End-to-End Explainability of Machine Learning Pipelines with Decision Predicate Graphs: A Financial Scenario Case Study

Leonardo Arrighi[1,2,*], Matheus Camilo Da Silva[2] and Sylvio Barbon Junior[2]

[1]Department of Mathematics and Geosciences, University of Trieste, Trieste, Italy
[2]Department of Engineering and Architecture, University of Trieste, Trieste, Italy

Abstract

Artificial Intelligence (AI) has become integral to numerous domains, with finance being a particularly prominent application area due to its high risk and the demand for uncovering patterns and insights from complex, heterogeneous data. However, the opacity of many AI models has raised serious concerns about interpretability and trust, especially in high-stakes settings such as financial decision-making. Moreover, real-world financial data is often vulnerable to the presence of outliers, which may arise from data entry errors, fraud, or rare but critical events. While eXplainable AI (XAI) methods have been developed to improve model transparency, they typically focus solely on explaining final predictions, overlooking the crucial influence of early pipeline stages such as data pre-processing. This oversight is particularly problematic, as decisions made during pre-processing (e.g., outlier detection) can significantly affect downstream model performance and regulatory compliance. In this work, we propose an end-to-end framework for explainability across the entire machine learning pipeline, applied to a financial risk assessment scenario. Leveraging Decision Predicate Graphs (DPG), we obtained DPG-based explanations of both Isolation Forest models used for outlier detection and Random Forest classifiers used for credit scoring. Through a real-world case study, we demonstrate how our approach delivers transparent insights into both the pre-processing and predictive components of financial pipelines, enhancing model accountability and fostering user trust.

Keywords

Ensemble Learning, Explainable Artificial Intelligence, Interpretability, Framework, Outliers

1. Introduction

Artificial Intelligence (AI) is increasingly applied across diverse fields, with finance emerging as a prominent domain due to AI's versatility, risk-management capabilities, and its power to detect patterns in heterogeneous data sources [1, 2]. Despite their remarkable performance, many AI models operate as "black boxes", making their internal decision processes opaque. Understanding these processes in finance and other high-stakes industries is essential for ensuring trust, compliance, and reliability. EXplainable AI (XAI) techniques are more used to reveal how complex models make decisions. Recent work has begun to address this need in loan default prediction. Bracke et al. [3] applied Quantitative Input Influence with SHapley Additive exPlanations (SHAP) [4] to quantify feature impacts and identify key drivers of mortgage defaults. Babaei et al. [5] integrated a SHAP-based feature selection into Random Forest (RF) models [6], yielding accurate and interpretable predictions for small and medium enterprises lending. More recently, Li and Wu [7] combined RF with SHAP to enhance predictive performance and highlight the most influential risk factors.

While most XAI research focuses on interpreting the final predictions of complex models [8], the entire ML pipeline—from data preprocessing to model evaluation—is critical for the integrity of results. Despite their importance, preprocessing steps, such as outlier removal, bias mitigation, and error

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
leonardo.arrighi@phd.units.it (L. Arrighi); matheus.camilodasilva@phd.units.it (M. C. D. Silva); sylvio.barbonjunior@units.it (S. B. Junior)
https://github.com/LeonardoArrighi/ (L. Arrighi)
0009-0006-2494-0349 (L. Arrighi); 0000-0002-1256-823X (M. C. D. Silva); 0000-0002-4988-0702 (S. B. Junior)

correction, are often overlooked in XAI discussions, even though in high-stakes fields like finance, these choices can make or break model outcomes. As Lipton et al. [9] emphasise, a truly transparent approach should explain every stage, starting with raw data preparation and extending through to prediction. This holistic view is also proposed by subsequent studies [10, 11], which contend that comprehensive explanations across the workflow enhance regulatory compliance and user confidence.

In our study, we used Decision Predicate Graphs (DPG), introduced by Arrighi et al. [12], to explain both the preprocessing stage—specifically, outlier detection and removal via an Isolation Forest (iForest) [13] in the DPG-GO extension of Ceschin et al. [14]—and the final RF classifier. According to [15], DPG is a *post-hoc*, *model-specific* XAI method that offers a *global* explanation of tree-based ensemble models' logic. The flexibility of DPG enables detailed explanations of the iForest, revealing how each feature influences outlier detection and exposing potential biases in the preprocessing stage. It also elucidates the RF classifier by surfacing its most informative metrics. Together, these capabilities form a modular framework that applies DPG-based XAI techniques across both critical phases of the classification pipeline: data preprocessing and model prediction. In a financial setting, this approach not only clarifies why a credit-scoring model flags an applicant as high risk but also traces how upstream cleaning steps influence the final outcome. This transparency enhances trust, especially in environments governed by regulatory frameworks such as the European Union's General Data Protection Regulation (GDPR), which mandates the right to explanation in automated decision-making [16]. Furthermore, DPG provide interpretable, logic-based representations that compliance officers and domain experts can readily review without deep technical expertise, ensuring our framework delivers true end-to-end explainability. The main contributions of our proposal are:

- Outlier Detection and Explanation: We use iForest to detect and remove outliers, then apply DPG-GO to produce a global, feature-level explanation of their detection, clarifying each feature's impact and uncovering potential biases during data cleaning.
- Model Training and Explanation: We train an RF model on the cleaned data and then apply DPG to generate a unified, ensemble-level explanation that visualises each tree's combined decision pathways.

2. Proposed Approach

To the best of our knowledge, this is the first framework that unifies XAI methods across both preprocessing and prediction stages, offering truly end-to-end interpretability in financial ML pipelines. Figure 1 illustrates the proposed framework's workflow, from data preprocessing to model prediction, using DPG in our financial case study context. The pipeline begins with a raw dataset, which is first processed by an outlier detection algorithm, specifically, iForest. iForest is an unsupervised algorithm that recursively partitions the feature space, exploiting the fact that anomalies, being both rare and distinct, require fewer random splits to be isolated and thus identified. During this preprocessing phase, a DPG-GO is generated to provide interpretable insights into the logic behind the removal of anomalous data points, ensuring transparency in the early stages of data handling. The resulting clean dataset is then passed to a model training phase, where an RF classifier is employed to perform credit risk assessment.

A second DPG is used to capture and explain the model's internal decision logic, highlighting the contribution of different features and their combinations to the final predictions. In fact, DPG leverages graph-theoretic metrics to provide insights into the tree-based ensemble model. In particular, it employs Betweenness Centrality (BC) and Local Reaching Centrality (LRC): BC pinpoints bottleneck nodes corresponding to crucial decision points, while LRC measures each node's influence by quantifying how its effects propagate through the graph.

Both DPGs (DPG-GO and DPG) for, respectively, preprocessing and classification, are represented as weighted directed graphs, where each node corresponds to a predicate and each edge is weighted by the frequency with which training samples satisfy the two predicates in sequence. A predicate is a

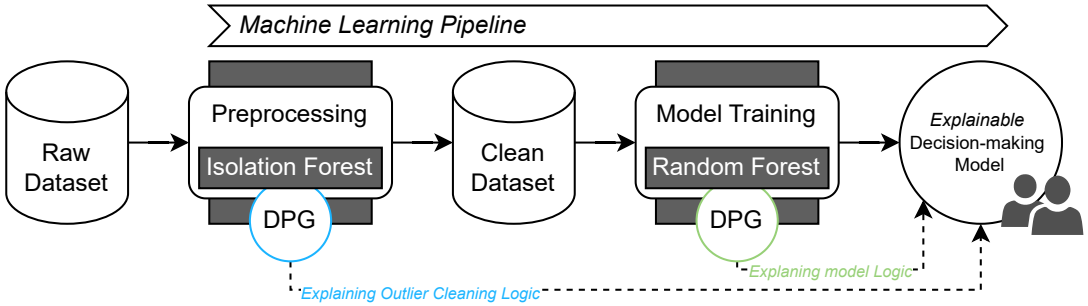

Figure 1: Overview of framework using Decision Predicate Graphs (DPG) within an ML pipeline. DPGs are applied at both the preprocessing stage (Isolation Forest) and the model training stage (Random Forest).

logical condition extracted from an internal tree node's split rule—i.e. a feature-value test of the form: "$f_i\ \sigma\ v$" (where f_i is some feature, σ is a symbol ($\leq$, $>$) and v the threshold used in that split).

It is important to mention that for outlier detection, the DPG-GO transforms an iForest into a weighted directed graph where predicates are represented as pairs (f, σ), abstracting from specific threshold values to allow generalisation across trees, unlike traditional DPG. Based on the Inlier-Outlier Propagation Score (IOP-Score), it is possible to compute the tendency of a predicate to lead toward the isolation of an outlier, thereby providing a nuanced view of how specific features influence the model's outlier identifications.

By composing predicates from both DPG-GO and DPG into graph structures, our framework captures the logical flow of decisions across the entire pipeline while also enabling graph-theoretic insights. Thus, the proposed approach addresses a key limitation of conventional XAI methods, which typically focus only on local explanations of model predictions, by offering a global and stepwise explanation that spans the entire ML pipeline.

3. Methods

Due to the privacy and accessibility constraints commonly associated with real-world financial datasets, we synthesised a dataset that simulates realistic consumer loan application scenarios while preserving reproducibility and interpretability. The dataset contains $n = 400$ samples, each characterised by four features and a binary `ApprovedLoan` label. The generation procedure was designed to reflect plausible demographic and financial distributions observed in credit risk assessment contexts[1]:

- Income (`Income`): Modelled using a log-normal distribution with a log-mean corresponding to \$60,000, and a moderate dispersion ($\sigma = 0.4$). This choice reflects the heavy-tailed nature of income in the population. Generated values were clipped to lie within the \$20,000–\$150,000 range to avoid extreme outliers.
- Credit Score (`CreditScore`): Sampled from a Gaussian distribution centered at 680 (standard deviation of 70) and truncated between the typical credit scoring bounds of 300 and 850. This ensures a realistic spread of creditworthiness scores.
- Marital Status (`MaritalStatus`): A categorical variable drawn from a discrete distribution with the following probabilities: Single (0.4), Married (0.5), and Divorced (0.1). These proportions are consistent with general demographic statistics in adult populations.
- Number of Dependent Children (`NumChildren`): Modelled using a Poisson distribution, with the expected number of children conditioned on marital status—lower for single applicants ($\lambda = 0.5$), higher for married ($\lambda = 1.2$), and moderate for divorced ($\lambda = 1.0$). Values were clipped to the $[0, 5]$ range to reflect typical household sizes.

A latent *loan approval score* was computed as a weighted sum of the input variables:

[1]The features appear in brackets as they were coded in the dataset and are then used in subsequent steps to improve readability.

$$Score = 0.00001 \cdot \text{Income} + 0.005 \cdot \text{CreditScore} + 0.3 \cdot \mathbb{1}_{\text{Married}} + 0.2 \cdot \mathbb{1}_{\text{Divorced}} - 0.15 \cdot \text{NumChildren} + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, 0.1)$ is an additive noise term. The approval probability for each application was then computed using the logistic sigmoid function, centred around the number of samples to calibrate approval rates:

$$Prob_{approval} = \frac{1}{1 + \exp\left(-(Score - 4.5)\right)}$$

The binary `ApprovedLoan` label was computed from a Bernoulli distribution using $Prob_{approval}$. After synthesising the dataset, we introduced 20 controlled outliers by randomly selecting observations and increasing only their credit-score values beyond the normal range, while leaving all other features unchanged to simulate a system error. The dataset is available here.[2] The iForest model is ran with 100 trees. The RF model is trained using 100 trees. To evaluate the impact of outlier removal, we trained two RF models on different datasets: one on the full training set (including iForest-identified outliers) and one on the training set after excluding those outliers. We had split the data so that 20% of the inliers served as a common test set, and we trained both models on the remaining 80%. We then evaluated both models on the same test set to compare their performance.

4. Results and Discussion

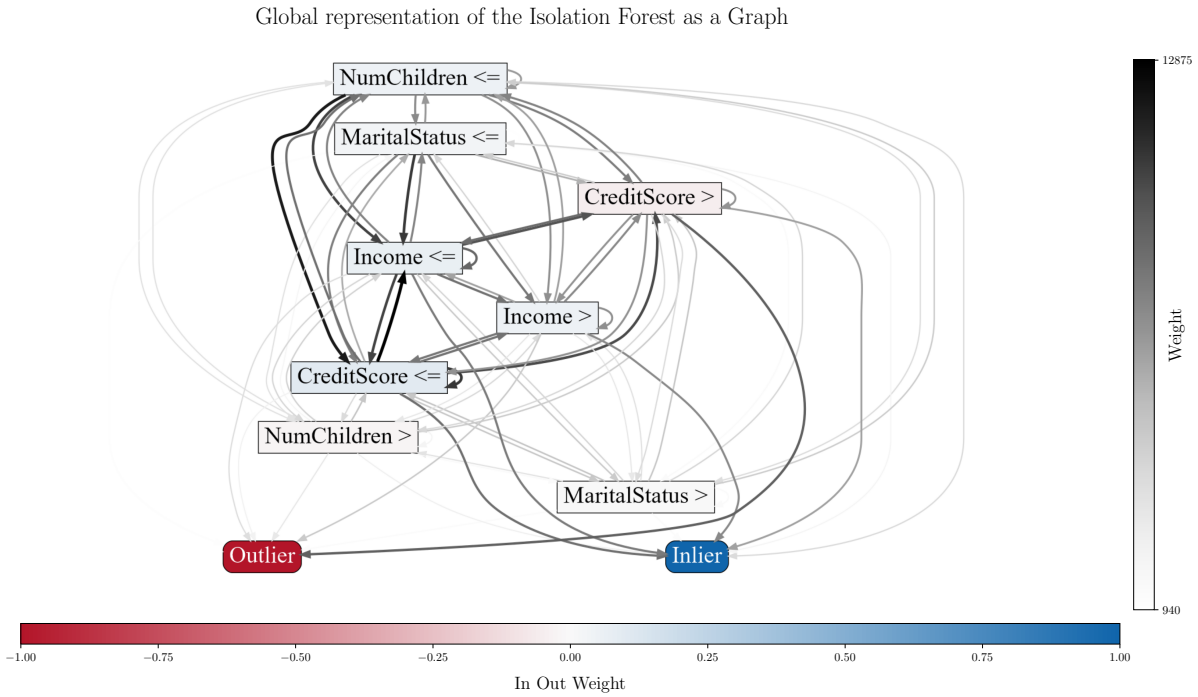

Figure 2: Global representation of the iForest model as a DPG-GO. The vertical bar on the right indicates the edge weights, while the horizontal bar at the bottom displays the IOP-Score of the nodes.

Following the described pipeline, we first ran iForest, obtaining a set of data points identified as outliers. Applying DPG-GO to explain the iForest model (Figure 2) reveals that *CreditScore* and *Income* are the most decisive features: thicker, darker edges correspond to frequently traversed paths that highlight influential features, while thinner, lighter edges indicate less significant splits. The IOP score, shown on the colour scale below Figure 2, quantifies each predicate's impact on outlier detection. Paths and nodes shaded in red denote a high likelihood of anomalous classification, whereas blue shading

[2] https://github.com/LeonardoArrighi/DPG-Pipeline

Table 1
Confusion matrices that depicted the performance evaluations of the RF models with 100 base tree learners, trained on the dataset with outliers (a) and on the cleaned dataset (b).

(a) With outliers: accuracy 88.16%

	Prediction	
Ground truth	Approved	Not Approved
Approved	25	6
Not Approved	3	42

(b) Cleaned dataset: accuracy 90.79%

	Prediction	
Ground truth	Approved	Not Approved
Approved	27	4
Not Approved	3	42

indicates strong association with the inlier class. Predicates such as `CreditScore >`, `NumChildren >`, and `MaritalStatus >` exhibit slightly negative IOP scores ($-0.047{,}5$, $-0.020{,}4$, and $-0.001{,}8$, respectively), indicating a mild inclination toward directing observations to the *Outlier* class. In contrast, the remaining predicates—`MaritalStatus <=` ($0.023{,}8$), `Income >` ($0.051{,}0$), `NumChildren <=` (0.0545), `Income <=` ($0.067{,}3$), and especially `CreditScore <=` ($0.095{,}8$)—present positive IOP scores, suggesting they predominantly govern splits that classify observations as inliers.

We classified the dataset as *Approved* or *Not Approved* using the RF models, as described in Section 3. We then evaluated both models on the same test set and presented their performance in the confusion matrices of Table 1.

The model trained without outliers achieved a modest but significant accuracy improvement—reaching 90.79%—which confirmed that removing anomalies could enhance predictive performance. Given RF's inherent robustness to outliers, the gain was small yet consistent.

Table 2
Top ten predicates ranked by their DPG metrics for a 100-tree RF model: (a) sorted by betweenness centrality (BC) on the left; (b) sorted by local reachability centrality (LRC) on the right.

(a) BC evaluation

Predicate	**BC**
`MaritalStatus <= 0.5`	0.166
`NumChildren > 0.5`	0.134
`NumChildren <= 0.5`	0.106
`MaritalStatus > 0.5`	0.090
`MaritalStatus > 1.5`	0.076
`NumChildren <= 1.5`	0.069
`MaritalStatus <= 1.5`	0.066
`CreditScore > 640.5`	0.058
`NumChildren > 1.5`	0.055
`NumChildren > 2.0`	0.043

(b) LRC evaluation

Predicate	**LRC**
`Income <= 76830.5`	9.122
`Income <= 99761.5`	9.116
`Income <= 103462.5`	9.032
`Income <= 75031.5`	8.864
`Income <= 72786.5`	8.531
`Income <= 77720.5`	8.524
`NumChildren <= 0.5`	8.504
`Income > 48467.0`	8.490
`Income > 50170.0`	8.439
`CreditScore <= 627.0`	8.188

After training the RF model, we explained it using DPG. As noted in [12], visualising complex ensembles with many tree-based learners is difficult, yet the derived metrics offer valuable insights into the model's decision mechanism.

The BC metric highlights potential bottleneck nodes, i.e., splits shared by many trees. As shown in Table 2a, the top features by BC are *MaritalStatus* and *NumChildren*. This result is expected, as both variables take on only a handful of discrete values, making them natural split points across the ensemble. In particular, the model relies mainly on splits at low values of these variables, underscoring that having children and being single strongly influence decisions. The third most central feature is *CreditScore*, whose split occurs near the data median; this threshold almost bisects the dataset into the two classes.

The LRC metric identifies the predicates that drive the RF's decision flow most critically. Table 2b ranks high-income splits on *Income* at the top, indicating that many model decisions depend on this feature. Unlike BC, *MaritalStatus* and *NumChildren* do not appear among the top LRC predicates—except for `NumChildren <= 0.5`, which still emerges as a key boundary—reflecting their lesser direct impact on the final classification.

5. Conclusion

This work introduced a framework for end-to-end explainability in ML pipelines, tailored to the financial domain. By integrating DPG with both outlier detection via IForest and classification using RF, we demonstrated how interpretable logic-based representations can clarify the influence of data preprocessing decisions and model predictions alike.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4o to check grammar and correct the structure of the text. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] B. B. Nair, V. P. Mohandas, Artificial intelligence applications in financial forecasting—a survey and some empirical results, Int. Dec. Tech. 9 (2015) 99–140.

[2] X. Li, A. Sigov, L. Ratkin, L. A. Ivanov, L. Li, Artificial intelligence applications in finance: a survey, Journal of Management Analytics 10 (2023) 676–692.

[3] P. Bracke, A. Datta, C. Jung, S. Sen, Machine learning explainability in finance: an application to default risk analysis, Bank of England working papers (2019).

[4] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., 2017, pp. 4768–4777.

[5] G. Babaei, P. Giudici, E. Raffinetti, Explainable FinTech lending, Journal of Economics and Business 125-126 (2023) 106126.

[6] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[7] H. Li, W. Wu, Loan default predictability with explainable machine learning, Finance Research Letters 60 (2024) 104867.

[8] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, et al., Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301.

[9] Z. C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery., Queue 16 (2018) 31–57.

[10] A. Holzinger, A. Carrington, H. Muller, Measuring the quality of explanations: The system causability scale (scs), KI - Künstliche Intelligenz 34 (2020) 193–198.

[11] C. V. G. Zelaya, Towards explaining the effects of data preprocessing on machine learning, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 2086–2090.

[12] L. Arrighi, L. Pennella, G. Marques Tavares, S. Barbon Junior, Decision predicate graphs: Enhancing interpretability in tree ensembles, in: World Conference on Explainable Artificial Intelligence, Springer Nature Switzerland, 2024, pp. 311–332.

[13] F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 413–422.

[14] M. Ceschin, L. Arrighi, L. Longo, S. B. Junior, Extending decision predicate graphs for comprehensive explanation of isolation forest, arXiv preprint arXiv:2505.04019 (2025).

[15] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, R. Ranjan, Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Computing Surveys 55 (2023) 194:1–194:33.

[16] A. Roos, The european union's general data protection regulation (GDPR) and its implications for south african data privacy law: An evaluation of selected 'content principles', The Comparative and International Law Journal of Southern Africa 53 (2020) 1–37.