

Beyond Correlation: Deploying Causal AI to Attribute E-Commerce Sales Loss at Scale

Deep Nayak*
Amazon
Bangalore, India
deepnyk@amazon.com

Harsh Goel*
Amazon
Bangalore, India
harsgoe@amazon.com

Mohit Kumar*
Amazon
Bangalore, India
mohitkrg@amazon.com

Sivaramakrishnan Kaveri
Amazon
Bangalore, India
kavers@amazon.com

Abstract

E-commerce stores struggle to understand the root causes of declining seller revenue, often relying on correlation-based analytics that leads to ineffective or even harmful interventions. This paper introduces the Revenue Loss Monitor (RLM), a system that uses Double Machine Learning (DML) to move from simple correlation to causal attribution of revenue loss. RLM successfully isolates the causal impact of seller-controllable factors, such as pricing, advertising, and inventory - from confounding market trends. Validated across thousands of seller accounts, RLM has demonstrably improved the accuracy of attributions. We discuss the end-to-end methodology, from our hybrid product and seller level modelling approach to the significant real-world challenges of data quality and model validation in a production environment. Online validation with domain experts shows that DML-based attributions are 19.8% more accurate than traditional correlation-based methods, with the system showing 24.9% month-over-month growth in active users. Furthermore, A/B experiments demonstrate that sellers acting on RLM-identified causal factors achieve up to 29% greater revenue recovery compared to a control group, highlighting RLM's effectiveness as a scalable and trustworthy causal business intelligence solution.

CCS Concepts

• Computing methodologies → Causal reasoning and diagnostics.

Keywords

Causal Inference; Double Machine Learning; E-Commerce; Attribution Modeling

ACM Reference Format:

Deep Nayak, Mohit Kumar, Harsh Goel, and Sivaramakrishnan Kaveri. 2026. Beyond Correlation: Deploying Causal AI to Attribute E-Commerce Sales Loss at Scale. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26)*, November 07–11, 2026, Rome, Italy.

* Authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

CIKM '26, Rome, Italy

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2539-5/2026/11

<https://doi.org/10.1145/3799682.3840136>

Rome, Italy. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3799682.3840136>

1 Introduction

A fundamental challenge in evidence based decision making is change attribution: determining which factors truly drive observed outcomes. From medical treatment effectiveness to marketing campaign impact, the ability to distinguish genuine causal relationships from spurious correlations determines whether interventions succeed or fail. Traditional approaches to attribution rely heavily on correlation analysis, identifying statistical associations between actions and outcomes. While correlation-based methods work reasonably well in controlled environments with limited confounding variables, they fundamentally fail in complex, real-world settings where multiple factors interact and influence both decisions and outcomes simultaneously.

The attribution challenge becomes particularly acute in dynamic business environments where decisions are made reactively based on anticipated changes in market conditions. Consider a scenario where a business increases advertising spend while simultaneously experiencing declining sales. A correlation-based attribution system would incorrectly conclude that increased advertising caused the sales decline, leading to the counterproductive recommendation to reduce advertising. In reality, the business likely increased advertising in response to early signals of an anticipated decline caused by external factors such as new competitor entry, seasonal shifts, or supply chain disruptions. This confounding between reactive decision-making and external drivers creates a systematic bias in correlation-based attribution that can lead to harmful interventions.

The e-commerce domain exemplifies these attribution challenges at scale. Online marketplaces operate in highly dynamic environments where sellers continuously adjust multiple levers - pricing, ad spend, inventory levels, promotional activities, and catalog breadth - often in response to market conditions that also directly influence sales outcomes. When sellers experience revenue declines, account managers managing these sellers must rapidly diagnose root causes and recommend interventions. However, the simultaneous nature of these adjustments creates a complex web of interactions where simple correlation analysis becomes not just inadequate but actively misleading. Existing approaches to business attribution fall into several categories, each with significant limitations: 1) **Correlation-based methods**, widely deployed in industry

analytics tools usually rely on feature importance measures [4] to identify statistical associations but cannot distinguish causation from correlation when confounding variables are present. These methods systematically misattribute effects when seller actions are reactive to market conditions. 2) **Experimental approaches**, such as A/B testing and randomized controlled trials [28], are the gold standard for establishing causation but are often impractical or risky in e-commerce, as randomizing seller experiences can erode customer trust and disrupt the marketplace. 3) **Traditional causal inference methods** require strong identifying assumptions (such as natural experiments or instrumental variables) [1] that are rarely satisfied in business settings, limiting their practical applicability.

Recent advances in causal inference, particularly Double Machine Learning (DML) [6], offer a promising solution. DML combines machine learning with the rigor of causal inference, enabling robust estimation of causal effects in high-dimensional settings with complex confounding relationships while being less sensitive to model misspecification than traditional approaches. However, despite its theoretical promise, there has been limited exploration of DML's application to real-world business attribution problems, particularly at the scale and complexity required for the e-commerce domain. The challenge involves not only implementing the statistical framework but also addressing practical concerns around data quality, computational scalability, model interpretability, and integration with existing business processes. Furthermore, the translation from causal estimates to actionable business recommendations requires careful consideration of statistical significance, business constraints, and implementation feasibility.

To address these challenges, we present **Revenue Loss Monitor (RLM)**, a production-scale system that applies DML to identify causal drivers of seller revenue decline in e-commerce, providing actionable insights in a complex, multi-dimensional operational setting. Our key contributions are: 1) A novel hybrid DML framework for revenue decline attribution, combining product and seller-level models to provide a comprehensive, multi-granularity diagnosis (item and portfolio). 2) The first large-scale DML system for this task in a production environment, processing data for millions of products and solving engineering challenges in scalability and real-time performance demonstrating that causal methods can be operationalized at industry scale. 3) Three complementary validation strategies: Achieving 19.8% improvement in attribution accuracy against correlation and causal baselines via expert evaluation with AMs, demonstrating that sellers acting on SHM-identified factors achieve on an average 29% higher revenue in the India marketplace and demonstrating robustness to key assumptions like unconfoundedness through synthetic datasets. The main customers of the proposed tool are experts (account managers) who verify all insights before sharing it with sellers.

2 Related Work

Our work is situated at the intersection of causal inference and large-scale business analytics, building on recent methodological advances to solve a long-standing challenge in e-commerce. The shift from correlation-based analysis to causal inference is critical for obtaining actionable insights in business and marketing, as outlined by Varian et al. [25]. A key methodological advance

enabling this shift is DML [6], which is particularly well-suited for e-commerce environments where outcomes are influenced by numerous factors. Recent extensions to the DML framework have further broadened its applicability, for instance, by enabling the estimation of effects from multiple, mixed-type treatments and their interactions [26], a scenario common in our domain where sellers adjust many levers simultaneously. Furthermore, DML has proven effective for intricate causal structures like mediation analysis with continuous treatments [7, 29], underscoring its power for dissecting complex causal pathways.

The attribution problem has been approached from several angles, each with distinct goals and limitations. A significant body of work is based on variations of Shapley values [4, 5, 11, 13, 23], providing ways to allocate credit. However, these frameworks are typically focused on defining attribution and often remain agnostic about the statistical estimation challenges posed by confounding, which is the primary obstacle in our setting. Similarly, other widely-used attribution methods borrowed from explainable AI, such as saliency maps [16, 18, 22, 24], rationales [2, 17], and attention mechanisms [10, 19] are not designed for causal estimation and have been shown to incorrectly identify feature importance, especially when confounding is present [31].

In e-commerce, causal methods have been successfully applied to targeted problems. For example, researchers have used machine learning to attribute the sales impact of specific slate recommendations [27] and applied causal models to measure the effect of isolated marketing campaigns [15]. While valuable, these studies address narrowly-scoped questions and do not provide the holistic, multi-factor attribution required to diagnose the root causes of an overall revenue decline.

More directly related to our work are recent methods for general causal attribution, though they differ in key assumptions and applicability. Quintas et al. [20] propose a multiply-robust estimator to decompose an outcome's change by modeling the problem as a causal mediation path. This framework is well-suited for decomposing changes between distinct populations (e.g., before and after a policy change) but is not directly applicable to our setting, which is primarily characterized by observed confounding from sellers' reactive decisions rather than a single intervention. Similarly, Sharma et al. [23] introduce a method for attributing a single observed change using a system's computational graph. While innovative, this approach is susceptible to confounding bias as it lacks a formal orthogonalization step, a core feature of DML, to disentangle the effects of correlated inputs. Furthermore, its focus on explaining a specific historical event, rather than estimating an average causal effect, makes it less suited for informing general business strategy.

The existing literature, while rich, leaves a critical gap that our work addresses. There is a lack of research on operationalizing a scalable, general-purpose causal attribution system in a dynamic, high-stakes production environment like e-commerce. While general surveys highlight the broad potential of machine learning in this domain [30], they do not focus on the specific challenges of revenue change attribution under confounding. Our contribution is therefore not just the application of DML for attribution, but the design, implementation, and validation of a production-scale system that provides robust, actionable insights for a complex and economically significant business problem.

3 Causal Attribution Methodology

Revenue change attribution in e-commerce involves identifying seller-controllable factors causally responsible for observed revenue loss over a given period. These span product-level actions (price changes, ad spend) and portfolio-level decisions (catalog expansion, fulfillment improvements). The key challenge is isolating causal effects from confounders such as seasonality, competition, and market trends. The system must reliably identify top k seller controllable factors responsible for revenue loss, accounting for the fact that sellers often make reactive decisions in response to anticipated market conditions rather than independent strategic choices and hence distinguishing true causal factors from spurious correlations to enable actionable recommendations. To address this, we develop RLM which uses DML to handle nonlinear relationships and decouple prediction from causal inference. The methodology has 3 components: 1) a core DML model for product-level attribution, 2) a hybrid framework for seller level attribution, 3) a significance filtering stage to ensure that the recommendations are reliable.

3.1 Assumptions

The causal effects estimated by RLM rely on three standard assumptions from observational causal inference. **Unconfoundedness** assumes that, after controlling for observed factors, treatment decisions are independent of potential outcomes. This is the most vulnerable assumption; we mitigate it by collaborating with domain experts to include all plausible confounders and validate sensitivity via a synthetic dataset (Section 4.5). **Overlap** requires sufficient variation in treatments across similar contexts, which we enforce through large-scale data and significance filtering (Section 3.4) that removes extreme or underpowered treatment changes. **Consistency** assumes treatments are well-defined and applied uniformly; this is the most robust assumption, as treatments correspond to unambiguous metric changes (e.g., listed price delta). Together, these assumptions frame our results as the best possible causal estimates given the available data.

3.2 Product-Level Causal Attribution

This section explains how we apply DML to attribute revenue loss to various seller-controllable factors. Traditional machine learning models often fail in this setting because seller decisions (e.g., pricing, advertising) are not made randomly; they are a response to market conditions that also affect sales, which introduces confounding bias. DML is specifically designed to address this challenge through a two-stage procedure that effectively de-biases the causal estimates of treatment effects. A complete workflow is provided in Fig 1.

3.2.1 Notation. Let's define the key variables at the level of an individual offer i (i.e., a specific product from a seller):

- Y_i : The outcome variable, representing the change in revenue this week compared to past 4 week avg for offer i .
- T_{ik} : The k^{th} treatment variable for offer i , such as the percentage change in price or the absolute change in ad spend. Note that each treatment in our case is the change in metric's value this week compared to past 4 week avg.
- X_i : A vector of control variables (confounders) for offer i that may influence both the treatment and the outcome, such as product category, market trends, or seasonality.

- $g(X_i) = E[Y_i|X_i]$: The conditional expectation of the outcome given the confounders.
- $m_k(X_i) = E[T_{ik}|X_i]$: The conditional expectation of the k^{th} treatment given the confounders.

3.2.2 DML Stages. DML follows a two-stage process:

- (1) **Nuisance Estimation and Residualization:** The goal of this stage is to remove the predictive effect of the confounders from both the outcome and the treatments. Using ML models we estimate the nuisance functions $\hat{g}(X_i)$ and $\hat{m}_k(X_i)$. We then compute the residuals:

- **Outcome Residual:** $\tilde{Y}_i = Y_i - \hat{g}(X_i)$
- **Treatment Residuals:** $\tilde{T}_{ik} = T_{ik} - \hat{m}_k(X_i)$

These residuals represent the variation in the outcome and treatments that is not explained by the confounders.

- (2) **Causal Effect Estimation:** This stage estimates the causal effect by training a model on the outcome residuals against the treatment residuals:

$$\tilde{Y}_i = f(\tilde{T}_{i1}, \tilde{T}_{i2}, \dots, \tilde{T}_{ip}) + \zeta_i$$

where ζ_i is the error term and p is the number of treatments. The estimated model f represents the estimated causal effects of each treatment on revenue loss, adjusted for the confounders.

3.2.3 Counterfactual Inference and Attribution. Once the DML model is trained, we attribute revenue loss using counterfactual inference. To isolate the impact of each metric k , we estimate what the revenue loss would have been had that metric's change been zero ($T_{ik}^{cf} = 0$), while all other factors remained same. The predicted counterfactual loss, $\hat{Y}_{ik}^{cf}$, is calculated using the trained model components:

$$\hat{Y}_{ik}^{cf} = \hat{g}(X_i) + \hat{f}(\tilde{T}_{i1}^{obs}, \dots, \tilde{T}_{ik}^{cf}, \dots, \tilde{T}_{ip}^{obs})$$

The attribution process is then straightforward:

- (1) **Calculate Contribution (C_{ik}):** For each metric k , its contribution to the loss is the difference between the observed loss and the counterfactual loss.

$$C_{ik} = Y_i - \hat{Y}_{ik}^{cf}$$

- (2) **Normalize and Rank:** All contributions are normalized to sum to the total observed loss, and metrics are ranked by these values to identify the primary drivers

Example: Consider a product with \$5,000 GMS loss and observed treatment changes: ad spend -\$200, price +5%, inventory -10 units. For each treatment k , we compute $\hat{Y}_{ik}^{cf}$ by setting $T_{ik}^{cf} = 0$ while holding others at observed values. Suppose the model predicts counterfactual losses of \$3,800 (ad spend), \$2,200 (price) and \$3,500 (inventory). Raw attributions are thus \$1,200, \$2,800 and \$1,500 respectively. After normalization to \$5,000, the final attribution is: Price (50.9%), Inventory (27.3%), Ad Spend (21.8%), guiding AMs to prioritize pricing adjustments. Fig 1 provides a complete workflow.

3.3 Seller-Level Attribution

While product-level attributions enable granular interventions, they are insufficient on their own, as account managers also need a portfolio-level view to identify key drivers of a seller's revenue decline before drilling down further. Simple aggregation to the

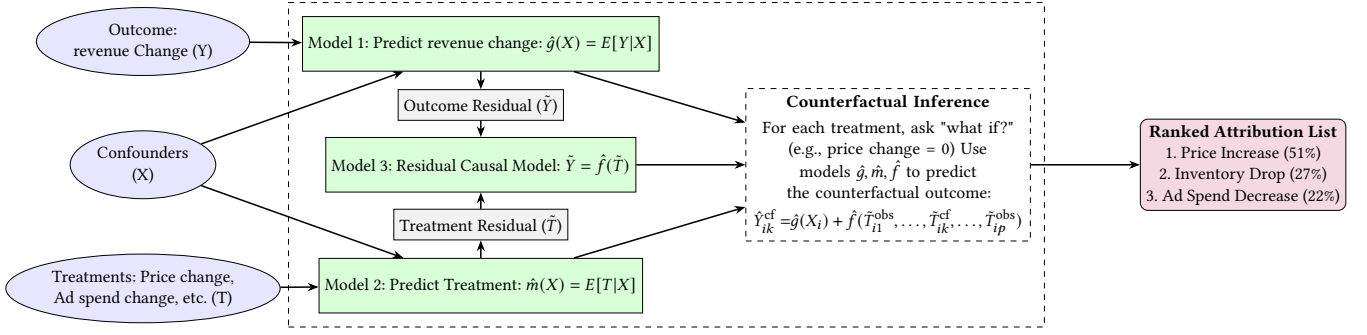


Figure 1: Workflow of RLM for product-level attribution using DML framework. Model 1 and 2 remove confounding effects, Model 3 estimates the true causal relationship from the residuals, and then the trained models are used for counterfactual inference, finally giving the ranked attribution.

seller level (e.g., total ad spend changes) obscures heterogeneous product effects, where gains and losses can cancel out and hide meaningful signals. For example, a \$100 reduction in ad spend on a high-performing product may cause significant losses, while \$100 increase on low-impact products may offset this in aggregate, masking true effects. Conversely, purely summing product-level attributions misses seller-level dynamics such as changes in catalog or fulfillment strategy. To address this, we propose a **hybrid attribution framework** combining product-level attributions with a seller-level causal model, providing a unified view of both localized and portfolio-wide drivers. The framework operates in three stages:

3.3.1 Stage 1: Aggregated Product-Level Attribution. We construct seller-level signals by aggregating product-level attributions. For seller s with products I_s , we retain only those products that have seen a decline in revenue and where the sign of $\hat{Y}_i$ (predicted revenue loss) aligns with Y_i . This ensures reliability by excluding products where attribution may be noisy or inconsistent. $I_s^* = \{i \in I_s : Y_i < 0, \text{sign}(Y_i) = \text{sign}(\hat{Y}_i)\}$. For each such product, we focus on loss-driving factors and normalize contributions to match observed loss: $\tilde{C}_{ik} = \frac{C_{ik}}{\sum_{k: C_{ik} < 0} C_{ik}} \cdot Y_i$, for $C_{ik} < 0$. Seller-level attribution for treatment k is obtained by aggregating across products: $\sum_{i \in I_s^*} \tilde{C}_{ik}$.

3.3.2 Stage 2: Dedicated Seller-Level Causal Model. This stage focuses on factors that are inherently defined at the seller portfolio level and cannot be decomposed to an individual product. These include systemic factors related to **selection mix and catalog-wide fulfillment strategies**. We build a separate DML model at the seller level, as per Section 3.2, to estimate the causal impact of these portfolio-level treatments on the seller's total revenue decline.

3.3.3 Stage 3: Hybrid Attribution. In the final stage, we combine insights from stage 1 and stage 2, prioritizing outputs from seller-level model as primary drivers. This aligns with how account managers diagnose performance - starting with a holistic view of the seller's business before drilling down to individual products and reflects the intuition that systemic, portfolio-wide factors often have a more fundamental impact on seller health than isolated product-level issues. Aggregated product-level attributions serve a complementary role, surfacing key drivers when the seller-level model finds no significant factors and ensuring that critical product-level issues

are not overlooked. This hybrid design enables a comprehensive and actionable attribution flow, guiding account managers from portfolio-level diagnosis to targeted product-level interventions.

3.4 Ensuring Actionable Recommendations

While DML provides causal effect estimates for each treatment T_{ik} , not all estimates are statistically reliable. For example, if a seller changes their promotional offer percentage by 2% but this change is based on only 50 transactions, the estimated causal effect may be unreliable due to insufficient sample size. A raw causal estimate may be statistically significant but too small to be business-relevant, or it may be based on too few data points to be reliable. To ensure our recommendations are both trustworthy and actionable, we apply a two-stage filtering mechanism:

1. Business Thresholding: We filter out changes below domain-expert defined minimum meaningful magnitudes. For e.g., price changes $< 1\%$ or advertising spend changes $< \$50$ are considered non-actionable and removed from attribution consideration.

2. Statistical Significance Testing: For treatments representing proportions (e.g., deal percentages, return rates), we calculate the required sample size to detect the observed treatment change reliably. We use the formula: $n_{ik}^* = \frac{z_{\alpha/2}^2 \cdot p_{ik}(1-p_{ik})}{T_{ik}^2}$, where p_{ik} is the baseline proportion and α is the significance level. We set $T_{ik}^{final} = 0$ if the actual sample size $n_{ik} < n_{ik}^*$, effectively filtering out statistically underpowered attributions and preventing recommendations based on statistical noise.

Both filters are applied during training and inference, ensuring the model learns from relevant data and attributes only significant changes. Additionally, account managers verify every attribution before sharing it with sellers, providing a human-in-the-loop safeguard against incorrect insights.

4 Experiments

Having established our causal attribution methodology, we now validate its effectiveness through experiments answering the following research questions. **RQ1:** Does RLM produce attributions that align with business logic and domain experts' expectations? **RQ2:** How does our causal attribution approach compare to correlation

based methods and other causal baselines in terms of recommendation quality? **RQ3:** Which ML architectures work best within our causal attribution framework? **RQ4:** Do our filtering mechanisms (business thresholds and statistical significance) improve attribution reliability? **RQ5:** How does the seller-level hybrid attribution framework compare to a dedicated seller-level causal attribution model trained using portfolio-level treatments?

4.1 Experimental Setup and Data

Our experiments use twelve months of weekly e-commerce marketplace data spanning millions of observations across diverse product categories and seller types. We adopt an out-of-time split, reserving the final four weeks for testing and using an 80–20 split of the remaining data for training and validation. The treatment set includes 14 seller-controllable factors such as pricing, advertising spend, promotions and inventory decisions for product-level attributions, and 4 portfolio-level factors capturing catalog-wide selection mix and fulfilment strategy for seller-level attributions; while 40+ confounders capture temporal trends, category dynamics, seasonality and external market conditions. Both treatment and outcome variables are defined as the difference between the current week’s values and their trailing 4-week averages. Experiments are conducted on cloud instances with 4 NVIDIA A10G GPUs and 96 vCPUs, using CUDA acceleration where possible, and models are tuned via Optuna with 30 trials per model. Our system uses only aggregated metrics with no propensity matching that could reveal competitive information. **Evaluation Metrics:** For the attribution task, where the model identifies the top k factors of revenue loss, we evaluate quality using feedback from domain experts who have been manually monitoring and sharing recommendations with sellers for years. We use **Precision** and **Recall**, calculated based on whether Account Managers confirmed the model’s attributions as aligning with their ground-truth business knowledge.

4.2 Baseline Methods & Architecture Evaluation

To validate our causal approach (RQ2 and RQ3), we implement both correlation-based and causal baselines, and evaluate multiple machine learning architectures across these frameworks.

Correlation-Based Attribution: Our primary correlation baseline uses SHAP (SHapley Additive exPlanations) attribution with various ML models. SHAP provides model-agnostic explanations by computing the marginal contribution of each feature to the prediction, with SHAP values summing to the difference between the prediction and expected model output. We implement two SHAP configurations: **(1) Treatment-Only Models:** Train regression models using only the 14 seller-controllable treatments as input features. **(2) Combined Models:** Train models using both treatments and 41 confounders as input features. For the combined configuration, we compute SHAP values for all features but filter out confounders during attribution ranking, promoting the next highest-ranked treatment factor to fill the top- k recommendations.

Causal Baselines: While our proposed approach uses DML, in order to compare against causal baselines, we use **Causal Forest**, an extension of random forests that directly estimates heterogeneous treatment effects by adaptively partitioning the feature space to maximize treatment effect heterogeneity. Causal Forests are trained

using both treatments and confounders to obtain counterfactual estimates for each controllable factor. We also include the **S-Learner**, a meta-learner that fits a single predictive model on outcomes as a function of both treatments and confounders, with counterfactuals generated by switching treatment values. This provides a simple yet widely used causal baseline.

Architecture Evaluation: We evaluate six architectures across correlation and causal frameworks: LightGBM [14], Ridge regression, ElasticNet [32], Extra Trees [8], Random Forest [3] and Multi-Layer Perceptron. Beyond these, we also test a recent transformer-based tabular SOTA, the Feature Tokenizer Transformer (FTT) [9], to evaluate whether attention-based representations can support attribution. FTT performs poorly on both axes. Predictive accuracy degrades sharply, with RMSE up by $\sim 78\%$, R^2 down by $\sim 95\%$ and directional accuracy down by 12 pp relative to the LightGBM combined baseline, and the resulting attributions are correspondingly weak. Training also requires 500+ epochs and substantially higher compute than the gradient-boosted and linear baselines. Attention-based attribution therefore offers no advantage over simpler regularised models in this setting. Our users (account managers) have easy access to LLMs and naturally attempt attribution via prompting. Hence we also evaluate several LLM configurations (Claude Sonnet 3.5/3.7/4 with both standard and Chain-of-Thought prompting, with prompts iteratively tuned via an LLM-as-a-judge loop and the same business guardrails applied to LLM outputs as to DML). Chain-of-Thought reasoning yields the largest gains, with Sonnet 3.7 + CoT being the best LLM configuration, but it still trails DML LightGBM by roughly 25 percentage points in precision and 7 in recall, indicating that explicit reasoning helps but a substantial gap to statistical causal methods remains. We also report predictive (revenue regression) performance for completeness, while emphasising that revenue prediction is not our primary objective: our goal is attribution quality, and a model that predicts revenue well can still produce poor attributions if it leans on confounders. We assess predictions using Root Mean Square Error (RMSE), R^2 and directional accuracy (the percentage of cases where the predicted sign of the revenue change matches the observed sign). DML requires strongly regularised models, with Ridge and LightGBM performing best, while Extra Trees and Random Forest exhibit RMSE inflation of 20–23% relative to the baseline due to the residualization step. Correlation models with confounders included achieve the best raw revenue-prediction metrics, but their attributions are dominated by confounders and so lack business relevance even after filtering them out at ranking time. Treatment-only correlation variants degrade prediction sharply.

4.3 Domain Expert Validation

To address RQ1, we conduct validation with account managers who have deep knowledge of seller dynamics and market conditions. This is the most critical evaluation component, as it directly tests whether the attributions are actionable for real business decisions. The validation follows a structured protocol in which account managers assess model attributions for sellers they have managed over time, across different treatment types. Evaluations consider both the magnitude and direction of seller-controllable factors, informed by historical performance data, treatment changes, and market

context. Table 1 reports attribution accuracy for product-level attributions, comparing our approach with correlation-based and causal baselines, as well as a heuristic that ranks treatments by observed change magnitude. Results are reported across multiple ML architectures and are statistically significant under a two-proportion z-test at the 5% significance level.

Table 1: Account Manager Workshop Validation: Model Attribution Accuracy Across Different Approaches ¹

Method	Precision ↑	Recall ↑	F1-Score ↑
DML LightGBM	-	-	-
Heuristics	-48.50%	-9.70%	-34.66%
Correlation Ridge	+0.00%	-35.80%	-23.40%
Correlation ElasticNet	+0.09%	-31.28%	-19.80%
Correlation LightGBM	+0.26%	-51.20%	-37.76%
Correlation Extra Trees	+0.83%	-52.40%	-38.96%
Correlation RF	-0.66%	-60.80%	-48.83%
Correlation MLP	-62.50%	-83.00%	-83.31%
DML Extra Trees	+0.00%	-55.20%	-42.16%
DML RF	+0.39%	-40.00%	-26.94%
DML Ridge	+0.44%	-37.10%	-24.42%
DML ElasticNet	+0.64%	-39.90%	-26.79%
DML MLP	-0.10%	-40.00%	-27.05%
Causal Forest	-36.60%	-16.90%	-27.88%
S-Learner (LightGBM)	-4.50%	-18.40%	-12.39%
Claude Sonnet 3.7 CoT	-25.20%	-6.70%	-16.76%

The validation answers RQ2 and reveals that correlation approaches achieve reasonable precision in some cases but suffer from poor recall, indicating they identify some genuine attribution patterns but miss many valid causal relationships. We also observe that correlation models tend to prioritize confounders in many cases which de-prioritizes important treatments leading to low recall. Answering RQ3, DML LightGBM achieves the best balance between precision and recall, demonstrating superior ability to capture both obvious and subtle factors that drive revenue declines. To answer RQ4, we implement guardrails and thresholds mentioned in Section 3.4. Statistical significance testing based guardrails provide improvements in precision (+5.8%) while reducing recall (-20.3%), reflecting the filtering of statistically underpowered attributions. This aligns with business preferences for high-confidence recommendations over comprehensive but potentially unreliable coverage. Business thresholds alone provide more modest improvements (+2.6% precision, -16.9% recall), proving that statistical power analysis adds value beyond domain expertise. The combination of both filtering mechanisms ensures that recommendations meet both business relevance criteria and statistical reliability standards.

4.4 Seller-Level Attribution Validation

As outlined in Section 3.3, seller-level attribution combines outputs from product-level causal models with outputs from dedicated seller-level causal models. Consistent with the product-level setup,

we use LightGBM as the base learner for DML and apply business thresholds along with statistical significance testing to filter out treatment changes that lack sufficient statistical power. To address RQ5, we follow the validation approach in Section 4.3, where model outputs are reviewed by account managers. We benchmark our hybrid framework against a baseline seller-level DML model that operates on aggregated portfolio signals (e.g., total ad spend changes or average price variation), which collapses product-level heterogeneity into a single representation. This provides a natural comparison to evaluate the benefit of preserving product-level granularity while attributing seller-level outcomes. Our hybrid framework significantly outperforms the baseline, achieving higher precision (+44.61%) and recall (+57.83%), with differences statistically significant under a two-proportion z-test at the 5% level, demonstrating superior accuracy and reliability.

4.5 Assumption Validation

Our approach is most sensitive to the unconfoundedness assumption as explained in Section 3.1. To validate robustness to hidden confounders, we conduct a structured synthetic experiment using a parameterized Structural Causal Model (SCM) with over 24 confounders and 12 treatments, where each hidden confounder has a significant effect on all treatments and the outcome. Ground-truth attributions are obtained via do-calculus interventions ($do(T_j = 0)$) with fixed noise terms. We vary the number of hidden confounders from 1 to 5 and train DML models on 50K samples, repeating the full training/inference loop with 5 random seeds for statistical significance. Since synthetic data provides true ground-truth factor rankings, we focus on ranking metrics: Mean Reciprocal Rank (MRR) and Normalised Discounted Cumulative Gain (NDCG@3). We observe that directional accuracy degrades only modestly from 0.92 to 0.88 as hidden confounders increase from 1 to 5, while NDCG@3 holds steady at 0.99 and MRR remains in the 0.88–0.91 range, indicating that ranking quality is robust to the presence of hidden confounders even when point estimates are mildly affected.

Empirical Overlap Diagnostics. To formally assess overlap in our 14-treatment continuous setting, we estimate the generalized propensity score (GPS) for each treatment conditional on all confounders using cross-validated gradient boosting, following the high-dimensional frameworks of Rassen et al. [21] and Ju et al. [12]. Traditional inverse-GPS weighting proves impractical: the mean effective sample size (ESS) ratio across treatments is only 9.6% (range 1.1–95.2%), with weight coefficients of variation up to 9.7, confirming that propensity-based reweighting is unstable with multiple correlated continuous treatments. In contrast, DML’s first-stage residualization reduces the mean absolute confounder–treatment correlation by 63.3% (from 0.020 to 0.007), with 99.1% of all confounder–treatment pairs achieving $|\rho| < 0.05$ post-residualization. A trimming sensitivity analysis, progressively removing observations in the 1st–10th GPS percentile tails jointly across all 14 treatments, shows that aggressive trimming discards 12–69% of observations while DML coefficient estimates remain stable in sign and significance for the retained samples. These diagnostics confirm that (a) overlap is empirically adequate for the vast majority of observations, (b) GPS-based reweighting is infeasible in this multi-treatment setting, and (c) DML residualization implicitly

¹Arrow notation: ↑ indicates higher values are better. Precision, Recall and F1-Score show absolute differences in percentage points compared to DML LightGBM baseline.

achieves the covariate balance that propensity methods target, with greater stability in high dimensions.

5 Results from Production Setting

RLM has been operating in a production setting for over a year with 24.9% month-over-month growth in active users, generating weekly insights that provide sufficient timeliness for account managers to recommend corrective actions to sellers. The system operates as a batch pipeline with an AWS-based inference workflow that automates end-to-end causal attribution. Raw data is ingested from the data lake through automated validation and alerting checks that ensure completeness across cross-team dependencies; the pipeline pauses when critical inputs are unavailable and automatically resumes once they are restored. The pipeline then computes treatment and confounder features as deviations from the trailing 4-week baseline, applying business thresholds and significance filters to retain meaningful changes. Trained seller- and product-level DML models perform parallelized causal inference for all declining entities at scale. Weekly attribution outputs are cached and served through an interactive dashboard for efficient access by account managers. The dashboard captures account manager feedback, including attribution validation, top-k driver selection, and missed-factor suggestions, enabling continuous monitoring and iterative model improvement through retraining.

5.1 Challenges and Lessons Learned

Causal inference system in a production-scale environment surfaced several key challenges:

Data quality is critical: Causal models are sensitive to data errors. Early issues like missing treatment values and edge-case mishandling led us to build a strong validation layer ensuring clean data.

Building trust with stakeholders: Moving from a descriptive to a causal framework required educating account managers. Validation workshops showcased the model’s value and built end-user trust, while the significance filtering layer emphasized prioritizing reliable recommendations over providing answers for every case.

Treatment identification: Choosing appropriate treatments was challenging. Missing or weak treatments risked misleading attributions, while too many added noise. We iteratively refined the treatment set with domain experts and account managers to balance coverage, interpretability, and statistical tractability, underscoring the value of domain knowledge alongside data-driven methods.

Feedback integration: Account manager feedback, though valuable, was often subjective. We implemented a structured dashboard loop for feedback, converting noisy opinions into labeled data to monitor performance, improve the model, and guide enhancements.

5.2 Results from A/B Testing

To evaluate RLM’s impact on revenue, we conducted a two-month A/B experiment comparing sellers receiving RLM-driven insights (test group) against a control group. Guided by business input, we focus on sellers most likely to receive account managers’ (AM) attention, assuming AMs review at least the top four sellers with the highest week-over-week revenue decline. Each week, these top four sellers per AM are randomly split into test and control groups, with RLM insights provided only to the test group. This ensures

Table 2: A/B Test Results: Incremental revenue recovery vs control. Mean $\pm$ 95% CI

Test Group	ATE - Revenue Recovery
RLM exposure (insights viewed)	7.94% $\pm$ 2.78%
RLM exposure + partial adoption (≥ 1 factor improved)	22.25% $\pm$ 3.23%
RLM exposure + full adoption (all factors improved)	29.14% $\pm$ 3.43%

evaluation is limited to cases where insights are most likely to be used. Newly appearing sellers entering this subset are assigned via stratified randomization based on revenue and active products, and remain fixed thereafter. This rolling design yields balanced cohorts, with final differences in revenue and product counts within 0.00424% across groups. We evaluate balance between groups across key metrics including trailing-4-weeks revenue, offer counts, and segment-wise revenue distribution. Using standardized mean difference for continuous variables and Cramér’s V for categorical variables (values < 0.05), we confirm strong randomization and no systematic bias. We measure revenue recovery as the percentage revenue loss recovered following adjustments to ML-identified factors over an eight-week forward window, where $t = 0$ denotes ML insight generation. The period $t = +1$ to $t = +4$ captures changes in recommended causal factors; empirically, the p90 time for implementation is approximately four weeks across most metrics. The window $t = +5$ to $t = +8$ measures resulting revenue recovery. To control for external influences such as seasonality and market trends, we estimate the average treatment effect (ATE) of RLM exposure and adoption using causal models (X-learner). As shown in Table 2, RLM exposure yields an incremental revenue recovery of 7.94% versus control. It increases to 22.25% when at least one ML-recommended factor is acted upon and to 29.14% when all recommendations are adopted.

6 Conclusion and Future Work

In this paper, we presented the Revenue Loss Monitor, a system that successfully applies Double Machine Learning to attribute e-commerce sales loss at scale. By moving from correlation to causation, our system provides account managers with more accurate, actionable, and trustworthy recommendations. We also presented a novel hybrid approach to attribute at Seller level and incorporated significance testing to identify actionable recommendations. Our framework, validated by domain experts, achieved 19.8% more accuracy than traditional correlation-based models. We detailed the practical challenges encountered while operating with the system in a production setting, emphasizing the importance of data integrity, stakeholder education, and combining statistical rigor with business logic. Future work will focus on expanding the range of treatments to include more complex, multi-step actions and exploring the use of more advanced deep learning models within the DML framework to capture even more nuanced relationships in the data. We also plan to handle hierarchical treatments and multiple time horizons like month-on-month and quarter-on-quarter attributions.

GenAI Usage Disclosure

Popular SOTA generative AI tools were used for minor rephrasing and editing of the text to improve the language and structure of the publication. The authors take full responsibility and ownership of the method proposed and its results.

References

- [1] Michael Baiocchi, Jing Cheng, and Dylan S Small. 2014. Instrumental variable methods for causal inference. *Statistics in medicine* 33, 13 (2014), 2297–2340.
- [2] Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. 2963–2977. doi:10.18653/v1/P19-1284
- [3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
- [4] Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. 2022. Causal structure-based root cause analysis of outliers. In *International conference on machine learning*. PMLR, 2357–2369.
- [5] Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. 2023. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence* 5, 6 (2023), 590–601.
- [6] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21, 1 (01 2018), C1–C68. arXiv:https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj001.pdf doi:10.1111/ectj.12097
- [7] Helmut Farbmacher, Martin Huber, Lukáš Laffèrs, Henrika Langen, and Martin Spindler. 2022. Causal mediation analysis with double machine learning. *The Econometrics Journal* 25, 2 (2022), 277–300.
- [8] Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning* 63, 1 (2006), 3–42.
- [9] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. *Advances in neural information processing systems* 34 (2021), 18932–18943.
- [10] Sarthak Jain, Sarah Wiegrefe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to Faithfully Rationalize by Construction. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:218470359
- [11] Dominik Janzing, Patrick Blöbaum, Atalanti A Mastakouri, Philipp M Faller, Lenon Minorics, and Kailash Budhathoki. 2024. Quantifying intrinsic causal contributions via structure preserving interventions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2188–2196.
- [12] Cheng Ju, Joshua Schwab, and Mark J van der Laan. 2019. On adaptive propensity score truncation in causal inference. *Statistical Methods in Medical Research* 28, 6 (2019), 1741–1760.
- [13] Alexander Jung. 2022. *Machine learning: the basics*. Springer Nature.
- [14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).
- [15] Henrika Langen and Martin Huber. 2023. How causal machine learning can leverage marketing strategies: Assessing and improving the performance of a coupon campaign. *PLOS ONE* 18, 1 (01 2023), 1–37. doi:10.1371/journal.pone.0278937
- [16] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS ONE* 10 (07 2015), e0130140. doi:10.1371/journal.pone.0130140
- [17] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. 107–117. doi:10.18653/v1/D16-1011
- [18] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [19] Pouya Pezeshkpour, Sarthak Jain, Sameer Singh, and Byron C. Wallace. 2021. Combining Feature and Instance Attribution to Detect Artifacts. *ArXiv abs/2107.00323* (2021). https://api.semanticscholar.org/CorpusID:235694304
- [20] Victor Quintas-Martinez, Mohammad Taha Bahadori, Eduardo Santiago, Jeff Mu, and David Heckerman. 2024. Multiply-robust causal change attribution. In *Proceedings of the 41st International Conference on Machine Learning (Vienna, Austria) (ICML '24)*. JMLR.org, Article 1701, 20 pages.
- [21] Jeremy A Rassen, Abhi A Shelat, Jessica Myers, Robert J Glynn, Kenneth J Rothman, and Sebastian Schneeweiss. 2012. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiology and Drug Safety* 21, S2 (2012), 69–80.
- [22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier (*KDD '16*). Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778
- [23] Amit Sharma, Hua Li, and Jian Jiao. 2022. The Counterfactual-Shapley Value: Attributing Change in System Metrics. arXiv:2208.08399 [cs.LG] https://arxiv.org/abs/2208.08399
- [24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
- [25] Hal Varian. 2016. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences* 113 (07 2016), 7310–7315. doi:10.1073/pnas.1510479113
- [26] Qingyan Xiang, Yubai Yuan, Dongyuan Song, Usman J. Wudil, Muktar H. Aliyu, C. William Wester, and Bryan E. Shepherd. 2025. Double machine learning to estimate the effects of multiple treatments and their interactions. arXiv:2505.12617 [stat.ME] https://arxiv.org/abs/2505.12617
- [27] Shenghe Xu, Yan Zhao, Sameer Kanase, Mitchell Goodman, Saad Khan, Brent Payne, and Patricia Grao. 2022. Machine learning attribution: Inferring item-level impact from slate recommendation in e-commerce. (2022). https://www.amazon.science/publications/machine-learning-attribution-inferring-item-level-impact-from-slate-recommendation-in-e-commerce
- [28] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.
- [29] Houssam Zenati, Judith Abécassis, Julie Josse, and Bertrand Thirion. 2025. Double Debiased Machine Learning for Mediation Analysis with Continuous Treatments. *arXiv preprint arXiv:2503.06156* (2025).
- [30] Xue Zhang, Fusen Guo, Tao Chen, Lei Pan, Gleb Beliaikov, and Jianzhang Wu. 2023. A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research. *Journal of Theoretical and Applied Electronic Commerce Research* 18 (12 2023), 2188–2216. doi:10.3390/jtaer18040110
- [31] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. 2022. Do feature attribution methods correctly attribute features?. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 9623–9633.
- [32] Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67, 2 (2005), 301–320.