Product Reviews. *Int. J. Inf. Manag.* **2023**, *70*, 102641. [\[CrossRef\]](#)

21. Ufeli, C.P.; Sattar, M.U.; Hasan, R.; Mahmood, S. Enhancing Customer Segmentation through Factor Analysis of Mixed Data (Famd)-Based Approach Using k-Means and Hierarchical Clustering Algorithms. *Information* **2025**, *16*, 441. [\[CrossRef\]](#)
22. El Attar, A.; El-Hajj, M. Explainable AI-Driven Customer Churn Prediction: A Multi-Model Ensemble Approach with SHAP-Based Feature Analysis. *Front. Artif. Intell.* **2026**, *9*, 1748799. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Imani, M.; Joudaki, M.; Beikmohammadi, A.; Arabnia, H.R. Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 105. [\[CrossRef\]](#)
24. Canhoto, A.I.; Clear, F. Artificial Intelligence and Machine Learning as Business Tools: A Framework for Diagnosing Value Destruction Potential. *Bus. Horiz.* **2020**, *63*, 183–193. [\[CrossRef\]](#)
25. Meng, F.; Jia, Y. Data-Driven Prioritization of User Requirements in Health E-Commerce: An Explainable Machine Learning Study. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *21*, 104. [\[CrossRef\]](#)
26. European Union. *Règlement (UE) 2024/1689 du Parlement Européen et du Conseil du 13 Juin 2024 Etablissant des Règles Harmonisées Concernant L'intelligence Artificielle et Modifiant les Règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 et les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828 (Règlement sur L'intelligence Artificielle) (Texte Présentant de L'intérêt Pour l'EEE)*; Publications Office of the European Union: Luxembourg, 2024.
27. Mantelero, A. AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment. *Comput. Law. Secur. Rev.* **2018**, *34*, 754–772. [\[CrossRef\]](#)
28. Almada, M. Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, Montreal, QC, Canada, 17–21 June 2019*; ACM: New York, NY, USA, 2019; pp. 2–11.
29. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
30. Brazilian E-Commerce Public Dataset by Olist. Available online: <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce> (accessed on 7 June 2026).
31. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.