

# ADIF: A Decision Intelligence Framework Integrating Predictive Driver Discovery, Causal Discovery, Causal Inference, Counterfactual Reasoning, and LLM-Assisted Decision Support for Retail Analytics

Greeshma Chanduri

Tagliatela College of Engineering,  
University of New Haven,  
Connecticut, United States of  
America

Sampath Kokkonda

Tagliatela College of Engineering,  
University of New Haven,  
Connecticut, United States of  
America

Moin Bhuiyan, PhD

Tagliatela College of Engineering,  
University of New Haven,  
Connecticut, United States of  
America

## ABSTRACT

Retail organizations increasingly require analytical systems that can move beyond prediction toward explanations of observed business changes and structured decision support. This paper presents ADIF, an end-to-end decision intelligence framework that connects predictive modeling, predictive-driver discovery, causal analysis, counterfactual reasoning, and large language model (LLM)-assisted communication in a single workflow. The framework combines XGBoost and SHAP for predictive modeling and feature attribution, the PC algorithm for causal-structure screening, DoWhy for treatment-effect estimation and refutation, counterfactual analysis for scenario evaluation, and Llama 3.1 through Ollama for natural-language decision support. ADIF was evaluated on the Online Retail II and Instacart datasets. On Online Retail II, Order Change and Quantity Change represented 63.42% and 36.58%, respectively, of the reported normalized feature-attribution contribution. The estimated effect of Quantity Change on Revenue Change was 0.59748, with a reported 95% confidence interval of approximately [0.5721, 0.6229]. On Instacart, the estimated effect of Orders on Reorder Rate was 0.006417, with a reported 95% confidence interval of approximately [0.00591, 0.00693]. Predictive evaluation produced a mean cross-validated  $R^2$  of 0.5883 for XGBoost on Instacart. The counterfactual stage was configured for 5%, 10%, 20%, and 50% treatment-change scenarios as model-based what-if analyses; no scenario outcome is claimed as a verified finding. Refutation experiments produced a negligible change under the random-common-cause test and a near-zero estimate under placebo treatment; these are treated as sensitivity checks rather than proof of causality. The LLM layer converts structured quantitative findings and limitations into decision-support narratives. The framework's contribution is the explicit evidence handoff between analytical stages while keeping predictive attribution, causal estimation, counterfactual scenarios, and language generation conceptually separate.

## General Terms

Decision Intelligence, Retail Analytics, Machine Learning, Causal Inference, Explainable AI

## Keywords

Decision Intelligence, Retail Analytics, XGBoost, SHAP, PC Algorithm, DoWhy, Counterfactual Reasoning, Large Language Models

## 1. INTRODUCTION

Retail analytics has traditionally emphasized descriptive

reporting and predictive modeling. Although these approaches can identify what happened and estimate what may happen next, business decision makers frequently require an additional layer of reasoning: which variables are associated with a change, whether an observed relationship is compatible with a causal interpretation, what would happen under an alternative intervention, and how the resulting evidence should be communicated.

These requirements motivate decision intelligence systems that connect multiple analytical stages rather than treating prediction, explanation, causal analysis, and decision support as isolated tasks. However, integrating these components introduces methodological challenges. Predictive feature importance is not equivalent to causal importance, causal discovery from observational data requires assumptions, and counterfactual estimates inherit the assumptions of the underlying causal model.

This paper presents an end-to-end workflow connecting five analytical stages: predictive modeling, predictive-driver discovery, causal analysis, counterfactual reasoning, and LLM-assisted decision support. The objective is not to claim novelty for XGBoost, SHAP, PC, DoWhy, or Llama individually. Instead, the methodological contribution is the explicit evidence-handoff design: predictive outputs are treated as candidate evidence, candidate drivers are subjected to a separately specified causal analysis, estimated effects are passed to scenario analysis, and the resulting quantitative evidence is passed to the LLM only as structured context.

The framework was evaluated using Online Retail II and Instacart data. The evaluation examines predictive performance, feature attribution, causal-effect estimates, refutation tests, counterfactual sensitivity, and the feasibility of evidence-grounded LLM communication.

The primary contributions are:

An end-to-end analytical workflow connecting prediction, predictive-driver discovery, causal analysis, counterfactual reasoning, and natural-language decision support.

- (1) An explicit separation between predictive feature attribution and causal-effect estimation.
- (2) A causal-analysis stage that exposes treatment/outcome definitions, causal assumptions, estimated effects, confidence intervals, and refutation results rather than treating feature importance as causality.

- (3) A counterfactual stage that translates an estimated treatment relationship into intervention-oriented scenarios while retaining the underlying assumptions.
- (4) An evidence-grounded LLM stage in which quantitative outputs are supplied as structured inputs rather than recomputed or inferred by the language model.
- (5) An empirical demonstration across two retail datasets, with limitations explicitly reported for observational causal analysis and LLM evaluation.

## 2. RELATED WORK

### 2.1 Predictive Retail Analytics

Machine learning has been extensively applied to retail forecasting, customer behavior prediction, demand estimation, and revenue analysis. XGBoost is particularly useful for structured tabular data because it combines boosting with regularization and efficient tree-based learning [1]. Random forests provide a nonparametric ensemble baseline for structured prediction [2]. However, high predictive performance does not by itself establish causal relationships between variables.

Shmueli [3] distinguishes predictive modeling from explanatory and causal modeling. This distinction is important for decision intelligence because a feature that strongly predicts an outcome may not be an appropriate intervention target.

### 2.2 Explainable Machine Learning

SHAP provides a game-theoretic approach for decomposing model predictions into feature-level contributions [4]. In ADIF, SHAP is used to identify variables associated with changes in model output. These contributions are treated as predictive evidence and are not interpreted directly as causal effects.

### 2.3 Causal Discovery and Inference

The PC algorithm is a constraint-based causal discovery method that uses conditional-independence relationships to estimate a candidate causal graph under assumptions such as causal sufficiency and faithfulness [5, 6]. Causal inference provides a complementary perspective to predictive modeling by defining a treatment, an outcome, and an estimand rather than relying only on predictive association.

The potential-outcomes framework provides a foundation for treatment-effect analysis [7], while modern causal-inference methods provide practical approaches for observational data [8]. DoWhy provides a workflow for specifying causal assumptions, identifying and estimating effects, and performing refutation tests [9].

### 2.4 Counterfactual Reasoning

Counterfactual reasoning asks how an outcome could differ under an alternative treatment or intervention. Wachter et al. [10] discuss counterfactual explanations as a mechanism for connecting machine-learning outputs with actionable changes. ADIF places scenario analysis after causal estimation so that counterfactual scenarios are explicitly tied to the treatment-outcome relationship.

### 2.5 Large Language Models for Decision Support

Large language models can transform structured analytical outputs into natural-language explanations and recommendations. The Llama 3 family provides open-weight language models suitable for locally executed inference [11]. In ADIF, the LLM does not independently estimate statistical

relationships. Instead, quantitative outputs and methodological limitations are supplied as structured context so that the language model functions primarily as an evidence-communication layer.

## 3. DATASETS

### 3.1 Online Retail II

The Online Retail II dataset contains transactions from a UK-based registered non-store online retailer between December 2009 and December 2011. The UCI repository reports 1,067,371 original instances and identifies Quantity, InvoiceDate, UnitPrice, CustomerID, and Country among the variables [12]. After the preprocessing used in this study, the analysis retained 1,041,670 transaction records, 5,878 customers, 4,917 products, and 43 countries, with total computed revenue of 20,972,594.57. Revenue was computed from transaction-level quantity and unit-price information. The preprocessing removed invalid or inconsistent records according to the study pipeline.

### 3.2 Instacart

The Instacart Online Grocery Shopping Dataset 2017 contains multiple related tables, including orders and order-product records. The prior order-product table contains 32,434,489 records in commonly distributed versions of the dataset [13]. The present study uses this transaction structure as a second retail evaluation environment and evaluates the Orders → Reorder Rate relationship.

### 3.3 Implementation and Evaluation Specification

The experimental implementation was developed in Python using Pandas and NumPy for data processing; scikit-learn for baseline modeling and evaluation; XGBoost for the primary predictive model; SHAP for feature attribution; causal-learn/PC for causal-structure screening; DoWhy for causal estimation and refutation; NetworkX for graph representation; Matplotlib and Seaborn for visualization; and Ollama with Llama 3.1 for the language-model decision-support layer.

For Online Retail II, the retained analytical dataset contains 1,041,670 records. The predictive comparison uses Linear Regression, Random Forest, and XGBoost. Online Retail II is evaluated using  $R^2$ , RMSE, and MAE, while the reported Instacart predictive result is the mean cross-validated  $R^2$ . Feature attribution is computed using SHAP, with the reported percentages obtained by normalizing absolute SHAP contributions among the displayed drivers.

The causal workflow consists of: (1) candidate-driver identification from predictive attribution; (2) PC-based conditional-independence screening; (3) treatment-outcome specification; (4) DoWhy effect estimation; (5) random-common-cause and placebo refutation for the reported Online Retail II effect; and (6) counterfactual sensitivity analysis at 5%, 10%, 20%, and 50% treatment-change scenarios. The LLM receives the structured analytical outputs rather than calculating predictive metrics or causal effects itself. The exact model parameters, split settings, conditional-independence configuration, treatment encoding, DoWhy estimator configuration, counterfactual transformation, and Llama/Ollama runtime settings are implementation-level settings recorded in the authors' implementation notebooks.

**Table 1. Online Retail II Predictive Model**

| Model             | $R^2$ | RMSE  | MAE   |
|-------------------|-------|-------|-------|
| Linear Regression | 0.214 | 0.314 | 0.231 |
| Random Forest     | 0.318 | 0.271 | 0.198 |
| XGBoost           | 0.362 | 0.252 | 0.184 |

## 4. ADIF FRAMEWORK

ADIF is organized as a sequential analytical workflow. Each stage has a distinct purpose and produces structured information for the subsequent stage. The workflow is intentionally designed so that an output from one stage is not silently reinterpreted as a stronger type of evidence in the next stage (Fig. 1).

Prediction identifies variables useful for prediction; driver discovery organizes feature-attribution results into candidate drivers; causal analysis evaluates specified treatment-outcome relationships under explicit assumptions; counterfactual reasoning evaluates hypothetical treatment changes; and the LLM communicates the structured evidence and limitations.

### 4.1 Predictive Modeling

XGBoost is used as the primary predictive model for structured retail data:

$$\hat{y} = f(x_1, x_2, \dots, x_p), \quad (1)$$

where  $y$  is the target business outcome and  $x_1, \dots, x_p$  represent explanatory features.

The predictive layer is evaluated using  $R^2$ , RMSE, and MAE where applicable. It is not interpreted as a causal model; its role is to identify variables that are useful for prediction and therefore suitable for further investigation.

### 4.2 Predictive-Driver Discovery

SHAP values are calculated to quantify the contribution of individual features to model predictions:

$$f(x) = E[f(x)] + \sum_{i=1}^n \phi_i \quad (2)$$

where  $\phi_i$  represents the contribution of feature  $i$ . The reported percentages are normalized absolute SHAP contributions among the displayed drivers.

For Online Retail II, Order Change and Quantity Change accounted for 63.42% and 36.58% of the reported contribution, respectively. These percentages describe predictive attribution within the fitted model and are not causal effects.

### 4.3 Causal Structure Screening

The implementation uses the PC algorithm to screen conditional-independence relationships and construct a candidate causal structure from the engineered retail variables. The procedure is used as a structure-screening stage before treatment-effect estimation; it is not treated as a definitive recovery of the true data-generating graph. The causal

interpretation depends on causal sufficiency, faithfulness, appropriate conditional-independence testing, and the adequacy of the variables supplied to the procedure [5, 6].

The principal treatment-outcome relationships evaluated in the experiments are Quantity Change  $\rightarrow$  Revenue Change for Online Retail II and Orders  $\rightarrow$  Reorder Rate for Instacart. These relationships are treated as analysis hypotheses carried from the predictive-driver stage into causal estimation rather than as automatically established causal discoveries.

## 4.4 Causal Effect Estimation

DoWhy is used to specify the causal model, identify the treatment effect, estimate the effect, and execute refutation tests. The implementation reports a treatment effect for each treatment-outcome pair together with an approximate 95% confidence interval. Because the treatment variables are engineered retail quantities rather than a universally defined binary intervention, the reported values are retained in the scale produced by the implementation and are not re-expressed as a generic  $T = 1$  versus  $T = 0$  contrast.

For Online Retail II, the reported effect of Quantity Change on Revenue Change was 0.59748, with a reported 95% confidence interval of approximately [0.5721, 0.6229]. For Instacart, the reported effect of Orders on Reorder Rate was 0.006417, with a reported 95% confidence interval of approximately [0.00591, 0.00693]. These intervals provide uncertainty information around the reported point estimates and must be interpreted jointly with the causal assumptions and treatment encoding.

## 4.5 Causal Assumptions

The causal interpretation depends on assumptions that cannot be completely verified from observational data. The primary assumptions include:

- consistency between observed treatment and potential outcomes,
- appropriate treatment and outcome definitions,
- absence of problematic unmeasured confounding conditional on the adjustment variables,
- adequate overlap in the relevant treatment ranges, and
- validity of the causal structure used for adjustment.

The PC algorithm additionally relies on assumptions including causal sufficiency and faithfulness. Therefore, the causal results are presented as conditional estimates rather than definitive proof of causality.

## 5. PREDICTIVE EVALUATION

The predictive layer was evaluated using Linear Regression, Random Forest, and XGBoost to examine whether the selected XGBoost model provided improved predictive performance relative to simpler baseline models. Online Retail II was evaluated using the coefficient of determination ( $R^2$ ), root mean squared error (RMSE), and mean absolute error (MAE). The reported results are summarized in Table 1 and Fig. 2.

XGBoost achieved the highest  $R^2$  value (0.362), followed by Random Forest (0.318) and Linear Regression (0.214). The absolute improvement in  $R^2$  was 0.148 over Linear Regression and 0.044 over Random Forest. XGBoost also produced the lowest

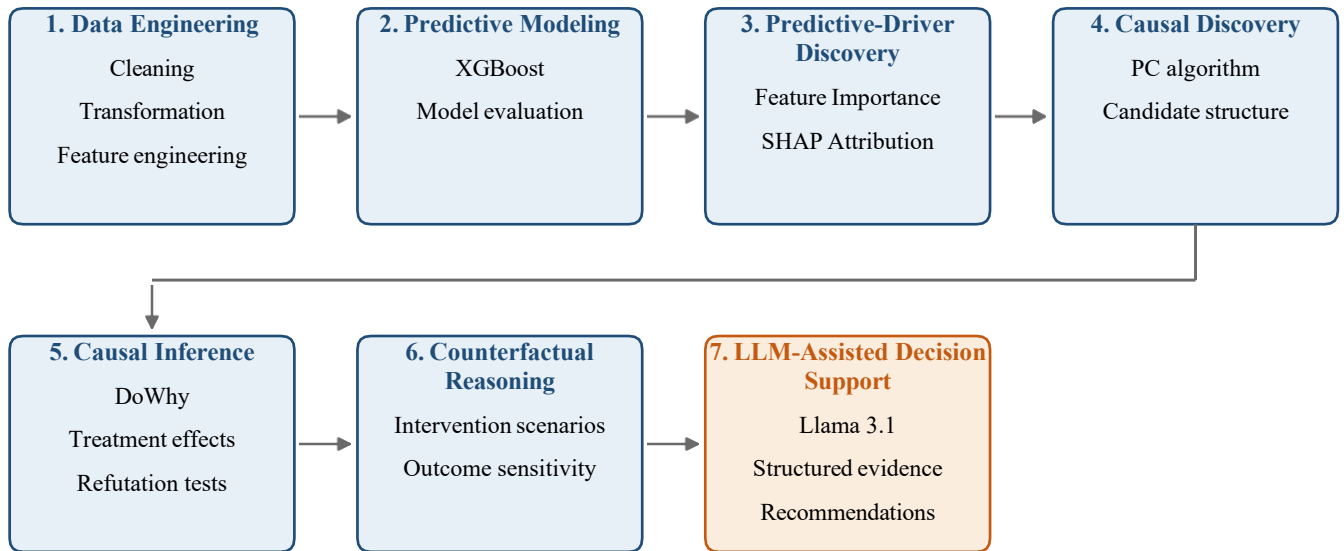


**Fig. 1. ADIF framework architecture showing the evidence-handoff workflow from data engineering and predictive analysis through predictive driver discovery, causal discovery, causal inference, counterfactual reasoning, and LLM-assisted decision support**

**Table 2. Instacart Predictive Baseline Results**

| Model             | $R^2$  |
|-------------------|--------|
| Linear Regression | 0.421  |
| Random Forest     | 0.527  |
| XGBoost           | 0.5883 |

error values. Its RMSE decreased from 0.314 for Linear Regression to 0.252, a reduction of approximately 19.7%, and from 0.271 for Random Forest, a reduction of approximately 7.0%. MAE decreased by approximately 20.3% relative to Linear Regression and 7.1% relative to Random Forest. The agreement between the  $R^2$ , RMSE, and MAE comparisons supports selecting XGBoost as the predictive component for the Online Retail II experiments. These differences describe the reported performance metrics and do not, by themselves, establish statistical significance or improved business outcomes.

For the Instacart dataset, the reported predictive comparison uses mean cross-validated  $R^2$  (Table 2).

XGBoost achieved a mean cross-validated  $R^2$  of 0.5883, compared with 0.527 for Random Forest and 0.421 for Linear Regression, an absolute difference of 0.0613 and 0.1673, respectively. Since corresponding Instacart RMSE and MAE values are not reported here, an error-based comparison cannot be made from the available results.

The two datasets provide evidence of the predictive performance of the selected model within their respective evaluation settings. In addition, predictive performance alone does not identify causal relationships or establish that a model-driven intervention will improve business outcomes. This distinction motivates the subsequent predictive-driver, causal, and counterfactual analyses in ADIF.

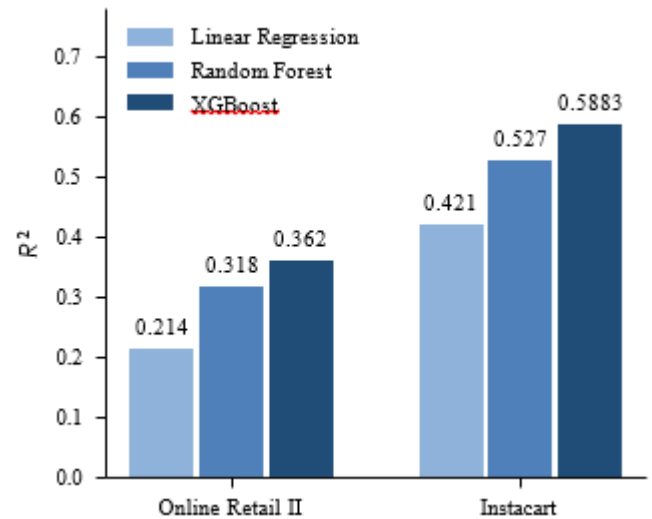


**Fig. 2. Predictive performance ( $R^2$ ) of the three models on each dataset. Instacart values are mean cross-validated  $R^2$ . The datasets differ in targets, features, and validation procedures and should not be compared directly.**

## 6. PREDICTIVE-DRIVER ANALYSIS

SHAP analysis was used to identify variables contributing most strongly to the predictive output (Fig. 3). For Online Retail II, Order Change and Quantity Change account for 63.42% and 36.58% of the reported normalized contribution. For Instacart, Orders and Products account for 93.16% and 6.84%, respectively. These figures summarize predictive attribution and should not be read as causal rankings.

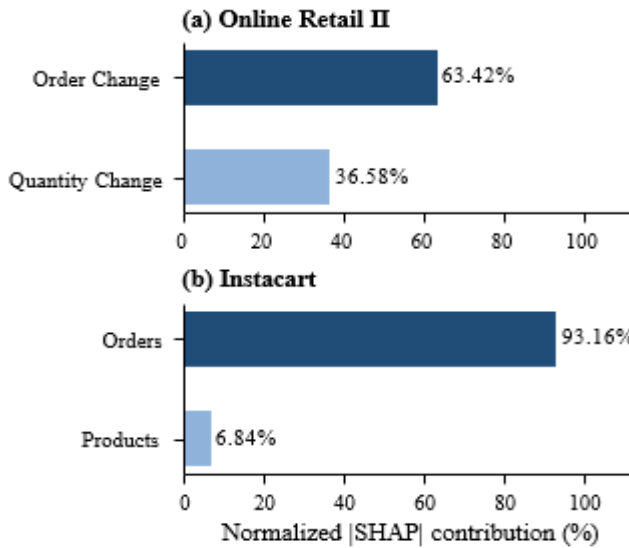


Fig. 3. Normalized absolute SHAP contribution among the displayed drivers: (a) Online Retail II; (b) Instacart. The values describe predictive attribution, not causal effects

## 7. CAUSAL ANALYSIS RESULTS

### 7.1 Online Retail II

The primary causal relationship evaluated on Online Retail II was Quantity Change  $\rightarrow$  Revenue Change. The estimated average treatment effect (ATE) was 0.59748, with an approximate 95% confidence interval of [0.5721, 0.6229] (Fig. 4a). The positive estimate and interval indicate a positive estimated effect under the specified model and assumptions.

However, transaction revenue is structurally related to quantity and unit price. Therefore, this result should not be interpreted as independent proof that increasing quantity necessarily causes proportional business growth. Instead, it demonstrates the application of ADIF's causal-estimation stage to a defined treatment-outcome relationship.

### 7.2 Instacart

For Instacart, the evaluated relationship was Orders  $\rightarrow$  Reorder Rate. The estimated ATE was 0.006417, with an approximate 95% confidence interval of [0.00591, 0.00693] (Fig. 4b). The positive estimate provides a second example of applying the workflow to a different retail behavioral outcome.

Both estimates are conditional on the treatment and outcome definitions, adjustment variables, and assumptions used in the causal models. The confidence intervals quantify uncertainty under the estimation procedure; they do not eliminate possible unmeasured confounding or validate the assumed causal structure. The estimates are therefore interpreted as assumption-dependent results rather than definitive proof of causality.

## 8. REFUTATION AND ROBUSTNESS

Refutation tests were performed using DoWhy to examine whether the reported Online Retail II effect was sensitive to controlled perturbations (Table 3, Fig. 5).

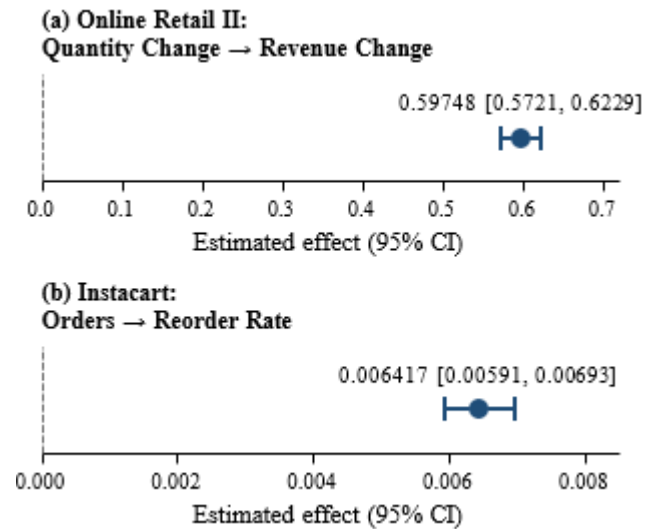


Fig. 4. Estimated treatment effects with 95% confidence intervals: (a) Quantity Change  $\rightarrow$  Revenue Change on Online Retail II; (b) Orders  $\rightarrow$  Reorder Rate on Instacart. The two panels use different scales.

Table 3. Causal Refutation Results for Online Retail II

| Test                | Original | Refuted     | <i>p</i> -value |
|---------------------|----------|-------------|-----------------|
| Random Common Cause | 0.59748  | 0.59727     | 0.90            |
| Placebo Treatment   | 0.59748  | $\approx 0$ | $< 0.001$       |

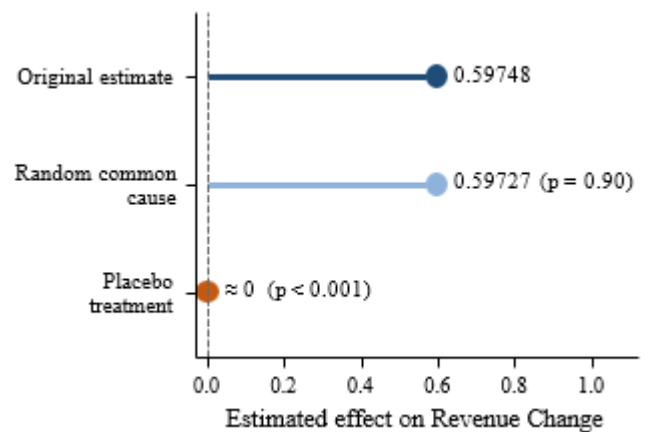


Fig. 5. DoWhy refutation results for the Online Retail II effect of Quantity Change on Revenue Change

The random-common-cause test changed the estimate only slightly under the tested configuration. The placebo-treatment test produced an approximately zero placebo estimate with  $p < 0.001$ . This is consistent with the expectation that a randomly assigned placebo should not reproduce the estimated treatment effect. The  $p$ -value is not interpreted as proof that the original causal estimate is correct. These tests provide sensitivity evidence but cannot eliminate unobserved confounding or establish that the assumed causal structure is correct. Passing them does not prove that the estimated effect is causal.

## 9. COUNTERFACTUAL ANALYSIS

After causal estimation, ADIF evaluates hypothetical changes in the treatment variable. For Quantity Change, the counterfactual stage was configured for hypothetical increases of 5%, 10%, 20%, and 50%. These scenarios are model-based what-if analyses: they inherit the assumptions of the underlying causal model, and this paper does not report or claim specific scenario outcomes. They should not be read as predicted or guaranteed business results.

## 10. END-TO-END WORKFLOW

For Online Retail II, the workflow proceeds as follows: predictive modeling identifies variables contributing to Revenue Change; SHAP identifies Order Change and Quantity Change as prominent predictive drivers; candidate drivers are passed to causal analysis, where conditional-independence structure is screened and a treatment-outcome relationship is specified; DoWhy estimates the treatment effect and confidence interval; refutation tests examine its sensitivity; counterfactual scenarios evaluate hypothetical treatment changes; and the structured findings, uncertainty information, and limitations are passed to the LLM layer. Each stage answers a different question: prediction identifies useful variables, attribution describes model behavior, causal analysis estimates a treatment relationship under assumptions, counterfactual reasoning evaluates hypothetical changes, and the LLM communicates the resulting evidence.

## 11. LLM-ASSISTED DECISION SUPPORT

The final ADIF layer uses Llama 3.1 through Ollama to convert structured analytical outputs into natural-language decision support. The LLM is not used to calculate ATEs, confidence intervals, SHAP values, or predictive metrics. These values are generated by the analytical components and supplied to the language model as structured context.

A representative decision-support input contains the identified predictive drivers, their attribution values, treatment and outcome definitions, estimated effects, confidence intervals, counterfactual scenarios, refutation results, and methodological limitations. The LLM then produces an explanation and recommendation based on those supplied values.

This design constrains the LLM to an evidence-communication role. Human review remains necessary before operational decisions are made.

### 11.1 LLM Evaluation Scope

This paper does not report a quantitative evaluation of LLM faithfulness: no independent labeled benchmark, evaluator protocol, or scoring rubric was defined for the present experiments. The LLM layer is therefore presented as a demonstration of evidence-grounded communication of structured analytical outputs, and quantitative faithfulness benchmarking remains future work.

## 12. CROSS-DATASET ANALYSIS

The Online Retail II and Instacart experiments demonstrate that the workflow can be applied to different transaction structures (Table 4).

These results should not be interpreted as a direct performance comparison between datasets because the datasets have different target variables, feature spaces, sample structures, and causal estimands.

**Table 4. Summary of Experimental Results**

| Dataset          | Predictive Result | Causal Result                            |
|------------------|-------------------|------------------------------------------|
| Online Retail II | XGBoost           | Quantity → Revenue, effect = 0.59748     |
|                  | $R^2 = 0.362$     |                                          |
| Instacart        | XGBoost CV        | Orders → Reorder Rate, effect = 0.006417 |
|                  | $R^2 = 0.5883$    |                                          |

## 13. DISCUSSION

The experiments illustrate the complementary roles of prediction, predictive-driver analysis, causal inference, counterfactual reasoning, and LLM-assisted communication. Predictive modeling provides a scalable mechanism for identifying variables associated with the target outcome, while SHAP provides an interpretable decomposition of model predictions. Neither establishes that changing an important feature will change the outcome.

Causal inference changes the question from “What predicts the outcome?” to “What is the estimated effect of changing the treatment under the specified model and assumptions?” Counterfactual reasoning extends this analysis by examining hypothetical treatment changes. The LLM layer serves a different purpose by converting structured analytical outputs into a format that can be consumed by decision makers.

The central contribution of ADIF is therefore the structured connection between these stages and the explicit preservation of evidence type and assumptions across stage boundaries, rather than the invention of a new predictive, causal-discovery, or language-model algorithm.

## 14. REPRODUCIBILITY

The implementation was developed in Python using the libraries listed in Section 3.3. Both datasets are publicly available from their sources [12, 13] and are not redistributed with this paper.

The analytical sequence is: data profiling and cleaning; KPI and feature construction; predictive comparison using Linear Regression, Random Forest, and XGBoost; SHAP-based predictive-driver attribution; PC-based causal-structure screening; DoWhy treatment-effect estimation; random-common-cause and placebo refutation; counterfactual scenario analysis; and evidence-grounded LLM communication. Exact settings, including XGBoost hyperparameters, baseline-model parameters, train/test or cross-validation split details, random seeds, SHAP sampling settings, the PC conditional-independence test and significance threshold, treatment/outcome encoding, DoWhy identification and estimator configuration, refutation parameters, the counterfactual transformation, the Llama 3.1 model variant, Ollama runtime configuration, prompt template, and decoding parameters, are implementation-level settings recorded in the authors’ implementation notebooks.

## 15. LIMITATIONS

Several limitations should be considered when interpreting the results. First, both datasets are observational, so causal estimates depend on assumptions and cannot automatically be interpreted as randomized experimental effects.

Second, the PC algorithm relies on assumptions such as causal sufficiency and faithfulness. Important unmeasured common causes may violate these assumptions and alter the candidate structure.

Third, the Quantity Change  $\rightarrow$  Revenue Change relationship is structurally connected because transaction revenue depends on quantity and unit price. Consequently, it should not be presented as independent evidence of causal business impact.

Fourth, causal and predictive results remain dependent on implementation choices such as treatment encoding, estimator configuration, validation splits, and conditional-independence settings.

Fifth, predictive importance and causal importance remain conceptually different. A high SHAP contribution does not establish that changing the corresponding feature will produce the same magnitude of change in the outcome.

Sixth, the LLM layer introduces risks including hallucination, omission, and unsupported recommendations. Structured quantitative inputs reduce these risks but do not eliminate them. The present paper does not claim a general quantitative LLM-faithfulness result.

Finally, the evaluation uses two retail datasets and therefore does not establish generalization to non-retail domains or to randomized decision settings.

## 16. CONCLUSION

This paper presented ADIF, an end-to-end decision intelligence framework that connects predictive modeling, predictive-driver discovery, causal analysis, counterfactual reasoning, and LLM-assisted decision support. The framework was evaluated using Online Retail II and Instacart data. XGBoost provided the predictive component, SHAP identified predictive drivers, the PC algorithm supported causal-structure screening, DoWhy estimated treatment effects and performed refutation analysis, counterfactual reasoning supported scenario analysis of hypothetical interventions, and Llama 3.1 translated structured analytical results into natural-language decision support.

For Online Retail II, the reported analysis identified Order Change and Quantity Change as major predictive contributors and reported an effect of 0.59748 for Quantity Change on Revenue Change. For Instacart, the Orders  $\rightarrow$  Reorder Rate analysis reported an effect of 0.006417. The predictive experiments reported a mean cross-validated  $R^2$  of 0.5883 for XGBoost on Instacart.

The results support the feasibility of connecting these analytical stages into a unified workflow. At the same time, the causal findings remain conditional on observational-data assumptions, the Quantity-to-Revenue relationship is structurally constrained, and the current evidence does not establish universal decision superiority or general LLM faithfulness.

Future work should evaluate ADIF using randomized or quasi-experimental data, publish the complete causal graphs and estimator configurations, use non-mechanical treatment-outcome relationships, compare alternative causal-discovery

methods such as NOTEARS, use leakage-resistant temporal or group-aware validation, and establish an independent human-rated benchmark for LLM faithfulness and decision-support utility.

## 17. ACKNOWLEDGMENTS

The authors would like to thank the University of New Haven for providing the academic environment and resources supporting this research.

## 18. REFERENCES

- [1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
- [2] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [3] G. Shmueli, "To Explain or to Predict?" *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.
- [4] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.
- [5] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed., MIT Press, 2000.
- [6] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, 2009.
- [7] D. B. Rubin, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, vol. 66, no. 5, pp. 688–701, 1974.
- [8] G. W. Imbens, "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economics and Statistics*, vol. 86, no. 1, pp. 4–29, 2004.
- [9] A. Sharma and E. Kiciman, "DoWhy: An End-to-End Library for Causal Inference," arXiv preprint arXiv:2011.04216, 2020.
- [10] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
- [11] A. Grattafiori et al., "The Llama 3 Herd of Models," arXiv preprint arXiv:2407.21783, 2024.
- [12] D. Chen, "Online Retail II," UCI Machine Learning Repository, 2019, doi: 10.24432/C5CG6D.
- [13] Instacart, "The Instacart Online Grocery Shopping Dataset 2017," 2017. Available: <https://www.instacart.com/datasets/grocery-shopping-2017>.