An End-to-End Pipeline for Causal ML with Continuous Treatments: An Application to Financial Decision Making

Javier Moral Hernández
javier.moral.hernandez@bbva.com
BBVA, AI Factory
Madrid, Spain

Clara Higuera-Cabañes
clara.higuera@bbva.com
BBVA, AI Factory
Madrid, Spain

Álvaro Ibraín
alvaroibrarod@gmail.com
BBVA, AI Factory
Madrid, Spain

Abstract

This paper presents an end-to-end causal machine learning (ML) pipeline designed for real-world applications with continuous treatments. The proposed framework consists of six sequential steps: dimensionality reduction, causal identification, positivity assumption violation handling, estimation, refutation and evaluation, and policy optimization. We introduce practical contributions not currently available in existing causal ML toolkits, specifically: (1) a method for detecting and quantifying positivity violations in continuous treatment settings (2) a novel, scalable two-stage dimensionality reduction framework tailored for causal inference with high-dimensional data; (3) the adaptation of sensitivity analysis and estimation methods originally designed for binary treatments to the continuous treatment space and (4) an end-to-end integration of these components into a modular, reproducible workflow. These innovations address real-world challenges in causal inference that are often not covered in theoretical frameworks but frequently encountered in industrial applications. The methodology is validated with a synthetic dataset inspired in a real-world financial debt collection use case, however its design can be applied to analogous problems across different industries. Results demonstrate that the proposed methodology offers a more computationally efficient approach and produces less biased estimates compared to standard methods for problems with continuous treatment and high-dimensional data. A fully functional GitHub repository with documented code and numbered notebooks is made available ensuring reproducibility and practical implementation. The pipeline presented is intended to contribute to closing the gap between academic approaches and practical application in industry contexts where causal ML can be highly beneficial such as the financial sector.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Causal reasoning and diagnostics**.

Keywords

Causal Machine Learning, Causality, Counterfactuals, Causal Inference, Causal ML in finance

KDD2025, Aug 03–07, 2025, Toronto, ON

ACM ISBN XXXXXXXX/25/07
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Javier Moral Hernández, Clara Higuera-Cabañes, and Álvaro Ibraín. 2025. An End-to-End Pipeline for Causal ML with Continuous Treatments: An Application to Financial Decision Making. In *Proceedings of 3rd Workshop on Causal Inference and Machine Learning in Practice (KDD2025).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Causal machine learning (causal ML) offers significant advantages over predictive ML, particularly in scenarios that require understanding cause-and-effect relationships, providing actionable insights, and avoiding spurious correlations. Unlike predictive ML, which identifies associations and correlations, causal ML assesses the true impact of interventions, generalize better to new scenarios, and ensure fairer, more reliable outcomes. This is especially valuable in domains where randomized control trials (RCTs) are impractical due to ethical, regulatory, operational, or economic constraints. [21][5].

The standard causal inference workflow includes: (1) Problem Formulation: Defining the causal question, treatment, and outcomes; (2) Identification: using causal discovery or expert-designed DAGs followed by estimand identification (e.g., backdoor, frontdoor) [22] [16]; (3) Estimation: applying suitable methods to estimate causal effects; (4) Validation: using refutation tests, balance and weights diagnostics and sensitivity analysis to ensure robustness[37] and (5) Policy Optimization, where estimated causal effects and counterfactuals are used to inform decision-making and optimize interventions.

Despite its theoretical advantages, causal ML remains underutilized in industry settings[19] due to significant technical challenges. These include complex data structures, the intricacies of business decision-making processes, and the limitations of current causal inference frameworks when applied to industrial-scale problems.

The first critical challenge stems from the high dimensionality of modern industrial datasets that introduce a range of interconnected challenges for causal discovery and effect estimation. Key issues include:

Computational Scalability: Constraint-based algorithms like PC [35] can be efficient under sparsity, but exhibit exponential worst-case complexity [14]. Score-based methods, such as Greedy Equivalence Search (GES) [4], are further constrained by the super-exponential growth in Markov equivalence classes [9], limiting their practical use as dimensionality increases.

Expert Knowledge Integration: Incorporating domain expertise in causal discovery is essential [18] but inherently subjective and labor-intensive. Experts must assess edge plausibility, resolve ambiguities, and identify spurious relationships—efforts that

scale quadratically with the number of variables. Iterative refinement compounds the complexity, requiring repeated evaluations of changing graph structures.

Validation via Testable Implications: The number of required conditional independence tests grows combinatorially with dataset size, affecting both computational feasibility and statistical reliability, especially as conditioning sets increase in dimension [32]. This also raises concerns over multiple testing and false discovery control.

Adjustment Set Selection for Estimation: Identifying valid adjustment sets becomes increasingly complex due to both the algorithmic difficulty of navigating large graphs and the exponential number of possible adjustment sets [13]. This creates a tension between statistical efficiency and practical feasibility [38].

While much of the causal inference literature focuses on binary treatments for simplicity, many real-world decisions involve continuous treatment variables, which presents the second major challenge when using standard causal inference tools. Popular estimators like T-learners and X-learners [16] are primarily designed for binary treatments and struggle to generalize effectively to continuous settings. Similarly, sensitivity analysis tools such as E-values [37] which help assess the robustness of causal estimates to unmeasured confounding, are typically not formulated to handle continuous interventions, limiting their applicability in practice.

The third challenge in industrial applications of causal inference is the violation of the positivity assumption [26]— the requirement that all individuals have a non-zero probability of receiving any treatment level. This assumption often fails in real-world settings where human decision-making limits exposure to certain treatments, and randomized control trials are not feasible. Early work by Petersen et al. [23] provided a comprehensive framework for diagnoses and remedies for positivity violations but it is mainly focused on binary treatments. More recent approaches by Guo et al. [10] propose solutions to continuous treatments using techniques such as local density estimation, trimming, and weighting to address regions of low treatment overlap.

Lastly, while existing Python frameworks provide valuable methods for addressing individual challenges in causal inference (e.g., DoWhy[33], causalml[3], EconML[25]), there remains a significant gap in comprehensive methodological frameworks that address the aforementioned challenges encountered in industrial applications. For instance, DoWhy is not prepared for continuous outcome or high dimensional data and Microsoft EconML library offers tools for heterogeneous treatment effect estimation but primarily focuses on estimation rather than the complete pipeline from data preparation to policy optimization. There exist limited work on pipelines that simultaneously handle high-dimensional data, continuous treatments, positivity violations and few data samples while providing practical guidance for implementation in real settings.

The financial sector stands to gain significantly from causal methods, which reveal cause-and-effect relationships beyond mere correlations. In [11] authors provide a comprehensive overview of the application of causal inference methods in the banking, finance, and insurance sectors. These methods have been applied in areas like investment management [1], fraud detection [39], and fair credit decisions [15]. However, despite growing interest, the field lacks standardized frameworks and example applications suited to finance. The complexity of financial data, such as heterogeneity in client profiles and sparsity in intervention histories—further, and the wide range of available techniques pose challenges to effective implementation in real settings.

This paper introduces a novel end-to-end causal machine learning framework designed to overcome the afore mentioned challenges. Validated on a financial debt collections use case and a synthetic dataset. The work aims to promote the use of causal methods in financial applications and other fields in which operational decision-making is critical. The paper is structured as follows. First it outlines the study's scope and main contribution, followed by description of the experimental design, presentation of results, and concluding with a discussion of findings, limitations, and future research directions.

The code for reproducing all experiments is publicly available at `github.com/javiermoralh/causal-pipeline`.

2 Problem statement / Case study

In banking, the debt collections department manages clients in loan default using strategies like refinancing, debt sales, or write-downs—partial debt forgiveness in exchange for immediate repayment. While some clients repay successfully after a write-down, others do not, making optimal decision-making crucial.

Traditionally, such decisions rely on expert judgment to balance repayment likelihood and loss minimization. While automation could enhance this process, prediction alone is insufficient— it demands a causal understanding of how different write-down levels affect repayment outcomes.

Though randomized controlled trials (RCTs) would offer ideal estimates of causal effects, they are impractical in this context due to ethical concerns and regulatory barriers. Practitioners must instead rely on observational data, which is subject to confounding bias: clients in worse financial health typically receive larger reductions, making it difficult to isolate causal effects.

Given the infeasibility of RCTs, the presence of confounding, and the need for personalized, counterfactual insights, causal inference methods become essential for optimizing debt collection strategies effectively and fairly.

3 Methods

In order to address the aforementioned challenges we propose a series of steps in a form of a pipeline that practitioners can use in order to solve domain agnostic problems as long as any of the above issues are present. The pipeline is comprised of six steps that can be run sequentially. These steps are depicted in Figure 4 in Appendix D.

(1) **Dimensionality Reduction**: A structured approach to identify potential confounders and outcome-only causes from high-dimensional data.
(2) **Identification**: A hybrid method combining algorithmic discovery with domain knowledge to get the final adjustment set.
(3) **Positivity Assumption Violation Handling**: A framework for detecting and quantifying regions of limited overlap in treatment assignment, alongside remediation strategies to address identified violations.

(4) **Estimation**: Methods for continuous treatment effects estimation.
(5) **Refutation and Evaluation**: Comprehensive evaluation techniques for estimate selection.
(6) **Policy Optimization**: Methodology for deriving optimal decision policies.

This work assumes that the causal estimand of interest is identifiable through the backdoor criterion [22]. Consequently, every stage of the pipeline – most notably the dimensionality-reduction and identification procedures – are tailored to selecting and balancing only those covariates that close backdoor paths (i.e., confounders). This focus implies that practitioners need control merely for these backdoor variables, simplifying the analysis while underscoring the importance of correctly detecting them to avoid residual bias.

In the following subsections each one of this blocks is covered in more detail.

3.1 Dimensionality Reduction

Applying causal inference to high-dimensional data requires a reduction in dimensionality that maintains causal integrity while addressing computational limits and potential positivity violations. Let $\mathcal{D} = (X_i, T_i, Y_i)_{i=1}^N$ denote the dataset, where $X \in \mathbb{R}^d$ are covariates, $T \in \mathbb{R}$ is the treatment, and Y is the outcome. We propose a two-stage selection framework to construct a reduced adjustment set $\mathbf{Z} = \mathbf{Z}_T \cup \mathbf{Z}_Y$, where $\mathbf{Z}_T$ captures treatment predictors and $\mathbf{Z}_Y$ includes outcome-relevant covariates. This reduction aims to provide a manageable set of candidate adjustment variables for the identification phase, which further refines them using causal criteria and domain expertise to find the backdoor variables.

In Stage 1, $\mathbf{Z}_T$ is identified using predictive machine learning-based feature selection to capture variables predictive of treatment assignment. Stage 2 uses dual partial-correlation analysis to identify covariates that influence the treatment–outcome relationship and covariates that affect only the outcome. The first type of covariates are retained as candidate *confounders*—they must be controlled for to obtain an unbiased effect under the backdoor criterion—whereas the second type, while not mandatory, are desirable because their inclusion can lower the variance of the causal estimate [2, 7]. The first subset in $\mathbf{Z}_Y$ is identified by computing the partial correlation between T and Y, conditional on each covariate X_j:

$$\rho(T, Y|X_j) = \text{Corr}(\text{Res}(T \sim X_j), \text{Res}(Y \sim X_j)) \tag{1}$$

where $\text{Res}(\cdot)$ denotes regression residuals. Covariates inducing substantial deviations from the baseline correlation $\rho(T, Y)$ are retained as potential confounders. For outcome-only predictors, variables strongly correlated with Y independent of T—determined via $\rho(X_j, Y|T)$—are included in $\mathbf{Z}_Y$.

The framework employs distinct approaches for treatment-related $\mathbf{Z}_T$ and outcome-related $\mathbf{Z}_Y$ selection, reflecting fundamental differences in how confounding and positivity violations affect these relationships. The framework distinguishes between treatment- and outcome-related features due to differing vulnerability to confounding and positivity violations. Standard predictive machine-learning feature selection can reliably capture treatment-related covariates because, in most observational settings, treatment assignment follows systematic decision rules that produce stable covariate–treatment links even where overlap is sparse [20, 30]. However, for outcome-related features, standard feature selection methods become unreliable due to the interplay between positivity violations and confounding. In regions where certain covariate-treatment combinations are unobserved, traditional feature selection techniques may identify spurious correlations that exist in the observed regions but do not represent genuine causal relationships. The dual partial correlation analysis circumvents these challenges by explicitly conditioning on covariates when examining confounders selection and conditioning on treatment when examining outcome-only predictor relationships.

This approach enables scalable, causally valid inference by reducing dimensionality while preserving essential confounding structures.

3.2 Identification

Following dimensionality reduction, the identification phase integrates algorithmic causal discovery with domain expertise validation. As Mäkelä et al. (2022) [18] highlight, causal discovery algorithms generate hypotheses rather than definitive conclusions, necessitating expert refinement. This phase employs an ensemble of methods—PC [35], FCI [36], and GES [4]—leveraging their complementary strengths to increase confidence in consistently identified relationships.

Algorithmic outputs serve as initial structural hypotheses, iteratively refined through domain expertise:

- Temporal Constraints: Ensuring relationships align with known variable orderings.
- Edge Validation: Assessing plausibility and temporal consistency.
- Causal Direction: Resolving algorithmic uncertainty with expert knowledge.
- Missing Relationships: Identifying overlooked causal links.
- Spurious Correlations: Removing statistically significant but non-causal associations.

Identifying and removing spurious correlations remains a key challenge, as statistical associations may lack theoretical justification. Experts assess these cases, refining the graph through additional statistical tests and confounder control. The complexity of this process scales quadratically with graph size, reinforcing the importance of prior dimensionality reduction.

Final causal graphs undergo rigorous validation via testable implications (conditional independence tests) until statistical and expert criteria align. The refined graph determines the adjustment set for causal effect estimation using identification algorithms such as the backdoor criterion [22], with outcome-related variables included to enhance precision [7].

3.3 Positivity Assumption Violation Quantification

We propose a three-step, model-agnostic procedure to detect regions of the covariate space where lack of overlap violates the positivity assumption for a continuous treatment based on Hirano et al. (2004) framework [12].

1. Treatment prediction. Fit a flexible regression $f : Z \to \mathbb{R}$ that predicts T and form residuals $\varepsilon = T - f(Z)$.

2. Residual-density estimation. Estimate the conditional density of the residuals $g(\varepsilon \mid Z)$ with any consistent method, e.g.

(1) kernel density estimators with data-driven bandwidth or plug-in rules [31, 34]
(2) a homoscedastic Gaussian with analytic variance
(3) conditional normalizing flows such as the GOALDeR continuous-GPS estimator, which learn $g(\varepsilon \mid Z)$ via invertible neural networks and naturally handle heteroscedasticity [6, 8].

The preferred model can be selected by held-out (pseudo)-log-likelihood or PSIS-LOO.

3. Overlap quantification. For any interval $[t_1, t_2]$ compute the generalised propensity score

$$P\big(T \in [t_1, t_2] \mid Z = z\big) = \int_{t_1}^{t_2} g\big(t - f(z) \mid z\big)\, dt.$$

A practical violation is flagged whenever this probability falls below a small tolerance ϵ. The scalar $P(T \in [t_1, t_2] \mid Z)$ offers a continuous measure of violation severity, highlights covariate regions with poor support, and guides trimming or re-weighting—without tying practitioners to a single residual-density specification.

Common remediation strategies following Petersen et al. [23] include (i) **covariate restriction** to drop variables that force non-overlap, (ii) **trimming** units whose generalised propensity scores lie in the tails, (iii) **weight stabilisation / truncation** to temper extreme inverse probabilities, (iv) **design modification or data augmentation** to avoid unlikely treatment levels, and (v) **model-based extrapolation** when strong substantive knowledge supports it. Each option trades bias for variance: trimming lowers power, reweighting inflates uncertainty, and extrapolation leans on unverifiable assumptions.

Before deciding on a remedy, analysts should check overlap quality by using some diagnostic tools as (a) checking Standardised Mean Differences of each covariate across treatment strata or after weighting [28]; (b) plotting the GPS density to flag regions with large density-ratios ; (c) inspecting the distribution of stabilised inverse-probability weights for extreme values; and (d) verifying a common support region [27] via overlayed treatment densities.

Given these challenges, the objective shifts from eliminating violations to minimizing bias while maintaining practical utility. This structured positivity violation framework bridges theoretical causal inference with real-world application, enabling practitioners to assess the validity and generalizability of causal estimates.

3.4 Estimation

With the adjustment set Z in place, the goal is to recover the conditional average *dose–response curves*—i.e. $\mathbb{E}[Y(t) \mid Z_i]$ for every feasible treatment level t. Unlike binary interventions, continuous treatments require modelling this entire surface, typically by evaluating potential outcomes at several grid points [29].

We benchmark three representative estimators (others can be swapped in):

(1) **Regression adjustment with interactions**. A linear (or logistic) model of Y on (T, Z) yields an interpretable baseline; T's coefficient and its interactions capture average and heterogeneous effects [22].
(2) **S-learner** [16]. A single flexible learner fits $Y = f(T, Z)$. This method can apply more advanced regularization techniques in order to cap- ture complex relationships while preventing overfitting, enabling the identification of non-linear interactions and heterogeneous effects.
(3) **Augmented IPTW for continuous** T [17]. This approach extends the classical AIPTW methodology to accommodate continuous treatment scenarios. Generalized Propensity Scores produce inverse-probability weights that, combined with an outcome model, form a *doubly-robust* estimator—consistent if either component is correct.

Together, these methods span a spectrum from simple, transparent baselines to more expressive and robust machine-learning approaches.

3.5 Refutation and Evaluation

This step is aimed at detecting potentially invalid estimates in order to avoid, as much as possible, spurious relationships that can remain after the aforementioned steps. The choice of refutation methods, such as, is open to the practitioner. However, in the scope of this work we only study the effect of the following two: (1) **Placebo Treatment Replacement test** [33]: With this test, actual treatments are substituted with random variables following identical distributions. Here, valid estimates should demonstrate no effect on outcomes, manifested as flat effect curves. (2) **Random Common Cause test** [33]: Evaluates estimate stability by introducing synthetic random confounders unrelated to both treatment and outcome variables. When successful, this test must yield consistent causal curves before and after the introduction of random confounders to the adjustment set.

Residual bias attributable to unmeasured confounders may persist even after the refutation procedures described above. To quantify the extent to which such bias could attenuate or overturn the estimated effects, we implement a sensitivity analysis based on the *E-value* [37] framework. The *E-value* quantifies the minimum strength of unmeasured confounding necessary to nullify the estimated effects. Higher E-values indicate greater estimate robustness. For continuous treatments, we extend the E-value methodology by computing a dose-response sensitivity curve $E(t)$, where for each treatment level t, we calculate the E-value required to nullify the estimated effect for a unit change in exposure centered at t.

Finally, estimator performance in ranking individuals by anticipated uplift is assessed with the Qini curve [24], which plots cumulative gain against the cumulative proportion of treated units ordered by the model. The area under this curve ($\mathrm{AUC}_{\mathrm{Qini}}$) condenses the curve into a single, sample-size-invariant metric, thereby enabling rigorous comparison across competing estimators.

Collectively, these three assessments—falsification tests, quantitative sensitivity metrics, and uplift-oriented performance curves constitute a concise yet rigorous validation suite for continuous-treatment causal inference.

4 Experimental setup

To evaluate the validity of the proposed method, we assess its performance within a financial context where the objective is to determine the optimal amount of debt write-downs that minimizes the total expected loss. Given the sensitive nature of financial data, a real publicly available version cannot be provided for reproducibility. Instead, a synthetic dataset has been generated that captures the essential characteristics and statistical properties of the original data, while not being an exact replication.

Specifically, the data generation process incorporates: (i) a continuous treatment variable representing debt loss percentages in the $[0, 100]$ range, (ii) non-linear and heterogeneous treatment effects through carefully crafted probability functions with interaction terms, (iii) systematic confounding bias through structured covariate relationships, (iv) high dimensionality with 410 features including both relevant causal variables and noisy ones, and (v) deliberate positivity assumption violations through a logistic-based assignment mechanism that creates regions of limited overlap in treatment assignment. The performance of the proposed method is then analyzed for the synthetic setting. More details on the generation process and its characteristics are explained in Appendix A.

Evaluation targets three aspects: (i) recovery of the true causal covariates, (ii) accuracy of conditional average dose-response curves which are shown in Figure 1, and (iii) computational efficiency.

Causal-variable retrieval is scored with precision and recall against the known adjustment set. Dose-response estimation bias is quantified by the root-mean-squared error (RMSE) between the true and estimated curves, averaged across 100 treatment levels on all validation cases. Specifically, for each individual i in the validation set, we compute both the "true" conditional average dose-response curve $f_i(t)$–coming from our generation process as $\mathbb{E}[Y(t) \mid Z_i]$–and the estimated curve $\hat{f}_i(t)$ at discrete treatment levels $t \in \{1, 2, \ldots, 100\}$, corresponding to the percentage of debt loss. The RMSE is then computed at each treatment level across all individuals to obtain the individual bias $\mathcal{B}(t)$:

$$\mathcal{B}(t) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i(t) - \hat{f}_i(t))^2} \tag{2}$$

The final metric, $\overline{\mathcal{B}}$, is obtained by averaging all individual biases over the entire evaluation set. Estimators that fail placebo or random-common-cause tests are discarded. Subsequently, the remaining methodologies are comparatively assessed through E-value sensitivity analysis and Qini-AUC in order to determine optimal performance characteristics.

Finally, we conduct an ablation study by disabling the dimensionality reduction and causal discovery components in the pipeline. This analysis isolates the marginal contribution of each step to the overall performance, providing insights into their impact on causal identification, estimation accuracy, and computational efficiency.

A comprehensive specification of the method's configuration, including correlation thresholds for dimensionality reduction, algorithmic choices for treatment-predictive feature derivation, and parameters for positivity violation detection, is provided in Appendix B. This documentation ensures methodological transparency and facilitates reproducibility.

Table 1: Comparison of Variable Identification Performance

Method	Precision	Recall	Covariates	Runtime (min)
Baseline	0.05	1	169	>600
Dim. Reduction	0.53	1	15	3
Dim. Reduction & Identification	0.80	1	10	4

5 Results

This section presents the results obtained at each stage of the proposed framework. We begin by examining improvements in causal variable selection, including a runtime analysis to assess computational efficiency. We then evaluate the enhancements in causal effect estimation and analyze the outputs that support informed decision-making in the proposed case study.

5.1 Causal identification improvement

The evaluation of the dimensionality reduction (3.1) and identification (3.2) phases reveals the effectiveness of the proposed approach in correctly identifying control variables for causal effect estimation within a high-dimensional feature space. Table 1 presents a comparative analysis between the proposed methodology and the baseline approach

The baseline approach achieved a precision of 0.05 with perfect recall (1.0), selecting 169 covariates—typical of traditional causal discovery methods applied to high-dimensional data. Among the three algorithms tested (PC, FCI, GES), only PC completed execution within 600 minutes, making it the basis for baseline results.

Our two-stage methodology significantly improved performance. Dimensionality reduction increased precision to 0.53 while maintaining recall at 1.0, reducing the covariate set to 15. The final identification phase further improved precision to 0.80, yielding a set of 10 covariates, including all 8 true causal controls and 2 treatment-related variables. While such variables can inflate estimator variance without reducing bias [7], their inclusion remains a known challenge in causal inference.

In terms of computational efficiency, the baseline required over 600 minutes for causal discovery, whereas the proposed approach reduced total runtime to just 4 minutes (3 minutes for dimensionality reduction and 1 minute for identification). This improvement is critical for industrial applications, enabling rapid iteration and expert feedback integration.

5.2 Estimates Performance

The evaluation of treatment effect estimation bias reveals significant improvements through the proposed methodology. Table 2 presents the comparative analysis of mean averaged bias across the three evaluated approaches. The baseline methodology utilizing the full feature set exhibits the highest estimation bias ($\overline{\mathcal{B}} = 0.292$, 95% CI $[0.279, 0.303]$), demonstrating the detrimental impact of

high-dimensional noise and spurious correlations. Restricting the adjustment set to the identified causal controls while omitting synthetic data augmentation reduces bias by 19% ($\overline{\mathcal{B}} = 0.236$, 95% CI [0.212, 0.257]), though residual bias persists due to unaddressed positivity violations. The complete proposed methodology achieves superior performance with $\overline{\mathcal{B}} = 0.117$ (95% CI [0.105, 0.131]), representing a 60% reduction in mean bias compared to the baseline.

Among the three estimation methodologies described in section 3.4, the S-learner emerged as the optimal approach. Initial refutation testing conducted with 100 bootstrapped samples revealed significant limitations in the AIPTW estimator, which failed to pass the placebo treatment replacement test (Figure 6 in Appendix D) at extreme debt loss percentages where the dose-response curve should theoretically approach zero.

Consequently, subsequent sensitivity analyses through E-values (Figure 7 in Appendix D) and cumulative gain curves (Figure 8 in Appendix D) focused exclusively on the S-learner and Linear Regression approaches. While both methodologies demonstrated comparable robustness in E-value analysis, the S-learner exhibited markedly superior performance in the AUC curve evaluation, ultimately justifying its selection as the primary estimation methodology for the comprehensive ablation study.

Table 2: Comparison of Estimation Bias Across Methodologies

Methodology	Mean Bias	95% Confidence Interval
Baseline	0.292	[0.279, 0.303]
Adjustment Set Only	0.236	[0.212, 0.257]
Proposed Methodology	0.117	[0.105, 0.131]

The treatment-level bias analysis, visualized in Figure 12 in Appendix D, demonstrates consistent superiority of the proposed methodology across the entire treatment domain. While all methods exhibit increased bias at extreme treatment values ($0 - 20\%$ and $80 - 100\%$ debt loss), the proposed approach shows particular robustness in these regions with maximum RMSE of 0.20 compared to 0.5 for the baseline. This enhanced performance is attributed to the monotonic synthetic data augmentation strategy that effectively mitigates positivity violations in low-density treatment regions. The methodology maintains stable estimation quality across mid-range treatments ($20 - 80\%$) with RMSE consistently below 0.20, significantly outperforming both baseline approaches.

This enhanced precision results from the synergistic combination of dimensionality reduction's variance minimization and synthetic augmentation's support expansion. The ablation study reveals that 32% of total bias reduction originates from proper adjustment set identification, while 68% derives from addressing positivity violations through domain-informed synthetic data generation.

These results suggest that while complete elimination of estimation bias remains challenging in practical applications with extreme positivity violations, the proposed framework achieves substantial improvements over conventional approaches.

5.3 Case Study Solution

The application of the proposed pipeline to the debt collections case study yields two primary analytical outputs that enable informed decision-making regarding optimal write-down levels. First, the estimated causal DAG (Figure 9 in Appendix D) reveals the underlying structural relationships between financial indicators, demonstrating that the debt treatment assignment (write-down percentage) is influenced by multiple customer characteristics. This structural understanding validates the necessity of controlling for these confounding variables to obtain unbiased treatment effect estimates.

Second, the conditional average dose-response curves (Figure 10 in Appendix D) illustrate the heterogeneous causal effects of different write-down percentages on repayment probability across the customer population. These curves represent counterfactual predictions for each customer under varying treatment intensities, enabling precise personalization of write-down offers. The substantial variation in curve shapes and slopes indicates marked heterogeneity in treatment effects, suggesting that uniform write-down policies would be suboptimal.

These estimated potential outcomes provide the foundation for subsequent policy optimizations, where institution-specific cost functions can be applied to determine optimal write-down levels that balance recovery probability against financial loss.

6 Discussion and Future Work

The proposed method addresses key challenges in applying causal machine learning to industrial contexts—challenges that are often overlooked in existing frameworks—including high-dimensional data, positivity assumption violations, and continuous treatments. It introduces an end-to-end, adaptable methodology, validated through a financial debt collection use case, and is designed for broader applicability across industries facing similar constraints

Despite its strengths, the pipeline has limitations that warrant further exploration: Dimensionality reduction currently relies on linear partial correlation, which may miss non-linear relationships and interaction effects; future improvements could incorporate mutual information or kernel-based methods. Positivity violation detection, based on GPS and kernel density estimation, could be enhanced to account for unobserved confounding. The data augmentation strategy assumes a monotonic treatment-response, which may not hold in all cases and could require adaptation for non-monotonic scenarios.

These limitations point to promising directions for future research, aimed at refining the framework and extending its utility in complex, real-world applications.

Acknowledgments

Authors would like to thank Dr. Ramin Ghelichi for his generous guidance, thoughtful insights, and unwavering support throughout this work. Authors also extend their gratitude to BBVA AI Factory for its continued support of research initiatives and Ricardo García for his valuable guidance from BBVA Advanced Analytics Discipline and Risk Analytics Department. Special thanks to GRM and Risk Collections teams for their contributions and expertise, which helped keep the methodology grounded in real-world needs.

References

[1] Susan Athey and Guido W. Imbens. 2019. Machine Learning Methods That Economists Should Know About. *Annual Review of Economics* 11 (2019), 685–725. https://doi.org/10.1146/annurev-economics-080218-025940

[2] M. Alan Brookhart, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Stürmer. 2006. Variable Selection for Propensity Score Models. *American Journal of Epidemiology* 163, 12 (2006), 1149–1156. https://doi.org/10.1093/aje/kwj149

[3] Huigang Chen, Totte Harinen, Albert Chen, Matt Teschke, Jeong-Yoon Lee, Victor Y. Yan, Mike Yung, Brian Cui, Robert Gray, Michael Zhang, Te-Lin Wu, and Frank Liu. 2020. CausalML: Python Package for Causal Machine Learning. *arXiv preprint arXiv:2002.11631* (2020). https://arxiv.org/abs/2002.11631

[4] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3 (2002), 507–554.

[5] George D. Demetri. 2021. Real-World Evidence and Advanced Soft Tissue Sarcoma: An Unbreakable Bond (ESMO Sarcoma and GIST Symposium 2020 Industry Satellite Symposium, Milan, February 4, 2020). *Oncology* 99, Suppl 1 (2021), 1–2. https://doi.org/10.1159/000515265

[6] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. 2019. Neural Spline Flows. In *NeurIPS*.

[7] Matheus Facure. 2023. *Causal Inference in Python.* O'Reilly Media, Inc., Sebastopol, CA, 89–93.

[8] Qian Gao, Jiale Wang, and Ruiling *et al.* Fang. 2025. A Doubly Robust Estimator for Continuous Treatments in High Dimensions. *BMC Medical Research Methodology* 25 (2025), 35.

[9] Steven B Gillispie and Michael D Perlman. 2001. Enumerating markov equivalence classes of acyclic digraph models. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 171–177.

[10] Ruoqing Guo, Fan Li, Lumley Thomas, et al. 2021. Violations of the Positivity Assumption in the Causal Analysis of Observational Data. *Statistics in Medicine* 40, 24 (2021), 5543–5560. https://doi.org/10.1002/sim.9157

[11] Anshul Gupta, Michael Koetter, Markus Pelger, and Marno Verbeek. 2023. Causal Inference for Banking, Finance, and Insurance: A Survey. *Journal of Financial Econometrics* 21, 4 (2023), 765–802. https://doi.org/10.1093/jjfinec/nbad023

[12] Keisuke Hirano and Guido W Imbens. 2004. The propensity score with continuous treatments. (2004).

[13] Daniel Israel, Aditya Grover, and Guy Van den Broeck. 2023. High Dimensional Causal Inference with Variational Backdoor Adjustment. *arXiv preprint arXiv:2310.06100* (2023).

[14] Markus Kalisch and Peter B"uhlmann. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* 8, 3 (2007).

[15] Niki Kilbertus, Marta Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination Through Causal Reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[16] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 4156–4165.

[17] Christoph F Kurz. 2021. Augmented Inverse Probability Weighting and the Double Robustness Property. *Medical Decision Making* 42, 2 (2021), 156–167.

[18] Jarmo Mäkelä, Laila Melkas, Ivan Mammarella, Tuomo Nieminen, Suyog Chandramouli, Rafael Savvides, and Kai Puolamäki. 2022. Technical note: Incorporating expert domain knowledge into causal structure discovery workflows. *Biogeosciences* 19 (2022), 2095–2099. https://doi.org/10.5194/bg-19-2095-2022

[19] Markets and Markets. 2024. *Causal AI Market by Offering, Application – Global Forecast to 2030.* Market Research Report 5805704. Markets and Markets, Global. 332 pages.

[20] Daniel F. McCaffrey, Greg Ridgeway, and Andrew R. Morral. 2004. Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods* 9, 4 (2004), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

[21] Mahnush Movahedi, Benjamin M. Case, James Honaker, Andrew Knox, Li Li, Yiming Paul Li, Sanjay Saravanan, Shubho Sengupta, and Erik Taubeneck. 2021. Privacy-Preserving Randomized Controlled Trials: A Protocol for Industry Scale Deployment. In *Proceedings of the 2021 on Cloud Computing Security Workshop* (Virtual Event, Republic of Korea) *(CCSW '21)*. Association for Computing Machinery, New York, NY, USA, 59–69. https://doi.org/10.1145/3474123.3486764

[22] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference* (2 ed.). Cambridge University Press, Cambridge.

[23] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. 2010. Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research* 21, 1 (2010), 31–54. https://doi.org/10.1177/0962280210386207

[24] Nicholas J. Radcliffe. 2007. Using control groups to target on predicted lift: Building and assessing uplift models. *Direct Marketing Analytics Journal* 3 (2007), 14–21.

[25] Microsoft Research. 2019. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. (2019). Available at https://github.com/microsoft/EconML.

[26] James M Robins. 1986. A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect. *Mathematical Modelling* 7, 9-12 (1986), 1393–1512.

[27] Paul R. Rosenbaum and Donald B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70, 1 (1983), 41–55. https://doi.org/10.1093/biomet/70.1.41

[28] Paul R. Rosenbaum and Donald B. Rubin. 1985. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician* 39, 1 (1985), 33–38.

[29] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66, 5 (1974), 688–701.

[30] Sebastian Schneeweiss, Jeremy A. Rassen, Robert J. Glynn, Jerry Avorn, Helen Mogun, and M. Alan Brookhart. 2009. High-Dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data. *Epidemiology* 20, 4 (2009), 512–522. https://doi.org/10.1097/EDE.0b013e3181a663cc

[31] David W. Scott. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, New York.

[32] Rajen D. Shah and Jonas Peters. 2020. The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. *Ann. Statist.* 48, 3 (2020), 1514–1538. https://doi.org/10.1214/19-AOS1857

[33] Amit Sharma and Emre Kıcıman. 2020. DoWhy: An End-to-End Library for Causal Inference. *arXiv preprint arXiv:2011.04216* (2020). https://arxiv.org/abs/2011.04216

[34] Bernard W. Silverman. 1986. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[35] Peter Spirtes and Clark Glymour. 1991. An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review* 9, 1 (1991), 62–72.

[36] Peter Spirtes, Christopher Meek, and Thomas Richardson. 1995. Causal Inference in the Presence of Latent Variables and Selection Bias. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (1995), 499–506.

[37] Tyler J. VanderWeele and Peng Ding. 2017. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine* 167, 4 (2017), 268–274.

[38] Janine Witte, Leonard Henckel, Marloes H. Maathuis, and Vanessa Didelez. 2020. On efficient adjustment in causal graphs. *Journal of Machine Learning Research* 21, 246 (2020), 1–45.

[39] Guibin Zhang, Shilong Wang, Yifan Duan, Xiaojiang Peng, Wang Ziqi, Junyuan Mao, Hao Wu, Xinke Jiang, and Kun Wang. 2024. CausalFD: Causal Invariance-Based Fraud Detection Against Camouflaged Fraudsters. *International Journal of Machine Learning and Cybernetics* (2024). https://doi.org/10.1007/s13042-024-02209-0

Appendix

A Synthetic Data-Generation

To validate the proposed methodological framework, we conducted experimentation using synthetic data. This approach enables rigorous evaluation of the pipeline's effectiveness under controlled conditions where the true causal relationships and treatment effects are known. The synthetic dataset was carefully designed to emulate the characteristics and challenges encountered in real-world financial applications while incorporating specific structural properties that test the pipeline's capabilities.

The dataset comprises covariates sampled from probability distributions approximating real financial indicators, including variables related to credit history and debt profiles. These covariates are categorized into three distinct groups: five confounders influencing both treatment assignment and outcome, three outcome-only variables directly affecting repayment probability but not treatment decisions, and two treatment-only variables predictive of debt loss levels but unrelated to the outcome. This structure ensures a realistic confounding scenario where treatment assignment is systematically biased by variables that also drive repayment outcomes.

The continuous treatment variable, representing the percentage of debt loss, is generated through a non-linear function of the confounders and treatment-only variables (3). The treatment assignment mechanism is deliberately constructed to systematically violate the positivity assumption. This is implemented through a logistic-based assignment function with non-linear interactions between covariates, followed by scaling to the [0,100] range. While modest random noise is incorporated via truncated normal distribution to maintain realism, the underlying structure ensures that specific regions of the treatment space become practically inaccessible for certain covariate profiles. This design choice authentically mirrors real-world scenarios where financial advisors assign write-down levels based on rigid policy rules tied to customers' financial health, creating covariate strata with no overlap across treatment ranges.

The outcome generation process was engineered to exhibit heterogeneous treatment effects as visually shown in Figure 1 and 2 in Appendix D, where the impact of interventions varies substantially across different covariate profiles. This heterogeneity is achieved through a carefully crafted probability function that incorporates both direct effects and interaction terms between treatments and covariates (6). Concretely, the inclusion of quadratic terms and multiplicative interactions between covariates guarantees heterogeneous treatment effects across individuals, with each customer exhibiting a unique dose-response curve. When combined with the treatment assignment mechanism, this generates a clear Simpson's paradox as demonstrated in Figure 3 in Appendix D, where the real average treatment-outcome relationship differs markedly from the observed outcome average at each treatment level.

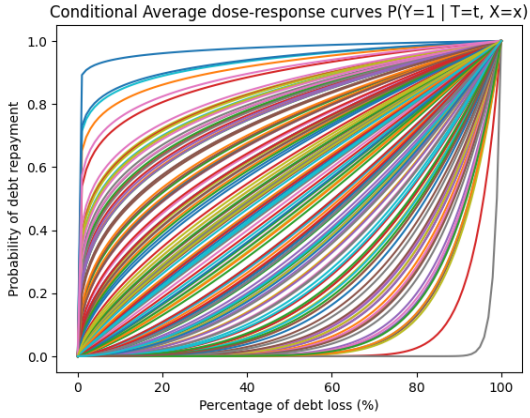

Figure 1: Heterogeneous treatment effects in synthetic generated data - P(Y=1|T=t, X=x)

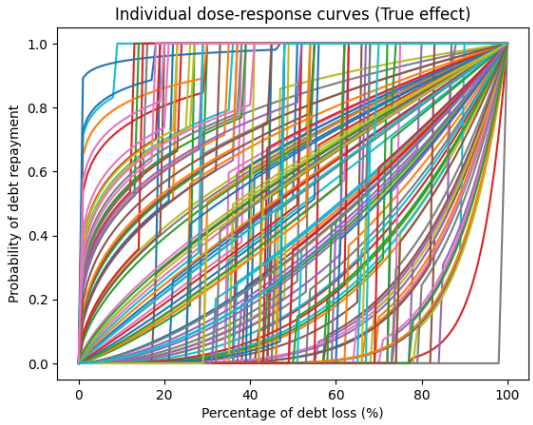

Figure 2: Heterogeneous treatment effects in synthetic generated data - Observed individual outcomes

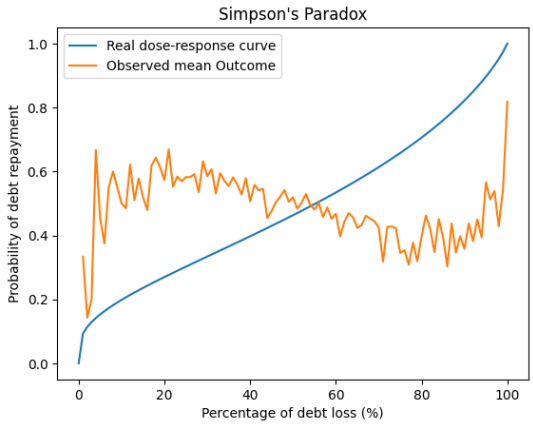

Figure 3: Simpson's paradox in synthetic generated data.

To further approximate real-world data complexity, the feature space was augmented with two additional categories of variables. First, 100 redundant features were generated by introducing controlled correlations with the existing causal variables, maintaining separate correlation structures for confounders, treatment-only, and outcome-only predictors. Second, 300 pure noise features were synthesized by sampling from various probability distributions commonly observed in financial data, ensuring these variables maintain no systematic relationship with either the treatment or outcome. This enriched feature space creates a challenging environment for causal discovery and effect estimation, testing the pipeline's capability to identify relevant variables and discard spurious relationships.

The synthetic data generation process thus encapsulates the core challenges outlined in Section 2: high dimensionality, positivity violations, continuous treatments, and heterogeneous effects. By design, it provides a controlled environment to evaluate the pipeline's ability to recover ground truth causal relationships, isolate confounding bias, and estimate personalized treatment effects—all while mirroring the statistical complexities of real financial data.

B Pipeline configuration

B.1 Dimensionality Reduction

The dimensionality reduction phase was implemented through the two-stage feature selection process explained in section 3.1. The first stage utilized a hybrid approach combining Fast Correlation-Based Filter (FCBF) with Sequential Forward Selection (SFS) to identify variables predictive of treatment assignment. The FCBF threshold was set to $\delta = 0.0001$, ensuring the elimination of redundant features while preserving variables with substantial pairwise correlations to the treatment. The subsequent SFS phase employed a CatBoost regressor configured with specific hyperparameters can be found in Appendix. A $K = 3$-fold cross-validation framework with root mean squared error (RMSE) minimization guided feature inclusion, terminating when marginal performance gains fell below $\Delta_{\text{RMSE}} < 0.00$. The second stage refined confounder detection and outcome prediction through dual partial correlation analyses. For confounder identification, variables altering the treatment-outcome relationship were retained using a threshold of $\rho_{\min} = 0.01$ on the partial correlation difference $|\rho(T, Y|X_j) - \rho(T, Y)|$. Concurrently, outcome predictors were selected under a stricter threshold of $\rho_{\min} = 0.1$ for $\rho(X_j, Y|T)$, ensuring robust associations independent of treatment effects. Both analyses applied a correlation threshold of 0.5 to eliminate multicollinear features, aligning with the synthetic data's redundancy structure. This dual-threshold approach balanced sensitivity to weak confounders with computational efficiency, critical for scalability in industrial applications.

B.2 Identification

While the methodology presented in Section 3.1 emphasizes the critical role of domain expertise in causal discovery and graph refinement for real-world applications, the synthetic nature of the experimental dataset necessitates a modified approach to maintain experimental validity. In this controlled setting, domain knowledge application is deliberately constrained to avoid inadvertently leveraging information from the known data-generating process, which would artificially inflate the methodology's performance metrics. Specifically, domain expertise is utilized solely for establishing temporal priors—enforcing the logical sequence where covariates and treatment assignment precede outcome observation—and for correcting directional inconsistencies in algorithms that do not inherently support temporal constraints. This restricted application notably excludes several components outlined in Section 3.1, including the identification of missing relationships, detection of spurious correlations, and general edge direction refinement through expert consultation. In cases where the initial causal discovery algorithms produce cyclic graphs, rather than employing the comprehensive domain knowledge-driven approach described for real-world applications, a simplified programmatic cycle-breaking procedure is implemented to maintain methodological integrity while avoiding reliance on the known underlying causal structure.

B.3 Positivity assumption violation

The experimental configuration for detecting and addressing positivity assumption violations employed a systematic approach based on the methodology described in Section 3.3. For violation detection, an epsilon threshold of $\epsilon = 0.001$ was established to identify regions of practical positivity violations, defined as areas where the conditional probability of treatment assignment falls below this threshold. The treatment space was partitioned into consecutive 2% intervals to facilitate granular analysis of violation patterns across the continuous treatment domain.

To mitigate the impact of identified positivity violations, a domain-knowledge-driven data augmentation strategy was implemented. This approach leverages a fundamental assumption from the banking domain: the monotonic relationship between debt loss and repayment probability. Specifically, if a customer repays their debt under a 60% writedown offer, it is reasonable to assume they would also repay under more favorable conditions (70%, 80%, ..., 100%) where a larger portion of the debt is assumed as a loss. Therefore, a synthetic sample of this customer can be added to the dataset where the treatment values is higher than the observed and the debt is also repaid. Conversely, if a customer defaults under a 40% writedown offer, they would likely default under less favorable conditions (30%, 20%, ..., 0%) where a smaller portion of the debt is assumed as a loss. We term this monotonicity-based augmentation approach "domain-knowledge data-propagation."

The synthetic data generation was calibrated to augment approximately 20% of the training sample size, targeting specifically the regions identified as violating the positivity assumption. The augmentation process was applied exclusively to the training dataset to maintain the integrity of the validation split for unbiased performance assessment. Sample generation was distributed randomly across the problematic regions, ensuring balanced coverage of various violation patterns.

This strategy aligns with the framework's broader philosophy outlined in Section 3.3, which acknowledges that the selection of positivity violation remediation approaches must balance practical constraints against methodological rigor. Given the inherent limitations of conventional strategies in real-world observational data—where restricting analysis to positivity-compliant regions

often results in prohibitively small samples, and limiting the adjustment set risks introducing significant confounding bias—our domain-knowledge-driven augmentation approach represents a pragmatic compromise between theoretical purity and practical utility.

B.4 Estimation

The experimental evaluation employed specific implementations for each estimation methodology. For the S-learner, a CatBoost classifier was utilized.

The AIPTW implementation comprised three distinct components: (i) a Generalized Propensity Score model utilizing aCatBoost regressor combined with kernel density estimation as detailed in Section 3.3, (ii) an outcome model employing a CatBoost classifier, and (iii) a final model consisting of a CatBoost regressor fitted on the pseudo-outcomes generated by the methodology, followed by logistic regression calibration to ensure well-calibrated probability estimates. To mitigate potential overfitting, all AIPTW component models were trained using out-of-sample predictions through cross-validation.

C Variables Definitions

$$
\begin{aligned}
X_{1i} &= \text{years since default} \\
X_{2i} &= \text{default debt amount} \\
X_{3i} &= \text{number of loans} \\
X_{4i} &= \text{external debt} \\
X_{5i} &= \text{number of cards} \\
X_{6i} &= \text{loss given default} \\
X_{7i} &= \text{number of refinances} \\
X_{8i} &= \text{customer history length} \\
X_{9i} &= \text{number of accounts} \\
X_{10i} &= \text{months since first payment}
\end{aligned}
$$

*all features are standardized before applying the formula

C.1 Treatment assignment formula

The treatment value for each individual i is generated through:

$$T_i = \text{clip}\left(100 \cdot \frac{1}{1 + e^{-\theta^\top X_i}} + \epsilon_i,\ 0,\ 100\right) \tag{3}$$

where the linear predictor $\theta^\top X_i$ is defined as:

$$
\begin{aligned}
\theta^\top X_i = {} & 0.5X_{1i} + 0.4\log(1 + X_{2i}) + 0.3X_{3i} + 0.3\log(1 + X_{4i}) + \\
& 0.2X_{5i} + 0.3X_{6i} + 0.2X_{7i} + 0.1X_{1i}\log(1 + X_{2i}) + 0.1X_{3i}^2
\end{aligned} \tag{4}
$$

and ϵ_i follows a truncated normal distribution:

$$\epsilon_i \sim \mathcal{TN}(0, \sigma^2 = 25, a = 0, b = 100) \tag{5}$$

The description of each variable X_{ji} can be found in the Appendix's variables definition.

C.2 Outcome generation formula

The outcome generation process comprises two distinct stages: an initial probability computation followed by a structured sampling procedure that incorporates domain-specific monotonicity constraints. For each individual i, the base probability of a positive outcome given treatment t is initially modeled as:

$$P(Y_i = 1|T = t, X_i) = \begin{cases} 0 & \text{if } t = 0 \\ 1 & \text{if } t = 100 \\ \left(\frac{t}{100}\right)^{\exp(\eta_i)} & \text{otherwise} \end{cases} \tag{6}$$

where the individual-specific coefficient η_i is computed through:

$$
\begin{aligned}
\eta_i = {} & 0.6X_{1i} + 0.5\log(1 + X_{2i}) + 0.5X_{3i} + 0.4\log(1 + X_{4i}) + 0.3X_{5i} - \\
& 0.4X_{8i} - 0.3X_{9i} - 0.2X_{10i} + 0.1X_{1i}\log(1 + X_{2i}) + 0.1X_{3i}^2
\end{aligned} \tag{7}
$$

where the description of each variable X_{ji} can be found in the Appendix's variables definition. These base probabilities undergo transformation through a binomial sampling process. The procedure first applies probability bounding:

$$\tilde{p}_i = \text{clip}(P(Y_i = 1|T = t, X_i), 0, 1) \tag{8}$$

followed by binomial sampling to determine the realized outcome:

$$Y_i \sim \text{Bernoulli}(\tilde{p}_i) \tag{9}$$

The final conditional average dose-response curves incorporate domain-specific monotonicity assumptions through the following formulation, where t_{obs} represents the observed treatment level and s the intervention :

$$P(Y_i = 1|T = s, X_i) = \begin{cases} 0 & \text{if } Y_i = 0 \text{ and } s \le t_{obs} \\ 1 & \text{if } Y_i = 1 \text{ and } s \ge t_{obs} \\ \tilde{p}_i & \text{otherwise} \end{cases} \tag{10}$$

D Figures

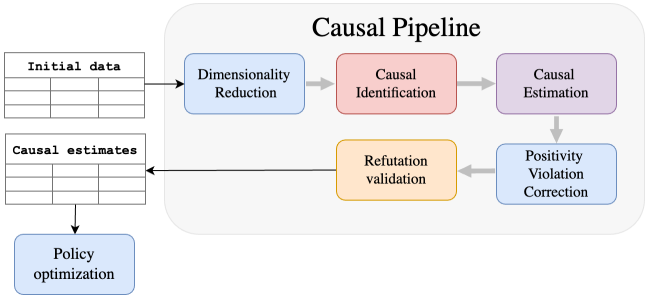

Figure 4: Diagram of the proposed causal pipeline.

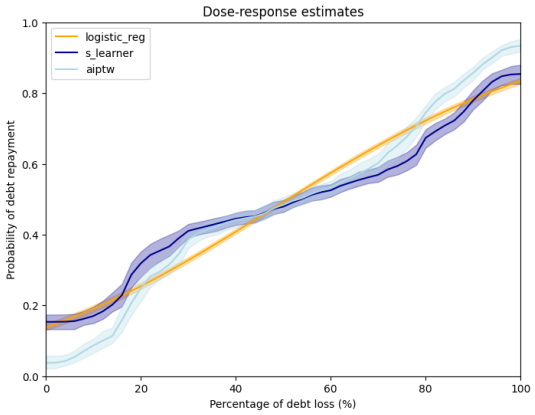

Figure 5: Estimates dose-response curves.

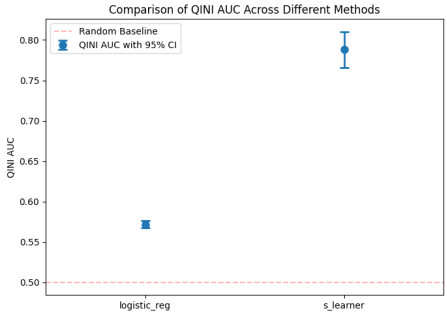

Figure 8: AUC Cumulative gain curves results

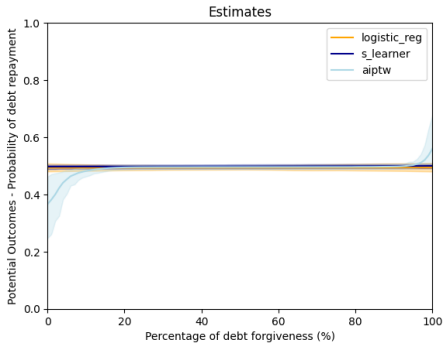

Figure 6: Placebo Treatment Replacement results

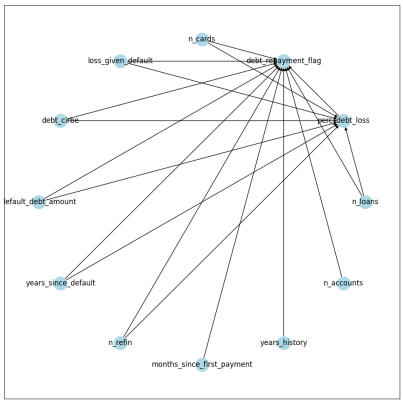

Figure 9: Estimated causal DAG.

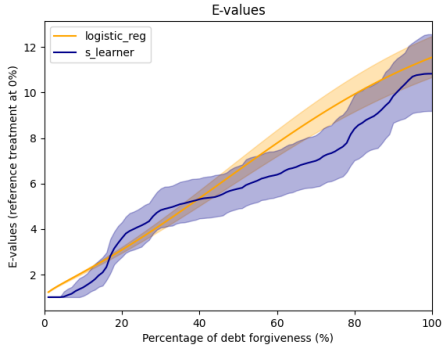

Figure 7: E-values sensitivity analysis results

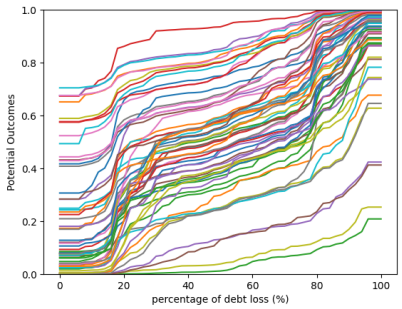

Figure 10: Individual dose-response estimated curves

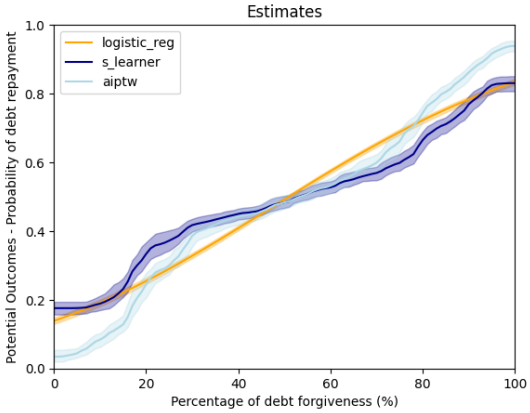

Figure 11: Random Common Cause test results.

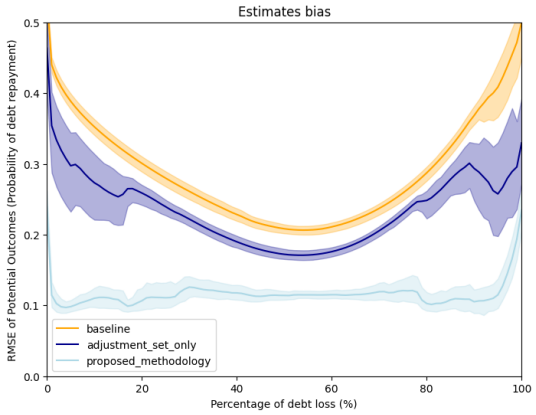

Figure 12: Estimates bias from adjustment set, baseline pipeline and proposed methodology across debt loss percentages

Received 15 July 2025