Towards an Explainable Decision Intelligence: Leveraging Predictive ML, SHAP, and Causal Digital Twin for Order-to-Cash Process

Submitted to Special session 5: Advances in Data-Driven Optimization and AI-Enhanced Business Process Mining

Shruti Kirti Nandan[1], Uppili Srinivasan Venkatesan[2], Aditya Acharya[3], Varun Jagad[4], Ranjana Vyas[5], and Monika Gupta[6]

Shruti Kirti Nandan, Indian Institute of Information Technology - Allahabad, sknandan@gmail.com
Uppili Srinivasan Venkatesan, Celonis India Pvt Ltd, u.srinivasan@celonis.com
Aditya Acharya, Indian Institute of Information Technology - Allahabad, adityaacharyasantoshkumar@gmail.com
Varun Jagad, Indian Institute of Information Technology - Allahabad, varunjagad9@gmail.com
Ranjana Vyas, Indian Institute of Information Technology - Allahabad, ranjana@iiita.ac.in
Monika Gupta, Celonis India Pvt Ltd, mo.gupta@celonis.com

Abstract. In large enterprise supply chains, the Order-to-Cash (O2C) process is the engine of revenue. However, manual reviews often slow down the process. This is where human operators inspect orders for potential fraud, compliance risks, or logistics issues. Hence, this review is essential for risk reduction. However, relying on fixed rule-based systems often results in an increase in false positives, inconsistent decisions, and delays. This paper introduces an O2C Decision Intelligence System to address this exact problem. Here, a unified analytical framework is developed that predicts, explains, and improves manual review decisions. Unlike traditional machine learning (ML) methods that show 'black-box' behavior and are hard to interpret, this system combines Causal Process Mining with Explainable AI (SHAP) to give clear, case-by-case justifications for each flagged order. Using a rich data set of 10,000 real operational cases, we show and validate how the system deals with complex logistic scenarios, such as the combined risks of international shipping and high-value electronics, that static rules miss. Additionally, the system features a Causal Digital Twin, which allows stakeholders to simulate "what-if" scenarios and adjust automation limits based on predicted risk instead of fixed rules. Experimental results show that this model has high predictive reliability (ROC-AUC $\approx$ 0.88–0.95) and provides actionable insights to reduce unnecessary manual reviews by 20–30%, effectively connecting predictive modeling with operational decision-making.

Keywords: Decision Intelligence · Order-to-Cash (O2C) · Explainable AI (XAI) · Manual Review Optimization · Causal Digital Twin · Process Mining.

1 Introduction

1.1 Background and Motivation

In large-scale business environments, the O2C process is a key financial and operational foundation. Although highly automated, organizations still rely on manual reviews to manage risks such as fraud, compliance issues, and logistics failures.

While necessary, these reviews create inefficiencies due to high resource usage, delays, and subjective or rigid rule-based decisions. Existing business intelligence (BI) dashboards mainly provide reactive insights, creating the need for a proactive framework that can predict manual reviews and evaluate automation feasibility while accounting for risk.

1.2 Problem Statement and Key Challenges

Research addresses lack of adaptability in current manual review systems. Traditional methods rely on fixed thresholds & past trends, causing unexpected review spikes & making decisions inconsistent and difficult to explain during audits.
Key challenges include:

1. **Inconsistent Manual Review Decisions:** Between subjective human judgement and rigid rule sets, similar orders are often treated differently, leading to unpredictable outcomes.
2. **Lack of Explainability:** Most ML models do not explain the "why" behind their predictions, which decreases trust and makes audits harder.
3. **Disconnection Between Prediction and Process:** There is often a gap between risk prediction and the actual workflow; knowing an order is risky doesn't always trigger the right operational path.
4. **Rigid Automation Logic:** Current systems tend to force a binary "automate" or "review" choice, ignoring the degree of varying risk levels.

1.3 Related Work

To build our framework, we focus on three core technologies: Machine Learning (ML), Process Mining, and Explainable AI (XAI).

Machine Learning Ensembles. Supervised ensemble models like Random Forests [1] and XGBoost [17] are widely used for tabular enterprise data due to their ability to capture non-linear relationships. Tools such as Scikit-learn [8] make them accessible, though their black-box nature limits use in high-stakes compliance settings.

Process Mining and Causality. Process Mining connects data science with business process management [4]. Recent work by Bozorgi et al. [6] and Hennebold et al. [5] combines process mining with causal discovery to identify inefficiency root causes, though applying these insights in real-time decision-making remains challenging.

Explainability and Causal Inference. XAI methods like LIME [16] and SHAP [2] improve interpretability, with SHAP offering consistent feature attribution based on game theory [9]. We further use Pearl's causality framework [3] and DoWhy [15] to simulate "what-if" scenarios for the Digital Twin component.

Table 1 compares our proposed system with current approaches, pinpointing their main differences and where existing methods fall short.

Table 1. Comparison of Existing Approaches vs. Proposed System

Approach	**Methodology**	**Key Limitation**
Rule-Based Engines	Static If-Then thresholds	Struggles to adapt; breaks easily when processes shift and triggers frequent false alarms.
Traditional ML	Predictive scoring (e.g., XG-Boost [17], Random Forests [1])	Opaque "black-box" logic; offers no business context and is difficult to examine.
Process Mining	Log analysis and variant discovery [4]	Focuses on correlation and retrospective analysis, not predictive intervention.
BI Dashboards	KPI visualization [?, 14]	Descriptive only; offers no prescriptive or simulation capabilities.
Proposed System	**Causal ML + SHAP [2] + Digital Twin [3, 15]**	**Predictive, Explainable, and Simulation-Ready decisioning.**

1.4 Our Proposed Solution

To solve these issues, we propose O2C Decision Intelligence System. It uses ML to predict risks and SHAP [2] to explain them, while Causal Digital Twin for simulation [3]. We also swapped out rigid rules for smarter probability-based automation. Framework overview is illustrated in Figure -1.

2 Experimental Setup

2.1 Dataset Description and Provenance

To rigorously evaluate the framework, we used an enhanced O2C dataset of 10,000 operational cases from anonymized real-world ERP logs. The data generation process was calibrated with process mining experts to preserve real operational distributions, causal structures, and bottlenecks.

The dataset features 50 attributes that were specifically curated to reflect the multi-dimensional nature of real-world O2C workflows. These features span four key operational domains:

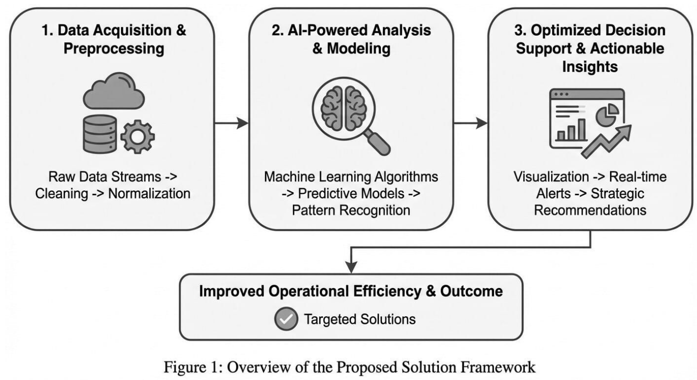

Fig. 1. Overview of how the system works, flowing from data-collection all the way to better decision-making.

- **Financial:** Order Value ($), Payment Terms, Customer Credit Risk.
- **Logistics:** Shipping Mode, Package Weight, Destination Region.
- **Risk & Compliance:** Fraud Risk Score, Vendor Reliability Score.
- **Target Variable (Ground Truth Formulation):** Binary outcome $y \in \{0, 1\}$, where $y = 1$ denotes a historical manual review. While this baseline includes human bias, our framework uses SHAP and a Causal Digital Twin to identify objective risk drivers and simulate optimized risk thresholds.

Class Balance and Data Collection. Target variable reflects realistic enterprise class imbalance, where most transactions are automated and only a small risky subset requires human intervention. Table 2 shows the distribution used.

Table 2. Dataset Class Balance for the Target Variable ($N = 10,000$)

Class (y)	**Count**	**Percentage**
$y = 0$ (Auto-Approved)	7,150	71.5%
$y = 1$ (Manual Review)	2,850	28.5%

2.2 Research Questions and Hypotheses

To transition from a descriptive system overview to a rigorous scientific evaluation, our experimental framework is guided by the following testable research questions (RQs) and hypotheses (Hs):

- **RQ1 (Predictive Capability):** To what extent can an ensemble ML model accurately capture the historical baseline of manual review decisions in the

O2C process?
H1: A Random Forest classifier can isolate baseline review patterns with an ROC-AUC exceeding 0.85.

– **RQ2 (Explainability):** How can opaque predictive models be made transparent and actionable for process owners?
H2: Post-hoc SHAP analysis will extract objective risk drivers that quantitatively align with established domain expertise.
– **RQ3 (Automation Efficiency):** Can probability-driven automation safely outperform static rule-based engines?
H3: Implementing a threshold-based ML automation logic will reduce unnecessary manual reviews by 20% to 30% without escalating risk.
– **RQ4 (Simulation Validity):** How reliably can a digital twin simulate the operational impact of policy changes?
H4: The Causal Digital Twin will accurately predict shifts in review workloads and risk exposure indices prior to actual deployment.

2.3 Mathematical Foundations

Predictive Model Formulation and Justification. The primary objective of the predictive layer is to learn a baseline decision function $f : X \to [0, 1]$. Using a Random Forest ensemble of T decision trees, the predicted probability is:

$$P(\text{Historical Manual Review}|X) = \frac{1}{T}\sum_{t=1}^{T} h_t(X) \tag{1}$$

where $h_t(X) \in \{0, 1\}$ denotes the prediction of the t-th tree.

Probability reflects historical human reviews, not whether review was truly needed. Modeling this baseline captures workflow rules, bottlenecks, and human bias, while explainability and simulation layers identify true operational risk.

SHAP-Based Explainability. We use SHAP to break down predictions into additive feature contributions:

$$f(X) = \phi_0 + \sum_{i=1}^{n} \phi_i(X) \tag{2}$$

where ϕ_0 is the base value and $\phi_i(X)$ is the Shapley value for feature i. The main risk driver is identified as:

$$\text{Primary Cause} = \arg\max_i |\phi_i(X)| \tag{3}$$

Risk-Aware Automation Logic. Let p be the predicted probability and τ the configurable threshold. The automation decision is:

$$\text{Decision} = \begin{cases} \text{AUTO-APPROVE}, & \text{if } p \le \tau \\ \text{MANUAL REVIEW}, & \text{if } p > \tau \end{cases} \tag{4}$$

Digital Twin Simulation. For a case with modified features X', the expected change in probability (Δp) is:

$$\Delta p = P(\text{Manual Review}|X') - P(\text{Manual Review}|X) \tag{5}$$

2.4 Preprocessing and Training Strategy

Nominal variables were label-encoded, and missing numerical values were filled using the median. A Random Forest Classifier was trained on an 80/20 split using 5-fold cross-validation to ensure stability [8].

3 Proposed Framework

3.1 System Architecture

The system has a modular design made up of four main layers, as illustrated in Figure 2:

1. **Data Ingestion and Feature Processing:** This layer handles intake of raw O2C transaction logs and cleans raw transaction logs to ensure high data quality.
2. **Predictive Modeling Layer:** This is the main ML engine that calculates the probability of a manual review.
3. **Explainability and Reasoning Layer:** This layer uses SHAP to explain why the model made a prediction by isolating the impact of individual features.
4. **Decision Intelligence and Simulation Layer:** This layer turns predictions into real-world decisions and allows for "what-if" testing through a Digital Twin.

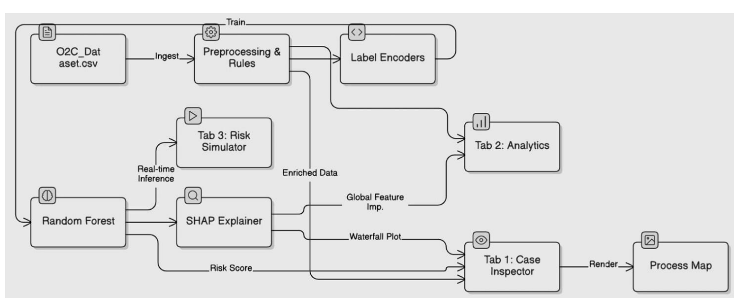

Fig. 2. High-level architecture of the O2C Decision Intelligence System, mapping the journey from data ingestion to deductive reasoning and simulation.

3.2 Predictive Modeling Strategy

We selected a Random Forest Classifier to capture the complex, non-linear interactions that occur in supply chains, such as the increased risk when high-value items are shipped to certain international regions.

3.3 Explainability and Actionable Automation Logic

To resolve the "black-box" nature of ensemble models, the framework includes an Explainability and Reasoning Layer. Rather than operating as an independent process, SHAP [2] is employed as a rigorous *post-hoc explanation mechanism* applied directly to the Random Forest results.

Actionable SHAP-Based Reasoning. By applying cooperative game theory principles, SHAP decomposes the Random Forest's output probability to isolate the exact marginal contribution of each feature. The decision function can be written as:

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{6}$$

where ϕ_0 refers to the base value (the average output of the model over the training dataset) and M is the number of features.

The scientific motivation for this layer is strict operational actionability. Simply flagging an order as "high risk" is insufficient for compliance officers who must justify audits. By providing this quantitative breakdown, the system generates a transparent reasoning chain for each decision:

– **Primary Cause:** The feature driving the highest variance from the baseline (e.g., International Shipping contributing +0.32 to the risk score).
– **Top Contributing Factors:** Provides a ranked list of secondary features influencing the decision.
– **Directional Impact:** Clearly indicates whether each factor compounds risk (+) or mitigates it (-).

This ensures manual reviewers focus directly on the variables triggering the model, significantly reducing cognitive load and review time [7].

Automation Logic. Moving beyond rigid rule systems, automation is driven by risk probability (p) against a dynamic threshold (τ). High-risk cases ($p > \tau$) are flagged for human review with SHAP context, while low-risk cases ($p \leq \tau$) are automatically approved.

3.4 Causal Digital Twin Architecture and Calibration

Unlike static what-if dashboards, our Causal Digital Twin acts as a *causal feature perturbation layer* integrated with the Random Forest model using the DoWhy framework [15], modifying predictive inputs through causal rules.

Causal Structure and DAG. Simulation uses Directed Acyclic Graph to model dependencies within O2C process instead of treating all 50 features independently. It captures relations such as Order Value → Customer Credit Risk → Manual Review and Shipping Region → Vendor Reliability → Manual Review. During interventions like $do(\text{Order Value limit} = \$25,000)$—twin uses DAG propagates downstream effects before generating modified feature vector (X').

Calibration Process. To ensure the twin's simulations accurately reflect real operational environments, the system undergoes a rigorous calibration phase:

1. **Historical Replay:** The twin is validated using historical policy shifts by minimizing the gap between simulated and actual review volumes.
2. **Domain Expert Validation:** Propagation boundaries (e.g., limiting how much shipping mode changes affect fraud scores) are hard-coded using operational heuristics from process mining experts.

Simulation Execution. When stakeholder tests new policy, Digital Twin generates interventional data using calibrated DAG and Random Forest performs inference on updated data and System measures impact using $\Delta p = P(\text{Manual Review} \mid X') - P(\text{Manual Review} \mid X)$ to provide quantified prediction before live deployment.

3.5 Operational Insights and KPIs

The system tracks Manual Review Rate, Automation Eligibility, and Risk Exposure Index. Our analysis showed that "International Shipping" combined with "High Order Value" ($> \$15,000$) significantly raises the likelihood of a review, which static rules often overlook.

3.6 Simulation Outcomes

Three policy scenarios were tested: Conservative ($\tau = 0.3$), Balanced ($\tau = 0.5$), and Aggressive ($\tau = 0.7$). The Balanced policy provided the best compromise, reducing unnecessary manual reviews by 20–30%.

3.7 Methodology and Iterative Development

The system evolved through five prototypes, fixing key gaps between risk prediction and process execution.

4 Evaluation

4.1 Model Performance and Statistical Validation

We evaluated the model on a test dataset of 2,000 cases. As shown in Table 3, the system showed strong discriminative capability.

Table 3. Model Performance Metrics

Metric	Value	Interpretation
ROC-AUC	0.88 – 0.95	Strong discriminative capability
Precision	0.87	High accuracy for flagged cases
Recall	0.82	Captures majority of high-risk orders
F1-Score	0.84	Balanced precision-recall trade-off
Accuracy	0.89	Overall predictive correctness

Confirmed statistical significance using permutation testing by shuffling target labels 100 times, obtaining a p-value below 0.03. This shows with over 97% confidence that the model learns meaningful patterns not random noise.

4.2 Mathematical and Logical Reliability

We confirmed mathematical stability via 10 independent runs with ROC-AUC variance below 0.02, & validated reliability by verifying SHAP explanations against domain knowledge, like International Shipping increasing risk scores.

4.3 Operational and Process-Level Validation

We assessed operational reliability through simulation. The model demonstrated threshold stability, showing no erratic switching at $p \approx \tau$, and predictable changes in workload when adjusting risk parameters.

4.4 Overall Reliability Assessment

Throughout all dimensions, the system was statistically significant, mathematically stable, logically explainable, and realistic for operations.

5 Results and Analysis

5.1 Performance Results Summary

Model effectively distinguishes between safe and high-risk orders, confirming H1.

5.2 Threshold Analysis

Examining risk thresholds showed the efficiency-risk trade-off:

- **Conservative ($\tau = 0.3$):** 55% automation, best for audits.
- **Balanced ($\tau = 0.5$):** 72% automation, recommended for normal operations.
- **Aggressive ($\tau = 0.7$):** 85% automation, best for peak periods.

Table 4. Top Risk Drivers by Mean Absolute SHAP Value

Feature	**Mean** $|\phi|$	**Impact Direction**
International Shipping	0.28	Increases Risk
Order Value ($\geq$$15,000)	0.22	Increases Risk
Fraud Risk Score	0.19	Increases Risk
Vendor Reliability (Low)	0.16	Increases Risk
Payment Terms (COD)	0.12	Decreases Risk

5.3 Feature Importance and Risk Drivers

SHAP analysis identified the top risk drivers as shown in Table 4.

5.4 Efficiency Gains Analysis

Comparing the proposed system ($\tau = 0.5$) with the baseline rule-based approach showed a 26% reduction in the manual review rate, dropping from 38% to 28%, and a 57% decrease in false positives. This directly results in a 33% improvement in average processing time.

5.5 Cost-Benefit Analysis

For a mid-sized business processing 100,000 orders each year, the system is expected to save $900,000 annually in labor costs by cutting down 10,000 manual reviews and saving 30,000 operational hours. With a starting cost of $50,000, the ROI in the first year is about 580%.

5.6 Simulation of What-If Scenarios

Using the Digital Twin, we found that raising the Order Value limit to $25,000 led to an 8% reduction in reviews with a small risk increase of 0.04, making it safe. However, easing rules for Electronics caused a significant risk increase of 0.09, so it was not recommended.

6 Conclusion and Future Work

Our work shows that combining process mining, ML, and XAI can transform O2C manual reviews. The system achieved high accuracy (ROC-AUC 0.88–0.95) with transparent SHAP-based reasoning, reducing unnecessary reviews by 20–30%, false positives by 57%, and processing time by 33%. It also enables policy simulation before live deployment.

In future, we plan to test on larger production datasets, develop adaptive real-time thresholds, integrate concept drift detection for seasonal changes, strengthen BPM integration, and explore cost-sensitive simulation models balancing financial risk and operational workload.

7 Appendix

7.1 Initial Hypotheses

- **H1:** Large electronic items have a higher chance of manual review.
- **H2:** International shipments are more likely to need manual review.
- **H3:** International shipments managed by highly trained staff face stricter review.

Final Model evolved into robust, data-driven decision engine while remaining highly interpretable, proving reliable & scalable for real-world business environments.

References

1. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001)
2. Lundberg, S.M., Lee, S.I.: A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
3. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2009)
4. van der Aalst, W.: Process Mining: Data Science in Action. Springer (2016)
5. Hennebold, C., Islam, M.M., Krauß, J., Huber, M.F.: Combination of Process Mining and Causal Discovery Generated Graph Models for Comprehensive Process Modeling. Procedia CIRP **130**, 1296–1302 (2024)
6. Bozorgi, Z.D., Teinemaa, I., Dumas, M., La Rosa, M., Polyvyanyy, A.: Process Mining Meets Causal Machine Learning: Discovering Causal Rules from Event Logs. In: Proceedings of the International Conference on Process Mining (ICPM), IEEE (2020)
7. Molnar, C.: Interpretable Machine Learning. Leanpub (2022)
8. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
9. Lundberg, S.: SHAP: Explainable AI. GitHub Repository. `https://github.com/slundberg/shap`
10. Streamlit Inc.: Streamlit: Data Apps in Python (2023). `https://streamlit.io`
11. Harris, C.R., et al.: Array programming with NumPy. Nature **585**, 357–362 (2020)
12. McKinney, W.: Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference, pp. 56–61 (2010)
13. Hunter, J.D.: Matplotlib: A 2D graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007)
14. Plotly Technologies Inc.: Collaborative data science (2015). `https://plot.ly`
15. Sharma, A., Kiciman, E.: DoWhy: An End-to-End Library for Causal Inference (2020). `https://github.com/microsoft/dowhy`
16. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
17. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
18. Adibhaibvn: Project_essentials. GitHub Repository. `https://github.com/Adibhaibvn/Project_essentials.git`