Unraveling Anomalies: Explaining Outliers with DTOR[⋆]

Riccardo Crupi[1,*], Daniele Regoli[1], Alessandro Damiano Sabatino[1], Immacolata Marano[2], Massimiliano Brinis[2], Luca Albertazzi[2], Andrea Cirillo[2] and Andrea Claudio Cosentini[1]

[1] *Data Science & Artificial Intelligence, Intesa Sanpaolo, Italy*
[2] *Audit Data & Advanced Analytics, Intesa Sanpaolo, Italy*

Abstract

Explaining outliers' occurrence and mechanisms is crucial across various domains, as malfunctions, frauds, and threats require valid explanations for effective countermeasures. With the increasing use of sophisticated Machine Learning techniques to identify anomalies, explaining their presence becomes more challenging. Our proposed Decision Tree Outlier Regressor (DTOR) addresses this challenge by providing rule-based explanations for individual data points using anomaly scores from a detection model. By leveraging a Decision Tree Regressor to compute estimation scores and extracting relative paths, DTOR illustrates its effectiveness across different anomaly detectors and diverse datasets, including those with numerous features.

Keywords

Outlier detection, Explainability, Decision Tree

1. Introduction

Internal audit in the banking sector is crucial for evaluating operational integrity and efficiency, assessing internal controls, risk management processes, and regulatory compliance. Anomaly detection techniques play a vital role in identifying atypical patterns and outliers within data populations analyzed for audit purposes, assisting in risk mitigation and fraud detection. However, ensuring the effective utilization of these techniques requires the ability to explain why certain records are considered anomalies, particularly for internal auditors with limited data analytics expertise [1, 2].

Among various anomaly detection techniques, Isolation Forest [3], One-Class SVM [4], and Gaussian Mixture Models [5] are prominent anomaly detection techniques widely employed in practical applications [6, 7]. These methods leverage diverse mathematical principles to detect anomalies efficiently. However, their interpretability may be limited, necessitating explainable

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

[⋆] The views and opinions expressed are those of the authors and do not necessarily reflect the views of Intesa Sanpaolo, its affiliates or its employees.

[*] Corresponding author.

riccardo.crupi@intesasanpaolo.com (R. Crupi)
0009-0005-6714-5161 (R. Crupi); 0000-0003-2711-8343 (D. Regoli); 0000-0002-1336-2057 (A. D. Sabatino)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

artificial intelligence (XAI) techniques to elucidate model decisions, ensure transparency, and enhance trust in AI-driven decisions [8, 9, 10].

To meet this requirement, we introduce a novel model-agnostic XAI framework specifically designed for anomaly detection in the banking sector. Unlike conventional XAI methods that primarily focus on feature importance (e.g., SHAP and DIFFI [8, 9]), our framework generates easily understandable rules to elucidate model predictions, thereby enhancing transparency and fostering trust in AI-driven decisions. Notable techniques such as LORE, RuleXAI, and Anchors [11, 12, 13] exemplify this approach.

Our approach aims to bridge the divide between interpretability and effectiveness in anomaly detection by offering human-understandable rules that clarify the rationales behind anomalous predictions. Relevant works in this domain include [14] and [15], focusing on online anomaly explanation and providing a survey of explainable anomaly detection methods, respectively. By harnessing rule-based explanations, our XAI framework ensures transparency and accessibility in the decision-making process of anomaly detection models for data scientists, domain experts, and colleagues in the banking industry.

2. Method

Our novel XAI method, inspired by the principles of the Isolation Forest algorithm, takes advantage of the concept of isolating anomalies with minimal cuts in the feature space. To provide clear explanations for anomaly detection decisions, we use decision tree regressors. In our approach, a decision tree regressor is trained to learn the anomaly scores assigned to each data point generated by the Anomaly Detector. Notably, during training, we introduce a weighted loss function that gives a significantly higher weight to the data point under consideration. This weighting scheme ensures that the decision tree regressor prioritizes accurate estimation of the anomaly score for the target data point, thereby improving the interpretability and reliability of the local explanation. After training the decision tree, extracting the path of the datapoint can provide an interpretable rule for the anomaly score (algorithm 1). The implementation of DTOR at the following link can be accessed online [1]

3. Experiments

This section delineates the configurations of three Anomaly Detector models trained on two public datasets and one private dataset from Intesa Sanpaolo (see Table 1), offering explanations using both Anchors and DTOR. The DTOR method and the experiments conducted on public datasets are available in the `GitHub` repository accessible via the following link: https://github.com/rcrupiISP/DTOR.

Algorithm 1: The DTOR approach generates explanations for a given instance.

```
def explain_instance:
    input  : $(x_\mathrm{e}, \hat{y}_\mathrm{e})$: the sample to be explained along its corresponding score from the
             AD;
             $(X_\mathrm{t}, \hat{y}_\mathrm{t})$: a train set and its corresponding scores from the AD;
             $\beta$: training weight associated to $x_\mathrm{e}$;
             h: list of parameters of the decision tree;
    output : a list of rules explaining the instance $(x_\mathrm{e}, \hat{y}_\mathrm{e})$
    $N \leftarrow$ len($X_\mathrm{t}$);
    model $\leftarrow$ DecisionTreeRegressor(h);
    /* append the sample $e$ in the train set                        */
    $\hat{X} \leftarrow$ concat $X_\mathrm{t}$ with $x_\mathrm{e}$;
    $\hat{y} \leftarrow$ concat $\hat{y}_\mathrm{t}$ with $\hat{y}_\mathrm{e}$;
    /* build the array of weights that gives more importance in the
       loss function to the sample $x_\mathrm{e}$                     */
    $\omega \leftarrow$ concat $\mathbf{1}_N$ with $\beta$;
    /* train the DT to the weighted configuration                    */
    model.fit(($\hat{X}, \hat{y}$), sample_wights=$\omega$);
    /* retrieve the path taken by $x_\mathrm{e}$ in the decision tree */
    rule $\leftarrow$ extract_path(model, $x_\mathrm{e}$);
    return rule
```

3.1. Datasets and AD models

Utilizing the novel XAI technique across various datasets aims to assess its effectiveness in explaining different types of anomalies learned by unsupervised Machine Learning models. The chosen Anomaly Detector models include IF, One-class SVM, and GMM [17]. Default parameters were opted for, as the primary objective of this study is to comprehend the explanation rather than optimize a performance metric specific to the dataset problem. Therefore, three distinct models were chosen to reason in different ways. The dataset was partitioned into training and testing sets. Specifically, the test set comprises 50 samples from each dataset, containing both anomalies and normal data points. The anomalies for GMM are defined to represent 5% of the training set, as well as for the isolation forest using the *contamination* hyperparameter set to 0.05. Default hyperparameters were retained for the SVM (kernel: radial basis function, ν = 0.5, representing the upper bound on the fraction of training errors), resulting in anomalies representing about 50% of the training set.

3.2. Rule-based XAI

We explore various explainability techniques, focusing on rule-based explanations due to challenges in interpreting feature importance methods like SHAP and DIFFI, especially with high-dimensional datasets. Initially, Anchors were used to explain the banking dataset, but we

[1] https://github.com/rcrupiISP/DTOR.

Table 1

Summary of Dataset Characteristics: Each item comprises information about a dataset, including its identifier (Dataset), dataset size (# instances), variables count (# columns), and a brief description (Description). The datasets were collected from the UCI Machine Learning Repository [16].

Dataset	# instances	# columns	Description
Banking (B)	100,000	26	Dataset obtained from Intesa Sanpaolo Bank was used for anomaly identification and improved client analysis to discover probable instances of fraud or criminal conduct.
Glass Identification (GI)	214	9	This information comes from the USA Forensic Science Service and includes six different glass kinds, each distinguished by its oxide composition.
Lymphography (L)	148	19	The lymphography dataset was obtained from the University Medical Center, Institute of Oncology, in Ljubljana, Yugoslavia.

found limitations, such as the inability to reason on regression tasks and constraints in model implementation, leading to the development of DTOR. In addition to Anchors and DTOR, we considered LORE and RuleXAI. However, LORE requires extensive hyperparameter tuning, increasing implementation complexity. Additionally, RuleXAI is not actively maintained, with outdated Python library requirements. For future work, we plan to compare DTOR with other explainability techniques.

We adopt a perspective of providing rule-based explanations to Data Scientists, summarizing examples in Table 2 with four key metrics: execution time, coverage, and rule length. For DTOR, we set specific hyperparameters tailored to the banking dataset, ensuring both quantity and quality of explanations. However, a dataset-specific approach is crucial to identifying the optimal anomaly detector and evaluating explanation quality effectively. The hyperparameters for DTOR are carefully chosen, with the `max depth` set to 8, the `min impurity decrease` to 10^{-5}, and the weight β for learning the rule to $0.1 * N$, suitable for unbalanced datasets with anomalies. DTOR estimates the anomaly score rather than a binary output, and the same threshold used in anomaly detection models is applied to determine anomalies. While not detailed here, each rule output by DTOR provides both precision and average anomaly score, enhancing informativeness.

Table 2

Examples of anomaly detection exlpaination on different datasets. The table includes information such as dataset name, example ID, anomaly detection (AD) model used, AD score, whether the instance is predicted by the AD model as an anomaly, coverage percentage, length of the detection rule, the detection rule itself, and the execution time in seconds.

Dataset	Example ID	AD model	AD score	Anomaly	Coverage (%)	Rule length	Rule	Execution time (s)
GI	1	SVM	0.32	True	19	2	Mg $\leq$ 3 AND K $>$ 0	2.1
GI	1	IF	-0.69	True	1.8	3	Si $\leq$ 71.3 AND Na $\leq$ 13.4 AND Na $>$ 12	3.9
GI	1	GMM	-650	True	0.6	2	K $>$ 7 AND Al $>$ 3.37	2.4
B	2	IF	-0.53	True	0.4	2	‘Appraisal time’ > 47 AND ‘Flag proposal’ = True	16
L	3	IF	-0.49	False	8.1	7	‘bl. of lymph. s’ $\leq$ 1.5 AND ‘lym.nodes enlar’ $>$ 2 AND ‘regeneration of’ $\leq$ 1.5 AND ‘dislocation of’ $>$ 1.5 AND ‘changes in stru’ $>$ 1.5 AND ‘by pass’ $>$ 1.5 AND ‘special forms’ $>$ 1.5	3.5

4. Discussion and conclusion

The findings derived from the DTOR algorithm provide significant insights into both anomaly detection and explainability methodologies. Notably, we observed a consistent trend towards shorter explanations for anomalies across various anomaly detection (AD) models and datasets, as evidenced by examples in Table 2, particularly instances with IDs 1 and 2. Conversely, instance ID 3 presents a lengthier explanation. This observation may align with the strategy employed by the Isolation Forest, which aims to isolate anomalies through a minimal number of steps. DTOR, by design, follows a similar path, leveraging the locally trained decision tree to isolate the sample. If the sample is an outlier, it can be easily separated with fewer steps, whereas non-outliers may require more complex separation. It's worth noting that our comparison was conducted against a surrogate classifier model, while our contribution introduces a surrogate regressor model. This distinction allows us not only to provide the rule but also to estimate the anomaly detection (AD) score, offering nuanced insights beyond binary classification tasks. Instance ID 1 showcases three distinct explanations, underscoring the variability introduced by different AD models and potential feature correlations. This phenomenon illustrates the Rashomon effect in explainability [18], where multiple plausible explanations coexist.

Although the execution time for generating explanations typically falls within seconds, it slightly increases for the banking dataset due to its larger sample size, necessitating additional computational resources. Looking ahead, further analysis is warranted to delve into these explanations in depth and compare them with state-of-the-art rule-based explainability techniques. Key metrics such as precision, coverage, and stability will be evaluated to assess the effectiveness of DTOR and its potential advantages over existing methods. For a more detailed analysis on the state of the art, performance and comparison experiments with Anchors, please refer to [19].

References

[1] J. Nonnenmacher, J. M. Gómez, Unsupervised anomaly detection for internal auditing: Literature review and research agenda., International Journal of Digital Accounting Research 21 (2021).

[2] A. Basile, R. Crupi, M. Grasso, A. Mercanti, D. Regoli, S. Scarsi, S. Yang, A. C. Cosentini, Disambiguation of company names via deep recurrent networks, Expert Systems with Applications 238 (2024) 122035. doi:10.1016/j.eswa.2023.122035.

[3] 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.

[4] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, Advances in neural information processing systems 12 (1999).

[5] D. A. Reynolds, et al., Gaussian mixture models., Encyclopedia of biometrics 741 (2009).

[6] Y. Zhao, Z. Nasrullah, Z. Li, PyOD: A python toolbox for scalable outlier detection, Journal of Machine Learning Research 20 (2019) 1–7. URL: http://jmlr.org/papers/v20/19-011.html.

[7] N. Kumar, D. Venugopal, L. Qiu, S. Kumar, Detecting anomalous online reviewers: An unsupervised approach using mixture models, Journal of Management Information Systems 36 (2019) 1313–1346.

[8] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable ai for trees, Nature machine intelligence 2 (2020) 56–67. doi:10.1038/s42256-019-0138-9.

[9] M. Carletti, M. Terzi, G. A. Susto, Interpretable anomaly detection with diffi: Depth-based feature importance of isolation forest, Engineering Applications of Artificial Intelligence 119 (2023) 105730. doi:10.1016/j.engappai.2022.105730.

[10] R. Crupi, A. Castelnovo, D. Regoli, B. San Miguel Gonzalez, Counterfactual explanations as interventions in latent space, Data Mining and Knowledge Discovery (2022) 1–37. doi:10.1007/s10618-022-00889-2.

[11] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems, arXiv preprint arXiv:1805.10820 (2018). URL: https://arxiv.org/abs/1805.10820.

[12] D. Macha, M. Kozielski, Ł. Wróbel, M. Sikora, Rulexai—a package for rule-based explanations of machine learning model, SoftwareX 20 (2022) 101209.

[13] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). doi:10.1609/aaai.v32i1.11491.

[14] R. P. Ribeiro, S. M. Mastelini, N. Davari, E. Aminian, B. Veloso, J. Gama, Online anomaly explanation: a case study on predictive maintenance, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2022, pp. 383–399.

[15] Z. Li, Y. Zhu, M. Van Leeuwen, A survey on explainable anomaly detection, ACM Transactions on Knowledge Discovery from Data 18 (2023) 1–54.

[16] A. Frank, Uci machine learning repository, http://archive. ics. uci. edu/ml (2010).

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

[18] M. G. M. M. Hasan, D. Talbert, Mitigating the rashomon effect in counterfactual explanation: A game-theoretic approach, in: The International FLAIRS Conference Proceedings, volume 35, 2022.

[19] R. Crupi, A. D. Sabatino, I. Marano, M. Brinis, L. Albertazzi, A. Cirillo, A. C. Cosentini, Dtor: Decision tree outlier regressor to explain anomalies, arXiv preprint arXiv:2403.10903 (2024).