



Temporal Validation of Case-Mix-Aware Credit-Card Complaint Response Analytics

Irina V. Pustokhina^{1,*} Denis A. Pustokhin²

¹ Department of Entrepreneurship and Logistics, Plekhanov Russian University of Economics, Moscow 117997, Russia

² Department of Logistics, State University of Management, Moscow 109542, Russia

Emails: PustokhinaIV@rea.ru · da_pustokhin@guu.ru

Received: July 20, 2026 Revised: September 05, 2026 Accepted: October 03, 2026 ★ Corresponding author

ABSTRACT

Complaint dashboards often pool unlike cases and report an overall resolution rate. For a financial-services operations team, that average can conceal which issues account for redress workload and whether a triage signal remains useful on later cases. We analyze 75,988 publicly reported U.S. credit-card complaints received in 2024. The outcome is the company's recorded "closed with monetary relief" category, not a verified payment amount or customer satisfaction. Intake issue, sub-issue, submission channel, and company identify a transparent logistic screening model trained on January–September records and evaluated on October–December records. Report-related issues comprise 31.9% of complaints but have a 0.63% monetary-relief rate, compared with 19.16% for other issues. An exact shift-share decomposition attributes 1.81 of the 2.64 percentage points of decline from the first to fourth quarter to changing issue mix. Issue alone achieves an out-of-time area under the receiver operating characteristic curve (AUC) of 0.819; adding sub-issue and channel raises it to 0.831. A cross-company specification reaches 0.873 (95% bootstrap interval 0.867–0.878) and captures 39.4% of recorded monetary-relief responses in its highest-scoring tenth of holdout cases. The company term captures institutional and routing differences, so that result does not measure issuer quality or the gain attainable within a single firm. Case-mix-aware reporting and temporal validation support staffing review queues without automating redress decisions.

Keywords: Business intelligence ▪ Complaint analytics ▪ Financial services ▪ Case mix ▪ Temporal validation

1. INTRODUCTION

A complaints manager can see a rising case count and still miss the more important change: the share of fee disputes may have fallen while cases involving credit-report information increased. Those two issues can lead to different recorded company responses. Staffing, specialist routing, and an explanation of quarter-to-quarter performance therefore require more than a portfolio-wide response rate. The relevant business intelligence task is to distinguish case composition from operational outcomes, then ask whether an analytic signal survives on complaints received later.

We study the public Consumer Financial Protection Bureau (CFPB) credit-card complaint portfolio for calendar 2024. The database exposes intake descriptors and company response categories, which makes it possible to test a specific operational question: how much of subsequent monetary-relief response can be anticipated from information available before closure? The study contributes a case-mix audit, a sequence of feature-ablation models, and a strictly later-quarter holdout. Its unit is a published complaint, not a credit-card customer or a dollar of compensation. We avoid translating a response category into welfare, fairness, or an optimal remedy.

2. RELATED WORK

Complaint data can inform operations as well as public oversight. Vairetti et al. [1] combine text analytics and multi-criteria decision methods to prioritize incoming complaints, whereas Dou et al. [2] show that banks learn from disclosed mortgage complaints about peers. Our study takes a narrower measurement route: structured issue fields support an inspectable model and a temporal test rather than a claim of optimal priority. Grosz and Raval [3] find that making a fraud-reporting website easier to use increases submitted complaints and detail, a reminder that published volume reflects the reporting channel as well as underlying events.

The response itself needs careful interpretation. Ali et al. [4] connect perceived justice in complaint handling with forgiveness and brand credibility; Roschk et al. [5] show that coalitions in online complaints can impair the effectiveness of a firm’s response. Neither customer sentiment nor justice is observable in a response label. Chen [6] finds that greater interpretability can temper negative reactions to algorithmic service failures, cautioning against opaque automated resolution. Chung et al. [7] study provider identity disclosure on customer-service platforms through two field experiments. These works motivate a decision-support design in which models flag cases for human review rather than assign remedies.

Finally, complaint portfolios are selected populations. Hayes et al. [8] document racial disparities in financial complaint volume and investigate their institutional context. The CFPB’s 2025 annual report notes that complaint counts are not counts of unique consumers and that some submissions are referred to other regulators rather than included in its totals [9]. We address the narrower within-database question of response-mix measurement. We do not infer an issuer’s treatment quality from a raw response proportion.

3. DATA AND STUDY DESIGN

The source is the CFPB Consumer Complaint Database. We retrieved the public “Credit card” product records received from January 1 through December 31, 2024, retaining original date, issue, sub-issue, company, channel, response, and complaint-ID fields. We removed one case whose date sent to the company fell after January 2026. Complaint IDs were unique; no narrative text, demographic inference, or geographic profiling enters the analysis. The resulting 75,988 cases include 10,066 coded “closed with monetary relief.” The CFPB’s 2025 annual report provides institutional context for complaint intake and response categories; our statistics are recalculated directly from the released records [9].

The first nine receipt months (56,021 cases) form the model-development sample; the last three (19,967) are held out intact. The split prevents a random train/test shuffle from overstating reliability under changing issue mix. We retain published issue and sub-issue categories and code absent text as “Unspecified.” Table 1 shows the response mix. The extract was retrieved on September 29, 2026, from a database that may revise older records; the packaged CSV fixes the observations analyzed here. A 2024 receipt date does not establish that every public field was visible during 2024 or by January 2026.

Table 1. Analytic cohort and temporal split

Period	Complaints	Relief	Relief (%)	Timely (%)
Jan–Sep, model development	56,021	7,512	13.41	99.81
Oct–Dec, temporal holdout	19,967	2,554	12.79	99.85
Full 2024 cohort	75,988	10,066	13.25	99.82

4. PROPOSED ANALYTICAL FRAMEWORK

4.1 Response Probability and Validation

Let $Y_i = 1$ if the published company response to complaint i is “closed with monetary relief,” and zero for every other recorded category. The baseline is the development-period prevalence. Four nested specifications use issue; issue plus sub-issue; those fields plus submission channel; and all three plus company. Each categorical level is one-hot encoded, with unseen holdout levels ignored. For feature vector x_i , the fitted probability is

$$\hat{p}_i = \{1 + \exp[-(\beta_0 + x_i^\top \beta)]\}^{-1},$$

$$\hat{\beta} = \arg \min_{\beta} \left[-\sum_{i \in T} \ell(Y_i, \hat{p}_i) + \frac{\lambda}{2} \|\beta\|_2^2 \right], \quad (1)$$

where T is the January–September cohort, ℓ is the Bernoulli log likelihood, and the regularization parameter is fixed by $C = 1$ in the supplied implementation. We do not tune on the holdout. Predictors are the fields in the final export; an operating firm must separately verify that each field is available at intake in its own workflow.

We report AUC, average precision (AP), and Brier score on the untouched October–December cohort. If S_q is the highest-scoring $\lceil qn \rceil$ cases, the observed response capture at review capacity q is

$$C(q) = \frac{\sum_{i \in S_q} Y_i}{\sum_{i \in H} Y_i}, \quad L(q) = \frac{\sum_{i \in S_q} Y_i / |S_q|}{\sum_{i \in H} Y_i / |H|}, \quad (2)$$

with H denoting the temporal holdout. Ties follow a fixed record order. The bootstrap interval resamples holdout cases 400 times with seed 1704, holding fitted models fixed. It measures finite-holdout variation, not a guarantee of performance in a new institution.

4.2 Issue Mix as a Portfolio Diagnostic

For portfolio c and issue k , let n_{ck} be its observed cases. We estimate an issue-specific rate outside the portfolio with a 20-case prior toward the outside-portfolio mean $\bar{y}_{-c}$. The expected monetary-relief count and percentage-point deviation are

$$E_c = \sum_k n_{ck} \frac{s_{-c,k} + 20\bar{y}_{-c}}{n_{-c,k} + 20}, \quad G_c = 100 \frac{O_c - E_c}{n_c}, \quad (3)$$

where $s_{-c,k}$ and $n_{-c,k}$ are relief counts and cases for issue k elsewhere, and O_c is the portfolio’s observed relief count. The leave-one-portfolio-out calculation avoids using a portfolio’s outcomes to set its own benchmark. It adjusts only for the observed issue field; differing institutions, selection, case severity, and response policy remain uncontrolled. The eight largest company portfolios are labeled P1–P8 for illustration, never as a league table.

Table 2. Most frequent credit-card complaint issues and recorded monetary-relief responses

Intake issue	Cases / share (%)	Relief / rate (%)
Purchase on statement	14,594 19.2%	4,611 31.60%
Incorrect report information	13,261 17.5%	78 0.59%
Investigation of prior problem	9,911 13.0%	59 0.60%
Card acquisition	9,683 12.7%	155 1.60%
Other terms or features	6,769 8.9%	1,043 15.41%
Fees or interest	6,343 8.3%	2,272 35.82%
Account closure	4,273 5.6%	351 8.21%
Payment problem	4,147 5.5%	907 21.87%
Advertising or promotions	2,766 3.6%	392 14.17%
Card-use problem	2,315 3.0%	130 5.62%

Issue labels are shortened for display; the analysis uses the original CFPB categories.

To distinguish quarter-to-quarter portfolio composition from changes within issue categories, we also use an exact, symmetric shift-share identity. Let w_{tk} denote issue k 's fraction of complaints in quarter t , and r_{tk} its observed monetary-relief proportion. Every issue occurs in each quarter. Relative to the first quarter,

$$P_t - P_1 = \sum_k (w_{tk} - w_{1k}) \frac{r_{tk} + r_{1k}}{2} + \sum_k (r_{tk} - r_{1k}) \frac{w_{tk} + w_{1k}}{2}, \quad (4)$$

where $P_t = \sum_k w_{tk} r_{tk}$. The first sum is the composition component; the second is the within-issue component. Their sum reproduces the observed percentage-point change without a regression or an assumed causal mechanism. The symmetric averaging removes the arbitrary choice of which period's weights to hold fixed.

5. RESULTS

5.1 Concentration of Recorded Monetary Relief

The central composition result is a sharp split inside one advertised product category. Six report-related issue categories account for 24,266 cases (31.9% of the sample) but only 154 monetary-relief responses (0.63%). The other 51,722 cases contain 9,912 monetary-relief responses (19.16%). This difference is descriptive: a credit-report correction can be meaningful even when it is coded as non-monetary relief. Among the ten most frequent issues in Table 2, fees or interest (35.82%) and purchases on statements (31.60%) are at the high end, while incorrect report information (0.59%) and investigation of an existing problem (0.60%) are at the low end. Their volumes make a pooled rate hard to interpret without the issue composition.

Figure 1 shows the monthly complaint load and response mix. The held-out quarter comprises 26.3% of the cohort. It has a 12.79% monetary-relief rate, compared with 13.41% during model development (Table 1); monthly swings are visible, so

Table 3. Quarterly shift-share decomposition of monetary-relief rates

Versus Q1	Cases	Rate (%)	Change (pp)	Mix (pp)	Within (pp)
Q2	18,971	12.87	-2.56	-1.79	-0.77
Q3	20,732	12.31	-3.12	-2.69	-0.42
Q4	19,967	12.79	-2.64	-1.81	-0.83

The comparison quarter is January–March 2024 (15.43%); “mix” and “within” sum to the observed change by Equation 4.

Table 4. Out-of-time performance on 19,967 October–December complaints

Specification	AUC / AP / Brier	Top 10% capture / lift
Constant baseline	0.500 0.128 0.112	10.0% 1.00
Issue only	0.819 0.306 0.094	25.8% 2.58
Issue + sub-issue	0.831 0.328 0.093	27.1% 2.71
Issue + sub-issue + channel	0.873 0.451 0.085	39.4% 3.93
All intake fields	0.873 0.451 0.085	39.4% 3.93

AP = average precision; AUC = area under the receiver operating characteristic curve. A lower Brier score is better. Capture and lift use a review capacity of 1,997 cases. The baseline capture of 10% is its expected value under random ordering.

the later quarter is a meaningful check rather than a random allocation of identical cases.

The first quarter's relief proportion is 15.43%, compared with 12.79% in the fourth. Of the 2.64 percentage-point decline, the exact decomposition assigns 1.81 points to changing issue shares and 0.83 points to changing rates within issues (Table 3). The mix term also accounts for 2.69 of the 3.12 points separating the third quarter from the first. These are accounting components, not estimates of changes caused by staffing or remedy policies. Nevertheless, they identify the scale of the interpretation error in a dashboard that attributes an aggregate decline entirely to worse case handling.

5.2 Out-of-Time Prediction and Capacity

The constant benchmark's AUC is 0.500 and AP equals the 12.79% holdout event rate. Issue alone yields AUC 0.819 and AP 0.306. Sub-issue raises AUC to 0.831; channel changes AUC little. Adding the company field increases AUC to 0.873 (95% bootstrap interval 0.867–0.878), with a 0.049–0.059 interval for its AUC gain over issue alone. The Brier score falls from 0.112 for a constant probability to 0.085 for the full model (Table 4). This additional discrimination across published company portfolios should not be represented as a within-issuer gain: issuer identity is constant for cases managed by one issuer.

At the same fixed review capacity, the issue/sub-issue/channel model finds 698 of the 2,554 holdout responses coded monetary relief (27.3%, lift 2.73); the cross-company model finds 1,005 (39.4%, lift 3.93). Figure 2 displays the capacity curve and grouped calibration. The full model's average predicted holdout rate, 13.38%, exceeds the observed 12.79%, so a queue-size rule is more defensible than an untested proba-

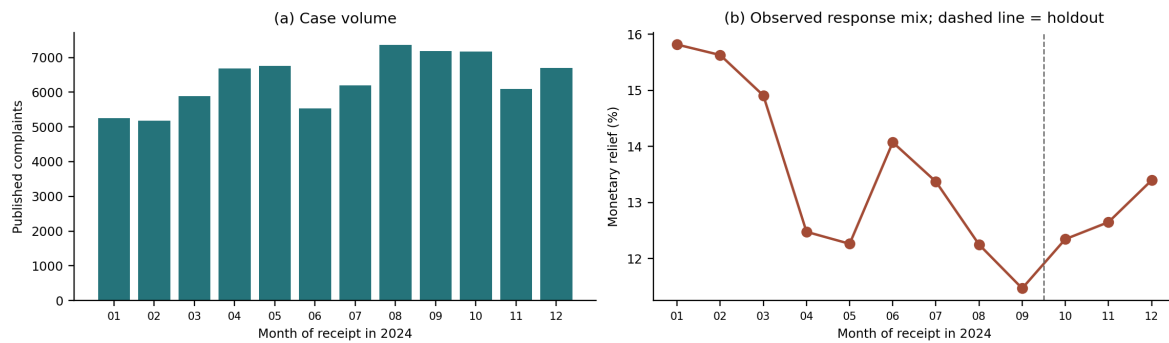


Figure 1. Two complementary views of the 2024 credit-card complaint portfolio. The dashed boundary separates the nine development months from the later three-month holdout. The line represents the share coded “closed with monetary relief,” not the value paid.

Table 5. Observed response capture at alternative review capacities

Review share	Cases	Portable capture	Full-model capture
5%	999	15.9%	21.9%
10%	1,997	27.3%	39.4%
20%	3,994	52.1%	63.6%
30%	5,991	72.9%	80.2%

Portable = issue, sub-issue, and channel; full model additionally includes company. Rankings use the October–December holdout.

bility threshold. These are *retrospective associations with a recorded response category*; prioritization must never pre-judge the appropriate remedy for a consumer.

Review capacity is a managerial choice, so Table 5 evaluates several budgets rather than treating the tenth as universal. With the portable model, screening 5% of the holdout portfolio identifies 15.9% of the subsequently recorded monetary-relief cases; a 20% queue identifies 52.1%. At 30% of cases, coverage reaches 72.9%. The larger cross-company figures describe the heterogeneous public portfolio and should not be quoted as expected gains for an individual issuer. Neither model estimates the effect of reviewing a case earlier: reallocation may alter the observed response itself, which only a prospective workflow evaluation could measure.

5.3 Robustness to Case Mix and Response Definition

Table 6 changes one design choice at a time for the portable issue/sub-issue/channel specification. Excluding the six report-related issues lowers October–December AUC from 0.831 to 0.738 and top-decile capture from 27.3% to 20.1%. The aggregate result therefore benefits materially from separating two very different issue families; it is not solely a fine-grained ranking among ordinary card-service disputes. Within the remaining 13,345 holdout cases, prevalence is 18.86%, and AP of 0.331 remains above that prevalence. The reduced AUC is also a reminder that AUCs computed on different case populations are not directly interchangeable measures of deployment value.

An earlier sequential split, with training through June and evaluation in July–September, yields AUC 0.845 and 29.1% top-decile capture. The result supports usefulness across two later quarters within 2024, but it cannot stand in for a 2025 prospective test. More consequentially, changing the target to either monetary or non-monetary relief produces a 37.69% holdout event rate, AUC 0.635, and 13.6% top-decile capture. The structured fields predict a narrow company response

Table 6. Sensitivity checks for the portable intake-field model

Validation scenario	Cases / rate (%)	AUC / AP	Top 10% capture
Main, portable	19,967	0.831	
	12.79	0.329	27.3%
No report-related issues	13,345	0.738	
	18.86	0.331	20.1%
	20,732	0.845	
Early-period test	12.31	0.339	29.1%
	19,967	0.635	
Any recorded relief	37.69	0.476	13.6%

Each row is fitted anew using its indicated development data. The early-period test trains on January–June and evaluates on July–September 2024; other rows use the October–December holdout. “Any recorded relief” combines monetary and non-monetary relief categories.

Table 7. Monthly calibration in the temporal holdout

Holdout month	Cases	Observed relief (%)	Predicted relief (%)
2024-10	7,176	12.35	12.86
2024-11	6,095	12.65	13.71
2024-12	6,696	13.40	13.63

Table 8. Holdout response mix by submission channel

Submission channel	Cases	Observed relief (%)	Predicted relief (%)
Phone	600	21.00	21.40
Postal mail	170	20.00	21.99
Referral	580	20.52	21.09
Web	18,617	12.22	12.80

classification more effectively than they predict any relief. A manager designing a workflow must therefore name the operational endpoint before selecting a model; an attractive AUC for one label does not validate a broad claim about complaint resolution.

5.4 Segment and Portfolio Audits

Observed relief in October, November, and December is 12.35%, 12.65%, and 13.40%, respectively; full-model means are 12.86%, 13.71%, and 13.63% (Table 7). Web submissions are much more numerous than the other channels, and the model’s slight overprediction is visible there as well (Table 8). Channel differences cannot be ascribed to channel treatment because complainants and issues are not assigned randomly to channels.

Table 9 illustrates why a raw cross-company percentage is a poor quality metric. The eight largest portfolios include three with close to zero recorded monetary relief and issue-mix

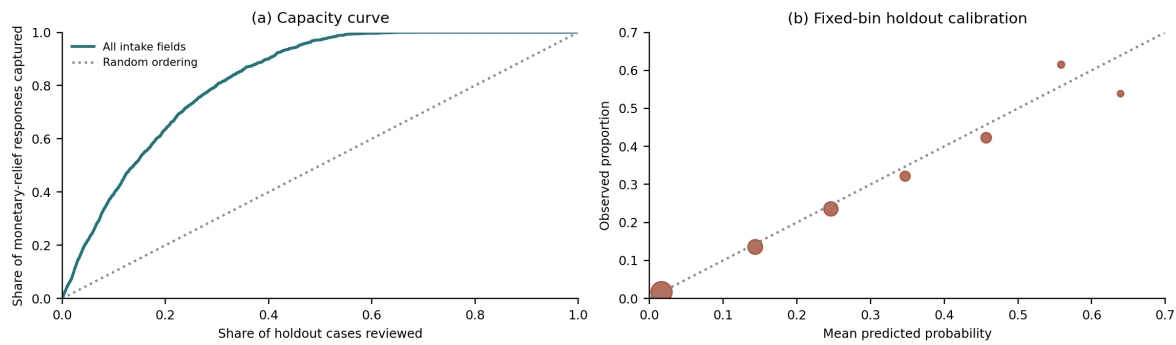


Figure 2. Holdout diagnostics for the model with issue, sub-issue, channel, and company. Panel (a) plots cumulative capture by ranked review capacity; panel (b) compares predicted and observed proportions in fixed ten-percentage-point bins, with point size indicating bin count, capped for legibility.

Table 9. Illustrative issue-mix audit of the eight largest company portfolios

Portfolio	Cases	Observed (%)	Mix expected (%)	Gap (pp)
P1	7,532	0.00	1.46	-1.46
P2	7,382	0.00	1.43	-1.43
P3	7,097	0.04	1.56	-1.52
P4	7,058	12.78	16.02	-3.24
P5	6,685	30.23	18.80	+11.43
P6	5,301	20.05	20.28	-0.23
P7	5,276	11.45	19.21	-7.76
P8	3,950	14.89	16.56	-1.68

Portfolio labels follow descending case volume. Expectations use Equation 3; gaps are observed minus expected percentage points.

expectations near 1.5%; P5 records 30.23% against a simple issue-mix expectation of 18.80%. Neither extreme establishes good or bad resolution: the company's identity, products and assignment process may differ, and the published category does not record whether consumers were made whole. The issue-mix gap is a prompt to audit case definitions and internal workflows, not a scorecard.

6. DISCUSSION

For a card-operations team, the practical output is a monthly case-mix panel: show intake issue and sub-issue shares, observed response classes, referral volume, and calibration of any review queue against later cases. The portable intake-field model can rank cases for specialist assessment across a common workflow; managers should separately record how many selected cases require fraud, billing, or credit-report expertise. The company-field model is useful for describing heterogeneity across the public database, but it is not a justified within-firm deployment model. Any operating implementation would require local timestamps, a verification of field availability, and an assessment of how review capacity changes resolution outcomes.

The analysis also identifies a measurement problem for board reporting. If the proportion of report-related cases rises, the aggregate monetary-relief rate can fall even while every issue-specific response rate stays fixed. Equation 4 makes that compositional question explicit. The portfolio benchmark in Equation 3 is intentionally modest: it adjusts for broad issue categories, not case severity, customer eligibility, company size, or transaction exposure. A lower monetary-relief category share may reflect a correction delivered without

payment, an explanation, or a different type of complaint. The CFPB's annual report notes that complaint counts are not unique-consumer counts and that some submissions are referred elsewhere; accordingly, our findings describe the published database rather than market-wide incidence [9].

Several boundaries matter. Outcomes may be recorded or revised after the receipt date; the export is a later extraction of 2024 cases, not a recreated historical database snapshot. The public response field is a classification, not a causal effect, consumer satisfaction score, or cash amount. The temporal test reduces one source of optimistic evaluation but does not prove stability in another year, market, or firm. Company identities are strong predictors partly because different institutions receive different types of complaints. Further validation should use prospective internal intake logs, resolution-cost data, and randomized or controlled workflow changes before any claim about a triage intervention's business value.

7. CONCLUSION

A single credit-card complaint rate conceals material differences in the case mix. In 75,988 published 2024 complaints, nearly one-third belong to report-related issue groups with a recorded monetary-relief rate below one percent; the remaining issues have a rate of 19.16%. Changing issue shares account arithmetically for 1.81 of the 2.64 percentage points separating the first and fourth quarters. A transparent issue-based model retains substantial discrimination on a later quarter (AUC 0.819), and a richer cross-company specification reaches 0.873, with an important limit on within-firm interpretation. Business intelligence for complaint operations should report issue composition, response-class mix, out-of-time calibration, and review-capacity capture together. Those measures can direct analyst attention to changing workloads and inconsistent processes while leaving the remedy to human review and the evidence in each case.

REFERENCES

- [1] Vairetti, C., Aránguiz, I., Maldonado, S., Karmy, J. P., and Leal, A., "Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making," *European Journal of Operational Research*, vol. 312, no. 3, pp. 1108–1118, 2024.
- [2] Dou, Y., Hung, M., She, G., and Wang, L. L., "Learning from peers: Evidence from disclosure of consumer complaints," *Journal of Accounting and Economics*, vol. 77, nos. 2–3, Art.

no. 101620, 2024.

- [3] Grosz, M., and Raval, D., “Amplifying consumers’ voice: The Federal Trade Commission’s Report Fraud website redesign,” *Marketing Science*, vol. 44, no. 3, pp. 525–545, 2024.
- [4] Ali, F., El-Manstrly, D., and Abbasi, G. A., “Would you forgive me? From perceived justice and complaint handling to customer forgiveness and brand credibility—symmetrical and asymmetrical perspectives,” *Journal of Business Research*, vol. 166, Art. no. 114138, 2023.
- [5] Roschk, H., Hosseinpour, M., and Breitsohl, J., “Coalitions and their negative consequences: An examination in service failure-recovery situations,” *Journal of Service Research*, vol. 26, no. 4, pp. 614–635, 2023.
- [6] Chen, C., “How consumers respond to service failures caused by algorithmic mistakes: The role of algorithmic interpretability,” *Journal of Business Research*, vol. 176, Art. no. 114610, 2024.
- [7] Chung, S., Jung, J., Park, J., Lee, C. H., and Ceran, Y., “Two-sided impacts of service provider’s identity disclosure in e-customer service platforms: Evidence from two field experiments,” *Information Systems Research*, vol. 36, no. 3, pp. 1631–1651, 2025.
- [8] Hayes, R. M., Jiang, F., Pan, Y., and Tang, H., “Racial disparities in financial complaints and the role of corporate social attitudes,” *Journal of Accounting Research*, vol. 63, no. 4, pp. 1289–1333, 2025.
- [9] Consumer Financial Protection Bureau, “2024 Consumer Response Annual Report,” 2025.