Explainable AI for Operational Research: A Defining Framework, Methods, Applications, and a Research Agenda

Koen W. de Bock, Kristof Coussement, Arno De Caigny, Roman Slowiński, Bart Baesens, Robert N Boute, Tsan-Ming Choi, Dursun Delen, Mathias Kraus, Stefan Lessmann, et al.

Title: Explainable AI for Operational Research: A Defining Framework, Methods, Applications, and a Research Agenda

Authors: Koen W. De Bock, Kristof Coussement, Arno De Caigny, Roman Slowiński, Bart Baesens, Robert N. Boute, Tsan-Ming Choi, Dursun Delen, Mathias Kraus, Stefan Lessmann, Sebastián Maldonado, David Martens, María Óskarsdóttir, Carla Vairetti, Wouter Verbeke and Richard Weber

This article was published in the ***European Journal of Operational Research***. Please cite as:

De Bock, K. W., Coussement, K., De Caigny, A., Slowiński, R., Baesens, B., Boute, R. N., , Choi, T.-M., Delen, D., Kraus, M., Lessmann, S., Maldonado, S., Martens; D., Óskarsdóttir, M., Vairetti, C., Verbeke, W. & Weber, R. (2023). Explainable AI for Operational Research: A Defining Framework, Methods, Applications, and a Research Agenda. *European Journal of Operational Research*.

DOI: https://doi.org/10.1016/j.ejor.2023.09.026

ScienceDirect URL: https://www.sciencedirect.com/science/article/pii/S0377221723007294

Explainable AI for Operational Research: A Defining Framework, Methods, Applications, and a Research Agenda

Koen W. De Bock[a], Kristof Coussement[b], Arno De Caigny[b], Roman Słowiński[c,d], Bart Baesens[e,f], Robert N. Boute[e,g], Tsan-Ming Choi[h], Dursun Delen[i,j], Mathias Kraus[k], Stefan Lessmann[l], Sebastián Maldonado[m,n], David Martens[o], María Óskarsdóttir[p], Carla Vairetti[q,n], Wouter Verbeke[e], Richard Weber[r,n]

[a]*Audencia Business School, 8 Route de la Jonelière, 44312, Nantes, France*
[b]*IESEG School of Management, Université de Lille, CNRS, UMR 9221 - LEM - Lille Economie Management, 3 Rue de la Digue, F-59000, Lille, France*
[c]*Poznan University of Technology, Piotrowo 2, 60-965, Poznań, Poland*
[d]*Systems Research Institute of the Polish Academy of Sciences, Newelska 6, 01-447, Warsaw, Poland*
[e]*Faculty of Economics and Business, KU Leuven, Naamsestraat 69, 2000, Leuven, Belgium*
[f]*Southampton Business School, University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom*
[g]*Technology and Operations Management Area, Vlerick Business School, Vlamingenstraat 83, 3000, Leuven, Belgium*
[h]*Centre for Supply Chain Research, University of Liverpool Management School, Chatham Street, Liverpool, L69 7ZH, United Kingdom*
[i]*Center for Health Systems Innovation, Spears School of Business, Oklahoma State University, 370 Business Building, Stillwater, OK 74078, United States*
[j]*Department of Industrial Engineering, Faculty of Engineering and Natural Sciences, Istinye University, Istanbul, Turkey*
[k]*Institute of Information Systems, FAU Erlangen–Nuremberg, Lange Gasse 20, Nuremberg, 90403, Germany*
[l]*School of Business and Economics, Humboldt University of Berlin, Unter den Linden 6, Berlin, 10099, Germany*
[m]*Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Diagonal Paraguay 257, Piso 20, Santiago, Chile*
[n]*Instituto de Sistemas Complejos de Ingeniería (ISCI), Santiago, Chile*
[o]*Department of Engineering Management, University of Antwerp, Prinsstraat 13, 2000, Antwerp, Belgium*
[p]*Department of Computer Science, Reykjavik University, Menntavegur 1, 102, Reykjavik, Iceland*
[q]*Facultad de Ingenieria y Ciencias Aplicadas, Universidad de los Andes, Santiago, Chile*
[r]*Department of Industrial Engineering, FCFM, University of Chile, Santiago, Chile*

Abstract

The ability to understand and explain the outcomes of data analysis methods, with regard to aiding decision-making, has become a critical requirement for many applications. For example, in operational research domains, data analytics have long been promoted as a way to enhance decision-making. This study proposes a comprehensive, normative framework to define explainable artificial intelligence (XAI) for operational research (XAIOR) as a reconciliation of three subdimensions that constitute its requirements: performance, attribution, and responsible analytics. In turn, this article offers in-depth overviews of how XAIOR can

be deployed through various methods with respect to distinct domains and applications. Finally, an agenda for future XAIOR research is defined.

Keywords: Decision analysis, XAI, explainable artificial intelligence, interpretable machine learning, XAIOR

1. Introduction

Through digitization, data have become resources that create value for the operational research (OR) domain (Duan et al., 2020). In turn, abilities to analyze, understand, and leverage data–or analytics competencies–have become critical success factors for OR projects (Conboy et al., 2020; Hindle et al., 2020; Vidgen et al., 2017). According to the official INFORMS definition, analytics reflect a scientific process of transforming data into insights, in ways that support better decision-making (INFORMS, 2015). As the trend of analytics articles published since 2010 indicates (Figure 1), attention to analytics in OR journals was limited until 2014, after which, and possibly in response to calls for papers (Hindle et al., 2020; Mortenson et al., 2015; Ranyard et al., 2015), analytics gained substantially more traction.

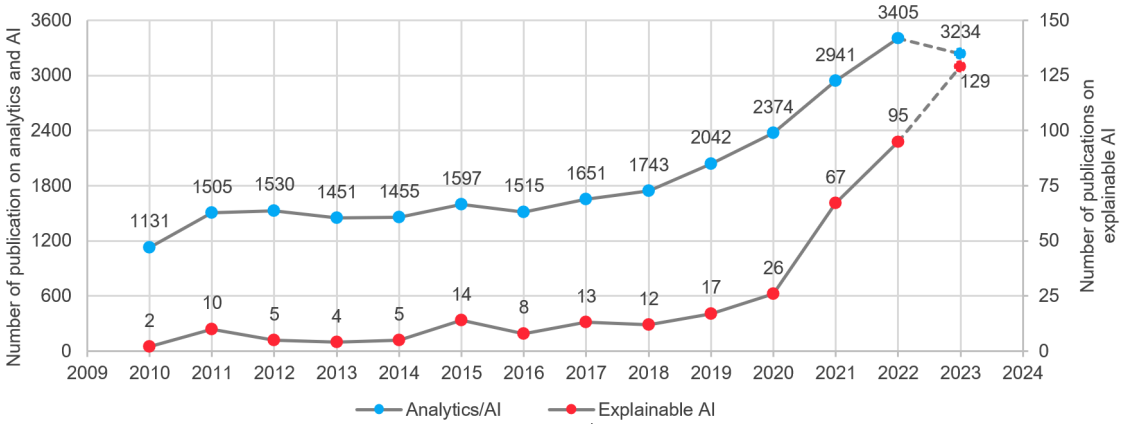


Figure 1: Counts of publications related to analytics/AI and explainable AI published in OR journals between 2010 and 2023 (until September 1st, 2023). Values refer to 30 scientific journals in Clarivate's *Operations Research & Management Science* category with the highest 5-year Journal Impact Factor, according to the 2022 Journal Citation Reports. These papers list analytics- or explainability-related keywords in their titles and abstracts. The search details are available on request.

Growing interest in analytics has also led to a plethora of methodologies and algorithms that claim the ability to solve increasingly complex tasks (Choi et al., 2018). In particular, pattern recognition (Nieddu and Patrizi, 2000), data mining (Olafsson et al., 2008), machine learning (Bengio et al., 2021), and deep learning (Kraus et al., 2020) have become predominant algorithmic paradigms. Such vastly growing complexity makes it difficult to understand the mechanisms by which predictions or decision outcomes emerge. In response, some researchers actively work to develop methods to increase the interpretability and decision transparency of algorithms (Molnar, 2022). For example, Goerigk and Hartisch (2023)

recently presented a framework for interpretable optimization algorithms. Thus, we confront a trade-off between operational performance and decision explainability that is salient for various AI-driven OR applications in fields such as healthcare (e.g., Davies et al., 2003) or finance (e.g., Baesens et al., 2003). Such trade-offs become even more acute when we attempt to account for the interrelated but sometimes conflicting requirements and expectations of internal and external stakeholders, such as:

- *Organizational perspective.* Organization decision-makers strongly stress the importance of adhering to algorithm-provided decisions rather than relying on business logic or intuition (Martens, 2008) and seek the power to take direct strategic, tactical, or operational action based on acquired insights (Coussement and Benoit, 2021).
- *Regulatory perspective.* The General Data Protection Regulation (GDPR) and Digital Markets and Digital Services Acts enforce customer privacy, data integrity, and security principles as legal regulations that curb companies' discretion to leverage and analyze customer data.
- *Ethical perspective.* Ethical considerations involving the environmental impact and fairness of analytics have inspired debates and policies (De-Arteaga et al., 2022; Martens, 2022). Offering the term *responsible analytics*, Vidgen et al. (2020) also proposes a business ethics canvas to help organizations plan and manage their analytical projects ethically. The results of a recent survey by Rao and Greenstein (2022) indicate that 98 percent of decision-making respondents planned to invest in responsible AI in 2022.

To encapsulate all these perspectives, expectations, and requirements, we adopt the term *explainable* artificial intelligence (XAI) herein. It represents an important challenge and opportunity for the OR community, especially considering how the volume of high-quality manuscripts related to explainable AI in the OR journals while growing, is still limited (see Figure 1).

To further define the term *explainable* AI in the OR domain, we ground our paper in the highly relevant review paper by Mortenson et al. (2015), which discusses the origin of the role of analytics in the operations management domain. They argue that OR decision-making must be based on data and evidence rather than on heuristics and intuition. Their view on analytics fits perfectly within the broader evolution of what is called *diaonetic management*. They argue in favor of preserving operations research as a unique management discipline and academic field where analytics is embraced to maximize impact during operational research decision-making. This study proposes an explainable AI framework tailored toward the OR domain that builds further on

1. review papers on analytics published in the OR field, but do not address explainable AI (Choi et al., 2018; Nieddu and Patrizi, 2000; Olafsson et al., 2008) or do so only in a limited fashion (Kraus et al., 2020), such that no OR-oriented definitions of analytics or explainable AI exist; and

2. review papers published outside the OR field that investigate explainable AI solely from a methodological (Barredo Arrieta et al., 2020; Linardatos et al., 2021) or general (Islam et al., 2022) perspective, with a domain-agnostic approach, focused primar-

ily on identifying the subdimensions of explainability and methods, without offering applications relevant for improving OR decision-making.

Noting these gaps in extant literature, we seek to define explainable AI for OR (XAIOR) by a framework which we define in Section 2, where explainable AI is a must-have besides other forms of analytics like performance and responsible analytics when successfully developing and deploying advanced analytics that turn data into insights for improved managerial decision-making. This paper thus takes a broader stance than solely looking at explainable AI but also reviews the most important aspects of performance and responsible analytics. We answer the following four broad questions related to explainable AI for OR (XAIOR) in subsequent sections:

- *What is XAIOR?* In Section 2, we provide a domain-specific definition of XAIOR and introduce a comprehensive, normative framework of XAIOR and its three dimensions: performance analytics (PA), attributable analytics (AA), and responsible analytics (RA).
- *How should XAIOR be implemented?* We present a non-exhaustive overview of critical XAIOR methodologies in Section 3, including experimental design and data selection, feature engineering and data preparation, algorithmic design and choice, post-hoc interpretation methods, and evaluation strategies and metrics.
- *Where should XAIOR be deployed?* In Section 4, we outline key applications of XAIOR, focusing on important OR domains such as forecasting, risk analysis, inventory control, marketing, and supply chain management.
- *What is the future of XAIOR?* We develop an agenda for further research in Section 5.

2. Defining Explainable AI for Operational Research

We define XAIOR as *the conceptualization and application of advanced methods for transforming data into insights that are simultaneously performant, attributable, and responsible for solving OR problems and enhancing decision-making.* This definition underlies a more elaborate framework of XAIOR, as presented in Figure 2. The framework comprises three dimensions, reflecting the three overarching principles that guide the conceptualization of XAIOR. They explain the inner workings of analytical methods and the reasons for any proposed decisions. These three dimensions align with the three types of analytics, as we explain next.

1. *Performance analytics* (PA). In the XAIOR framework, the final solution can make valid, reliable decisions in a scalable manner.

2. *Attributable analytics* (AA). For companies around the world that are building analytical competencies and skills, a need arises to transform their heuristic, experience-based, and often subjective decision-making strategy into a data-driven approach. Therefore, decision-makers need to understand how the methods function in a way that enables them to intuit concrete action points.

3. *Responsible analytics* (RA). Organizations and decision-making instances must comply with legal, ethical, and financial requirements for analytics development.

We zoom in on these definitions, underlying dimensions, target audiences, their organizational priority, and scope of the three types of analytics, as represented in Figure 2.

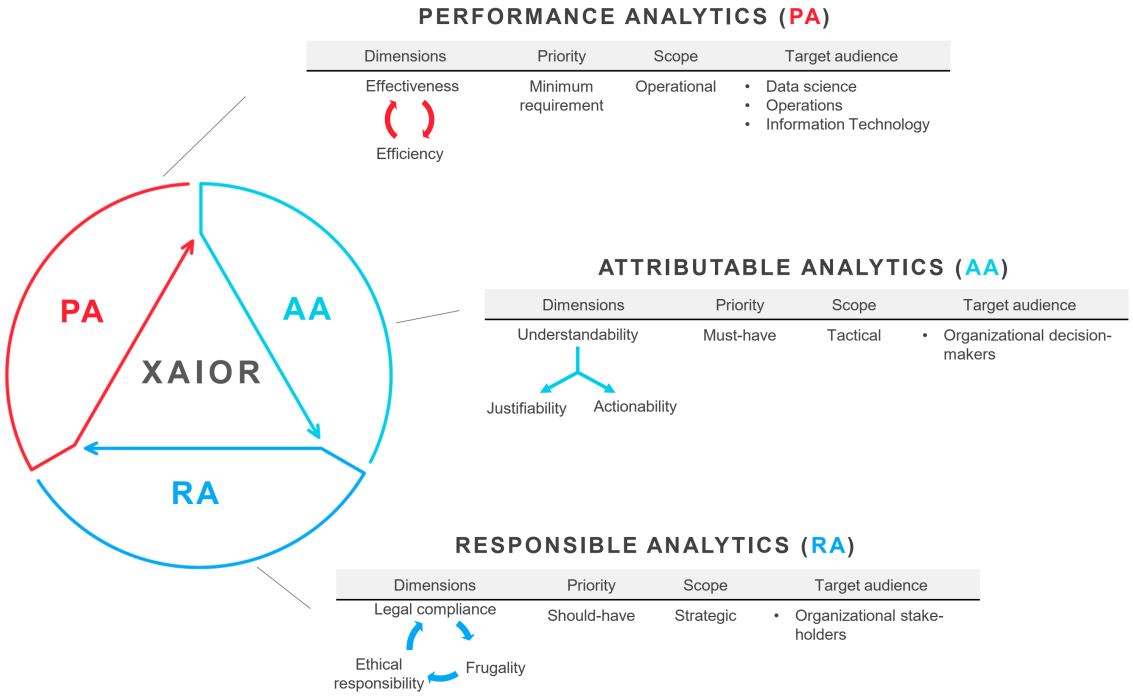


Figure 2: Defining XAIOR

2.1. Performance Analytics

Turning to the first dimension of the XAIOR framework in Figure 2, we define PA as the development or improvement of advanced methods for transforming data into insights to solve operational or managerial problems effectively and efficiently and thus enhance decision-making. The OR community is inherently interested in ways to boost the performance of methods and solutions. A minimum requirement of a XAIOR solution is optimized performance, which often is the responsibility of operational departments like data science, operations, or IT. To achieve such optimization, they mainly focus on improving *effectiveness* and *efficiency*. Effectiveness depends on the proportion of recommended solutions to a given OR problem that are either correct or consistent with the preferences of the decision-maker. An analytical solution is efficient if the run times to produce the solutions do not increase drastically with more observations and variables in the input data set. Coussement and Buckinx (2011), in their evaluation of a new probability-mapping approach for calibration (i.e., the process of adjustment of posterior probabilities output by a classification algorithm toward the true prior probability distribution of the target classes), use a log-likelihood metric to gauge the effectiveness of the calibration approaches. Fleszar (2022) proposes a new mixed-integer linear programming (MILP) model and two heuristics for a bin packing problem with conflicts and item fragmentation. The proposed model produces better and faster solutions than any other benchmark. He assesses the effectiveness of the final proposed model according to average percentage (optimality) gaps and its efficiency as

the average and maximum computation time in seconds. If effectiveness represents consistency between the model and the decision-makers' preferences, it likely requires preference learning from decision examples (Corrente et al., 2013). But efficiency is still important in this context because the decision-maker must be in the loop of the learning process and receive understandable feedback about any model changes without undue delay.

2.2. Attributable Analytics

The second dimension in the XAIOR framework, AA, refers to the development or improvement of advanced methods that can transform data into insights, establish clear reasoning for decision-making, and achieve understandability, justifiability, or actionability. Such analytics are required for any XAIOR solution to bridge the gap with organizational decision-makers. The arrows in Figure 2 suggest the conditional relations among understandability, justifiability, and actionability dimensions; each preceding dimension works as a precondition of each subsequent dimension. Furthermore:

- *Understandability* represents a basic level and refers to the analytical solution's ability to allow human users to understand the method's functioning and the decisions reached. For our study context, we use this term interchangeably with comprehensibility, interpretability, and transparency. When Mitrović et al. (2018) examines which features and feature types to retain to achieve the best solutions from prepaid and postpaid churn prediction models, they showcase not only which features are important but also how they relate to customer churn behavior. Similarly, De Caigny et al. (2018) propose a hybrid, segmented modeling approach based on logistic regression and decision trees that can clarify for marketing managers why customers churn based on insights into the main churn drivers in each segment.
- *Justifiability* pertains to whether the outcomes produced by the method are in line with the intuition of domain experts. It helps ensure that the decision-maker trusts the models developed. With their RULEM method, Verbeke et al. (2017) produce monotonic, ordinal rule-based classification models, which they subject to two justifiability evaluation metrics to determine the degree to which a classification model aligns with domain knowledge, expressed in the form of monotonicity constraints. Błaszczyński et al. (2021) also derives monotonic decision rules from bank data, seeking to explain fraudulent behaviors by customers in a way that makes sense to lenders.
- *Actionability* implies that a method can pinpoint, for the decision-maker, how and where to allocate resources to solve the problem. For instance, da Costa et al. (2023) propose a *dynamic traveling maintainer problem with alerts* that always approximates the optimal policy to act upon when given access to complete condition information to avoid downtime of industrial assets. Another example, Baykasoğlu and Özbakir (2007) proposes MEPAR-miner, a multi-expression program for association rule mining, that can discover effective and actionable IF-THEN classification rules, which in turn improve decision accuracy while also giving problem domain experts a helpful means to extract knowledge from the data and take related action.

2.3. Responsible Analytics

This third and final dimension in the XAIOR framework, RA, is defined as the development or improvement of advanced methods for transforming data into insights in pursuit of compliance with societal expectations, such as ethical, legal, or frugal norms. A recent, growing trend in the OR domain embraces corporate social responsibility (CSR), prompting much more OR research as well. Even if RA is a recommended, rather than a required, dimension of XAIOR solutions, it is extremely beneficial to create solutions that external stakeholders trust. Liu et al. (2022b) cite the impact of CSR leadership in a multi-tier supply chain setting, for example. The increased pressure from public and private organizational stakeholders for firms to comply with ethical, legal, and frugal standards defines the dimensions of this third type of analytics in the XAIOR framework, as we detail further here:

- *Ethical responsibility.* Solving the ethical challenges resulting from the development and deployment of new methods to support decision-making generally requires RA. Growing interest centers particularly on the ethical aspects of method development, often related to method fairness (De-Arteaga et al., 2022). Research in this domain investigates biased decision-making in an effort to understand and prevent it. Fair analytics avoid imposing any discrimination during the method development and decision-making process, regardless of the potential origin of that discrimination (e.g., age, gender, race, sexual orientation, religion). A well-known example involves the gender bias created by the algorithm used to allocate credit limits for the credit card issued by Apple (Satell and Abdel-Magied, 2020). Customers apply online, during which they receive an automated offer for a certain credit limit; it quickly emerged that men were being offered significantly higher credit limits than women, even if they had identical financial positions and credit risks. Among the various articles that investigate and propose measures of algorithm fairness to detect and avoid these unfair decisions, Kozodoi et al. (2022) revisit statistical fairness metrics and empirically investigate their adequacy for credit scoring decisions.
- *Legal compliance.* Advanced methods and analytical solutions must comply with the law. Credit scoring professionals working under BASEL or IFRS 9 regulations must provide clear insights into the probability of default, loss-given default, or exposure at default. Accordingly, Drenovak et al. (2017) proposes a mean capital requirement portfolio optimization method that incorporates the capital requirements for market risk established by BASEL 2.5. When their optimization features the Basel 2.5 formula in the objective function, the results are superior to those obtained using the old (Basel II) formula in stress scenarios. The General Data Protection Regulation (GDPR; implemented May 25, 2018) also establishes that every individual consumer has the right to receive an explanation of any decision made by an algorithm, as well as the right to privacy. Li (2018), seeking to build an online invitation response prediction model, proposes a novel, privacy-friendly mixture cure model with Bayesian networks. The predictive accuracy improves by 24% but still accounts for privacy considerations in relation to the input data.
- *Frugality.* The field of deep learning has become well-embedded in the OR domain, applied to various uses, such as credit scoring (e.g., Stevenson et al. (2021)), order picking (e.g.,

van der Gaast and Weidinger (2022)), and bankruptcy prediction (e.g., Mai et al. (2019)). However, optimizing deep learning architectures requires substantial resources, such that many organizations consider the environmental impacts of their use of analytics. This focus on frugal criteria during method development informs some new ways to build RA. A prominent example comes from transfer learning; a method built for a given application might work for another application, as when De Moor et al. (2022) use a deep Q-network to manage perishable inventories and, rather than starting to train the method from scratch, employ existing heuristics as a starting point to ensure the stability of their transfer learning approach.

3. Implementing XAIOR

In this section, we provide an overview of methodological options that can be deployed to contribute to XAIOR and its three dimensions. Figure 3 depicts the structure of our discussion.

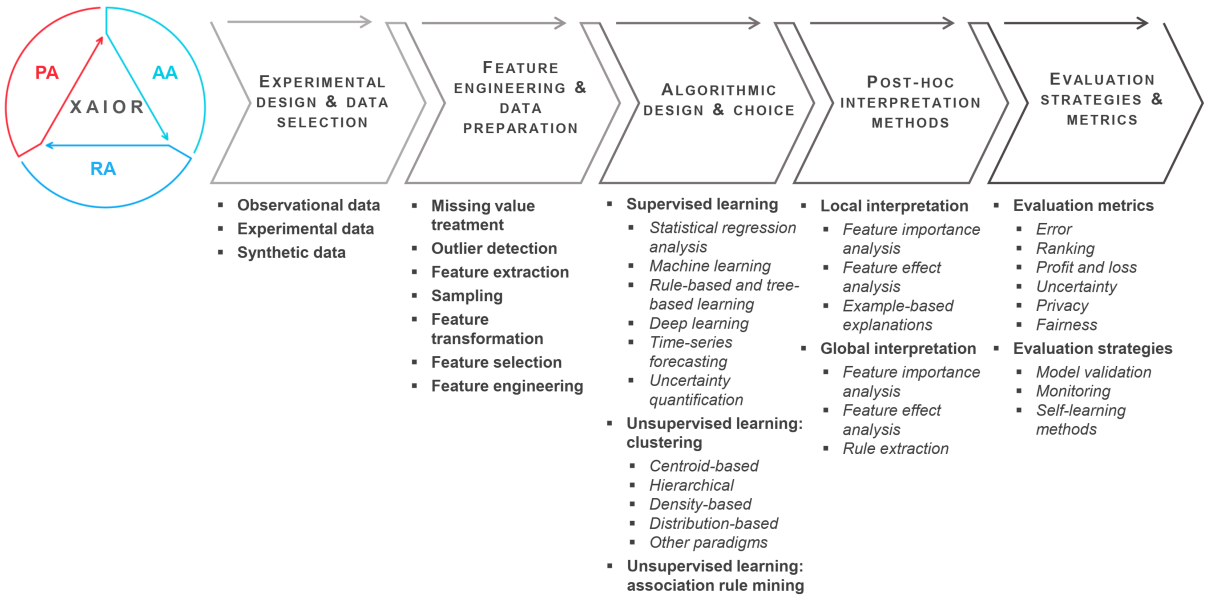


Figure 3: Implementing XAIOR

3.1. Experimental design & data selection

The deployment of analytics in OR includes various types of data, depending on their availability and relevance. In some cases, data scientists depend solely on structured (tabular) data; others leverage unstructured (e.g., images, video, audio, network) data to optimize operational decision-making. The nature of the collected data determines the methodological options available in subsequent steps. For example, unstructured data require adapted data preparation methods (Section 3.2) and learning algorithms (Section 3.3.1).

Beyond the type of data, the experimental design is pertinent; in some cases, an analytical project must gather existing data, but in others, the creation of new data is necessary to support any subsequent analysis. We distinguish three types of data that might be collected:

- *Observational data*. These data are readily available, stemming from the adoption of information systems and technology to administer or automate operations. Although typically abundant and inexpensive, these data also can be biased and insufficiently representative of the population of interest. They also might not be independent or identically distributed. Using observational data to obtain insights and drive decision-making may hinder the achievement of PA, AA, and RA. In credit risk modeling, for example, frequent rejections of loan applications by customers with low creditworthiness create biased observational data sets, such that using those data to develop a credit application scorecard would lead to poor performance and questionable attributability (Banasik et al., 2003). Moreover, the use of observational data can maintain or even reinforce prejudices and thus raise ethical issues. Data that reflect historical job hiring decisions possibly suffer from bias, for example (Tambe et al., 2019).
- *Experimental data*. The active collection of experimental data often involves surveys or experiments, such as randomized, controlled trials that have been purposefully designed and carefully executed. The level of control over the data collection process and the precise considerations typically taken when designing an appropriate experiment imply that experimental data can support straightforward analyses and achieve satisfactory performance. In many settings, experimental data are collected explicitly to explain something (Shmueli, 2010) or establish exact relations between independent and dependent variables rather than to predict or prescribe operational decision-making. Collecting data purposefully also appears critically important for ensuring that analytical models are simultaneously performative, attributable, and responsible. For example, direct feedback loops, as arise when "models directly influence the selection of their own future training data" (Sculley et al., 2015), can be addressed with experimental data. In marketing, it is common practice to create control groups and then collect data to gauge the effect of marketing campaigns (Radcliffe, 2007). Similarly, financial institutions can conduct experiments in which they accept loan applications that normally would be rejected for the purpose of collecting data to improve their credit risk model development (Kozodoi et al., 2020). Obtaining experimental data can be prohibitively costly though, and in some settings, experiments may be infeasible, whether due to practical limitations or ethical considerations. For example, in organ transplant settings, recipients cannot be selected randomly, as would be required for experimental validity, due to medical and ethical considerations (Berrevoets et al., 2020). In such settings, the only data that are available are observational, so more advanced analytical methods are needed.
- *Synthetic data*. A viable alternative to experimental or observational data relies on (semi-)synthetic data, generated by a simulator that needs to be representative to ensure the eventual result of the analysis is useful. This approach is very common in certain OR domains and also expanding, particularly in scientific fields, due to its ability to accommodate privacy concerns and achieve reproducibility. For example, when working with

Table 1: Data preparation method categories and their relations to XAIOR dimensions

Method Category	Examples	XAIOR Support PA	AA	RA
Missing value treatment	Mean/median/mode imputation, (k-) Nearest Neighbors imputation, MICE, VC-DRSA	X		
Outlier detection	Winsorization, Isolation Forest	X		
Feature extraction	PCA, ICA, kernel PCA	X		
Sampling	Oversampling, Undersampling, SMOTE, ADASYN	X		
Feature transformation	Normalization, Standardisation, Box-Cox transformation, Logarithmic transformation, One-hot encoding	X		
Feature selection	Filter-based (e.g., ReliefF), Fair feature selection, Wrappers (e.g., Stepwise procedures)	X	X	X
Feature engineering	RFM, Embeddings, SAFE-ML	X	X	X

Notes: PA = performance analytics; AA = attributable analytics; RA = responsible analytics. References that describe these methods in detail can be found in Table A.1 in Appendix A.

small or necessarily imbalanced data samples, adding (semi-)synthetic data can improve model performance (e.g., SMOTE, ADASYN).

3.2. Feature engineering & data preparation

Data preparation refers to the process of cleaning and transforming raw data prior to processing and analysis. Table 1 lists several data preparation method categories and reveals how they relate to the XAIOR dimensions. Many methods reflect a narrow focus on increasing effectiveness or efficiency; that is, they primarily support PA. Yet the potential for increased performance through effective data preparation is well-acknowledged in OR (Coussement et al., 2017; Crone et al., 2006), particularly in relation to the following:

- *Missing value treatment* with univariate imputation methods, such as mean, median, or mode imputation; multivariate imputation methods, such as nearest neighbor imputation or multiple imputations by chained equations (MICE); or robust learning methods such as VC-DRSA that are able to handle missing values directly (Szeląg et al., 2017).
- *Outlier detection* like winsorization or isolation forests.
- *Feature extraction* methods such as principal component analysis (PCA), independent component analysis (ICA), kernel PCA, and neural network-based embedding methods.
- *Sampling* methods such as random over- and undersampling, as well as related methods that generate synthetic data, such as SMOTE and ADASYN. These options are especially relevant in the case of imbalanced data sets, for which predictive methods often struggle to learn the patterns of the minority class, so resampling methods exert great impacts on the models' performance, as demonstrated using imbalanced data sets in both churn prediction and credit scoring settings (Soares De Melo Junior et al., 2019; Zhu et al., 2018).
- *Feature transformation*, such as scaling through normalizing or standardization, as well as the one-hot encoding for categorical features.
- *Feature selection*, such as ReliefF, a filter-based feature selection method.

In contrast, the creation and selection of some features can serve and enable PA, AA, and RA simultaneously. In particular, we highlight *feature selection* and *feature engineering*. First, feature selection entails identifying and removing redundant features through

filter- or wrapper-based methods, which can improve model performance and facilitate interpretation. Moreover, the removal of sensitive features, such as gender, race, and other features that correlate strongly with them, is critical to fair machine learning. Regarding fair credit scoring, Kozodoi et al. (2022) provides a systematic overview of fairness techniques and compares different fairness processors; they identify nine fairness preprocessing processors. Biswas and Rajan (2021) also demonstrates the impact of various data preprocessing methods, including PCA, SMOTE, and scaling, with the finding that certain methods cause models to exhibit unfairness. For example, data filtering and missing value removal change the data distribution and thereby introduce biases. Unbalanced data demand a means to ensure that all minority classes are adequately represented. Finally, feature selection can support model frugality by reducing the computational costs of analytical learning methods.

Second, feature engineering relies on raw data to enhance the performance, attributability, and responsibility of analytical models. It can be manual, automated, or hybrid:

- *Manual feature engineering* relies on domain knowledge, so it contributes to understandability, justifiability, and actionability. Recency, frequency, and monetary value (RFM) features are often obtained from transactional data, for example (Cheng and Chen, 2009), and product usage trends can be obtained from customer lifetime value modeling (Glady et al., 2009). Óskarsdóttir et al. (2022) augment tabular data about insurance fraud with manually identified features that can characterize clients' network data on the basis of centrality measures and link-based features.
- *Automated feature engineering* is becoming more common, especially for dealing with unstructured data, such as text, images, and network data. The popular feature engineering method TF–IDF counts the frequency of words in a text; word embedding models transform raw text into dense real-valued vectors while encoding the meaning of the words (e.g., word2vec). Powerful language models such as BERT can even gauge customer satisfaction expressed in text samples (Aldunate et al., 2022). Network data also can be automatically feature-engineered to extract structural information about observations in the network in relation to neighbors and positions based on learning of the node embeddings with neural network-based methods, such as node2vec and graphSAGE (Óskarsdóttir et al., 2020; Van Belle et al., 2022).
- *Hybrid feature engineering* methods combine manual and automated feature engineering. Gosiewska et al. (2021) propose SAFE ML: A complex supervisor model that engineers interpretable features that subsequently get added to an easily interpretable model.

3.3. Algorithmic design & choice

In this section, we present methods to deploy XAIOR, distinguishing between methods for supervised versus unsupervised learning.

3.3.1. Supervised learning

Supervised learning methods represent efforts to learn about a model or function that maps input features to one or more outcome variables of interest. In OR, supervised learning techniques are commonly deployed for classification and regression purposes when the target variable is categorical or numerical, respectively. A comprehensive discussion of the many

available taxonomies is beyond the scope of this paper. Instead, we outline six major methodological families that are especially relevant to supervised learning in XAIOR: (1) Statistical regression analysis, (2) machine learning (ML), (3) rule-based and tree-based learning, (4) deep learning, (5) time-series forecasting and (6) methods for uncertainty quantification. Classes (3) to (6) can be characterized as subclasses of (1) and (2) but are discussed separately due to their relevance to OR and XAIOR. In each category, our overview includes both black-box and white-box methods. Briefly, *black-box* methods provide great predictive performance (i.e., focus on PA), but their inner functioning is not readily interpretable. *White-box* or glass-box methods instead, are inherently interpretable and tend to prioritize AA or RA over PA. Table 2 provides algorithm examples from each family as they relate to XAIOR.

Table 2: Supervised learning methods and their relation to the XAIOR framework.

			XAIOR Support		
Method category	**Subtype**	**Method examples**	**PA**	**AA**	**RA**
Statistical regression analysis	Linear	GLM	X	X	
	Nonlinear	Nonlinear regression, GAM, GPR with local explanation	X	X	
	Regularization	LASSO regression, ridge regression, elastic net	X	X	
Machine learning	Classic methods	kNN, SVM, Naive Bayes	X		
	Ensemble learning	Bagging, AdaBoost, RF, RotF, XGBoost	X		
	Hybrid methods	RuleFit, LLM, SRE, PLTR	X	X	
	Domain-specific	ProfLogit, ProfTree, B2Boost, uplift LLM	X	X	
	Fair ML	Fair regression, generative adversarial networks	X	X	X
Rule-based and tree-based learning	Decision rules	RIPPER, CORELS	X	X	
	Decision trees	ID3, CART, C4.5, C5.0	X	X	
	Monotonic rules	Dominance-based Rough Set Approach (DRSA)	X	X	
		VP-DRSA, VC-DRSA, RULEM, AntMiner+	X	X	
	Monotonic trees	MID, VC-MDT, ICT, REMT	X	X	
Deep learning	Architectures	MLP, CNN, RNN, GNN, LSTM, Transformers	X		
	Interpretable ANN	Neural additive models	X	X	
	Post-hoc analysis	Integrated gradients, layer-wise relevance propagation	X	X	
	Bias reduction	Fair adversarial debiasing, FADE	X	X	X
	Transfer learning	BERT, RoBERTa, VGG16	X	X	X
	Knowledge distillation	Adversarial, Multi-teacher, Cross-modal	X	X	X
Time-series forecasting	Econometric and statistical methods	ARIMA, Holt-Winters, Box-Jenkins	X	X	
	Deep learning	LSTM, Temporal fusion transformers	X	X	
	Ensemble learning	Forecast combination, forecast reconciliation	X	X	
	Plots	Residuals, Seasonal decomposition		X	
		ACF/PACF, Confidence intervals		X	
Uncertainty quantification	Regression analysis	GPR, quantile regression		X	
	Ensemble learning	Deep ensembles	X	X	
	Deep learning	Monte Carlo drop-out, Bayesian neural networks	X	X	

Notes: PA = performance analytics; AA = attributable analytics; RA = responsible analytics. References that describe these methods in detail can be found in Table A.2 in Appendix A.

Statistical regression Analysis

Statistical regression methods are popular choices for supervised learning. They rely on strong assumptions about data distributions and functional model forms. As a result, these methods tend to be highly interpretable (i.e., white-box). Notable representatives of this family include generalized linear models (GLMs). Formally, given a set of p predictor

variables $X \in \mathbb{R}^p$, and an outcome variable Y, the model takes the form:

$$g(E(Y|X)) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{1}$$

where g is the link function, and its argument is a linear predictor, reflecting the weighted sum of the input features determined by the coefficients β_j associated with variable X_j, with β_0 as the offset or intercept. In GLM, associations between input features and the outcome variable are additive and linear, which ensures monotonicity and facilitates interpretation. Two influential example configurations of GLMs are *linear regression* for continuous outcome variables ($g(\mu) = \mu$) and *logistic regression* for binary classification ($g(\mu) = logit(\mu)$).

Some notable extensions relax data distribution and linearity assumptions or penalize the loss function in pursuit of stronger generalizability and interpretability, including:

- Nonlinear regression and generalized additive models (GAM), which are viable when the relationship between the variables and the outcome is nonlinear (Hastie et al., 2008).
- Gaussian process regression (GPR) and its variants, such as a GPR with local explanation derived from sample-wise feature weights (Yoshikawa and Iwata (2021)).
- Penalized regression methods impose shrinkage by adding a constraint to the loss function. Prominent examples include LASSO regression, ridge regression, and elastic net regularization.

Machine learning

Machine learning methods learn as effectively and efficiently as possible. As a result, this category features black-box methods that have been widely adopted in OR, such as:

- Classic nonparametric methods such as k-nearest neighbors (k-NN), support vector machines, and naive Bayes;
- Neural networks and deep learning, which are particularly prominent in modern OR applications, as we discuss in Section 4;
- Ensemble learners such as bagging, random forest (RF), rotation forest (RotF), AdaBoost, and extreme gradient boosting (XGBoost).

As illustrated by their emphasis on model accuracy, these black-box methods primarily enhance the PA dimension. In addition, a wide array of purposefully established white-box machine learning methods exists, which are highly interpretable and seek AA or RA explicitly, including:

- Hybrid versions, such as a logit leaf model (LLM) (LLM; De Caigny et al., 2018) that combines clustering and classification, rule ensembles (RuleFit) and penalized logistic tree regressions (PLTR) that reconcile rule-based learning and regularized regression, and spline-rule ensembles (SRE; De Bock and De Caigny, 2021) that combine rule ensembles with penalized cubic regression splines to enhance performance as well as understandability.
- Context-specific or domain-optimized methods, such as cost-sensitive ensemble learning (De Bock et al., 2020), profit-driven classifications, or uplift models (Devriendt et al.,

2021). Instead of optimizing a statistical measure of model fit, methods such as ProfTree or ProfLogit maximize the average profit that drives the implementation of a classifier.
- Fair machine learning methods (De-Arteaga et al., 2022; Mehrabi et al., 2021).
- Rule-based and tree-based methods, which we discuss separately, due to their particular relevance to OR (see Section 3.3.1).

Rule-based and tree-based learning

Decision rules and trees are frequently used in OR, particularly in multi-attribute classification problems. Multiple reviews summarize these models and their learning (e.g., (Bodria et al., 2021)). They are suitable for both PA and AA. According to Semenova et al. (2022), they also provide simple, accurate models that can be superior to accurate, more complex models in terms of understandability. We listed some notable examples of these algorithms in Table 2.

For this section, we focus specifically on decision rules and trees learned from *ordinal* data. In OR, the assessment of alternative decisions usually involves multiple attributes with ordinal or cardinal scales. For this reason, in OR, data submitted to analytics are usually ordinal, which explains our focus. Multi-attribute assessments entail a multidimensional decision problem and involve a dominance relation in the set of alternative decisions. This relation is the only objective information derived from the statement of a multidimensional decision problem. The dominance relation makes, however, a weak partial order in the set of alternatives, thus leaving some alternative decisions incomparable, especially if assessments across multiple dimensions are conflicting (i.e., improvement to one dimension causes deterioration in others). Incomparability prevents unambiguous recommendations for optimization, classification, or ranking, which are the main classes of decision problems considered in OR. Thus, the decision-aiding methodologies developed within OR mainly focus on aggregating multiple dimensions into a *preference model*, which makes the alternatives more comparable in light of users' preferences.

Modeling users' preferences is essential to decision-aiding in OR. In the framework of ordinal data analytics it proceeds through learning preference patterns from holistic preference information about users' judgments. The preference patterns explain users' past decisions and predict future ones. They imply a monotonic relationship between conditions and decisions (e.g., an alternative that is better on considered attributes is higher in quality).

The best-known preference patterns are *monotonic decision rules* (Greco et al., 2001), composed of logical statements that relate conditions on particular attributes with some decision, such as, "if $g_i(a) \succeq r_i$ & $g_j(a) \succeq r_j$ & ... $g_k(a) \succeq r_k$, then alternative $a \rightarrow$ Class t or better" for classification, or else, "if $g_i(a) \succeq^{\geq h(i)} g_i(b)$ & $g_j(a) \succeq^{\geq h(j)} g_j(b)$ & ... $g_p(a) \succeq^{\geq h(p)} g_p(b)$, then $a \succeq b$" for best choice or ranking, where $\succeq$ is a weak preference relation; $r_i, r_j, \ldots, r_k$ are threshold values on selected attributes $\{g_i, g_j, \ldots, g_k\} \subseteq \mathcal{G}$ induced from data; $\mathcal{G}$ is the set of all considered attributes; $\succeq^{\geq h(\cdot)}$ is a weak preference relation with intensity in degree at least $h(\cdot)$; and $h(i), h(j), \ldots, h(p)$ are degrees of preference intentsity for cardinal attributes $\{g_i, g_j, \ldots, g_p\} \subseteq \mathcal{G}$, also induced from the data.

In addition, the rule model of preferences has been compared at an axiomatic level with two earlier preference models (Słowiński et al., 2002):

- Multiple attribute utility theory (MAUT) (Keeney and Raiffa, 1979), according to a value function that assigns, to each alternative $a \in \mathcal{A}$, a real value, such as the weighted sum of performances $U(a) = \sum_{i \in \mathcal{G}} k_i \times g_i(a)$, or a more general additive function $U(a) = \sum_{i \in \mathcal{G}} u_i[g_i(a)]$, where u_i are marginal value functions, or non-additive integrals that can handle interactions among attributes, such as the Choquet integral for cardinal attributes or the Sugeno integral for ordinal attributes (Grabisch, 1996).
- Outranking models (Roy, 2005) that use systems of binary relations, including the outranking relation $S = \{\sim, \succ^w, \succ^s\}$, where $\sim$ means indifference, $\succ^w$ indicates weak preference, and $\succ^s$ is strong preference, such that relation $a \succeq b$ reads: "alternative a is at least as good as alternative b."

The comparison then establishes that the rule model requires the weakest axioms, which means that the value function or outranking model can represent particular preferences if and only if the rule model can (see also Greco et al. (2004)). Moreover, the rules identify values that drive users' decisions; each rule represents an intelligible scenario of a causal relationship between performance on a subset of attributes and a comprehensive judgment.

The rules are induced from preference information obtained from users, in the form of decision examples (i.e., users' past judgments or judgments elicited by request). Yet decision examples may be inconsistent with the dominance principle that is commonly accepted for multi-attribute decision problems. Such inconsistency arises, e.g., in the case of ordinal classification, if alternative a has been assigned to a worse decision class than alternative b, but a is at least as good as b on all the considered attributes (i.e., a dominates b). Inconsistency has many sources, including missing attributes in the descriptions of the alternatives, unstable preferences, or conflicts between users. Handling these inconsistencies is critical to preference learning; they cannot be dismissed as noise or error that needs to be eliminated from data, nor should they be amalgamated with consistent data through the use of some averaging operators. They need to be identified and presented as uncertain patterns.

The concept of a rough set (Pawlak, 1982) is useful for handling data inconsistency, though originally, it was limited to inconsistency with respect to the indiscernibility principle. To deal with inconsistencies pertaining to the dominance principle, as are typical for ordinal data, Greco et al. (2001) generalized the original rough set concept by substituting the indiscernibility relation with a dominance relation in a rough approximation of preference-ordered decision classes. The resulting methodology, the dominance-based rough set approach (DRSA), is able to infer users' preferences in the form of monotonic decision rules induced from data structured by the dominance-based rough approximations.

Depending on which classification examples support the induced rules, they can be characterized by different values of the adopted interestingness measures. Greco et al. (2016) consider some recommended rule interestingness measures according to Bayesian and likelihood confirmation assessments. The interestingness measures then can help to classify new alternatives according to whether those alternatives are matched by no rule, exactly one rule, or several rules (even if they are contradictory). Such a classification scheme has been proposed by Błaszczyński et al. (2007).

Two parametric versions of DRSA also have been proposed, which relax the original

definition of rough approximations in various ways. They include variable-precision DRSA (Inuiguchi et al., 2009) and variable-consistency DRSA (Błaszczyński et al., 2009). Statistical interpretations of these two parametric DRSA from an empirical risk minimization perspective (as is typical of machine learning) are available from Kusunoki et al. (2021). A stochastic DRSA model also has been presented by Kotłowski and Słowiński (2008) and Kotłowski et al. (2008).

Algorithms for inducing decision rules from rough approximations include a *minimal-cover* strategy that offers a minimal set of rules that represents the users' preferences in the most concise way (Błaszczyński et al., 2011). A recent trend integrates several rule classifiers, called base classifiers, into ensembles or committees of classifiers (Kotłowski and Słowiński, 2009). Various methods of generating differentiated base classifiers for their integration into the ensemble classifiers were proposed. The best known are bagging (Błaszczyński et al., 2010) and boosting (Dembczyński et al., 2010), which modify the set of alternatives by sampling or weighting particular examples and use the same learning algorithm to create base classifiers.

Ordinal data analytics involving monotonic decision rules induced by DRSA for other multi-attribute decision problems was described in (Słowiński et al., 2020). For a characterization of other methods of learning monotonic decision rules and trees, see Cano et al. (2019).

Deep learning

In the past decade, artificial neural networks (ANN) have achieved promising results for various OR applications, often outperforming traditional ML models in terms of predictive performance (Kraus et al., 2020). In particular, the flexible design of deep learning architectures supports the derivation of models that process input data, especially unstructured data, in a natural way. Some well-established architectures include the following:

- *Convolutional neural networks (CNNs)* exploit spatial relations across adjacent inputs, such as pixels in images (He et al., 2022).
- *Recurrent neural networks (RNNs)* sequentially process data and keep a memory of processed time series, which is commonly needed in finance (Krauss et al., 2017).
- *Graph neural networks (GNNs)* naturally process graph data.
- *Transformer neural networks* learn to attend to specific parts of sequential data, such as in text (Kriebel and Stitz, 2022).

Within the XAIOR framework, the design of highly complex neural networks with millions or billions of parameters is closely linked to PA. Yet certain characteristics of neural networks also can be used to account for AA, such as (1) the differentiability of common neural networks, which supports assessments of gradient information; (2) their stacked architectures, so it is possible to follow the propagation of information through the model, layer by layer; and (3) the general idea of connecting neurons, which implies architectures that are intrinsically interpretable.[1], as we discuss in Section 3.4.

[1]Model-agnostic interpretability methods, such as SHAP, are also widely used to analyze deep learning models

We present three strategies that exploit these characteristics to gain insights into the functioning of deep learning models and contribute to AA. First, the effect of a change in an input feature on the ANN output can be determined by using information about the partial derivatives of the model with respect to the inputs. Let f be the optimized ANN and $x = (x_1, \ldots, x_p)$ be its inputs. Then the partial derivative $\frac{\delta f}{\delta x_i}(\hat{x}_i)$ describes the rate of change for model input x_i at feature value $\hat{x}_i$. The assumption that inputs with large partial derivatives are the ones most relevant for the ANN output does not hold though; the effects of inputs quickly saturate. That is, the effects of inputs on the neural network output increase sharply within a small range of inputs but remain constant outside of this range. As a remedy, integrated gradients can assess the effect of an input feature on the ANN output (Kosasih and Brintrup, 2022). The partial derivatives between a base vector b (e.g., black image, zero-valued vector) and the actual input vector $\hat{x}$ get integrated, such that

$$(\hat{x}_i - b_i) \int_{\alpha=0}^{1} \frac{\delta f(b + \alpha\,(\hat{x} - b))}{\delta x_i} d\alpha. \tag{2}$$

Second, it is also possible to exploit the layered architecture of ANNs to explain model predictions and propagate the effects from the model output to the inputs. This approach helpfully propagates more interpretable patterns, usually learned by neurons in later layers, to the input. Notably, layer-wise relevance propagation sends the model output backward to the inputs, using different propagation strategies for early, middle, and later layers in the ANN. Relevance R is defined and initialized with the model output's activation, such that $R_{\text{total}} = f(\hat{x})$. For the layer connected to the output, relevance then gets distributed among the r neurons, with respect to their activation, $a_1, \ldots, a_r$, and the weights that connect the neurons, $w_{\text{out},1}, \ldots, w_{\text{out},r}$. For neuron k, relevance can be computed as:

$$R_k = \frac{|a_k\, w_{\text{out},k}|}{\sum_{j=1}^{r} |a_j\, w_{\text{out},j}|}, \tag{3}$$

which describes the share of information that each neuron adds to computing the model output's activation. By propagating relevance from the model output to its input, layer by layer, the data scientist obtains relevance scores for the inputs, which then explain the model prediction.

Third, another option is to design neural networks in such a way that their intrinsic functioning can be assessed easily without any additional post hoc analyses. The resulting family of models is called generalized additive models, and they have been applied successfully in OR (Djeundje and Crook, 2019). By removing interactions between input features, these models take the following form:

$$f(x) = f_1(x_1) + f_2(x_2) + \ldots + f_m(x_m), \tag{4}$$

where each f_i describes a subneural network that maps the i-th input to the output. These neural additive or explainable neural networks (Yang et al., 2021) offer the advantage of eliminating the need for a post hoc analysis because the effect of an input x_i on the output is fully described by the subneural network f_i.

With its clear focus on PA, research into deep learning only partially addresses challenges pertaining to RA. However, the versatile design of neural network architectures and the optimization problem can contribute to addressing such challenges. For example, Kozodoi et al. (2022) shows that adversarial debiasing using neural networks can increase fairness in credit scoring. Adversarial debiasing involves training a neural network with the following objective (Zhang et al., 2018):

$$\min_f L(f(x), y) + \alpha \, \text{PI}, \tag{5}$$

where $L(f(x), y)$ is a general loss function, α denotes regularization strength, and PI represents the prejudice index that quantifies the degree of unfairness. Incorporating additional regularization terms can be appealing, but they are soft constraints and cannot fully prevent unwanted bias in model learning.

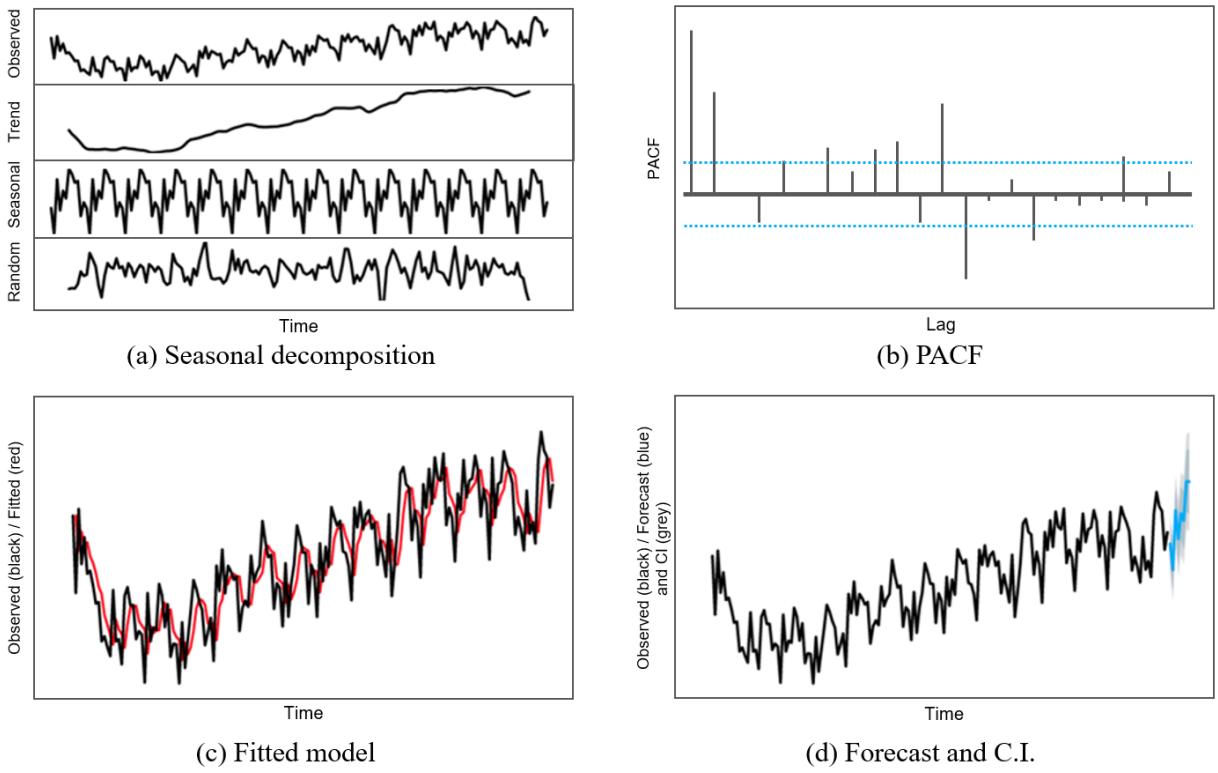
Optimizing powerful deep learning models requires large amounts of data and many optimization steps due to the high number of parameters. This consumes significant energy, resulting in notable environmental impacts. To evaluate neural networks with frugal criteria like emissions, some firms have adopted new RA approaches. For example, for transfer learning, an already optimized neural network model serves as the starting point, which gets optimized for the desired task. Thus, Kriebel and Stitz (2022) use a so-called language model as a starting point and then optimize it to predict credit defaults from user-generated text in peer-to-peer lending. Such transfer learning reduces optimization time, and environmental impacts, and improves model performance by leveraging knowledge from the original language model to solve specific problems more accurately.

In a similar vein, researchers propose techniques to reduce the resources required to evaluate and infer deep learning models, such as after deployment. Knowledge distillation implies transferring knowledge from a large model to a simpler one. Although large models (such as deep learning models) have a much greater knowledge capacity than small models, they might not be fully utilized, such that a small model could provide similar predictive performance at a much lesser computational cost. In addition to being less expensive to evaluate, smaller models can be deployed on less powerful hardware (e.g., mobile phones).

Time-series forecasting

Time-series forecasting is a particular form of regression in which the covariates are lagged variables of the outcome. This intrinsically interpretable task, even when using machine learning models, plots the main series along with the fitted model and the forecast in a two-dimensional chart (Hyndman and Athanasopoulos, 2021). Beyond the fitted model and the forecast, seasonal decomposition plots, partial autocorrelation function (PACF) plots, and confidence intervals represent useful visualization tools for decision-making, too (Hyndman and Athanasopoulos, 2021). Figure 4 illustrates some examples, revealing, for example, why the seasonal decomposition and PACF plots are useful tools for understanding seasonal and autoregressive patterns, whereas the confidence intervals of the forecast give insights into the accuracy and uncertainty of the prediction.

Time-series forecasting methods include:



(a) Seasonal decomposition (b) PACF

(c) Fitted model (d) Forecast and C.I.

Figure 4: Four different visualization methods for decision-making in time-series forecasting using the NYC-Births data set

- *Econometric and statistical methods*, such as ARIMA and Box-Jenkins. Variations such as seasonal autoregressive integrated moving average (SARIMA) and triple exponential smoothing model from the Holt-Winters family (Hyndman and Athanasopoulos, 2021) can depict three main aspects of a series: trend, seasonality, and autoregressive patterns.
- *Machine learning methods* capable of modeling nonlinear interactions. In terms of visualization capabilities, they are limited by their inability to derive confidence intervals and the variables' contributions. Therefore, recent forecasting literature seeking to address these limitations proposes Bayesian models to account for uncertainty and to model confidence intervals (Zeng and Li, 2021). Alternatively, novel regularization strategies, such as the group LASSO, might be designed to identify the contributions of lagged variables in high-dimensional time-series (Nicholson et al., 2020).
- *Ensemble methods*, which combine multiple forecasting models, are also used increasingly to enhance forecast accuracy (Cang and Yu, 2014; Kang et al., 2022; Winkler and Makridakis, 1983), taking forms such as:
  - *Forecast combination.* Multivariate, often high-dimensional, time-series settings can extract time-series features that determine the weights of candidate forecasting models in a subsequent combination step (Ma and Fildes, 2021; Montero-Manso et al., 2020).
  - *Forecast reconciliation.* This method combines multivariate time-series forecasting with

a hierarchical dependency structure, such as sales data related to products in the same category or consumption/demand data grouped by geographic region (Panagiotelis et al., 2021).
  - *Other approaches.* Some studies relax requirements for the relatedness of the time series and demonstrate the superiority of a global model to forecast multiple time series over developing local, time-series-specific models even when the set cannot be considered related (Montero-Manso and Hyndman, 2021). Given an effort to extract information from one set of time series to forecast another set, it is worth noting the connections between forecast reconciliation and transfer learning, which has become a de facto standard for natural language processing (NLP) and computer vision (Lecun et al., 2015).
- *Deep learning* methods, and in particular, architectures such as transformers and long short-term memory (LSTM) enhance forecasting due to their ability to model data as a sequence of information. Understandable approaches depict and interpret attention weights in ways that produce valuable decision-making insights (Ding et al., 2020).
- *Probabilistic methods* for forecasting estimate the full conditional distribution of the target variable rather than a specific moment (e.g., conditional mean) (Gneiting and Katzfuss, 2014). They extend the confidence intervals to support improved decision-making (AA and actionability). These models often rely on Bayesian methods (Frazier et al., 2019), though more recently, approaches based on deep learning, such as DeepAR, also have been proposed.

Uncertainty quantification

A special category of methods that we opt to identify separately is suitable for uncertainty quantification. Beyond generating predictions, such methods quantify the confidence of models in the prediction. The enhanced interpretability and actionability of these estimates contribute to AA. As seen in Table 2, these methods emerge from some of the methodological families outlined above. Examples are Gaussian process regression, quantile regression, ensemble learning, and practices such as Monte Carlo drop-out found in deep learning.

3.3.2. Unsupervised learning

The objective of unsupervised learning is to recognize patterns or structural properties in data without an associated label. An aspect of this process involves *clustering*, which aims to group similar observations in the same cluster whereas dissimilar observations should belong to different clusters. In this subsection, we outline some categories of key algorithms and how they support XAIOR (Table 3). A more comprehensive overview is available from (Saxena et al., 2017). Classical clustering methods include:

- *Centroid-based clustering* minimizes the distances between observations and class prototypes. Examples are k-means clustering and partitioning around medoids (PAM).
- *Hierarchical clustering* entails two major variations: agglomerative and divisive (Saxena et al., 2017). The former starts with each observation as its own cluster and iteratively merges clusters until one cluster emerges, comprised of the entire set of observations. The latter approach starts with the entire data set as one cluster and divides them in each

Table 3: Unsupervised learning methods and their relation to the XAIOR framework.

		XAIOR Support		
Method category	**Method examples**	**PA**	**AA**	**RA**
Centroid-based clustering	K-means, PAM		X	
Hierarchical clustering	Agglomerative clustering		X	
	Divisive clustering		X	
Density-based clustering	DBSCAN, OPTICS		X	
Distribution-based clustering	GMM/EM		X	
Other clustering paradigms	SVC		X	
	Dynamic clustering	X	X	
	Semi-supervised clustering		X	X
	Clustering under uncertainty	X	X	X
	Subspace clustering	X	X	X
Association rule learning	A priori, FP-growth	X	X	

Notes: PA = performance analytics; AA = attributable analytics; RA = responsible analytics. References that describe these methods in detail can be found in Table A.3 in Appendix A.

iteration. Several enhancements to hierarchical clustering have been proposed (see, e.g., Saxena et al., 2017).

- *Distribution-based clustering*, which assumes a model that can describe the observations' distributions and optimizes the respective model's parameters, such as the EM algorithm for estimating a Gaussian mixture model (GMM) (Xu and Wunsch, 2005).
- *Density-based clustering* includes methods such as DBSCAN and OPTICS, which assign instances in high-density spatial areas to clusters.
- *Support vector clustering (SVC)*, perhaps the most relevant method for maximum margin clustering, determine support vectors situated on the margin of each cluster.

Clustering supports AA by nature. Centroid-based clustering methods clarify a potentially huge set of high-dimensional observations by determining clusters' centers. Such information is very useful for customer segmentation; the centers provide an interpretation of the respective segments. Clustering paradigms that have more particular relevance for PA and RA include:

- *Clustering under uncertainty*, such as probabilistic, fuzzy, possibilistic, rough, and granular clustering, as reviewed by (D'Urso, 2017).
- *Dynamic clustering*, which reveals changes to a cluster solution, is useful when timely reactions are necessary. Several methods employ dynamic clustering cycles (Saltos et al., 2017), such that a methodology to update a cluster solution augments the base clustering algorithm. A taxonomy of dynamic clustering is presented by (Peters and Weber, 2018).
- *Semi-supervised clustering* uses background knowledge to guide otherwise unsupervised learning processes used in traditional clustering. Adding constraints leads to *constrained clustering*. For example, in geographical information systems (GIS), certain geographical elements, like streets and rivers, may not be assigned to the same cluster, regardless of their proximity in the feature space (Ruiz et al., 2010).
- *Subspace clustering* determines clusters in subspaces of the original data space, so it provides a means to treat high dimensionality effectively, ensure the interpretability of results, and provide scalability and usability (Agrawal et al., 2005).

Subspace clustering and dynamic clustering are particularly relevant for PA, given their focus on efficiency. Subspace clustering in particular creates more efficient predictive models because it uses fewer variables per cluster (Wang et al., 2015). Dynamic clustering updates a cluster solution iteratively, instead of starting each time from scratch, which increases efficiency. For example, dynamic rough-fuzzy support vector clustering (dynamic RF-SVC) provides a base method within a dynamic clustering cycle to explain changing cluster structures. Adequate treatment of uncertain phenomena has an especially important role in dynamic clustering because the changes lead to uncertainty. The dynamic RF-SVC can detect modifications such as the creation, elimination, movement, merging, and splitting of clusters, as well as the traceability of outliers (Saltos et al., 2017).

Unsupervised learning also can accomplish RA. First, because semi-supervised clustering adds constraints, it represents a means to include explicit ethical or legal considerations. The detection of fake reviews represents such an application (Rathore et al., 2021). Second, subspace clustering can preserve privacy, in that it identifies segments without using all available features (Wang et al., 2015). Community detection among the social networks of criminals, combined with topic modeling of victims' narratives, could offer useful hints for prosecution. An example is the methodology developed by Troncoso and Weber (2020), which has been applied to detect criminal associations within a network of suspects. Third, clustering using uncertainty modeling deployed for outlier detection could address ethical issues, such as unfair exclusions of minority populations (Deepak and Abraham, 2021), as well as legal concerns, such as fraud detection (e.g., Carcillo et al., 2021).

It should be noted that unsupervised learning includes tasks beyond clustering. One is *association rule mining* which aims to uncover relationships across variables. Notable algorithms for this task are the Apriori and FP-growth algorithms. Another problem addressed by unsupervised learning is *dimensionality reduction*, which includes feature extraction methods such as PCA and ICA (see Section 3.2).

3.4. Post-hoc interpretation methods

Post-hoc explanation methods are meant to explain the predictions of existing supervised learners. Such methods contribute to AA directly since they aim to make model predictions and decisions understandable. This understandability enables other subdimensions of AA as well as RA. Any explanation has three defining components: explaining (a) a prediction (a score or a decision), (b) made by a prediction model, (c) on some set of instances.

- Starting with what is being explained: a prediction score or a predicted class. Data scientists often operate in environments where the threshold applied to convert prediction scores to decisions is dynamic. Consider, e.g., credit scoring. During uncertain times like the start of the COVID-19 pandemic or during a war, banks will become more cautious in their lending practices. This results in them lowering the threshold for credit scores that determine whether someone is approved or rejected for a loan. There will, therefore, be a preference of data scientists to evaluate and explain prediction scores rather than decisions (which come from applying a threshold to a prediction score), which sheds light on the popularity of post-hoc methods such as LIME and SHAP, which are explained in more detail next. However, this does not necessarily reflect the requirements of the end-users:

a loan applicant is more interested in understanding why credit was denied rather than explaining why a certain score was given.

- The second defining component refers to the prediction model itself. Some explanation methods use this model as a black-box model that, given an input, provides an output, while other methods explicitly make use of a particular inner structure or defining characteristics of the prediction model, such as the architecture of a neural network, the support vectors in an SVM, or the gradient of the scoring function. The former, model-agnostic ones, can easily be applied to any black-box model. In contrast, the latter, model-specific ones, are tailored towards specific models, often with a superior performance yet less broad applicability.
- The last defining component looks at whether an individual prediction is to be explained, leading to instance-based or local explanation methods, or whether an explanation is needed over the complete set of predictions, known as global explanation methods. A taxonomy of methods according to this component has been proposed by Martens (2022), which also looks at the dimension of what the explanation looks like: does it provide the importance of features, does it provide plots of feature values, or does it provide rules. Before detailing the primary approaches, note the irony of these post-hoc explanation techniques: to explain complex models, we are adding more complex algorithms to explain the predictions made by the initial models. This irony has led to some researchers arguing for the importance of inherently comprehensible models (Rudin, 2019), which conflicts with the use of well-performing black-box models as trained by popular deep learning and ensemble methods.

The following overview primarily involves methods designed to explain supervised learning models. Many of them originate from, or build upon, the broader literature stream on *sensitivity analysis* (Borgonovo and Plischke, 2016), aimed at generating insights in model mechanisms and output iton response to changes in model inputs.

Table 4: Post-hoc explanation methods and their relation to the XAIOR framework

			XAIOR Support		
Scope	**Method category**	**Method examples**	**PA**	**AA**	**RA**
Local	Feature importance analysis	LIME, SHAP, LRP		X	
	Feature effect analysis	ICE		X	
	Example-based explanations	Counterfactuals, Anchors		X	
Global	Feature importance analysis	PI, SHAP, Sobol' indices , Shapley effects		X	
	Feature effect analysis	PDP		X	
	Rule extraction	RIPPER, ANN-DT, DeepRED	X	X	

Notes: PA = performance analytics; AA = attributable analytics; RA = responsible analytics. References that describe these methods in detail can be found in Table A.4 in Appendix A.

3.4.1. Local explanation methods

LIME, SHAP, LRP, and ICE are four popular explanation methods that explain an individual instance's prediction score.

- *LIME* (Ribeiro et al., 2016) does so by creating a set of artificial data points around

the instance to be explained and having the black-box model provide a prediction score. Next, a linear regression model is trained on this data. As we now have an inherently interpretable model, the coefficients of this linear model are shown, ranked by their absolute value, to indicate the most important features for that instance's prediction score.
- *SHAP* further expands on this by ensuring that the importance weights correspond to Shapley values (Lundberg and Lee, 2017).
- *Layer-wise Relevance Propagation (LRP)* (Binder et al., 2016) similarly provides a weight to each input but is made explicitly for explaining predictions made by artificial neural networks on image data. The result is a heat map, where the color of each pixel indicates its importance for the prediction.
- *Individual Conditional Expectation (ICE)* plots indicate how the prediction score changes as the value of a certain feature is changed while keeping the values for the other features constant (Goldstein et al., 2015).

Notice again that the previous instance-based methods explain a prediction score, not a decision. A *counterfactual explanation* of a classification of a data instance provides an irreducible set of evidence present in the data instance to be explained such that removing that evidence would change the decision (Martens and Provost, 2014). For example, an explanation of why a Facebook app user in the US is targeted for a display advertisement for the Democrat party could be: *If the user would not have liked the Facebook pages NBC, Barack Obama, and Greenpeace, then the user's inferred political leaning would change from democrat to neutral*. Chen et al. (2017) argue that these counterfactual explanations can help decide which Facebook likes should be cloaked to suppress the prediction.

Terminology-wise, the counterfactual is the data instance that leads to a different classification (for example, a resume with certain words removed), while the explanation is the difference between the data instance to be explained and the counterfactual (for example, the words to be removed in the resume). Counterfactuals had been used in philosophy for a long time (Schock, 1962) and were introduced in the predictive modeling domain by Martens and Provost (2014) for textual data and further popularized by Wachter et al. (2017) for tabular data. The counterfactual approach has gained lots of attraction, as it explains a decision, which arguably is what end-users most often care about, and does so without disclosing the entire model (Barocas et al., 2020).

3.4.2. Global explanation methods

Global explanation methods explain a model's prediction over an entire data set. A commonly used approach is to look at what features are most 'important' for the model prediction. Breiman (2001) arguably first popularized these *permutation-based feature importance scores*, or simply permutation importances (PI), in his seminal paper on random forests. Randomly changing a feature's value across the entire dataset and assessing the impact on the model's predictive accuracy gives an indication of how important that feature is to the prediction. Local methods such as SHAP can also be used as global feature importance methods, by averaging instance-level values over the data set at hand. Related methods proposed for global feature importance analysis are Sobol indices and Shapley effects.

Table 5: Evaluation metrics and evaluation strategies and their relation to the XAIOR dimensions

Method category	Subtype	Method examples	XAIOR Support PA	AA	RA
Evaluation metrics	Error	PCC, MSE, MAD	X		
	Ranking	AUC, Lift, Pearson correlation	X		
	Profit and loss	Maximum profit criterion, EMPC, Expected misclassification cost	X	X	
	Uncertainty	Calibration curve		X	X
	Privacy	K-anonymity, L-diversity, T-closeness			X
	Fairness	Statistical parity, Demographic parity, Equalized opportunity			X
Model evaluation strategies	Model validation	Out-of-sample, Out-of-period, Out-of-universe validation	X	X	X
	Monitoring	Field testing	X	X	X
	Self-learning methods	Online learning methods, Bandit algorithms, Reinforcement learning	X	X	X

Notes: PA = performance analytics; AA = attributable analytics; RA = responsible analytics. References that describe these methods in detail can be found in Table A.5 in Appendix A5.

Partial dependency plots (PDP) further elaborate on such explanations by providing two-dimensional plots (Friedman, 2001). The marginal (average) effect on the prediction score is given on the vertical axis at the feature value shown on the horizontal axis. Such plots illustrate the relationship between a feature and the output score over the entire range of possible feature values and can be used to visualize interaction effects.

Finally, rule extraction provides a set of rules that mimic how the black box model makes its predictions (Craven and Shavlik, 1996; Martens et al., 2009). In its basic form, one can apply any rule induction technique on the original training data, with the class labels changed to the black box predicted labels. Examples are RIPPER, ANN-DT and DeepRED. Substituting a black box model with one obtained through rule extraction results in efficiency gains and thus contributes to PA.

3.5. Evaluation strategies & metrics

A critical step prior to deploying an analytical solution concerns a comprehensive evaluation across the relevant dimensions of the XAIOR framework in Figure 2. To this end, an evaluation strategy is to be designed that aligns with the applicable user requirements by selecting appropriate metrics and procedures, depending on the problem characteristics and context. In Table 5, various types of evaluation approaches are classified in terms of the relevant dimensions in the XAIOR framework.

As mentioned in Section 2.1, the OR community has inherently been interested in boosting the performance of methods and solutions. Consequently, various evaluation procedures and metrics have been proposed and adopted for assessing and optimizing performance. For each task, e.g., classification, regression, or clustering, a range of performance measures allows for assessing the ability of the obtained solution to optimize decision-making. Specialized, application-dependent measures often exist that allow fine-tuning the evaluation to take into account problem-specific characteristics, such as a highly skewed class distribution (e.g., in fraud detection (Baesens et al., 2015)) or error-dependent and stochastic costs (e.g., in churn prediction (Verbraken et al., 2013)).

The AA dimension in the XAIOR framework identifies three dimensions, i.e., understandability, justifiability, and actionability. Whereas the understandability of an analytical solution typically depends on the analytical method that is applied (e.g., a decision tree and logistic regression typically yield interpretable models, whereas deep learning does not), the justifiability of a solution is to be evaluated. To this end, domain knowledge can often be expressed in terms of constraints that apply. For instance, based on domain knowledge, a positive relation could be expected between a predictor and a target variable in a binary classification model. Given a logistic regression model, it is straightforward to evaluate whether this constraint is satisfied by inspecting the sign of the coefficient of the predictor, which should be positive. To assess the uncertainty of model outcomes, calibration curves can be deployed. For more complex models, e.g., decision trees, rule sets, or ordinal classification models, more advanced metrics may be adopted for evaluating justifiability, as proposed in, e.g., Verbeke et al. (2017). Assessing the actionability of an analytical solution is a highly complex task and is typically done qualitatively.

Finally, as to RA, evaluation metrics may be applied *to assess the privacy of the dataset, in terms of k-anonymity, l-diversity or t-closeness, or the fairness to sensitive groups using measures such as statistical parity, or of the model with metrics such as demographic parity or equalized opportunity* (Martens, 2022, pp.175-176). Additionally, robustness and sustainability may be assessed quantitatively, although no agreement exists in the literature on standard evaluation approaches and metrics.

Prior to deployment, in addition to adopting commonly used methods to simulate the future performance of the solution, such as out-of-sample evaluation or cross-validation, a small-scale field test or experiment, such as an A/B test, may be set up to assess real-world performance. Once an analytical solution is deployed, the operational performance typically needs to be monitored continuously. To this end, the same evaluation metrics may be adopted during the development process, resulting in an out-of-time or out-of-universe validation. Monitoring may be performed at three levels depending on the problem characteristics and the solution's architecture:

- At the first level, the population stability can be monitored (e.g., in terms of the population stability index or deviation index (Baesens et al., 2016)), which involves a comparison of the sample that was used to develop the model and the current population on which the solution is applied. If the sample is no longer representative of the population, the solution may need to be updated.
- A second level involves using the estimates produced by the solution for decision-making, e.g., in the case of binary classification, to classify entities in groups, which is called the discrimination power of the solution.
- A third level concerns the calibration of the estimates, e.g., in binary classification, whether the estimated probabilities match the realized proportions.

Note that some solutions have built-in monitoring procedures and can continuously learn from new data, such as online learning methods, bandit algorithms, and reinforcement learning.

4. Deploying XAIOR

This section provides a non-exhaustive overview of analytical applications in the most important OR domains and their link to the XAIOR framework. Figure 5 gives an overview of the most important deployment areas in OR, i.e., forecasting, risk analysis, inventory control, marketing, and supply chain management.

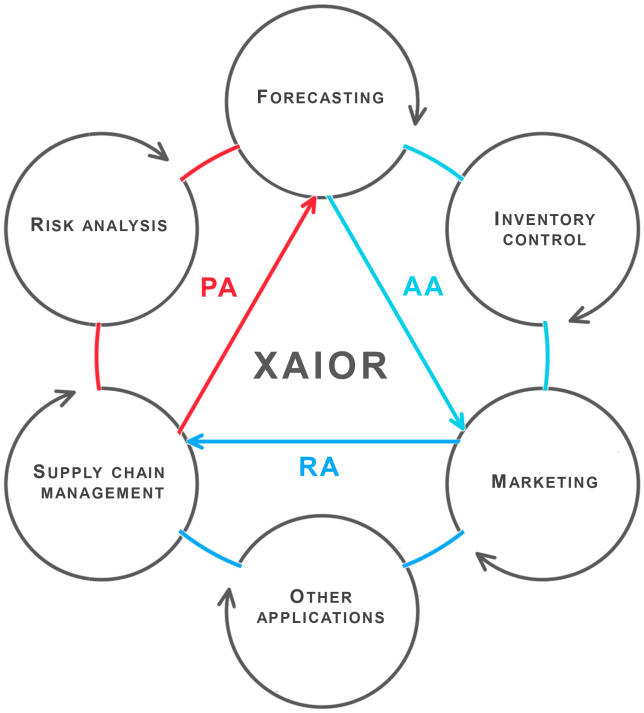


Figure 5: Deploying XAIOR

4.1. Forecasting

As discussed in Section 3.3.1, time-series forecasting is the use of a model to predict future values based on previously observed time-stamped values. It is a crucial part of operational decision-making (Ma and Fildes, 2021). This section discusses the state of the XAIOR dimensions by exemplifying time-series forecasting applications in OR.

- *Performance analytics.* PA is well covered with many OR papers examining the merits of developing new methods for improved accuracy across various applications like demand and sales forecasting (Seyedan and Mafakheri, 2020) or financial market modeling (Sezer et al., 2020). Recent examples include recurrent neural networks or tree-based algorithms (Fischer and Krauss, 2018). Further, the quest for higher performance has also inspired adapted model evaluation criteria through, e.g., asymmetric loss functions.
- *Attributable analytics.* AA has experienced considerable coverage in extant OR literature across various domains like transportation (Li Long et al., 2021), energy (Gürses-Tran et al., 2022), health care (Yang, 2022), and risk management (Bastos and Matos, 2022).

We further zoom into the understandability, justifiability, and actionability aspects of this XAIOR dimension.

- Traditional time-series forecasting methods naturally address *understandability* by charting the actual and predicted target over time (see Fig. 4). Methods to decompose the forecast error into interpretable components provide further insight (Nikolopoulos et al., 2007). It is noteworthy that many time series forecasting methods are intrinsically interpretable. For instance, the famous Box-Jenkins methodology (Hyndman and Athanasopoulos, 2021) designs a forecasting model such that an auto-regressive part, a seasonal and/or trend component, exogenous predictors, and their influence on the target are explicitly discounted. Further, time-series causality methods such as the *Neural Granger* causality model (Tank et al., 2022) also provide rich insights into co-movements and the dependency structure of time-series.
- Forecasting literature has paid much attention to *justifiability*. Extant literature often examines the interplay between statistical forecasts and organizational stakeholders' opinions in the form of human expert adjustments to statistical forecasts (Perera et al., 2019). Many studies offer insight under which conditions human adjustments are effective (Khosrowabadi et al., 2022) and guide how to incorporate expert knowledge in statistical forecasts (Hewage et al., 2022) to address justifiability concerns.
- It is fair to say that *actionability* deserves more attention in the forecasting literature. This dimension is only covered in specific applications such as spare parts and intermittent demand forecasting (Boylan and Syntetos, 2016). It is well known that the large fraction of zero values in an intermittent (demand) time series complicates forecasting and requires a tailor-made methodology (Goltsos et al., 2022). Several papers addressing this requirement stress the interplay between the forecasting method and inventory management optimization (Ye et al., 2022). Studies on the calculation of inventory levels based on the forecast errors and their distribution (Teunter et al., 2017; Turrini and Meissner, 2019) exemplify this research stream and, more generally, how a holistic methodology for decision support - encompassing all steps from past data, over a demand forecast, to a concrete recommendation of how to act - may be crafted. Some scholars coin this paradigm as *predict-and-optimize* and contrast it with the more traditional approach of addressing forecasting and optimization independently, that is *predict-then-optimize* (Elmachtoub and Grigas, 2022). Recent advances in causal forecasting (Grecov et al., 2022) are a promising step in this direction, offering a higher degree of decisional guidance.

- *Responsible analytics.* RA has received the least recognition in the forecasting literature. Requirements concerning RA are much more likely to occur in the context of a concrete application setting. Studies on financial risk management, as reviewed in Section 4.2, are a good example. Another explanation for the scarcity of RA in forecasting is that many popular applications do not involve (personal) data of human subjects. This reduces the necessity of regulatory oversight.

4.2. Risk analysis

Risk analysis is the process of identifying and assessing factors that negatively impact the success of critical organizational projects. A plethora of OR techniques has been developed and studied for qualifying, estimating, and managing various types of risk, such as credit risk (Baesens et al., 2016), fraud risk (Baesens et al., 2015), market risk (Drenovak et al., 2017), operational risk (Mitra et al., 2015), and marketing risk (De Caigny et al., 2018). We kindly refer the reader to Doumpos et al. (2023) for a recent review on the usage of AI in risk analysis and banking as a whole. As we illustrate below, many of these developments reported in OR literature almost organically grew in time throughout the dimensions of the XAIOR framework.

- *Performance analytics.* Extant literature has heavily focussed on PA, with many early-stage developments centered around maximizing performance metrics such as accuracy, recall, precision, top decile lift, or the area under the Receiver Operating Characteristics (ROC) curve. More recent research has re-focused on including profit around three major themes:

    1. The development of tailored performance metrics that especially focus on the profit dimension of the risk type considered. For example, in Verbraken et al. (2013), the Expected Maximum Profit for Churn (EMPC) was introduced, which was later extended to a credit risk setting in Verbraken et al. (2014).
    2. Profit performance metrics were subsequently adopted directly in optimizing the analytical techniques themselves rather than optimizing business irrelevant cost functions. For instance, ProfLogit (Stripling et al., 2018) and ProfTree (Höppner et al., 2020) are extensions of logistic regression and decision trees, respectively, both directly optimizing the EMPC measure in a churn risk context. Other examples of profit-driven analytical techniques are cslogit (based on logistic regression) and csboost (based on gradient tree boosting), both optimizing an instance-dependent cost measure in a fraud risk context (Höppner et al., 2022).
    3. Researchers conducted various benchmarking studies contrasting recently introduced analytical techniques (e.g., deep learning, XGBoost) with *traditional* methods (e.g., regression or decision trees) in terms of both statistical as well as profit-driven measures (Gunnarsson et al., 2021; Lessmann et al., 2015). One striking finding of many studies is that, often, traditional methods still perform very competitively with their newer counterparts both in terms of statistical as well as profit-based performance metrics.

- *Attributable analytics.* AA has gained substantial importance in risk analysis in recent years. For example, in a credit risk setting, regulatory guidelines issued by central banking authorities (e.g., the Basel Accords, IFRS 9) require the adoption of white box, interpretable analytical models such that credit decisions can always be properly explained and justified to both customers and regulators. Further, fraud detection models should also be complemented with explanatory facilities such that well-targeted fraud prevention mechanisms can be put in place. Interpretability in risk analysis is obtained in two ways. The

first option is to use white-box techniques like regression or decision trees. A second way is to use a complex algorithm (e.g., XGBoost or deep learning) and complement it with explanatory post-hoc facilities. Examples are partial dependence plots, ICE plots, LIME or Shapley values (see Section 3.4 for an overview). Using these post-hoc interpretability techniques will contribute to making analytical risk models not only interpretable and justifiable but also actionable.

- *Responsible analytics.* RA is under-investigated in extant literature, but it is more relevant than ever. Various new data sources have emerged to better quantify different types of risk such as online behavioral data originating from Google, Facebook, or Twitter, call detail record (CDR) data from telecommunication providers, or Internet of Things data from smartwatches or telematics devices. These new data sources are very interesting and predictive for credit risk and (insurance) fraud risk prediction (see, e.g., Óskarsdóttir et al., 2019). However, the collection and crunching of these data obviously come with ethical, fairness, and legal challenges which are the topic of debate to many researchers, regulators, and governments nowadays. In fact, predictive models might result in algorithmic bias, yielding outcomes that reinforce inequalities in society, as discussed in Kordzadeh and Ghasemaghaei (2022). Kozodoi et al. (2022) empirically study the profit-fairness trade-off in credit scoring. Figure 6 provides empirical evidence of this trade-off between profit (Y-axis) and separation as a fairness metric (X-axis) on seven credit scoring data sets using the concept of Pareto frontiers. Figure 6 reveals that the unfairness can be substantially reduced at a relatively low cost. For instance, according to Figure 6, reducing the difference in error rates below 0.2 is possible while sacrificing less than €0.01 profit per EUR issued.

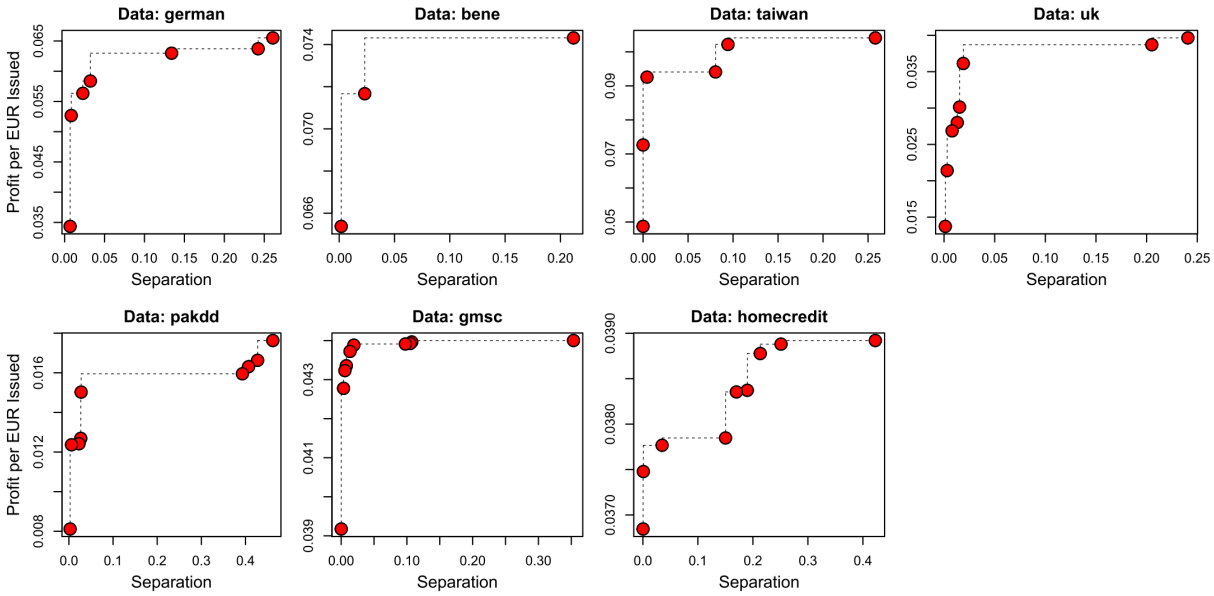


Figure 6: Profit-fairness trade-off (Kozodoi et al., 2022)

Further research is needed about the frugal aspect. This is important to consider, especially with the emergence of powerful analytical techniques with a heavy ecological carbon

footprint in terms of both model estimation and deployment.

4.3. Inventory control

Inventory control is the problem faced by a firm that must decide how much to order in each period to meet the demand for its products while minimizing costs. Using (data) analytics in inventory control is not new (Erkip, 2022). The classical approach for solving data-driven inventory decisions is "predict, then optimize". Here, the model and/or demand parameters are estimated in the first stage, and then, its predictions are utilized in an optimization problem for decision-making in the second stage. The prediction can rely on statistical modeling or more advanced supervised machine learning algorithms (Bastani et al., 2022).

An alternative approach directly prescribes (i.e., predicts and subsequently optimizes) the inventory decisions using data. One such technique in this category is reinforcement learning (RL). RL is different from (un)supervised learning: rather than describing or predicting an outcome, it directly *prescribes* which decision or action to take, based on the current state of the system, while taking the future impact of these decisions into account. Mathematically, it formulates a problem as a Markov decision process, in which an action taken in a given state transitions the system to a new state and generates a reward (or cost). RL requires further training for the algorithm to learn how to optimize its actions. However, instead of comparing the output directly to the 'correct' answers (as in supervised learning), training an RL algorithm relies on trial and error by simulating sequences of states, actions, and rewards. These simulations can be fed by either observational data or simulated data, conditional on an accurate data generation engine (Boute and Udenio, 2021). Just like neural networks are now well-established in (deep) supervised learning, they are also applied in RL, known as deep reinforcement learning (DRL), which can also be applied to inventory control (Boute et al., 2022). In this section, we will explain the dimensions of the XAIOR framework in light of (D)RL applications in inventory control.

- *Performance analytics.* PA has been extensively addressed in inventory control literature as it comes closest to its heart, i.e., effectively solving inventory problems and enhancing inventory decision-making. For instance, Gijsbrechts et al. (2022) provides a rigorous performance evaluation of DRL for the lost sales, dual sourcing, and multi-echelon inventory management problem. In contrast, van Jaarsveld (2020) focuses on the lost sales inventory problem. They demonstrate that their DRL algorithms can outperform the performance of state-of-the-art heuristics and other approximate dynamic programming methods. Liu et al. (2022a) apply a multi-agent DRL-based framework to 50,000 product references for Alibaba, the largest e-commerce platform in China. They present evidence that their DRL algorithms outperform human buyers in reducing out-of-stock rates and inventory levels. Moreover, their algorithms are more effective and robust, including during unexpected extreme situations such as COVID-19 outbreaks and lockdowns.
- *Attributable analytics.* AA, in search of understandability of the DRL policies, is especially relevant to gain intuition behind the inventory policies obtained through DRL. Unfortunately, although neural network policies are flexible and performant, they are notoriously

difficult to interpret. This sharply contrasts with the often highly intuitive character of inventory policies obtained via classical analytical methods. For instance, Vanvuchelen et al. (2020) provides an attempt to gain intuition behind the DRL policies by visualizing the inventory decisions in each situation and comparing them against the optimal ones (for small-scale problems) and benchmark heuristics (for realistic-sized problems). They demonstrate how the algorithm approaches the optimal policy structure compared to the benchmark heuristics. Future research is needed to help explain and interpret DRL policies. When models provide managers with the intuition behind the action, the adoption of DRL in practice will be fostered. Likewise, we could use DRL to learn the structure of well-performing solutions, which may lead to new heuristic policies for challenging problems that have, until now, resisted precise or approximate analysis (Boute et al., 2022).

- *Responsible analytics.* The value of DRL stems from its ability to (semi-)autonomously process data to produce inventory control prescriptions. These are typically used to optimize operational parameters, such as customer service levels or inventory costs. The same characteristics also make DRL a powerful tool to improve other objectives, notably sustainability development goals. For instance, Gijsbrechts et al. (2022) apply their DRL algorithm on a real data set of a consumer goods company to combine multiple transport modes in parallel, where part of the shipment is shipped using a slow but more carbon-friendly transport mode such as rail- or waterways, and part of the shipment is shipped using a more responsive mode such as road or air freight. Their results can be helpful to stimulate a modal shift to low-emission transport modes without adverse impact on service levels or costs. De Moor et al. (2022) apply transfer learning from existing, well-performing heuristics to stabilize the training process and improve the performance of DRL in inventory control. They apply potential-based reward shaping to improve the performance of DRL to manage the inventory of perishable goods. Examples are fresh foods or drugs with an expiry date. The optimal inventory policy is notoriously complex, as it is a function of both the inventory position and the age distribution of the inventory. When the latter is ignored, it will result in more waste. Transferring knowledge embedded in existing heuristic inventory policies improves DRL performance and, consequently, reduces the waste of perishable inventory. Whereas these works focus on the environmental aspects of sustainability, Vanvuchelen et al. (2022) use DRL to improve the social dimension. They use DRL to improve the accessibility to malaria medicines in Zambia's public pharmaceutical supply chain. They show how lateral trans-shipments between health facilities can further reduce the variation of service levels across facilities and improve the equity of access to essential medicines in Zambia. It shows that DRL, as a tool to improve inventory control, can foster environmental and social improvements.

4.4. Marketing

The marketing field analyses customer data to describe and predict customer behavior in various stages of the customer journey, i.e., the acquisition, development with cross- and upselling activities, and retention stage. This section highlights noteworthy research in these areas across the three dimensions of the XAIOR framework.

- *Performance analytics.* Two important business characteristics explain the prevalence of PA in marketing.

    1. The evolutive nature of business contexts. For instance, marketing budgets and target class distributions may vary over time. E.g., digital ad targeting is a function of evolving factors such as product lifecycle stage, available budget, expected conversion rate, etc. Such changing contexts imply that the decision threshold to target someone with an ad also will vary. This example motivates the ongoing popularity of performance curves in marketing, such as ROC curves in assessing the performance of predictive models, which show the performance across the entire range of decision thresholds (Brook and Arnold, 2019).
    2. Marketing accountability. The benefits and costs of marketing actions are often available. For instance, the cost of sending out a marketing offer and the reward of accepting an offer is often known. This facilitates using expected profit and profit curves as evaluation metrics and predictive models that directly optimize these. For instance, (Martens et al., 2016) provides profit curves for response modeling in a banking setting, while (Verbeke et al., 2012) discusses a profit-driven approach for churn prediction.

- *Attributable analytics.* The marketing field has focused heavily on making customer analytics models attributable. First, global explanation methods like rule extraction were proposed previously for churn prediction and response modeling (Verbeke et al., 2017). Furthermore, there is a stream of hybrid modeling approaches where homogeneous segments are first identified in the customer base. Subsequently, segment-specific models are trained. This approach was found to enhance both predictive performance and understandability. For instance, Table 6 visualizes the logit leaf model (LLM) approach proposed by De Caigny et al. (2018) on the publicly available *cell2cell* customer churn prediction dataset. The LLM consists of two steps. In the first step, customer segments are identified using decision rules, and in the second phase, a logistic regression model is created for every leaf of this tree. The authors show in an extensive benchmarking experiment that LLM's predictive performance is competitive to SOTA benchmark algorithms. At the same time, the interpretability is drastically increased through the identification of segment-specific churn drivers, as seen in column "2nd step: logistic Regression" in Table 6.

  Furthermore, it is worth noting that the marketing domain often uses textual and behavioral data characterized by high dimensions and sparseness (Ramon et al., 2020). Traditionally, interpretable models, such as linear ones, become black boxes due to the massive dimensionality of the features. This motivates the use of post-hoc instance-based explanation approaches for marketing applications, such as LIME, SHAP, and counterfactuals (Ramon et al., 2020). These methods automatically map the model to the few relevant features for the model's prediction.

Table 6: Visualization of the *cell2cell* logit leaf model for customer churn prediction (based on De Caigny et al., 2018)

1st step: decision tree						**2nd step: logistic regressions**										
						Intercept	**Shared features**							**Segment-specific features**		
Seg.	**Rule 1**	**Rule 2**	**Rule 3**	**Rule 4**	**Churn rate**		***retcalls***	***changem***	***changem dummy***	***recchrge***	***creditde***	***custcare***	***setprcm***	***eqpdays***	***directas***	***outcalls***
1	*eqpdays* <= -0.31				14.3%	0.46	1.12	-0.07	0.96		-0.28		-0.63	0.23	-0.52	
2	*eqpdays* > -0.31	*eqpdays* <= -0.6			16.7%	0.19	0.98	-0.24	1.59		-0.5		0.51			
3	*eqpdays* > -0.06	*webcap* <= 0			51.9%	0.23	0.17			-0.19	-0.35	-0.2				
4	*eqpdays* > -0.06	*webcap* > 0	*callwait* <= -0.51	*changem* <= -0.05	14.6%	0.11	0.18	-0.29	1.66					0.42	-0.38	-0.2
5	*eqpdays* > -0.06	*webcap* > 0	*callwait* <= -0.51	*changem* > -0.05	12.7%		-0.3	0.19	0.29		-0.14				-0.2	
6	*eqpdays* > -0.06	*webcap* > 0	*callwait* > -0.51		30.2%	0.21	0.59	-0.1	1.37	-0.11	-0.36	-0.09		0.16	-0.21	

- *Responsible analytics.* OR research that relates to RA and its subdimensions in marketing is scarce. This is surprising since marketing practices are typically very visible and impactful to companies and customers. For example, ethical concerns may arise in advertising targeting. Advertising networks offer transparency in their targeting practices. For example, Google's AdChoices allows end users to investigate why an ad is served to them. Another illustration is the widely discussed *Target* case, which has shown us all the fallout that can come in predicting pregnancy (Martens, 2022). Even if end users consent to predict (baby) product interest, and even if the model performs accurately, the sensitivity of pregnancy prediction cautions against it.

4.5. Supply chain management

Supply chain management (SCM) involves different functions in the multi-echelon system and is related to managing the flows of goods and services and all processes that transform raw materials into finalized products. We discuss the evolution towards analytics-driven SCM for the three dimensions of the XAIOR framework.

- *Performance analytics.* Ample work in SCM literature has focused on improving effectiveness and efficiency. In particular, Bayesian decision theory has been widely used to incorporate information into the decision-making process to enhance accuracy. For example, Iyer and Bergen (1997) adopt the Bayesian conjugate pair theory to explore responsive supply chain operations. The authors quantify the impact of information updating in the supply chain and discuss how to achieve Pareto improvement in the channel. Aronis et al. (2004) explore inventory management in a supply chain with Bayesian information updating. They assess the Bayesian prior distribution for the failure rates of different spare parts and subsequently develop the algorithm to analytically update the inventory policy's parameters using information. Choi et al. (2006) extend Iyer and Bergen (1997)'s analysis to the case with two different Bayesian models, namely the Bayesian conjugate models with known and unknown variance, respectively. The authors highlight the importance of having a more sophisticated Bayesian model as well as the proper choice of the observation target.
- *Attributable analytics.* Extant SCM literature has focused on interpretability, justifiability, and actionability. With the advance of computational power and the popularity of data analytics, the Bayesian (belief) network approach (BNA) has received growing interest over the past decade in supply chain risk analysis. The Bayesian network is, in fact, a "probabilistic graphical model" that can help analytically assess the probabilistic relationships among the variables under investigation. For instance, Garvey et al. (2015) study via the BNA risk propagation in supply chains. In their proposed model, inter-dependencies among different categories of "risks" are modeled. The authors also derive the risk measures. Model performance and interpretability is demonstrated by conducting simulation experiments. Liu et al. (2021) choose the robust "dynamic Bayesian network approach" (DBNA) to explore supply chain disruptions. The authors consider the "worst-case probability" situation and build a mathematical optimization model to provide analytically explainable logic in finding the optimal solution. Sakib et al. (2021) study the supply

chains for oils and gases. The authors introduce the BNA-based models to help forecast and analyze challenges in the supply chain. They highlight the critical factors that affect supply chains. The BNA is very performant in supply chain risk analyses. Since the BNA also uses the Bayesian approach, many details are analytically explainable, at least partially. For supply chains in the *Industry 4.0* era, analytics-driven insights play a critical role. However, most AI decision-making tools in supply chains are based solely on automated processing and profiling. Thus, they are generally not explainable, and hence, they are treated as black boxes. It makes them difficult to understand for supply chain operations managers. In the OR literature, only very few studies discuss the use of AA in supply chains, and we review them as follows. Senoner et al. (2022) investigate the use of AA for electronics supply chains. The authors focus on process quality improvement in semiconductor manufacturing and build a novel data-driven decision-making (DDM) model. To cope with real-world situations, the DDM model must be able to process data sets that are highly complex. The authors propose a model that includes SHAP to analytically explain how the systems parameters and the manufacturing process quality are correlated. This helps to streamline the manufacturing process with quality in mind.

- *Responsible analytics.* Relatively few studies are present in the current OR literature of SCM that address RA. One exception is Westerski et al. (2021), who explores the use of explainable AI in detecting the ethical problems with procurement fraud. The authors model different categories of procurement fraud with the use of proper statistical measures. They establish a decision-making framework to rank the fraud's severity and assess the score for each procurement transaction. This helps to improve the auditing function. In both Senoner et al. (2022) and Westerski et al. (2021), the intelligence systems' techniques and resulting logics are understandable. This fosters the trust of the supply chain managers in using them and facilitates further extensions in future research. In fact, for Industry 5.0, in which the focus is on human-machine reconciliation, the importance of having a balance between "machines" and "humans" (and human society) is well-advocated. Choi et al. (2022) propose an analytical framework with a feedback loop for achieving sustainable social welfare (SSW), which includes human welfare, the environment, and company benefits in using disruptive technologies for supply chain operations. One important highlight of their proposal is the importance of policymakers in deciding the carrot-and-stick policy to ensure companies have the right incentive to achieve SSW.

4.6. Other applications

In this section, we summarize other relatively less-known applications of analytics as part of OR. We particularly focus on the following OR domains: healthcare, litigation, and educational analytics, and discuss their link with the building blocks of the XAIOR framework.

- *Healthcare.* Various applications have focused on optimizing the decision-making strategy in the context of improving the health and well-being of people, such as in the context of organ transplantation operations. The use of data analytics methods, as opposed to intuition and experience-based utility functions, for optimal allocation of the organs to potential recipients, optimizes the allocation process and thereby saves more lives (Al-Ebbini

et al., 2017). Not only do analytics predict the prognostics of these significant events, but also explain the reasoning behind the prescribed actions. AA is used synergistically to develop and deploy powerful mathematical models as screening mechanisms for the future onset of diabetes complications. For instance, machine learning models developed on the electronic health records (EHR) database are used to predict diabetic retinopathy (Piri et al., 2017), a leading cause of blindness among working-aged adults. Such an analytics model is used as a screening tool by medical professionals to urge diabetic patients to get it confirmed and treated so that they maintain their eyesight. Such automated early warning mechanisms are especially useful in rural settings where specialist like ophthalmologists is scarce (Wang et al., 2021). Another healthcare domain where the use of an EHR database along with AA makes a significant impact is in the analyses of relatively rare chronic diseases (Reddy and Delen, 2018). Such data-driven analysis leads to better understanding, diagnosis, explanation, and management of these diseases. Often, these data-driven explanatory analytics studies discover patterns that pave the way for novel clinical and biological investigations toward better diagnostic and treatment regiments (Reddy et al., 2019).

- *Litigation.* A particular application is analytics for drug courts. The purpose behind the establishment of drug courts was to create an alternative to traditional criminal courts to transform the traditional punitive jurisprudence into a therapeutic one. Under this new philosophy, the eligible offenders are considered individuals in need of rehabilitative treatments and are persuaded to undergo a regimen that seeks to return them to the community as productive contributors rather than sending them to prison. This initiative, if performed properly, has proven to be effective in lowering costs to the community and improving social outcomes. To enable better management of resources and improvement of outcomes, advanced analytics models are developed using large real-world data obtained from drug courts to predict and explain who would or would not graduate from these treatments (Zolbanin et al., 2020), who would be a returning offender (i.e., recidivism) (Delen et al., 2021), and to prescribe a set of guidelines (presented as characteristics of the offenders) that can help jurisdictions and drug court administrators to make more effective and efficient decisions.
- *Educational analytics.* Lastly, we look at the college student attrition problem. Student retention is an essential part of any college enrollment management system. It affects a university's rankings, reputation, and financial well-being. Therefore, student retention has become one of the top priorities for decision-makers in higher education institutions. Improving student retention starts with a thorough understanding of the reasons behind attrition. Such an understanding is the basis for accurately predicting at-risk students and appropriately and responsibly intervening to help them to stay in school. To go beyond the intuitionist approaches to understanding the underlying causes of attrition and to make the outcomes more actionable, in a series of exemplary studies, researchers have used multiple years of institutional data along with several machine learning techniques to develop analytical models to predict and explain the reasons behind student attrition (Delen, 2010). Explanatory capabilities of these prediction models provide the much-needed guideline to approach an at-risk student with a specific regiment plan to improve

his/her possibilities of returning to school for the sophomore year. Because more than half of the attrition happens in the freshmen year, better management of freshmen student attrition translates to better retention and graduation rates.

5. Discussion and setting an agenda for future research

In this paper, we present a framework for XAIOR and provide a review of existing methods and applications according to the three main dimensions of XAIOR. In what follows, we summarize our main findings for PA, AA and RA across methods and applications and establish an agenda for future research.

- *Performance analytics.* In terms of methods and applications, the PA dimension of the XAIOR framework is well-established. This is not surprising, as performance is necessary for any analytical OR solution.
- *Attributable analytics.* AA has also received much attention from the OR community. A prominent example is post-hoc interpretation, enabled through methods specifically developed for AA. Such methods allow for deriving insights from so-called "black box" models. The level of advancement of AA within applications often depends on domain-specific requirements. Some applications, for example, are more advanced in dimensions of AA, such as forecasting, risk management, or marketing, while other application domains, such as inventory control and supply chain management, are still lagging.
- *Responsible analytics.* Only recently, RA became an important aspect in many applications, leading to the development of methods or adjustments to methods to deal with RA specifically. Despite recent advancements, RA, however, is still a dimension that needs further research within OR.

Figure 7 presents a research agenda that is linked to the XAIOR framework. Research topics cover a single dimension of the XAIOR framework, or combine PA, AA, and/or RA dimensions. Based on the current state of research on XAIOR, we propose five promising research themes, across methods and applications, that will advance the XAIOR domain in the near future. For each of these themes, highlighted with a specific icon, we list some exemplary research questions in Figure 7 that will further inspire readers.

- *Data innovation.* Data enrichment and data augmentation studies are important in OR, showing the importance of innovative, unstructured, or structured data sources, such as textual, image, or social network data, to improve models. In the near future, new data sources such as data issued by generative AI models, geospatial data, data linked to IoT applications, or data from the Metaverse might show value for various domains. Despite the importance of data augmentation studies, they traditionally focus mainly on the PA dimension of the XAIOR framework. It is, however, relevant to link new data sources across all dimensions of the XAIOR framework. Research may then question, for example, whether all applications require massive amounts of data to train a model for a marginal gain in performance, but at permanent maintenance and energy costs. Another innovation may focus on the use of synthetic data from simulators, which found some applications in the OR domain already (Brailsford et al., 2019) Similarly, state-of-the-art

models and platforms to artificially generate data, such as Stable Diffusion or chatGPT, offer promising new research venues. Hence, the full potential of artificially generated data to improve OR applications in terms of PA, AA, and RA is yet to be explored.

- *Deep learning.* Despite the fact that deep learning does not always perform better than traditional machine learning algorithms, especially when well-designed features are available (Gunnarsson et al., 2021), there is still much potential for further research. Most existing research focuses on the PA aspect of deep learning, while AA and RA aspects would require more attention. In line with the AA dimension, an interesting path is to explore how "black box" deep learning algorithms could be opened. Most attempts to do so are post-hoc evaluation methods, such as SHAP. There are also attempts to improve model interpretability of deep learning models by, for example, inducing decision rules, although this remains a difficult task. Next, transfer learning is another promising way to further advance OR applications. In NLP, once large language models are trained, they can be fine-tuned for specific applications, which allows to boost performance and reduces the training time for the specific application. Hence, exploring other ways of transfer learning would be interesting. This is important when considering frugal aspects in deep learning, which requires more research to reduce the cost of deploying and operating deep learning networks. For example, frugal algorithms can be used to reduce the number of parameters that are required for a deep learning network, which can reduce the amount of computation and storage required. Similarly, algorithms can be designed to make better use of training data. Additionally, frugal architectures can be used to reduce the number of layers in a deep learning network, which can also reduce the cost of deploying and operating a deep learning network. Finally, other learning paradigms are advancing such as zero/one/few-shot learning, reinforcement learning, or semi-supervised learning. Such approaches can become important within the OR field as well.
- *Integrated XAIOR.* Two aspects are important to consider, being the development of new metrics and the optimization of algorithms along metrics. First, there exist streams in OR that focus on the development of better evaluation metrics to replace purely statistical metrics, such as profit metrics in marketing (Verbeke et al., 2012) or credit scoring (Verbraken et al., 2014). Most metrics focus on PA, although some recent developments try to include AA and RA as well (Kozodoi et al., 2022). Yet, more research is needed to create better metrics for all dimensions of the XAIOR framework. Second, analytical solutions should ideally be evaluated over all PA, AA, and RA dimensions in line with multi-criteria evaluation literature. A challenge is that all aspects of the traditional data processing pipeline might have an impact, so multi-criteria evaluation should not only focus on the algorithm but also consider the broader solution. So far, limited research evaluated algorithms across these different dimensions.

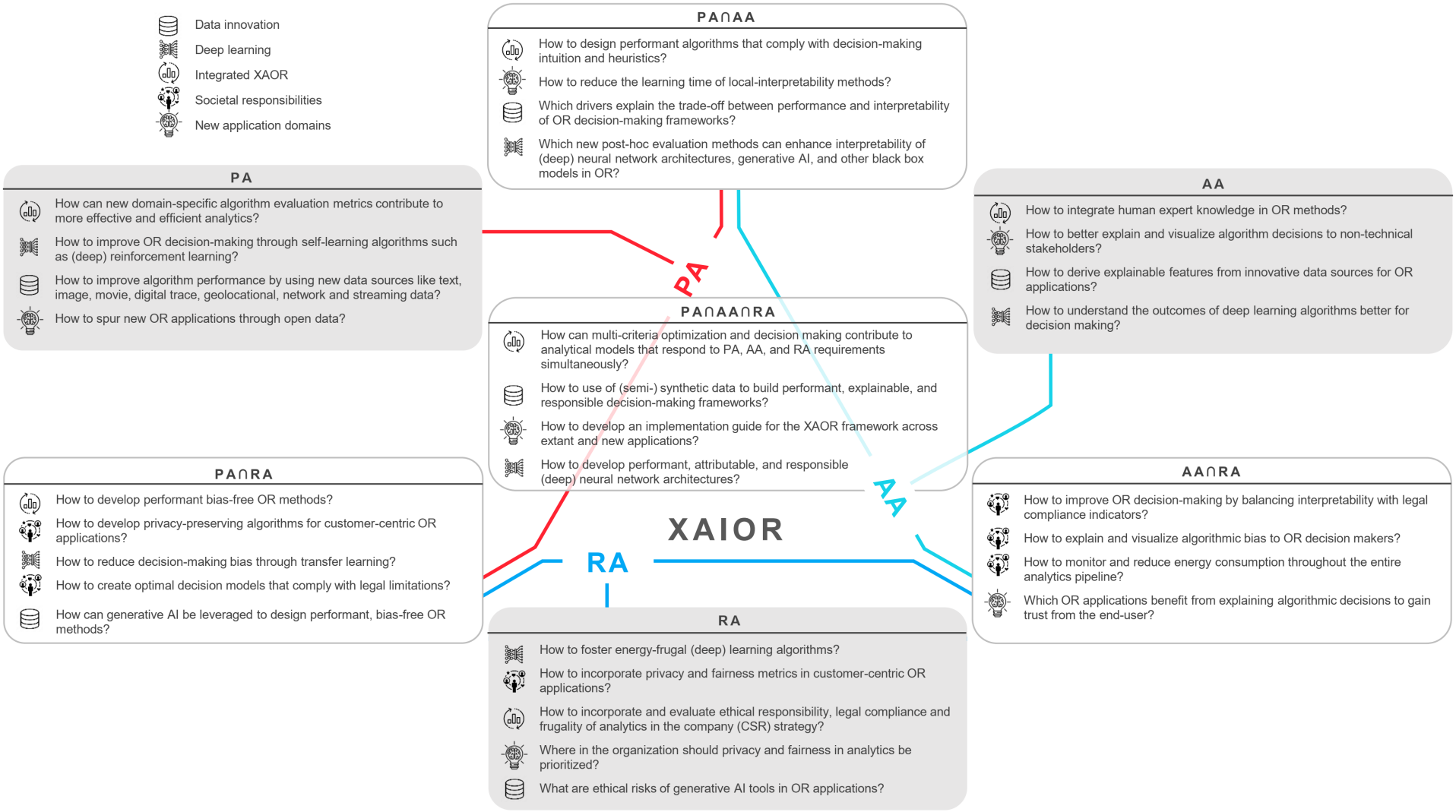


Figure 7: Future research themes and agenda for XAIOR

- *Societal responsibility.* Stakeholders in society become more aware of potential risks linked to algorithm-assisted- and automated decision-making tools, especially in times when generative AI models such as large language models (LLMs, e.g., OpenAI's ChatGPT) are being integrated with various solutions at a rapid pace. There is, for example, an increased sensitivity towards algorithmic biases, which makes that solely considering performance might not suffice to implement a solution. Despite the attention to such issues, more research is needed on how to detect, prevent, and mitigate algorithmic biases within OR applications, requiring domain-specific research. Therefore, it is important to explain algorithmic decisions, as already required in certain industries such as credit scoring. Also, awareness about the ecological costs of saving, storing, and analyzing data is pushing towards more frugal analytics. Research could further explore how models can become more efficient to achieve this goal. As a final topic, privacy and data protection are important concerns for organizations. In Europe, there is an all-encompassing law regulating the acquisition, storage, or use of personal data after the introduction of Regulation (EU) 2016/679 (the General Data Protection Regulation, or *GDPR*), published in May 2016 with enforcement starting in May 2018. In the US, data protection is partly regulated by the Privacy Act of 1974, which establishes a code of fair practice to govern the collection of personal data, the Health Insurance Portability and Accountability Act of 1996 (HIPAA) to protect health information privacy rights, and the Electronic Communications Privacy Act (ECPA) of 1986 that establishes sanctions for interception of electronic communication. As a response, research in OR mainly focuses on the input side (Li, 2018) to protect data, but other aspects could be considered as well. Hence, more research is still needed on the development of privacy-preserving solutions.
- *New application domains.* New application domains are likely to arise as a result of increased data availability, new data sources, new technologies, new industries, and new societal challenges. Domains like sports analytics, health analytics or analytics linked to robotics are likely to become more important. Embedding all aspects of XAIOR within these new applications is an inspiring challenge. Sometimes the domain as such is a driver for a certain dimension of the XAIOR framework. Indeed, using analytics that assists in a modal shift to low-emission transport contributes to RA goals (De Moor et al., 2022; Gijsbrechts et al., 2022) by the application itself. Innovative applications that spring from RA challenges are therefore also an important future research direction.

6. Conclusions

There is an increasing need to explain analytical solutions, originating from expectations of internal and external stakeholders, yet this is not fully captured in the OR literature. Despite some review papers focusing on explainability, existing research falls short in (i) proposing an OR-oriented definition of explainable AI for the operational research domain and (ii) zooming in on specific methods and application requirements. In this paper, we first define and characterize XAIOR, i.e., explainable AI for Operational Research. Specifically, XAIOR is defined as an interplay of three dimensions, i.e., performance analytics, attributable analytics, and responsible analytics.

We subsequently discuss the implementation of XAIOR across the data analytics pipeline. In particular, we discuss state-of-the-art methodologies for experimental design & data selection, feature engineering & data preparation, algorithmic design & choice, post-hoc interpretation methods, and evaluation strategies & metrics, and we link these with our XAIOR dimensions. We find that further research is still needed to integrate all XAIOR dimensions, especially AA and RA. In an overview of applications of XAIOR, we discuss prior work on XAIOR and its subdimensions in 6 crucial OR domains. These include forecasting, risk analysis, inventory control, marketing, supply chain management, and other applications. We find that the maturity of PA, AA, and RA depends on the application domain with an under-representation of AA and RA.

Based on these overviews, we identify critical avenues for future research linked to five research themes, i.e., data innovation, deep learning, integration of the XAIOR framework's dimensions and subdimensions, responding to societal changes, and new innovative applications. We propose specific research questions linked to these five research themes and in relation to the XAIOR framework that might inspire researchers to apply and contribute to XAIOR.

Acknowledgements

The authors acknowledge all researchers who, through their work, have advocated and accelerated the adoption of (explainable) analytics in OR. The research of Roman Słowiński was supported by TAILOR, a project funded by the EU Horizon 2020 (research and innovation funding) programme (EC GA number 952215). Richard Weber acknowledges financial support from FONDECYT Chile (1221562), Fondef (IT23I0061), ANID PIA/PUENTE (AFB220003), and NeEDS, a project funded by the EU Horizon 2020 programme (EC GA number 822214).

References

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P., 2005. Automatic subspace clustering of high dimensional data. Data Mining and Knowledge Discovery 11, 5–33.

Al-Ebbini, L., Oztekin, A., Sevkli, Z., Delen, D., 2017. Predicting quality of life for lung transplant recipients: A hybrid genetic algorithms-based methodology. 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies, AEECT 2017 2018-January, 1–6.

Aldunate, Á., Maldonado, S., Vairetti, C., Armelini, G., 2022. Understanding customer satisfaction via deep learning and natural language processing. Expert Systems with Applications 209, 118309.

Aronis, K.P., Magou, I., Dekker, R., Tagaras, G., 2004. Inventory control of spare parts using a bayesian approach: A case study. European Journal of Operational Research 154, 730–739.

Baesens, B., Roesch, D., Scheule, H., 2016. Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS (Wiley and SAS Business Series). Wiley.

Baesens, B., Setiono, R., Mues, C., Vanthienen, J., 2003. Using neural network rule extraction and decision tables for credit-risk evaluation. Management Science 49, 312–329.

Baesens, B., Vlasselaer, V.V., Verbeke, W., 2015. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques. Wiley.

Banasik, J., Crook, J., Thomas, L., 2003. Sample selection bias in credit scoring models. Journal of the Operational Research Society 54, 822–832.

Barocas, S., Selbst, A.D., Raghavan, M., 2020. The hidden assumptions behind counterfactual explanations and principal reasons, in: FAT∗2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. pp. 80–89.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion 58, 82–115.

Bastani, H., Zhang, D.J., Zhang, H., 2022. Applied machine learning in operations management, in: Springer Series in Supply Chain Management. Springer Nature, United States. volume 11 of *Springer Series in Supply Chain Management*, pp. 189–222.

Bastos, J.A., Matos, S.M., 2022. Explainable models of credit losses. European Journal of Operational Research 301, 386–394.

Baykasoğlu, A., Özbakir, L., 2007. Mepar-miner: Multi-expression programming for classification rule mining. European Journal of Operational Research 183, 767–784.

Bengio, Y., Lodi, A., Prouvost, A., 2021. Machine learning for combinatorial optimization: A methodological tour d'horizon. European Journal of Operational Research 290, 405–421.

Berrevoets, J., Jordon, J., Bica, I., Gimson, A., Van Der Schaar, M., 2020. Organite: Optimal transplant donor organ offering using an individual treatment effect. Advances in Neural Information Processing Systems 2020-December, 20037–20050.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R., Samek, W., 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9887 Lncs, 63–71.

Biswas, S., Rajan, H., 2021. Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline, in: ESEC/FSE 2021 - Proceedings of the 29th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp. 981–993.

Błaszczyński, J., de Almeida Filho, A.T., Matuszyk, A., Szeląg, M., Słowiński, R., 2021. Auto loan fraud detection using dominance-based rough set approach versus machine learning methods. Expert Systems with Applications 163, 113740.

Błaszczyński, J., Greco, S., Słowiński, R., 2007. Multi-criteria classification - a new scheme for application of dominance-based decision rules. European Journal of Operational Research 181, 1030–1044.

Błaszczyński, J., Greco, S., Słowiński, R., Szeląg, M., 2009. Monotonic variable consistency rough set approaches. International Journal of Approximate Reasoning 50, 979–999.

Błaszczyński, J., Słowiński, R., Stefanowski, J., 2010. Variable consistency bagging ensembles, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, Berlin. volume 5946 Lncs, pp. 40–52.

Błaszczyński, J., Słowiński, R., Szeląg, M., 2011. Sequential covering rule induction algorithm for variable consistency rough set approaches. Information Sciences 181, 987–1002.

Bodria, F., Giannotti, F., Guidotti, R., Naretto, F., Pedreschi, D., Rinzivillo, S., 2021. Benchmarking and survey of explanation methods for black box models. URL: `http://arxiv.org/abs/2102.13076`.

Borgonovo, E., Plischke, E., 2016. Sensitivity analysis: A review of recent advances. European Journal of Operational Research 248, 869–887.

Boute, R.N., Gijsbrechts, J., van Jaarsveld, W., Vanvuchelen, N., 2022. Deep reinforcement learning for inventory control: A roadmap. European Journal of Operational Research 298, 401–412.

Boute, R.N., Udenio, M., 2021. Ai in logistics and supply chain management. SSRN Electronic Journal .

Boylan, J.E., Syntetos, A.A., 2016. Intermittent Demand Forecasting - Context, Methods and Applications. Wiley, Hoboken.

Brailsford, S.C., Eldabi, T., Kunc, M., Mustafee, N., Osorio, A.F., 2019. Hybrid simulation modelling in operational research: A state-of-the-art review. European Journal of Operational Research 278, 721–737.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Brook, R.J., Arnold, G.C., 2019. 07. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1st ed., O'Reilly Media, Inc.

Cang, S., Yu, H., 2014. A combination selection algorithm on forecasting. European Journal of Operational Research 234, 127–139.

Cano, J.R., Gutiérrez, P.A., Krawczyk, B., Woźniak, M., García, S., 2019. Monotonic classification: An overview on algorithms, performance measures and data sets. Neurocomputing 341, 168–182.

Carcillo, F., Le Borgne, Y.A., Caelen, O., Kessaci, Y., Oblé, F., Bontempi, G., 2021. Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences 557, 317–331.

Chen, D., Fraiberger, S.P., Moakler, R., Provost, F., 2017. Enhancing transparency and control when drawing data-driven inferences about individuals. Big Data 5, 197–212.

Cheng, C.H., Chen, Y.S., 2009. Classifying the segmentation of customer value via rfm model and rs theory. Expert Systems with Applications 36, 4176–4184.

Choi, T.M., Kumar, S., Yue, X., Chan, H.L., 2022. Disruptive technologies and operations management in the industry 4.0 era and beyond. Production and Operations Management 31, 9–31.

Choi, T.M., Li, D., Yan, H., 2006. Quick response policy with bayesian information updates. European Journal of Operational Research 170, 788–808.

Choi, T.M., Wallace, S.W., Wang, Y., 2018. Big data analytics in operations management. Production and Operations Management 27, 1868–1883.

Conboy, K., Mikalef, P., Dennehy, D., Krogstie, J., 2020. Using business analytics to enhance dynamic capabilities in operations research: A case analysis and research agenda. European Journal of Operational Research 281, 656–672.

Corrente, S., Greco, S., Kadziński, M., Słowiński, R., 2013. Robust ordinal regression in preference learning and ranking. Machine Learning 93, 381–422.

Coussement, K., Benoit, D.F., 2021. Interpretable data science for decision making. Decision Support Systems 150.

Coussement, K., Buckinx, W., 2011. A probability-mapping algorithm for calibrating the posterior probabilities: A direct

marketing application. European Journal of Operational Research 214, 732–738.
Coussement, K., Lessmann, S., Verstraeten, G., 2017. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems 95, 27–36.
Craven, M.W., Shavlik, J.W., 1996. Extracting tree-structured representations of trained neural networks, in: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, USA. pp. 24–30.
Crone, S.F., Lessmann, S., Stahlbock, R., 2006. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. European Journal of Operational Research 173, 781–800.
da Costa, P., Verleijsdonk, P., Voorberg, S., Akcay, A., Kapodistria, S., van Jaarsveld, W., Zhang, Y., 2023. Policies for the dynamic traveling maintainer problem with alerts. European Journal of Operational Research 305, 1141–1152.
Davies, R., Roderick, P., Raftery, J., 2003. The evaluation of disease prevention and treatment using simulation models. European Journal of Operational Research 150, 53–66.
De-Arteaga, M., Feuerriegel, S., Saar-Tsechansky, M., 2022. Algorithmic fairness in business analytics: Directions for research and practice. Production and Operations Management 31, 3749–3770.
De Bock, K.W., Coussement, K., Lessmann, S., 2020. Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach. European Journal of Operational Research 285, 612–630.
De Bock, K.W., De Caigny, A., 2021. Spline-rule ensemble classifiers with structured sparsity regularization for interpretable customer churn modeling. Decision Support Systems , 113523.
De Caigny, A., Coussement, K., De Bock, K.W., 2018. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 269, 760–772.
De Moor, B.J., Gijsbrechts, J., Boute, R.N., 2022. Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management. European Journal of Operational Research 301, 535–545.
Deepak, P., Abraham, S.S., 2021. Fairlof: Fairness in outlier detection. Data Science and Engineering 6, 485–499.
Delen, D., 2010. A comparative analysis of machine learning techniques for student retention management. Decision Support Systems 49, 498–506.
Delen, D., Zolbanin, H.M., Crosby, D., Wright, D., 2021. To imprison or not to imprison: an analytics model for drug courts. Annals of Operations Research 303, 101–124.
Dembczyński, K., Kotłowski, W., Słowiński, R., 2010. Beyond sequential covering - boosted decision rules, in: Koronacki, J. (Ed.), Studies in Computational Intelligence. Springer, Berlin. volume 262, pp. 209–225.
Devriendt, F., Berrevoets, J., Verbeke, W., 2021. Why you should stop predicting customer churn and start using uplift models. Information Sciences 548, 497–515.
Ding, Y., Zhu, Y., Feng, J., Zhang, P., Cheng, Z., 2020. Interpretable spatio-temporal attention lstm model for flood forecasting. Neurocomputing 403, 348–359.
Djeundje, V.B., Crook, J., 2019. Identifying hidden patterns in credit risk survival data using generalised additive models. European Journal of Operational Research 277, 366–376.
Doumpos, M., Zopounidis, C., Gounopoulos, D., Platanakis, E., Zhang, W., 2023. Operational research and artificial intelligence methods in banking. European Journal of Operational Research 306, 1–16.
Drenovak, M., Ranković, V., Ivanović, M., Urošević, B., Jelic, R., 2017. Market risk management in a post-basel ii regulatory environment. European Journal of Operational Research 257, 1030–1044.
Duan, Y., Cao, G., Edwards, J.S., 2020. Understanding the impact of business analytics on innovation. European Journal of Operational Research 281, 673–686.
D'Urso, P., 2017. Informational paradigm, management of uncertainty and theoretical formalisms in the clustering framework: A review. Information Sciences 400-401, 30–62.
Elmachtoub, A.N., Grigas, P., 2022. Smart "predict, then optimize". Management Science 68, 9–26.
Erkip, N.K., 2022. Can accessing much data reshape the theory? inventory theory under the challenge of data-driven systems. European Journal of Operational Research .
Fischer, T., Krauss, C., 2018. Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research 270, 654–669.
Fleszar, K., 2022. A MILP model and two heuristics for the bin packing problem with conflicts and item fragmentation. European Journal of Operational Research 303, 37–53.
Frazier, D.T., Maneesoonthorn, W., Martin, G.M., McCabe, B.P., 2019. Approximate bayesian forecasting. International Journal of Forecasting 35, 521–539.
Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29, 1189–1232.
van der Gaast, J.P., Weidinger, F., 2022. A deep learning approach for the selection of an order picking system. European Journal of Operational Research 302, 530–543.
Garvey, M.D., Carnovale, S., Yeniyurt, S., 2015. An analytical framework for supply network risk propagation: A bayesian network approach. European Journal of Operational Research 243, 618–627.
Gijsbrechts, J., Boute, R.N., Van Mieghem, J.A., Zhang, D.J., 2022. Can deep reinforcement learning improve inventory management? performance on lost sales, dual-sourcing, and multi-echelon problems. Manufacturing and Service Operations Management 24, 1349–1368.
Glady, N., Baesens, B., Croux, C., 2009. Modeling churn using customer lifetime value. European Journal of Operational Research 197, 402–411.
Gneiting, T., Katzfuss, M., 2014. Probabilistic forecasting. Annual Review of Statistics and Its Application 1, 125–151.
Goerigk, M., Hartisch, M., 2023. A Framework for Inherently Interpretable Optimization Models. European Journal of Operational Research 310, 1312–1324.
Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking inside the black box: Visualizing statistical learning with

plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24, 44–65.
Goltsos, T.E., Syntetos, A.A., Glock, C.H., Ioannou, G., 2022. Inventory – forecasting: Mind the gap. European Journal of Operational Research 299, 397–419.
Gosiewska, A., Kozak, A., Biecek, P., 2021. Simpler is better: Lifting interpretability-performance trade-off via automated feature engineering. Decision Support Systems 150, 113556.
Grabisch, M., 1996. The application of fuzzy integrals in multicriteria decision making. European Journal of Operational Research 89, 445–456.
Greco, S., Matarazzo, B., Słowiński, R., 2001. Rough sets theory for multicriteria decision analysis. European Journal of Operational Research 129, 1–47.
Greco, S., Matarazzo, B., Słowiński, R., 2004. Axiomatic characterization of a general utility function and its particular cases in terms of conjoint measurement and rough-set decision rules. European Journal of Operational Research 158, 271–292.
Greco, S., Słowiński, R., Szczech, I., 2016. Measures of rule interestingness in various perspectives of confirmation. Information Sciences 346-347, 216–235.
Grecov, P., Prasanna, A.N., Ackermann, K., Campbell, S., Scott, D., Lubman, D.I., Bergmeir, C., 2022. Probabilistic causal effect estimation with global neural network forecasting models. IEEE Transactions on Neural Networks and Learning Systems , 1–15.
Gunnarsson, B.R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., Lemahieu, W., 2021. Deep learning for credit scoring: Do or don't? European Journal of Operational Research 295, 292–305.
Gürses-Tran, G., Körner, T.A., Monti, A., 2022. Introducing explainability in sequence-to-sequence learning for short-term load forecasting. Electric Power Systems Research 212, 108366.
Hastie, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning - Data Mining, Inference, and Prediction (2nd Edition). Springer.
He, J., Liu, X., Duan, Q., Chan, W.K.V., Qi, M., 2022. Reinforcement learning for multi-item retrieval in the puzzle-based storage system. European Journal of Operational Research .
Hewage, H.C., Perera, H.N., De Baets, S., 2022. Forecast adjustments during post-promotional periods. European Journal of Operational Research 300, 461–472.
Hindle, G., Kunc, M., Mortensen, M., Oztekin, A., Vidgen, R., 2020. Business analytics: Defining the field and identifying a research agenda. European Journal of Operational Research 281, 483–490.
Höppner, S., Baesens, B., Verbeke, W., Verdonck, T., 2022. Instance-dependent cost-sensitive learning for detecting transfer fraud. European Journal of Operational Research 297, 291–300.
Höppner, S., Stripling, E., Baesens, B., vanden Broucke, S., Verdonck, T., 2020. Profit driven decision trees for churn prediction. European Journal of Operational Research 284, 920–933.
Hyndman, R.J., Athanasopoulos, G., 2021. Forecasting: Principles and Practice. volume 3rd. OTexts.
INFORMS, 2015. Definition of analytics. URL: https://www.informs.org/About-INFORMS/News-Room/O.R.-and-Analytics-in-the-News/Best-definition-of-analytics.
Inuiguchi, M., Yoshioka, Y., Kusunoki, Y., 2009. Variable-precision dominance-based rough set approach and attribute reduction. International Journal of Approximate Reasoning 50, 1199–1214.
Islam, M.R., Ahmed, M.U., Barua, S., Begum, S., 2022. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences (Switzerland) 12.
Iyer, A.V., Bergen, M.E., 1997. Quick response in manufacturer-retailer channels. Management Science 43, 559–570.
van Jaarsveld, W., 2020. Deep controlled learning of dynamic policies with an application to lost-sales inventory control. CoRR abs/2011.15122.
Kang, Y., Cao, W., Petropoulos, F., Li, F., 2022. Forecast with forecasts: Diversity matters. European Journal of Operational Research 301, 180–190.
Keeney, R.L., Raiffa, H., 1979. Decisions with multiple objectives: Preferences and value trade-offs. IEEE Transactions on Systems, Man and Cybernetics 9, 403.
Khosrowabadi, N., Hoberg, K., Imdahl, C., 2022. Evaluating human behaviour in response to ai recommendations for judgemental forecasting. European Journal of Operational Research 303, 1151–1167.
Kordzadeh, N., Ghasemaghaei, M., 2022. Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems 31, 388–409.
Kosasih, E.E., Brintrup, A., 2022. A machine learning approach for predicting hidden links in supply chain with graph neural networks. International Journal of Production Research 60, 5380–5393.
Kotłowski, W., Dembczyński, K., Greco, S., Słowiński, R., 2008. Stochastic dominance-based rough set model for ordinal classification. Information Sciences 178, 4019–4037.
Kotłowski, W., Słowiński, R., 2008. Statistical approach to ordinal classification with monotonicity constraints, in: Fürnkranz, J., Hüllermeier, E. (Eds.), ECML/PKDD 2008 Workshop on Preference Learning. Proc. ECML/PKDD 2008 Workshop, p. 16 pages.
Kotłowski, W., Słowiński, R., 2009. Rule learning with monotonicity constraints, in: Proceedings of the 26th International Conference On Machine Learning, ICML 2009. Omnipress, ACM International Conference Proceedings Series, vol. 382, art. no. 67, Montreal, pp. 537–544.
Kozodoi, N., Jacob, J., Lessmann, S., 2022. Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research 297, 1083–1094.
Kozodoi, N., Katsas, P., Lessmann, S., Moreira-Matias, L., Papakonstantinou, K., 2020. Shallow self-learning for reject inference in credit scoring, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer. pp. 516–532.

Kraus, M., Feuerriegel, S., Oztekin, A., 2020. Deep learning in business analytics and operations research: Models, applications and managerial implications. European Journal of Operational Research 281, 628–641.
Krauss, C., Do, X.A., Huck, N., 2017. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500. European Journal of Operational Research 259, 689–702.
Kriebel, J., Stitz, L., 2022. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. European Journal of Operational Research 302, 309–323.
Kusunoki, Y., Błaszczyński, J., Inuiguchi, M., Słowiński, R., 2021. Empirical risk minimization for dominance-based rough set approaches. Information Sciences 567, 395–417.
Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research 247, 124–136.
Li, L., 2018. Predicting online invitation responses with a competing risk model using privacy-friendly social event data. European Journal of Operational Research 270, 698–708.
Li Long, C., Guleria, Y., Alam, S., 2021. Air passenger forecasting using neural granger causal google trend queries. Journal of Air Transport Management 95, 102083.
Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2021. Explainable ai: A review of machine learning interpretability methods. Entropy 23, 1–45.
Liu, J., Lin, S., Xin, L., Zhang, Y., 2022a. Ai vs. human buyers: A study of alibaba's inventory replenishment system. SSRN Electronic Journal .
Liu, M., Liu, Z., Chu, F., Zheng, F., Chu, C., 2021. A new robust dynamic bayesian network approach for disruption risk assessment under the supply chain ripple effect. International Journal of Production Research 59, 265–285.
Liu, W., Wei, W., Choi, T.M., Yan, X., 2022b. Impacts of leadership on corporate social responsibility management in multi-tier supply chains. European Journal of Operational Research 299, 483–496.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, pp. 4766–4775.
Ma, S., Fildes, R., 2021. Retail sales forecasting with meta-learning. European Journal of Operational Research 288, 111–128.
Mai, F., Tian, S., Lee, C., Ma, L., 2019. Deep learning models for bankruptcy prediction using textual disclosures. European Journal of Operational Research 274, 743–758.
Martens, D., 2008. Building acceptable classification models for financial engineering applications. ACM SIGKDD Explorations Newsletter 10, 30–31.
Martens, D., 2022. Data Science Ethics: Concepts, Techniques, and Cautionary Tales. Oxford University Press.
Martens, D., Baesens, B., Gestel, T.V., 2009. Decompositional rule extraction from support vector machines by active learning. IEEE Transactions on Knowledge and Data Engineering 21, 178–191.
Martens, D., Provost, F., 2014. Explaining data-driven document classifications. MIS Quarterly: Management Information Systems 38, 73–99.
Martens, D., Provost, F., Clark, J., de Fortuny, E.J., 2016. Mining massive fine-grained behavior data to improve predictive analytics. MIS Quarterly: Management Information Systems 40, 869–888.
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys 54.
Mitra, S., Karathanasopoulos, A., Sermpinis, G., Dunis, C., Hood, J., 2015. Operational risk: Emerging markets, sectors and measurement. European Journal of Operational Research 241, 122–132.
Mitrović, S., Baesens, B., Lemahieu, W., De Weerdt, J., 2018. On the operational efficiency of different feature types for telco churn prediction. European Journal of Operational Research 267, 1141–1155.
Molnar, C., 2022. Interpretable Machine Learning: A Guide For Making Black Box Models Explainable.
Montero-Manso, P., Athanasopoulos, G., Hyndman, R.J., Talagala, T.S., 2020. Fforma: Feature-based forecast model averaging. International Journal of Forecasting 36, 86–92.
Montero-Manso, P., Hyndman, R.J., 2021. Principles and algorithms for forecasting groups of time series: Locality and globality. International Journal of Forecasting 37, 1632–1653.
Mortenson, M.J., Doherty, N.F., Robinson, S., 2015. Operational research from taylorism to terabytes: A research agenda for the analytics age. European Journal of Operational Research 241, 583–595.
Nicholson, W.B., Wilms, I., Bien, J., Matteson, D.S., 2020. High dimensional forecasting via interpretable vector autoregression. Journal of Machine Learning Research 21, 1–52.
Nieddu, L., Patrizi, G., 2000. Formal methods in pattern recognition: A review. European Journal of Operational Research 120, 459–495.
Nikolopoulos, K., Goodwin, P., Patelis, A., Assimakopoulos, V., 2007. Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. European Journal of Operational Research 180, 354–368.
Olafsson, S., Li, X., Wu, S., 2008. Operations research and data mining. European Journal of Operational Research 187, 1429–1448.
Óskarsdóttir, M., Ahmed, W., Antonio, K., Baesens, B., Dendievel, R., Donas, T., Reynkens, T., 2022. Social network analytics for supervised fraud detection in insurance. Risk Analysis 42, 1872–1890.
Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., Baesens, B., 2019. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing Journal 74, 26–39.
Óskarsdóttir, M., Cornette, S., Deseure, F., Baesens, B., 2020. Inductive representation learning on feature rich complex networks for churn prediction in telco, in: Studies in Computational Intelligence, Springer. pp. 845–853.
Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., Hyndman, R.J., 2021. Forecast reconciliation: A geometric view with

new insights on bias correction. International Journal of Forecasting 37, 343–359.
Pawlak, Z., 1982. Rough sets. International journal of computer & information sciences 11, 341–356.
Perera, H.N., Hurley, J., Fahimnia, B., Reisi, M., 2019. The human factor in supply chain forecasting: A systematic review. European Journal of Operational Research 274, 574–600.
Peters, G., Weber, R., 2018. dynxcube – categorizing dynamic data analysis. Information Sciences 463-464, 21–32.
Piri, S., Delen, D., Liu, T., Zolbanin, H.M., 2017. A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble. Decision Support Systems 101, 12–27.
Radcliffe, N.J., 2007. Using control groups to target on predicted lift: Building and assessing uplift model. Direct Marketing Analytics Journal , 14–21.
Ramon, Y., Martens, D., Provost, F., Evgeniou, T., 2020. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c. Advances in Data Analysis and Classification 14, 801–819.
Ranyard, J.C., Fildes, R., Hu, T.I., 2015. Reassessing the scope of or practice: The influences of problem structuring methods and the analytics movement. European Journal of Operational Research 245, 1–13.
Rao, A., Greenstein, B., 2022. PwC 2022 AI business survey. URL: `https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-business-survey.html`.
Rathore, P., Soni, J., Prabakar, N., Palaniswami, M., Santi, P., 2021. Identifying groups of fake reviewers using a semisupervised approach. IEEE Transactions on Computational Social Systems 8, 1369–1378.
Reddy, B.K., Delen, D., 2018. Predicting hospital readmission for lupus patients: An rnn-lstm-based deep-learning methodology. Computers in Biology and Medicine 101, 199–209.
Reddy, B.K., Delen, D., Agrawal, R.K., 2019. Predicting and explaining inflammation in crohn's disease patients using predictive analytics methods and electronic medical record data. Health Informatics Journal 25, 1201–1218.
Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144.
Roy, B., 2005. Paradigms and challenges, in: Figueira, J., Greco, S., Ehrgott, M. (Eds.), International Series in Operations Research and Management Science. Springer, Boston. volume 78, pp. 3–24.
Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206–215.
Ruiz, C., Spiliopoulou, M., Menasalvas, E., 2010. Density-based semi-supervised clustering. Data Mining and Knowledge Discovery 21, 345–370.
Sakib, N., Ibne Hossain, N.U., Nur, F., Talluri, S., Jaradat, R., Lawrence, J.M., 2021. An assessment of probabilistic disaster in the oil and gas supply chain leveraging bayesian belief network. International Journal of Production Economics 235, 7.
Saltos, R., Weber, R., Maldonado, S., 2017. Dynamic rough-fuzzy support vector clustering. IEEE Transactions on Fuzzy Systems 25, 1508–1521.
Satell, G., Abdel-Magied, Y., 2020. Ai fairness isn't just an ethical issue. Harvard Business Review , 1–5.
Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., Er, M.J., Ding, W., Lin, C.T., 2017. A review of clustering techniques and developments. Neurocomputing 267, 664–681.
Schock, R., 1962. A note on subjunctive and counterfactual implication. Notre Dame Journal of Formal Logic 3, 289–290.
Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D., 2015. Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems 2015-January, 2503–2511.
Semenova, L., Rudin, C., Parr, R., 2022. On the existence of simpler machine learning models, in: ACM International Conference Proceeding Series, ACM Proc. pp. 1827–1858.
Senoner, J., Netland, T., Feuerriegel, S., 2022. Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. Management Science 68, 5704–5723.
Seyedan, M., Mafakheri, F., 2020. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. Journal of Big Data 7, 53.
Sezer, O.B., Gudelek, M.U., Ozbayoglu, A.M., 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing Journal 90, 106181.
Shmueli, G., 2010. To explain or to predict? Statistical Science 25, 289–310.
Słowiński, R., Greco, S., Matarazzo, B., 2002. Axiomatization of utility, outranking and decision rule preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle. Control and Cybernetics 31, 1005–1035.
Słowiński, R., Greco, S., Matarazzo, B., 2020. Rough sets in decision-making, in: Meyers, R.A. (Ed.), Encyclopedia of Complexity and Systems Science. Springer, Berlin, Heidelberg, pp. 1–50.
Soares De Melo Junior, L., Nardini, F.M., Renso, C., Fernandes De MacEdo, J.A., 2019. An empirical comparison of classification algorithms for imbalanced credit scoring datasets, in: Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019, Ieee. pp. 747–754.
Stevenson, M., Mues, C., Bravo, C., 2021. The value of text for small business default prediction: A deep learning approach. European Journal of Operational Research 295, 758–771.
Stripling, E., vanden Broucke, S., Antonio, K., Baesens, B., Snoeck, M., 2018. Profit maximizing logistic model for customer churn prediction using genetic algorithms. Swarm and Evolutionary Computation 40, 116–130.
Szeląg, M., Błaszczyński, J., Słowiński, R., 2017. Rough set analysis of classification data with missing values, in: Et al., L.P. (Ed.), Lecture Notes in Computer Science. Springer, pp. 552–565.
Tambe, P., Cappelli, P., Yakubovich, V., 2019. Artificial intelligence in human resources management: Challenges and a path forward. California Management Review 61, 15–42.

Tank, A., Covert, I., Foti, N., Shojaie, A., Fox, E.B., 2022. Neural granger causality. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 4267–4279.
Teunter, R.H., Syntetos, A.A., Babai, M.Z., 2017. Stock keeping unit fill rate specification. European Journal of Operational Research 259, 917–925.
Troncoso, F., Weber, R., 2020. A novel approach to detect associations in criminal networks. Decision Support Systems 128, 113159.
Turrini, L., Meissner, J., 2019. Spare parts inventory management: New evidence from distribution fitting. European Journal of Operational Research 273, 118–130.
Van Belle, R., Van Damme, C., Tytgat, H., De Weerdt, J., 2022. Inductive graph representation learning for fraud detection. Expert Systems with Applications 193, 116463.
Vanvuchelen, N., De Boeck, K., Boute, R., 2022. Cluster-based lateral transshipments for the zambian health supply chain.
Vanvuchelen, N., Gijsbrechts, J., Boute, R., 2020. Use of proximal policy optimization for the joint replenishment problem. Computers in Industry 119, 103239.
Verbeke, W., Dejaeger, K., Martens, D., Hur, J., Baesens, B., 2012. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218, 211–229.
Verbeke, W., Martens, D., Baesens, B., 2017. Rulem: A novel heuristic rule learning approach for ordinal classification with monotonicity constraints. Applied Soft Computing Journal 60, 858–873.
Verbraken, T., Bravo, C., Weber, R., Baesens, B., 2014. Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research 238, 505–513.
Verbraken, T., Verbeke, W., Baesens, B., 2013. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering 25, 961–973.
Vidgen, R., Hindle, G., Randolph, I., 2020. Exploring the ethical implications of business analytics with a business ethics canvas. European Journal of Operational Research 281, 491–501.
Vidgen, R., Shaw, S., Grant, D.B., 2017. Management challenges in creating value from business analytics. European Journal of Operational Research 261, 626–639.
Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harvard Journal of Law & Technology 31, 841–887.
Wang, R., Miao, Z., Liu, T., Liu, M., Grdinovac, K., Song, X., Liang, Y., Delen, D., Paiva, W., 2021. Derivation and validation of essential predictors and risk index for early detection of diabetic retinopathy using electronic health records. Journal of Clinical Medicine 10, 1473.
Wang, Y., Wang, Y.X., Singh, A., 2015. Differentially private subspace clustering, in: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 1000–1008.
Westerski, A., Kanagasabai, R., Shaham, E., Narayanan, A., Wong, J., Singh, M., 2021. Explainable anomaly detection for procurement fraud identification–lessons from practical deployments. International Transactions in Operational Research 28, 3276–3302.
Winkler, R.L., Makridakis, S., 1983. The combination of forecasts. Journal of the Royal Statistical Society. Series A (General) 146, 150.
Xu, R., Wunsch, D., 2005. Survey of clustering algorithms. IEEE Transactions on Neural Networks 16, 645–678.
Yang, C.C., 2022. Explainable artificial intelligence for predictive modeling in healthcare. Journal of Healthcare Informatics Research 6, 228–239.
Yang, Z., Zhang, A., Sudjianto, A., 2021. Gami-net: An explainable neural network based on generalized additive models with structured interactions. Pattern Recognition 120, 108192.
Ye, Y., Lu, Y., Robinson, P., Narayanan, A., 2022. An empirical bayes approach to incorporating demand intermittency and irregularity into inventory control. European Journal of Operational Research 303, 255–272.
Yoshikawa, Y., Iwata, T., 2021. Gaussian process regression with interpretable sample-wise feature weights. IEEE Transactions on Neural Networks and Learning Systems .
Zeng, Z., Li, M., 2021. Bayesian median autoregression for robust time series forecasting. International Journal of Forecasting 37, 1000–1010.
Zhang, B.H., Lemoine, B., Mitchell, M., 2018. Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA. p. 335–340.
Zhu, B., Baesens, B., Backiel, A., Vanden Broucke, S.K., 2018. Benchmarking sampling techniques for imbalance learning in churn prediction. Journal of the Operational Research Society 69, 49–65.
Zolbanin, H.M., Delen, D., Crosby, D., Wright, D., 2020. A predictive analytics-based decision support system for drug courts. Information Systems Frontiers 22, 1323–1342.