KU LEUVEN

Essays on data-driven decision support: Applications in HRM and methodological advances

Dissertation presented to obtain the degree of Doctor in Business Economics at KU Leuven

by

Simon De Vos

Number *000* September 2025

Since the theses in the series published by the Faculty of Economics and Business are the personal work of their authors, only the latter bear full responsibility.

Doctoral Committee

Promotor

Prof. Dr. Wouter Verbeke	KU Leuven

Doctoral committee

Prof. Dr. Johannes De Smedt	KU Leuven
Prof. Dr. Marijke Verbruggen	KU Leuven
Prof. Dr. Chris Wuytens	KU Leuven
Prof. Dr. Stefan Lessmann	Humboldt University of Berlin

Chair

Prof. Dr. Robert Boute	KU Leuven

Acknowledgements

Left empty for acknowledgements

Simon De Vos

page for dedication

Summary

Data plays an increasingly important role in managerial decision-making. Improved data availability, greater computing resources, and advances in machine learning (ML) enable organizations to derive insights from historical data through pattern recognition, helping them anticipate future outcomes. Analytical methods must match the decision problem and context at hand. This holds especially for applications in human resource management (HRM), where decisions impact individuals directly and legal and ethical considerations are critical.

This dissertation builds on three types of analytics: *descriptive* to summarize past patterns, *predictive* to estimate future outcomes, and *prescriptive* to recommend actions. Each serves a different role in decision-making and shapes the contributions in both parts of the dissertation.

Part I covers HR analytics applications, focusing on internal mobility, employee turnover, and internal job matching. The work was developed in close collaboration with Acerta, an HR services provider, through an iterative process of discussion, implementation, and validation.

Part II presents methodological contributions that generalize beyond HR to bring ML closer to the complexities of real-world decision-making. It introduces robust techniques for instance-dependent cost-sensitive classification, fairness-aware learning for resource allocation, and a predict-then-optimize framework for uplift modeling with continuous treatments.

By combining applied and methodological work, the dissertation helps close the gap between technical developments and practical needs. It offers concrete tools for organizations aiming to make better decisions with data.

Contents

Prologue

1
Introduction

Data-driven approaches play an increasingly significant role in managerial decision-making [1]. Improved data availability, increased computing resources, and advances in machine learning (ML) enable organizations to derive insights from historical data, identify patterns, anticipate future outcomes, and inform actions [2]. Other than effective data management — covering aspects such as collection practices, assuring data quality, storage, and governance — it is essential to rely on analytical methods suited to the specific problem at hand. Reliable insights require a precise alignment between problem characteristics and analytical approaches. Given the growing enthusiasm for ML and AI — and the accompanying risk of inflated expectations — it is crucial to carefully develop, evaluate, and select methods that genuinely address organizational needs. The alignment of analytical methods with domain-specific goals, constraints, and challenges is particularly important in sensitive fields like human resource (HR) management, where data is often more sensitive, decisions have direct implications for individuals, and privacy and fairness concerns are heightened compared to other managerial decision-making domains.

1.1 Decision support: From descriptive to predictive and prescriptive analytics

Data-driven decision-making can be supported by three progressively sophisticated categories of analytics:

1. Descriptive analytics summarizes historical data, uncovering patterns that explain past events. It addresses questions like: *'What happened in the past, and what are the key trends or patterns in the data?'* Formally, given dataset $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_k\}_{i=1}^{N}$, descriptive analytics employs an operator $S : \mathcal{D} \rightarrow P$, mapping raw data into a set of summary statistics or patterns P. For Chapter 2, this includes discovering employee mobility patterns from historical HR data.

2. Predictive analytics goes a step further by using statistical and ML methods to forecast future outcomes based on historical data. It aims to answer the question *'What is likely to happen in the future based on historical data?'*. Here, an outcome Y is approximated by a predictive model as $Y = f(X)$, where input features X predict outcome Y. The function f is estimated based on historical data. Chapter 3 employs predictive analytics to estimate employee turnover probability. Chapters 5 and 6 methodologically integrate additional managerial considerations — cost-sensitivity and fairness, respectively — to better align model predictions with organizational objectives.

3. Prescriptive analytics recommends actionable decisions by optimizing outcomes given input features and constraints. It aims to answer the question *'What actions should be taken to achieve a desired outcome?'* and is formalized as:

$$Y = f(X, T), \quad T^* = \arg\max_{T \in \mathcal{T}} Y(X, T),$$

where $\mathcal{T}$ represents the feasible set of decisions and T^* represents an optimal decision. The complexity of the $argmax$ operator can vary strongly depending on the use case's context. Chapter 4 focuses on recommending a next career step by predicting a job-employee fit score and selecting the highest-ranking option as a recommendation. In contrast, Chapter 7 also considers cost-sensitivity and fairness in the decision-making step, making the $argmax$ operator more complex. In such cases, more advanced optimization techniques are required to effectively identify an optimal decision.

From analysis to decisions

The analytical methods differ in how they support decision-making: descriptive analytics informs without making predictions or prescribing actions; predictive analytics forecasts outcomes, indirectly influencing decisions; prescriptive analytics explicitly integrates decision-making into analytical models, directly guiding optimal actions.

Regardless of these distinctions, translating analytical outputs into managerial decisions remains challenging. This difficulty primarily arises from uncertainty: future states and their associated outcomes often cannot be precisely estimated or fully quantified. Analytical insights — whether descriptive, predictive, or prescriptive — are therefore always conditional on assumptions, approximations, and data quality. Effective use of analytics thus demands rigorous problem formulation, explicit alignment of analytical methods with organizational goals and constraints, and careful consideration of how uncertainty impacts decisions.

Research objectives and structure

This dissertation addresses two central research questions:

1. How can analytics effectively support decision-making within HR management, particularly regarding internal mobility and employee turnover?
2. How can analytical methods be adapted — specifically by incorporating fairness and cost-sensitivity — to better align with managerial objectives?

To systematically approach these questions, this dissertation is organized in two parts. Part I focuses on HR analytics applications, showcasing contributions to optimize HR-related decisions. Part II introduces methodological advances on robustness, fairness, cost-sensitivity, and continuous-valued treatments to better align analytics with the intricacies of real-world managerial contexts.

Together, these contributions in applied and methodological research are intended to empower organizations to make effective, efficient, and ethical decisions that support their managerial objectives.

1.2 Part I: HR analytics applications

As defined by [3]: *'HR analytics involves the systematic application of analytical techniques — including statistical modeling, data mining, and ML — to address challenges in HR management and establish measurable business impact through data-driven decisions.'* In this dissertation, HR is explicitly considered a strategic organizational partner rather than merely an administrative function [4]. Driven by intensified competition for talent and recognition of human capital as a crucial organizational asset [5], HR analytics leverages employee-related data to support decisions on, for example, hiring, retention, internal mobility, and talent development. As discussed in Section 1.1, because of improved data availability, increased computational resources, and methodological advancements, also the potential of HR analytics has been transformed from traditional descriptive reporting toward sophisticated predictive and prescriptive approaches [6].

Despite these opportunities, implementing HR analytics remains challenging. Employee data often resides in fragmented and incompatible systems, complicating integration and meaningful analysis [5]. Additionally, ethical debates and regulations such as GDPR and the European AI Act impose significant compliance requirements regarding fairness, transparency, and bias mitigation in data-driven HR decisions [7], [8]. Furthermore, HR outcomes are often intangible, delayed, difficult to measure, or context-dependent, hindering straightforward modeling and evaluation. Consequently,

while some organizations successfully apply predictive and prescriptive analytics, many continue to rely primarily on descriptive methods [3], [5]. A notable gap thus persists between academic research — which often prioritizes methodological innovation — and industry practice, emphasizing interpretability, practical usability, and immediate applicability.

This dissertation bridges these perspectives by presenting practical HR analytics applications in Part I, focusing on internal mobility patterns (Chapter 2), predicting employee turnover (Chapter 3), and optimizing internal job recommendations (Chapter 4). Collectively, these chapters demonstrate concrete analytical approaches that organizations can adopt, emphasizing both methodological rigor and practical relevance.

1.3 Part II: Methodological advances

Standard analytical methods typically optimize traditional performance metrics, such as predictive accuracy or precision [2]. However, organizational decision-making frequently demands incorporating additional considerations beyond these metrics, such as cost-sensitivity and algorithmic fairness [9], [10]. Cost-sensitive analytics and algorithmic fairness are relevant in many domains, including HR, where decisions carry direct ethical implications and financial consequences [11], [12]. To tackle these requirements, Part II of this dissertation introduces methodological advances specifically focused on cost-sensitive analytics and fairness-aware decision-making. Although these methods are illustrated in application contexts other than HR — such as fraud detection, marketing, healthcare, or lending — the contributions presented are fundamentally methodological and remain broadly applicable.

Cost-sensitive learning is an ML paradigm that explicitly accounts for costs or benefits associated with prediction errors or correct decisions [13]. Unlike conventional approaches that typically maximize accuracy-based metrics, cost-sensitive learning aims to minimize total incurred costs — or, conversely, maximize overall benefits — associated with model predictions or decisions. This is particularly valuable when misclassification consequences are asymmetric or when specific outcomes entail higher costs than others. For instance, in fraud detection, the primary objective is not simply maximizing the detection rate but minimizing total financial losses arising from fraudulent activities [14]. Prioritizing the investigation of high-value transactions, even at the cost of overlooking less consequential cases, better aligns modeling objectives with real-world business priorities. The cost-sensitive approach can be incorporated either directly during the training phase (*predict-and-optimize*, Chapter 5) or through a separate post-training optimization step (*predict-then-optimize*, Chapter 7).

In a similar vein, cost-sensitive learning is relevant to HR analytics applications. In a predictive context, it enables modeling employee turnover with greater nuance. Although Chapter 3 addresses turnover prediction as a binary classification task, not all employee departures are equally undesirable — some turnover instances may even have beneficial effects [15]. Cost-sensitive methods, as the methodological focus of Chapter 5, incorporate these nuances by assigning differentiated costs or benefits to turnover cases. In a prescriptive setting, cost-sensitive analytics guides resource allocation toward higher-impact HR decisions. For example, retention strategies might prioritize functions that are difficult or costly to replace, while recruitment efforts may target candidates promising the highest returns, as illustrated methodologically in Chapter 7.

Fairness in algorithmic decision-making addresses the ethical concern that data-driven methods, like ML models, should not perpetuate or amplify discriminatory biases present in historical data [16]. In HR applications, fairness is particularly relevant because these systems often involve sensitive personal data and have direct implications for individuals. For example, biased algorithms can lead to unfair hiring or promotion practices, which is evident from the infamous Amazon hiring case, which was discriminatory toward women [17]. Moreover, beyond the ethical aspects, there are also legal implications. The European AI Act [18] categorizes most AI applications in HR as high-risk. This means that organizations using such tools in the EU for decision-making have *'strict obligations'* to *'minimize risks of discriminatory outcomes'* [18].

By incorporating fairness considerations into their modeling process, organizations can ensure that their decisions are not only effective in terms of standard metrics like accuracy or profit but, also align with broader business objectives and adhere to fairness principles. However, fairness in algorithmic decision-making is inherently complex, making the selection of an appropriate fairness concept challenging [19]. Many commonly used fairness metrics, like demographic parity and equal opportunity, are incompatible with each other unless under very specific conditions [20]. Although the literature on algorithmic fairness presents a wide range of fairness notions and metrics [16], [21], we focus specifically on fairness through independence concepts. This dissertation focuses on fairness in a predictive setting (Chapter 6), where we consider a *predict-and-optimize approach* to include fairness during the model training process, and in a prescriptive setting (Chapter 7), where fairness is considered a constraint in a *predict-then-optimize* setting.

1.4 Publications and outline

We conclude this introduction by highlighting the main contributions of this dissertation and outlining its structure. The dissertation's structure is presented in Figure 1.1. The main body consists of Chapters 2-7, organized into Part I and Part II.

Chapter 2 presents a case study with Acerta, where internal mobility is examined using process analytics. Specifically, process discovery techniques are applied to HR event logs to generate employee journey maps that visualize career paths within the organization. These maps reveal that internal mobility is often more complex than assumed, highlighting discrepancies between expected career trajectories and actual mobility patterns. By highlighting uncommon growth paths, stepping stones to certain functions, and describing hard-to-fill positions, the study provides a descriptive tool for managing internal employee mobility. This chapter has been published as [22]:

S. De Vos, J. De Smedt, C. Wuytens, *et al.*, "Leveraging process mining to optimize internal employee mobility strategies," in *Business Process Management Cases Vol. 3: Implementation in Practice*, Springer, 2025, pp. 15–28

Chapter 3 addresses employee turnover using predictive analytics. Employee turnover imposes substantial costs on organizations, not only due to recruitment and training expenses but also through the loss of expertise and reduced productivity, making the prediction of which employees are at risk of leaving crucial for improving retention strategies and ensuring workforce stability. To address the fragmented nature of existing research in this area, the contribution of this chapter is twofold. It first presents a scoping review that highlights inconsistencies in existing research and provides a comprehensive consolidation of existing literature. Next, it provides a benchmarking experiment with 14 classification methods on 9 datasets. The combination of both contributions provides a structured overview of the current state-of-the-art, identifies key challenges in employee turnover prediction, and provides a focal point on employee turnover prediction research. This chapter has been published as [23]:

S. De Vos, C. Bockel-Rickermann, J. Van Belle, *et al.*, "Predicting employee turnover: Scoping and benchmarking the state-of-the-art," *Business & Information Systems Engineering*, pp. 1–20, 2024

Chapter 4 explores how prescriptive analytics can support internal mobility within organizations, taking the process perspective as introduced by Chapter 2 as a starting point. It presents a data-driven recommender system

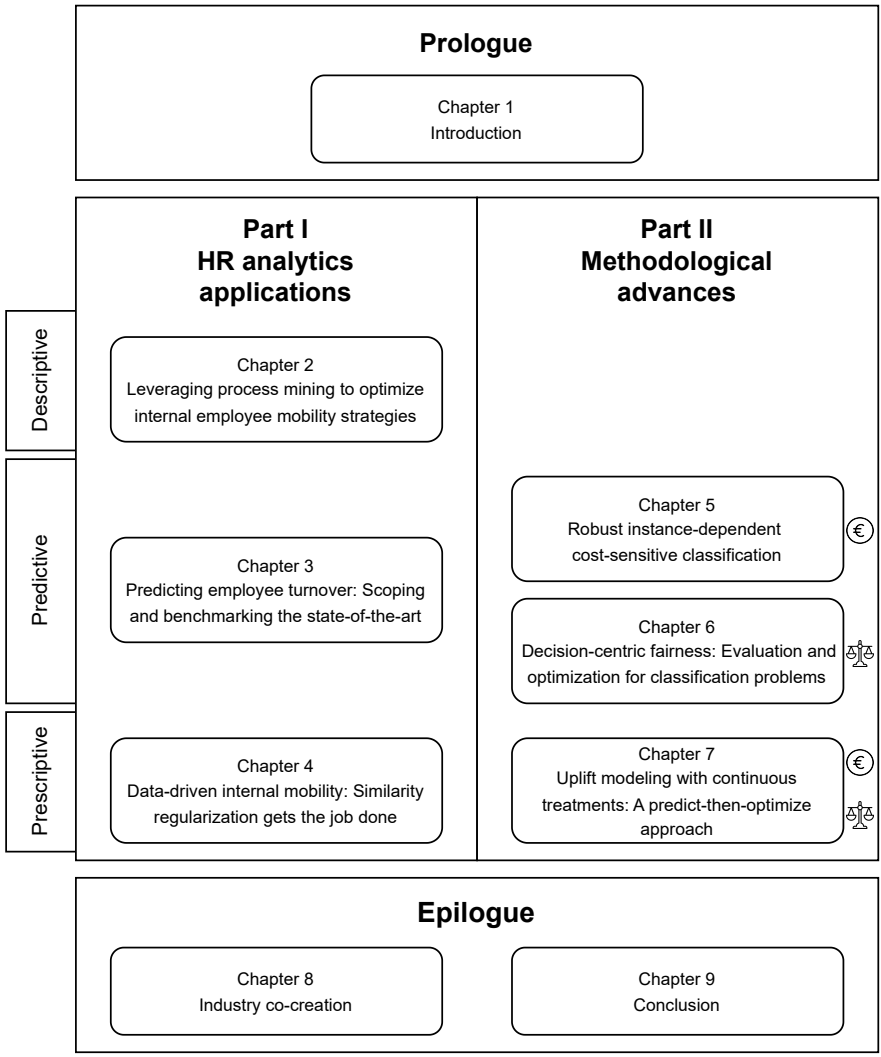


Figure 1.1: Overview of this dissertation's structure. The main body consists of Chapters 2-7, organized into Part I and Part II.

that matches employees to internal job opportunities. The system builds on collaborative filtering techniques, enhanced with a similarity-based regularization term incorporating employee characteristics. This addition addresses the cold start issue often faced in internal placement systems. The approach is evaluated using three real-life datasets, showing strong performance compared to established benchmarks. In addition to the technical contributions, we publish a new HR dataset for further research. This chapter has been published as [24]:

S. De Vos, J. De Smedt, M. Verbruggen, *et al.*, "Data-driven internal mobility: Similarity regularization gets the job done," *Knowledge-Based Systems*, vol. 295, p. 111 824, 2024

Chapter 5 presents a methodological contribution to cost-sensitive predictive analytics. Specifically, we focus on improving instance-dependent cost-sensitive (IDCS) learning methods. While such methods are useful for binary classification tasks with varying misclassification costs, we experimentally show that they are often not robust to noise and outliers concerning cost information. This chapter introduces an end-to-end method to enhance their robustness: detecting outliers, correcting their cost data if needed, and integrating the adjusted information into the IDCS method. When implemented with cslogit, this approach results in r-cslogit, a more resilient method that achieves better performance across different noise levels. This chapter has been published as [25]:

S. De Vos, T. Vanderschueren, T. Verdonck, *et al.*, "Robust instance-dependent cost-sensitive classification," *Advances in Data Analysis and Classification*, vol. 17, no. 4, pp. 1057–1079, 2023

Chapter 6 provides a methodological contribution to predictive analytics by proposing a decision-centric approach to fairness in binary classification problems. Recognizing that traditional fairness metrics do not adequately capture the nuances of practical decision-making contexts, this chapter introduces an approach that explicitly targets fairness within actionable decision regions. This ensures that fairness considerations align closely with real-world business practices, maintaining model performance and versatility across varied decision thresholds within the specified decision area. The effectiveness of the proposed method is demonstrated empirically, highlighting substantial benefits for applications where fairness is critical, such as various HR applications. This chapter is under review at *European Journal of Operational Research*. A preprint has been published online as [26]:

S. De Vos, J. Van Belle, A. Algaba, *et al.*, "Decision-centric fairness: Evaluation and optimization for resource allocation problems," *arXiv preprint arXiv:2504.20642*, 2025

Chapter 7 introduces a novel methodological framework within prescriptive analytics, specifically addressing uplift modeling with continuous-valued treatments. It extends traditional binary-treatment approaches by proposing a predict-then-optimize method that first estimates conditional average dose responses using causal ML techniques, followed by solving a dose-allocation problem through integer linear programming. This flexible optimization approach integrates fairness constraints and instance-dependent costs, enabling efficient, utility-maximizing resource allocation. Experimental evaluations demonstrate the framework's versatility and effectiveness, showcasing its practical applicability and advantages across diverse fields such as healthcare, lending, and HR management. This chapter is currently under revision at *European Journal of Operational Research*. A preprint has been published online as [27]:

S. De Vos, C. Bockel-Rickermann, S. Lessmann, *et al.*, "Uplift modeling with continuous treatments: A predict-then-optimize approach," *arXiv preprint arXiv:2412.09232*, 2024

Chapter 8 reflects on the co-creation process with Acerta, involving iterative feedback loops between industry and academia. It discusses how this collaboration shaped the research in Part I and evaluates its practical value, while also exploring the potential for further implementation of the methodological advances from Part II in an applied setting.

Chapter 9 concludes this dissertation with a general discussion, offering a reflection on our contributions and their limitations. We also outline potential future research directions to extend and build upon our findings.

Part I

HR analytics applications

2
Leveraging process mining to optimize internal employee mobility strategies

The significance of human resource (HR) analytics in facilitating data-driven decision-making for managing internal employee mobility has been emphasized by the recent increasing competition in attracting and retaining the best employees referred to as the war for talent. Existing HR analytics methods typically provide support for operational and tactical decision-making. However, there is a need for long-term strategic decision support. Additionally, as current methods for managing internal mobility are being challenged, the development of new appropriate HR analytics methods is necessary.

In collaboration with KU Leuven, Acerta Consult implemented process mining techniques to address this issue. Specifically, process discovery techniques were applied to the event logs of HR data to generate employee journey maps (EJMs) that depict the different historic paths employees have taken within an organization.

These EJMs demonstrated the difference between idealized career paths and the actual complexity of employee mobility. These discrepancies have the potential to reshape the incorrect assumptions held by HR managers. The data-driven insight gained through these EJMs can assist HR professionals by providing decision support for a wide range of cases including the identification of infrequent growth paths, analyzing hard-to-fill positions, and better understanding the causes of turnover.

The process perspective on internal mobility provides valuable insights for HR managers and was able to shed light on the general complexity of careers. As a result, this perspective can serve as a foundation for further analyses, including predictive and prescriptive modeling, while taking into account HR-specific constraints and challenges.

2.1 Introduction

In recent years, many organizations have focused on utilizing the vast amount of data available to support decision-making in their daily activities in order to gain a competitive advantage. Like other business areas, HR departments are now attempting to use data to support their operational, tactical, and strategic decisions.

As reported by the Financial Times, the global trend of increasing competition for talent in the job market is evidenced by the all-time low unemployment rates in the eurozone (6.6% of the workforce) and the high number of job openings in the US (roughly two per unemployed worker) [28]. In response, companies are offering higher wages and benefits to attract and retain employees, leading to increases in the cost of goods and services. Employers are implementing strategies to address high turnover rates, which include incentives such as bonuses and career development opportunities. To support the development of these strategies, HR analytics is more relevant than ever in terms of attracting and retaining good employees.

Organizations like Acerta are facing these challenges in HR analytics and investigate questions such as *How can companies optimize their employees' career paths?* or *What paradigm should they start from?* This business case takes a process perspective on internal mobility, specifically the implementation of process mining techniques, to support HR managers at Acerta Consult. Together with their research partners at KU Leuven, they are internally exploring the capabilities of applying process mining techniques to HR data. With the insight from this case, they aim to further broaden the range of HR analytics services offered to their clients.

Acerta is a major HR service provider in Belgium with 25 offices across the country. They serve a diverse range of customers, including starters, self-employed, SMEs, and large companies, managing administration for one million employees in the Belgian market. Acerta has 1600 employees, a 20% market share, and services 350,000 self-employed workers and 40,000 companies in Belgium, resulting in a yearly turnover of 260 million euros. Acerta Consult, the company's consulting branch with 450 employees, offers services such as recruitment and selection, outplacement, legal, and training.

Specifically, we applied process mining techniques to longitudinal career data to discover employee journey maps (EJMs) and as such examine internal employee mobility. The technical details of this approach are explained in Section 2.3.1 The behavioral nature of careers requires the use of dynamic methods for modeling, making process mining techniques a suitable method for analyzing internal mobility. These techniques can be effectively combined with existing human experience to provide insight based on data to support, rather than automate, decision-making.

This article summarizes how to conceptualize and implement considerations of internal mobility as a process, through some example HR-related use cases from Acerta. Section 2.2 first provides an introduction to the field of HR analytics and explains why data-driven methods are needed for strategic decision support. Next, it presents the specific problem statement concerning the current internal mobility management methods and discusses why a process perspective is beneficial. Section 2.3 describes the actions taken, the methods used to implement the process perspective envisioned for managing internal mobility, and the data structure. The results are discussed in Section 2.4 and the challenges faced and lessons learned from the case are discussed in Section 2.5.

2.2 Situation faced

To maintain its position as one of Belgium's leading providers of HR services, Acerta Consult stays current with the latest developments in data-driven support for strategic HR decisions and maintains a focus on employee mobility. The first part of this section explains why we need a solution for strategic decision support and discusses the growth and evolution of HR tech, accompanied by the challenges of adoption. To highlight the usefulness of our process analytics approach as a solution, the second part outlines the issues with current methods of internal mobility management in relation to the specific case study of Acerta.

2.2.1 Need for strategic decision support

The use of technology within the HR field is not a new phenomenon, but the HR tech domain has evolved significantly in recent years and is becoming active in an ever-expanding range of applications. In the 1990s, HR processes were successfully integrated within enterprise resource planning systems (ERPs), whereas we now see HR-specific information systems (HRISs) often being used. This has led to the development of solutions for specific operational domains, such as recruitment, performance management, onboarding, training and development, payroll administration, compensation and benefits, alternative workforce planning, reporting, and internal communication. In short, operational HR applications and their corresponding tech industry are booming.

However, the adoption of HR tech for strategic decision support remains limited [29], [30]. Based on a comprehensive review of the relevant literature [29], [30] and the extensive experience of Acerta as an HR services provider, we have identified three challenges that contribute to the lagging adoption of data-driven decision support in HR compared to other fields.

First, there is a self-evident gap in the knowledge and abilities of the data science teams who create analytical tools and the HR departments that are to use them. This disparity often leads to challenges in both the development and adoption of these tools. Second, employee data is sensitive and the data that is available on employees is often limited, making it difficult to apply advanced methods that require large amounts of data, for example, machine learning. For example, to construct a system for automated CV screening, a vast amount of training data is necessary. However, on account of the sensitive data and limited availability, this requirement might not be fulfilled in practice. Third, HR professionals are typically charged with making tactical and strategic decisions over the medium to long term, with outcomes that are difficult to quantify and observe. These decisions are typically made based on unstructured and varied sources of information, as well as expertise and experience. In contrast, data-driven methods are typically used to support repetitive operational decisions with clear short-term outcomes, for which homogeneous and structured data sources are readily available. As a result, HR professionals may be difficult to support with the arsenal of tools that is typically used by data scientists in other domains.

In response to these challenges, we propose applying process mining techniques to longitudinal HR data. The EJMs offer three key advantages that align with HR practices and resources. First, the process mining tooling in the form of dashboards is easy to use and allows HR practitioners to work with analytical instruments by bridging the gap between the data-related knowledge and skills of the HR departments and the data science teams. Second, as we were not searching for the best process model but using EJMs for data representation that will be interpreted by an HR professional, process mining techniques can be applied to smaller amounts of data. Third, by analyzing longitudinal data, this approach, complemented by the expertise of HR professionals, is well suited to supporting medium- to long-term decision-making with difficult-to-quantify outcomes. In summary, the proposed approach offers a useful solution that addresses three typical challenges that we have identified in the use of data-driven techniques for strategic HR decision-making.

2.2.2 Problems in current internal mobility management

Currently, internal mobility at Acerta is managed using a certain amount of HR professional expertise and intuition in combination with evidence-based techniques from more traditional academic HR literature. However, with this management approach, we uncovered two additional challenges specific to internal mobility. In addition to the general issues in HR tech previously

mentioned, these two challenges highlight the need for data-driven insight into internal mobility.

The first challenge we pinpointed was the recent changes in managerial formats. Like many organizations, Acerta is increasingly moving away from rigid organizational structures and delayering hierarchy. Therefore, the traditional career ladders that steered the internal flows of employees have become less evident and less standardized. In addition, a shift toward focusing on the skills of individuals and the management of their competencies is taking place, reinforcing the aforementioned trend toward less standardization in internal mobility. As a result, traditional methods of managing employee pathways are being increasingly challenged.

The second challenge that we identified is known as the *fragmented process knowledge* problem [31]. This problem arises when domain experts, such as HR professionals, have a limited understanding of the overall process of managing mobility within an organization. However, HR professionals typically have a very detailed understanding of their specific responsibilities. The HR-employee ratio can vary depending on factors such as the size and complexity of an organization, but is roughly 1:50, i.e., 1 HR professional per 50 employees, in Acerta's case. Furthermore, Acerta's HR data reveals that the average tenure of an HR professional is around 4 years, based on their specific role and seniority level. The fragmented process knowledge of internal mobility among various HR professionals hampers centralized control, hindering its effective management within the organization.

2.3 Action taken

KU Leuven and Acerta took action by considering internal employee mobility as a process. In this section, we first explain how HR concepts can be translated into process concepts and how we used process discovery techniques. Then we discuss the data required from a technical and practical perspective.

2.3.1 Process mining as a solution

Process mining incorporates the process perspective into data mining and machine learning and helps organizations further comprehend their business processes [32]. It involves analyzing event logs to uncover the underlying process models, bridging the gap between traditional model-based process analysis and data-centric analysis methods. By analyzing historical event data, process owners like HR professionals can gain insight into business processes, such as, in this case study, internal mobility, where career paths are the individual traces.

In order to understand the behavior contained within an event log, automatic process discovery techniques were utilized to construct a model. Various approaches could have been employed for this purpose [32]. However, implementing process mining in real-world scenarios is a challenging task. In the case of internal mobility, we see many process variants and the data is typically censored because process instances, i.e., careers, take a long time. Additionally, traces for internal mobility processes usually contain few activities compared to those found in other settings [33]. As a result, its implementation at Acerta was a highly iterative process and we emphasized the need for close collaboration between process analysts and business experts.

This case study deploys process discovery techniques for data representation and visualization rather than for building the best model. Moreover, in the context of employee journey mapping, it is not only important to identify common paths taken by employees but also important to consider less frequent and obvious paths. Thus, directly-follows graphs (DFGs) were utilized as the preferred method of representation as they are the de facto standard in the industry [34]. However, as there are some potential risks associated with using DFGs and frequency-based simplification, it is important for practitioners to have a thorough understanding of how these process models are generated [34]. To address this, Acerta arranged information sessions to provide active guidance during deployment.

We generated EJMs on internal mobility by translating concepts from the field of process mining into HR concepts as follows:

- A case is conceptually equivalent to an employee. Over time, the employee can transition from one function to another, undertaking a journey.
- An activity translates to occupying a function within an organization. It is carried out by an employee and is characterized by a timestamp indicating when the activity starts and ends. In our particular HR context, there is no overlap or parallelism in the activities of an employee as each employee can only have one function at a time.
- A trace covers all the activities performed in a particular process instance by a specific case. Therefore, it is equivalent to an employee journey. Each trace is defined as the ordered set of subsequent functions that an employee occupies throughout their career within an organization.
- Applying process discovery techniques to an event log consisting of longitudinal HR data is conceptually the same as discovering EJMs. The event log is defined as the collection of traces and, therefore, contains information on all employee journeys. An EJM is the aggregate of all the paths employees have taken within an organization.

Table 2.1: Synthetic example data $\mathcal{D}$ in the format of an event log.

u	t^s	t^e	v	x_1	x_2	x_3	x_4	x_5	y	
1	10/2014	06/2016	Job 1	MSc	Physics	F	1975	1	0.4	$\mathcal{D}_1$
1	07/2016	02/2019	Job 3	MSc	Physics	F	1975	1	0.8	
1	03/2019	07/2022	Job 5	MSc	Physics	F	1975	1	0.5	
2	09/2009	02/2016	Job 1	BSc	Finance	M	1981	0.8	0.9	$\mathcal{D}_2$
2	03/2016	07/2022	Job 4	BSc	Finance	M	1981	0.8	0.3	
3	06/2016	03/2019	Job 2	PhD	Electronics	M	1977	1	0.8	$\mathcal{D}_3$
3	04/2019	07/2022	Job 3	PhD	Electronics	M	1977	1	0.3	
4	...	...	...	...	...	...	...	...	...	

2.3.2 Data requirements

Typically, an event log in the setting of HR has the structure as visualized in Table 2.1. An event in a career path is a tuple $(u_i, t_i^s, t_i^e, v_i, \mathbf{x}_i)$ where u_i is a unique identifier for an employee, holding the function v_i from time t_i^s to time t_i^e. On top of control flow information, employee u_i has features $\mathbf{x}_i$ which can include amongst others degree, branch of study, gender, date of birth, and the full-time equivalent.

An employee profile $\mathcal{D}_i \subset \mathcal{D}$ of the person u_i consists of the combination of $|\mathcal{D}_i|$ tuples where $|\mathcal{D}_i|$ is the number of functions that this employee has occupied within this organization. Consequently, an employee profile $\mathcal{D}_i$ consists of the visited jobs and personal information $\mathbf{x}_i$ of person u_i.

In order to gain access to this career data in event log format, Acerta faced several challenges when undertaking the HR analytics project. The first hurdle to overcome was the issue of data protection. Several stakeholders were involved in the project and each had their own interests to defend. For example, HR professionals were enthusiastic about moving forward with the project, while the legal department was, understandably, more cautious about sharing confidential data. In order to comply with the General Data Protection Regulation (GDPR), employment contracts had to be amended to allow for the use of employee data for analytical purposes. Second, there was the issue of data collection. Despite the substantial amount of available data, it was dispersed across several sources. For instance, data on previous employment was stored in different locations to the personal employee information and assessment interview results. As a result, this required an additional step of pooling the data together. Third was the issue of data cleaning. Processes in other contexts tend to be relatively short compared to career path data, which was set to a time frame of 10 years. As this period is longer than other typical data analysis projects, impurities entered the data as a result of, among other things, internal restructurings, changes in job titles, and the merging or division of positions.

2.4 Results achieved

In this section, we discuss the outcomes attained using our proposed approach at Acerta. First, the achieved results are discussed in their most general form. Our approach allows for the automatic discovery of EJMs, which describe and quantify internal mobility within an organization in a data-driven way. This leads to new insight for HR professionals. Then, we demonstrate how such EJMs can be used to solve specific HR cases.

2.4.1 General results

Our approach takes the descriptive character of EJMs as a starting point. These maps are mainly useful for control flow analysis, i.e., the internal movement of employees. Additionally, extra filtering techniques can be applied to personal features like educational details, office location, and salary bracket.

Figures 2.1 and 2.2 present two EJMs, represented by DFGs, that contain the activity *func 177*. To generate these visuals, we use the tool Disco by Fluxicon. They are anonymized examples of internal employee mobility at Acerta and demonstrate the usefulness of a good EJM. In addition to other functions, we include two special activities: *before_ data_ capture* indicates that the individual was already an employee of the organization prior to the start date of data capturing, and *leave* represents an employee leaving the organization and occurs instantaneously. The transition between a function and the *sink* suggests that the individual is still active in this function at the end of the data-capturing period.

Figure 2.1 shows what internal mobility, according to HR professionals, looks like by design. In this figure, we filter on traces containing the activity *func 177* and display only 15% of the most frequent activities. Hence, the EJM only shows a selection of the most frequent traces. When we showed this to the HR professionals at Acerta, they agreed that this was a plausible map of the internal mobility in that company. To add nuance to this, although HR professionals often recognize that this representation is a simplification of reality, they find it challenging to pinpoint what exactly the more complex reality looks like. This figure shows a typical traditional, linear career path from junior to senior to expert positions. These are typical vertical movements with little variation.

However, in reality, Figure 2.2 shows that paths are more complex. We filter on traces with the activity *func 177* and displayed 50% of the most frequent activities. The data reveals a wide range of career paths observed within the organization with less linear and more varied behavior. Most employees follow unique paths, frequently transitioning between unexpected

and less obvious activities, rather than progressing through the traditional career ladders depicted in Figure 2.1.

Growth paths do not always align with expectations, as Figure 2.2 shows a more diverse set of transitions than Figure 2.1. In this way, EJMs can help break HR managers' incorrect assumptions. It should be noted that even Figure 2.2 is still a simplification of reality as it only displays the 50% most frequent activities. The full EJMs are too detailed to be included as a figure in this article.

When both a process model and an event log are provided, the discrepancy between the two highlights the value of data-driven techniques. This discrepancy can be linked to conformance checking, where the goal is to find similarities and differences between the behavior modeled and the behavior observed.

After making this connection with conformance checking, process redesign became the logical next step. The differences between Figures 2.1 and 2.2 allowed the HR professionals to face the facts and deal with the as-is situation. This, in turn, allows them to implement well-thought-out and well-informed changes to their strategic HR management concerning internal mobility by, for example, focusing on active efforts to stimulate transitions between management functions and more technical functions. There are two main reasons why redesigning a business process can be beneficial, as discussed by [31]. The first reason is that the nature of organizations is often organic, and business processes naturally evolve over time, in general, becoming more complex, which leads to a decline in performance. This also applies to policies related to internal mobility that change over time. The second reason is that business environments are constantly changing, and HR strategies must be flexible in response.

2.4.2 Concrete HR Cases

In this subsection, we present a condensed overview of the insight obtained from the analysis presented in the previous sections in relation to the implementation at Acerta and their expert input. We make this tangible with a selection of examples.

1. *Detection of infrequent and less obvious paths.* In process analytics, the optimal workflow from the first to the final step is often the most efficient and effective. Process managers are interested in identifying and analyzing this workflow in order to optimize the process. In contrast, HR professionals are regularly interested in the opposite, in identifying infrequent and less obvious paths to explain questions such as unexpected success stories of internal mobility, transferability of skills, and atypical growth paths.

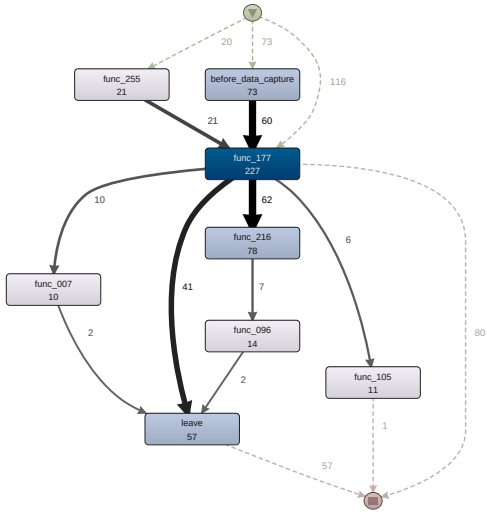


Figure 2.1: A simplified EJM with career paths that contain the activity *func 177*.

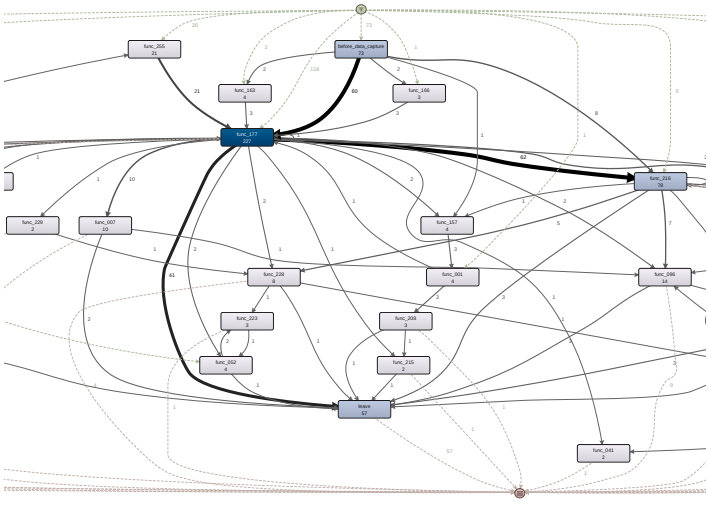


Figure 2.2: An EJM displaying what career paths containing the activity *func 177* truly look like.

2. *Hard-to-fill positions.* Analysis of the in- and outflows of selected positions can identify favorable employee properties or profiles to attract talent in the future. For instance, when it is difficult to hire someone for a business intelligence team leader role, our longitudinal approach can look at the past experiences and future plans of employees who previously held that role. This will help us identify the qualities of employees who are likely to stay in the position for a long time.
3. *Vertical vs. lateral movements.* Both types of career moves can be beneficial for employees. Acerta stimulates both the exploration of different paths and the level of interaction between them. Examples of how our method can help here include detecting the prevalence of vertical movement and the typical positions between which lateral movement occurs.
4. *Paths that lead to turnover.* Internal mobility can impact turnover behavior in various ways. If employees have opportunities to move within a company and find more fulfilling or suitable roles, they may be less likely to leave the company altogether. Furthermore, if internal mobility opportunities are limited or employees are not satisfied with the available roles, they may be more likely to seek opportunities elsewhere, resulting in higher turnover rates. Our approach helps Acerta understand and quantify the link between internal mobility and employee turnover behavior.

2.5 Lessons learned

This story of co-creation between KU Leuven and Acerta is unique in terms of the field of application of process mining. To close this chapter, we discuss our learnings and the challenges we faced when adopting process-driven HR analytics and then conclude with a short closing word.

2.5.1 Challenges

Data The data in this case study on process discovery techniques in HR analytics is very sparse as it covers 1600 employees, 250 job titles, and only around 4000 job transitions over a 10-year period. This means that the majority of employees have only one or two jobs in their internal career, resulting in very short traces. Additionally, the trace length distribution is highly skewed, which can further complicate the analysis. Hence, we emphasize that DFGs are used for data visualization and representation purposes rather than for finding the best process model. Another difficulty is the observability of the data. While 10 years of event data may seem like a long time, careers are often much longer, with some starting before the

data capture began. This resulted in a large number of unobserved prefixes in the traces, which can potentially result in incorrect EJMs. Conversely, many cases will still be ongoing when the data collection ends, causing these cases to be censored. A third challenge is the availability and quality of the data. In order to accurately represent employee journeys, a long period of data capture is required. However, data from this far back in time is often not available, and even when it is, it may contain inconsistencies because of organizational changes and the splitting or merging of certain functions.

In addition to these technical challenges, there are also legal considerations to take into account. The GDPR sets strict constraints on the use of personal data and the upcoming European AI Act will impose additional restrictions on the use of AI-driven applications in HR, such as employment, management of workers, and CV-sorting software [8]. These constraints must be carefully considered in order to ensure compliance and avoid legal issues.

Adoption Process discovery techniques are effective for analyzing internal mobility but implementing them can be challenging on account of stakeholder interests. In any organization, there are numerous stakeholders with different interests and concerns. This can make it challenging to gain consensus and move forward with projects, particularly when it comes to sharing sensitive data. In this case study, the HR manager was eager to get started but the legal department was understandably cautious about sharing confidential data. Only after several discussions and explanations of the tool and its capabilities did the stakeholders agree to move forward. This experience highlights that simply having a method available is not enough; stakeholders must understand it and see its value. A solution here is to start with a specific problem and make the deliverable tangible.

2.5.2 Insights gained

It is important to note that there may be differences in the focus and interests of *traditional* process analysts and HR professionals when it comes to these techniques. While process analysts may be more interested in the technical aspects of the processes themselves, HR professionals may be more focused on the implications for employees and the organization as a whole.

These techniques have been shown to partly support the existing knowledge and practices at Acerta. However, they have also provided new insight in the sense that they were able to correct several incorrect assumptions the managers had. This highlights the complementary nature of process discovery techniques and existing knowledge and expertise in HR. However, it is also important to recognize that the use of analytics in HR is not without its challenges and constraints.

As such, it is crucial to carefully manage expectations when using process discovery techniques and other analytical approaches in HR. It is important to make it clear beforehand what these techniques can and cannot do and, thereby, avoid making overly optimistic or unrealistic claims about their capabilities. While these techniques can provide valuable insight based on existing data, they should not be seen as a substitute for creative and thoughtful HR policies. Instead, they should be viewed as a tool that can help HR professionals better understand and optimize their existing processes.

2.5.3 Conclusion

KU Leuven and Acerta Consult collaborated on the development of an HR analytics method by treating internal employee mobility as a process and providing a more objective approach to strategic decision-making. The article discussed the method's development, challenges, and lessons learned, highlighting its positive reception at Acerta and potential for future development as a service. The method can provide valuable insight for managing employee mobility in HR analytics where the adoption of data-driven techniques is currently lagging behind other fields.

3
Predicting employee turnover: Scoping and benchmarking the state-of-the-art

Employee turnover presents a significant challenge to organizations. High turnover rates impose substantial costs on organizations, e.g., direct costs resulting from rehiring efforts and training new employees, and indirect costs resulting from the loss of expertise and declining organizational productivity. Hence, predicting employee turnover is an important task for human resource departments and organizations as a whole, as it can help to proactively approach employees at risk of churning to improve retention and workforce stability. With ever more data at hand and increasing competition in the labor market, analytical tools are inevitable to improve workforce management and aid human resource managers in their decision-making. Yet, the existing literature on predictive analytics for employee turnover is scattered and fails to present a coherent and holistic view. To find common ground in the established literature, the paper provides a scoping and benchmarking of the state-of-the-art. The scoping concludes that established research results are difficult to compare due to inconsistent methodologies and experimental setups. To overcome these issues, an extensive benchmarking experiment is conducted including 14 classification methods and 9 datasets. The results provide a unique focal point for research on employee turnover prediction and aim to benefit academic research and industry practitioners. The code and public datasets are available on Github to facilitate further extension of the research.

3.1 Introduction

Employee turnover, describing the unplanned departure of employees [35], presents a substantial challenge to organizations. High or unforeseen turnover rates can impose substantial costs through recruitment, hiring, and training expenses for new employees. These costs are typically estimated to surpass annual salary costs [36]. Losing employees can have several other detrimental effects, such as a decline in organizational productivity and competitiveness due to a decrease in employee morale, reduced engagement, and a loss of skills and expertise [37]. Hence, identifying employees at risk of leaving [38] is a critical task for human resource (HR) departments as it enables organizations to counter employee turnover through targeted retention efforts, and provides essential information for strategic workforce planning, talent management, and succession planning, all of which are critical aspects of modern HR management.

Existing research on employee turnover prediction has utilized various approaches. Contrary to relying solely on expert-based approaches [39], also traditional statistical methods [40], and machine learning techniques [41] have been used to develop predictive models for employee turnover.

Predictive analytics seamlessly integrates into traditional business intelligence (BI), often forming a part of Human Resource Information Systems (HRIS) [42]. By incorporating analytics components, the aim is to enhance decision-making support within systems. HRIS have a rich history, where efforts to develop balanced scorecards are examples of elementary predictive systems [42]. As decision-makers observe and manage a larger number of subjects and incorporate ever-increasing amounts of data, the shift from qualitative expert-based decision-making to the use of quantitative predictive models is an important advancement. This advancement, reflecting the ongoing evolution of technology within organizational contexts [43], has spurred research in the scientific literature on employee turnover prediction. To date, this research has explored the predictiveness of different types of data and has investigated both different classification methods – ranging from logistic regression to neural networks – and performance metrics to evaluate predictive accuracy [40].

Comparing the results of previous studies is challenging due to several reasons, including the use of different datasets and classification methods, inconsistent data preprocessing, different approaches to hyperparameter tuning, and selective reporting of performance metrics. This makes it difficult to draw meaningful conclusions by meta-analyzing previous research findings. Existing research on employee turnover prediction is lacking a well-defined common ground with respect to the applicability of analytical methods, the robustness of performance, and the procedure to evaluate performance. The state-of-the-art is yet to be determined.

This paper presents an overview of recent studies in employee turnover prediction, involving two key components: a scoping review and a benchmarking experiment. This is a widely adopted approach with studies published in research areas such as predictive process monitoring [44], energy quantification [45], and customer-centric decision support [46]. First, the scoping review aims to identify current research topics and limitations in existing studies. By conducting a structured scoping of the existing literature, we identify the limitations and inconsistencies in established methodologies and experimental setups. Second, the benchmarking experiment enables the evaluation of established methodologies for predicting employee turnover and an identification of the state-of-the-art. We follow a comprehensive and standardized methodology for evaluating and comparing the performance of 14 different classification methods on 9 datasets, including synthetic and real data, and report on a wide range of metrics. Statistical tests are carried out to assess the significance of performance differences between classifiers. Additionally, we test the impact of various class rebalancing methods and feature selection.

With this paper, we aim to target professionals and academics at the intersection of data science and human resource management. The scoping review and benchmarking experiment, the main components of this study, are not standalone. Rather, they are complementary to each other. Our goal is to establish a distinct focal point in predictive analytics for employee turnover. Our contributions to the literature on employee turnover prediction are the following:

- We consolidate the scattered literature on employee turnover prediction and provide a structured scoping of previous research.
- Based on established methodologies for employee turnover prediction, we provide a structured original experimental setup that enables fair comparison of methods along established and novel datasets, as well as a range of established performance metrics. Our code and public datasets are available on Github to facilitate further extension of our research.
- We guide both industry practitioners and academic researchers in identifying targeted and effective methodologies for decision-making and future research, respectively.

The remainder of this paper is organized as follows. In Section 3.2 we define and summarize the research field of employee turnover prediction and discuss immediately related works. Section 3.3 presents our scoping review, spanning a total of 56 studies. The setup of our benchmarking experiment is presented in Section 3.4, of which results are presented and discussed in Section 3.5. We conclude in Section 3.6 by formulating suggestions and key takeaways for academics and practitioners.

3.2 Related literature

Employee turnover in the traditional HR literature.
Employee turnover, defined as employees' voluntary termination of their employment relationships [38], has captivated the interest of scholars and practitioners alike for well over a century. It is a dynamic and ever-evolving field [47]. To gain comprehensive insights into this phenomenon, researchers have provided overviews in reference works such as [38], [47].

The literature has recognized a range of demographic and job-related factors that impact employee turnover [48]. Nevertheless, there is frequent disagreement regarding the direction and extent of these factors' effects on turnover. Consider, for example, the impact of age on turnover. Direct effects vary across studies, with some suggesting that turnover intention first increases and then decreases with age [49], while others observe a negative relationship between age and turnover intention [38], [40], [50]. For teachers, the relationship between age and turnover has been found to be U-shaped [51]. On top of that, turnover can also be influenced by indirect effects like job satisfaction and organizational commitment. These effects are complex, with some studies suggesting a positive linear relationship between job satisfaction and age [52], [53], while others find a U-shaped pattern [54].

In traditional turnover theories, broad principles are often employed to address entire populations. For instance, these theories commonly suggest that job satisfaction reduces turnover, while embeddedness ties employees to their organizations [55]. However, a compelling question arises: do turnover theories manifest differently across distinct segments of employee populations? Consider the possibility that rational decision-making models may be more relevant to long-term employees in stable industries than to short-term employees in highly dynamic work environments [47]. Additionally, could we take this a step further and apply these theories customized to the individual employee level?

Recent investigations have steered away from a one-size-fits-all perspective on turnover, favoring instead theories that specify the conditions under which particular factors become more or less influential in employees' decisions to quit [47]. These conditions can range from a focus on more specific subgroups based on factors like sector (e.g., construction [56], pharmacists [57], hospitality [58]) to an individual level focus (e.g., the Proximal Withdrawal States Theory (PWST) [59] and the unfolding model [60]).

While there is a growing trend to tackle the problem of employee turnover at a higher level of granularity, the above studies are all still explanatory in nature [61]. That is, they are all focused on theory building and rather rely on data and statistical modeling to test (multiple) causal hypotheses within the theoretical framework that is being developed or scrutinized.

With increasingly more and richer HR data becoming available, the adoption of data-driven models for predicting employee turnover is becoming more widespread. This alternative approach is predictive in nature, and applies statistical or machine learning methods to data for the purpose of predicting new or future observations [61]. Consequently, this approach naturally focuses on the individual employee level. This data-driven modeling approach to employee turnover prediction constitutes a key component of HR analytics.

HR Analytics.
Within the value stream of an employee journey [39], [62], HR Analytics serves as an umbrella term encompassing various potential applications. These include recruitment and selection [11], [63], [64], internal mobility management [24], learning and development, performance management and prediction [65], skills, talent, and succession management [66]. Numerous different definitions of HR Analytics have been proposed and a good overview can be found in reference work such as [3].

HR Analytics can be broadly categorized into three levels: descriptive, predictive, and prescriptive [67]. In this context, these three levels can address the following questions:

1. Descriptive analytics reveal and describe current and historical data patterns and relationships. E.g.: What is the turnover rate within different departments, and what are the distinctive attributes of employees who have departed from the organization?
2. Predictive analytics extend this by inferring predictions about the future which are derived from the current and historical data. E.g.: What is the probability that a specific employee will leave the organization?
3. Prescriptive analytics offer guidance on how to achieve desired outcomes. E.g.: What actions should be taken to minimize the risk of an employee leaving?

Predictive analytics for employee turnover.
In the HR setting, we stress the importance of steering clear of decision automation. Instead, our emphasis lies in outlining potential scenarios and evaluating the risk of employee turnover in a predictive framework, refraining from prescribing specific actions. In contrast, descriptive analytics depict past occurrences. Meanwhile, prescriptive analytics suggests optimal actions or decisions based on the outcomes of predictive analytics, to optimize results or attain specific goals under a particular policy [63]. In this sense, although prescriptive analytics has proven its worth in other similar domains, such as customer churn [68], predictive modeling serves as a prerequisite for prescriptive modeling. Therefore, in this study, we focus on predictive analytics for employee turnover.

As discussed above, the use of data-driven methods for predicting employee turnover at the individual employee level is becoming more widespread as increasingly more and richer HR data becomes available. As opposed to the theory-driven explanatory modeling paradigm, this modeling approach is predictive in nature as it starts from available organization-specific historical data (vs. a theoretical model) to predict employee turnover for unseen observations. For a detailed discussion on the differences between explanatory and predictive modeling, one may refer to [61]. Hence, it offers valuable insights for effective turnover management within individual organizations as it can be used by HR practitioners to identify (and characterize) employees who are at risk of leaving voluntarily and unexpectedly, so that they can try to prevent turnover in a targeted manner. Moreover, data-driven methods often provide more reliable information about the underlying (data) mechanisms [69].

Our study focuses on the application of data-driven methods to predict employee turnover at the individual employee level, moving away from expert-driven and explanatory modeling approaches.

3.3 Scoping predictive analytics for employee turnover

To understand the current state of predictive analytics solutions for employee turnover, we are adopting a scoping review methodology. This enables the identification and mapping of the key concepts, sources, and types of evidence available in a given research area [70]. We will present the methodology to our review in Section 3.3.1, present findings in Section 3.3.2, and distill gaps in the research field in Section 3.3.3.

3.3.1 Review methodology

Our approach to conducting a scoping review is primarily inspired by the framework outlined by Arksey and O'Malley [70] and Munn, Peters, Stern, *et al.* [71]. As presented in Table 3.1, our search strategy involves six distinct stages: database selection, topic selection, record selection, abstract review, full-paper review, and data extraction. Each stage is designed to ensure comprehensive coverage of the literature on predictive analytics for employee turnover while restricting the scope to relevant studies.

Step 1: Database selection.
We conduct an exhaustive search of the Scopus and Web of Science (WoS) databases. These databases are renowned for their extensive coverage of

Table 3.1: Search strategy and number of records. n.m.: Not meaningful

Filtering Step	Decision	Number of records	
		Scopus	Web of Science
Database selection	- Scopus - Web of Science	n.m.	n.m.
Topic selection	- "Prediction" ‖ "Predicting" ‖ "Forecasting" ‖ "Prognosis" - "Employee" ‖ "Worker" ‖ "Laborer" ‖ "Jobholder" - "Turnover" ‖ "Attrition" ‖ "Churn" ‖ "Departure" - "Analytics" ‖ "Data" ‖ "Machine learning"	459	220
Time frame	- 2008 - 2022	377	191
Outlets	- Journals - Conference Proceedings	326	183
Fields	- Business - Economics - Mathematics - Computer Sci. - Engineering	221	126
Language	- English	217	125
Duplicates	- Excl. double counting		276
Abstract review	- Excl. irrelevant records		97
Full paper review	- Excl. irrelevant records		56

diverse scholarly publications and records, rendering them highly esteemed and commonly adopted for academic investigations [72].

Step 2: Topic selection.
The field of predictive analytics for employee turnover lacks a standardized terminology (cf. Section 3.2). To streamline our search, we established four essential components that a study must possess to be considered a candidate: (i) predicting (ii) employee (iii) turnover via a (iv) data-driven approach. We have included synonyms and related search terms for each of these elements in our search strategy, as outlined in Table 3.1. Our screening process involves analyzing titles, abstracts, and keywords from both databases considered. The complete search queries are presented in Table A.1 in the Appendix.

Step 3: Record selection.
The third step of our methodology includes substeps aimed at refining our retrieved studies by imposing additional criteria. As analytical research has made significant advancements in recent years, we will focus on publications from 2008 onwards. Further, to filter any non-analytical studies, we will focus on manuscripts published in scientific journals and conference proceedings, ensuring the quality of records and comparability in terms of the intended audience and writing style. Additionally, we will filter manuscripts by their research field to filter any non-analytical studies. The research fields encompass *business*, *computer science*, *economics*, *engineering*, and *mathematics*. We only consider studies written in English. Adhering to these criteria reduced the number of eligible papers to 276.

Step 4: Abstract review.
Before analyzing the full-text manuscript, we conduct a feasibility check by screening the abstracts of the remaining manuscripts. To this end, we employ the following exclusion criteria, resulting in 97 papers:

- Papers that are not HR-related. This group can be divided into two subcategories: Firstly, papers addressing closely related topics, such as customer churn prediction, that may have passed through previous filters. Secondly, papers that employ homonymous terms, such as *turnover* in a financial context, rather than an HR-related context.
- Papers that adopt a macroeconomic approach. Examples include papers related to labor economics, those that examine the labor market or adopt a non-micro perspective.
- Papers that are not specifically focused on predicting employee turnover as a concrete problem, but rather concentrate on a particular aspect of it or provide only a high-level overview of a framework. Examples of such papers include those on 'the conceptual exploration of...', 'the ethical implications of...' and 'the managerial implications of...'.

Step 5: Full-text review.
Finally, we performed a full-text screening on the remaining papers and excluded any manuscript that: (i) contains clear technical mistakes, such as an improper train/test split or reporting results on training data; (ii) lacks a clear selection of predictive methods; (iii) does not approach employee turnover prediction as a binary classification problem. This includes papers that predict job satisfaction or turnover intention rather than employee turnover itself. Following these screening criteria, we identified a final 56 papers that meet our inclusion criteria.

Step 6: Data extraction.
The relevant data from the selected papers is extracted and organized. The results are presented in the following section.

3.3.2 The scope of predictive analytics for employee turnover

The results of our scoping review are summarized in Table 3.2. For each article, we report the dataset(s) used, the methodology applied (including class balancing, feature selection, and hyperparameter tuning), and the type of evaluation metrics used. We distilled five key insights from our scoping review:

1. The number of different datasets used across studies is limited. Most manuscripts only consider a single dataset for assessing classifier performance and only three studies have considered more than one dataset.

Table 3.2: Overview of established literature. *n.a.*: Not applicable

	Data					Methodology				Evaluation	
Ref.	Number datasets	Origin	Public (Name)	Num. Obs.	Num. Var.	Num. Classifier	Balanc.	Feature Select.	Para. tuning	Thresh.	AUC
[64]	1	High-tech		3,825	8	1					
[73]	1	Software		150	13	5				✓	
[74]	1	Manufacturing		881	44	1		✓		✓	
[41]	1	*n.a.*		1,575	25	3				✓	
[75]	1	IT		130	19	3				✓	
[76]	1	Call center		1,037	10	2				✓	✓
[77]	1	Call center		3,543	6	2				✓	✓
[78]	1	Automotive		*n.a.*	14	7				✓	
[79]	1	*Synthetic*	✓ (Kaggle1)	15,000	10	5				✓	
[80]	1	*Synthetic*	✓ (IBM)	1,470	30	6		✓		✓	
[81]	1	Call center		479	11	3		✓	✓	✓	
[82]	1	Kaggle1	✓ (Kaggle1)	15,000	10	6		✓		✓	✓
[83]	1	Telecom, Indonesia		16,649	11	3				✓	
[84]	1	*Synthetic*	✓ (IBM)	1,470	32	3	✓	✓	✓	✓	
[85]	1	*Synthetic*	✓ (IBM)	1,470	32	1				✓	
[86]	1	IT, India		1,650	22	1		✓		✓	
[48]	1	HR services, Belgium		13,484	13	1				✓	✓
[87]	1	*Synthetic*	✓ (IBM)	1,470	32	5		✓			
[88]	1	*Synthetic*	✓ (Kaggle1)	15,000	10	5		✓		✓	
[89]	1	*Synthetic*	✓ (Kaggle1)	15,000	10	5	✓	✓	✓	✓	✓
[90]	1	Communications, China		2,000	30	5		✓	✓	✓	✓
[91]	1	*Synthetic*	✓ (IBM)	1,470	32	5	✓	✓		✓	✓
[92]	1	*Synthetic*	✓ (Kaggle1)	15,000	10	4				✓	✓
[93]	2	*Synthetic* Banking, USA	✓ (IBM)	1,470 9,089	31 19	10		✓		✓	✓
[94]	1	High-tech, China		66,911	*n.a.*	8	✓		✓	✓	✓
[95]	1	Social network, China		47,257	9	6		✓	✓	✓	✓
[96]	1	Glassdoor		5,550	45	5		✓		✓	✓
[97]	1	*Synthetic*	✓ (IBM)	1,470	32	7				✓	
[98]	1	*Synthetic*	✓ (Kaggle1)	15,000	10	3				✓	
[99]	1	Social network, China		287,229	22	16		✓		✓	✓
[100]	1	Metallurgy		1,866	22	5	✓			✓	✓
[101]	1	Healthcare		12,000	24	6		✓		✓	✓
[102]	1	*Synthetic*	✓ (IBM)	1,470	32	9				✓	✓
[103]	1	*Synthetic*	✓ (IBM)	1,470	32	1	✓		✓	✓	
[104]	3	*Synthetic* *Synthetic* *n.a.*	✓ (IBM) ✓ (Kaggle1)	1,470 15,000 450	32 10 16	9		✓	✓	✓	
[105]	1	*Synthetic*	✓ (IBM)	1,470	32	6			✓	✓	
[106]	1	Packaging		6,552	*n.a.*	4	✓		✓	✓	✓
[107]	1	*Synthetic*	✓ (Kaggle1)	15,000	10	6		✓		✓	
[108]	1	*Synthetic*	✓ (IBM)	1,470	32	3	✓	✓	✓	✓	✓
[109]	1	*Synthetic*	✓ (IBM)	1,470	32	6		✓		✓	
[110]	1	*Synthetic*	✓ (IBM)	1,470	32	5				✓	✓
[111]	1	Consumer goods		365	7	3			✓	✓	
[112]	1	*Synthetic*	✓ (IBM)	1,470	32	6	✓	✓	✓	✓	✓
[113]	2	*Synthetic* *Synthetic*	✓ (IBM) ✓ (Kaggle1)	1,470 15,000	32 9	22		✓		✓	✓
[114]	1	*n.a.*		330	8	6		✓		✓	✓
[115]	1	Social network, China		287,229	24	7		✓		✓	✓
[116]	1	*Synthetic*	✓ (IBM)	1,470	32	3	✓	✓	✓	✓	✓
[117]	1	*Synthetic*	✓ (IBM)	1,470	32	9	✓		✓	✓	
[118]	1	*Synthetic*	✓ (IBM)	1,470	32	6		✓	✓	✓	✓
[119]	1	*Synthetic*	✓ (Kaggle1)	15,000	10	5		✓		✓	
[120]	1	Retail		1,186	9	6			✓	✓	
[121]	1	*Synthetic*	✓ (IBM)	1,470	32	3	✓	✓		✓	✓
[122]	1	*Synthetic*	✓ (IBM)	1,470	32	4	✓	✓	✓	✓	✓
[123]	1	*Synthetic*	✓ (IBM)	1,470	32	4		✓		✓	✓
[124]	1	*Synthetic*	✓ (IBM)	1,470	32	8	✓	✓	✓	✓	✓
[125]	1	*Synthetic*	✓ (IBM)	1,470	32	6		✓		✓	✓
This paper	9	cf. Table 3.4				14	✓	✓	✓	✓	✓

In addition, many studies have focused solely on synthetic data, with the public IBM and Kaggle1 datasets being used 24 and 10 times, respectively. This has led to both a lack of research diversity and has constrained the potential for generating broadly applicable insights that can be used in real-world applications.

2. The majority of studies use a limited set of classification methods, with most studies utilizing five or fewer methods. Six studies consider only one classification method. This lack of comparison limits the potential to draw robust conclusions.
3. The modeling pipelines in the studies are inconsistent. We see a variety of different approaches to, for example, preprocessing (in terms of training data balancing), feature selection, or hyperparameter tuning. These steps are essential to the performance of methods [126]. Inconsistency limits comparability between studies. For example, more than half of the selected studies do not perform hyperparameter tuning, and only about one-fourth of studies test the effect of rebalancing training data [127].
4. Studies do not report comparable performance metrics. Metrics can be categorized into different types based on the notion of classifier performance that they embody. Most studies report on threshold metrics where predictions are rounded to 1 if their propensities exceed a certain threshold and to 0 otherwise. Consequently, the stated performance strongly depends on this threshold, which is often not specified (and presumably equal to the default threshold of 0.5). Additionally, the metrics are not consistent across studies (for an overview of metrics, see Section 3.4). Threshold-independent area under the curve (AUC) metrics, such as area under the receiver operating characteristic curve (AUC-ROC) and Gini, which are popular in related fields like customer churn prediction [128], fraud detection [129] and credit risk modeling [130], are used in only about half of the studies. The difference in performance metrics does not allow for a coherent comparison across studies.
5. A notable observation with regard to the literature is the lack of attention that is given to statistical hypothesis testing. Specifically, a mere 5% of studies examine the statistical significance of performance differences among various classification methods. Moreover, when evaluating classifier performance across multiple datasets, relying on multiple pairwise t-tests or ANOVA tests with their parametric assumptions may not be the most suitable approach. These tests assume normality and homogeneity of variances, which cannot be guaranteed when assessing the performance of analytical models over diverse datasets [131]. Additionally, conducting paired t-tests for all possible classifier

pairs is suboptimal because when so many tests are made, a certain proportion of the null hypotheses is rejected due to random chance. In such situations, non-parametric tests, such as the Friedman test along with appropriate post-hoc tests, are a more fitting choice (cf. Section 3.4.5). They make limited assumptions and are safer for comparing classifier accuracy since they do not rely on normality or homogeneity assumptions.

In addition to these five key findings, we will provide further insight into the different methods and datasets used in the selected studies in the description of our experimental setup (cf. Section 3.4).

3.3.3 Research gaps in predictive analytics for employee turnover

Based on the above analysis, we define three gaps in the literature on predictive analytics for employee turnover.

Data: Most studies in the literature are using one dataset. Moreover, mostly synthetic datasets are used with Kaggle1 and IBM being most popular. Consequently, the literature on predictive analytics for employee turnover lacks a common set of benchmark datasets to be used for method performance evaluation. As case-specific conditions might apply, testing a method on a single dataset is likely not sufficient. This hypothesis also remains to be tested.

Modeling pipeline: To compare the performance of methods on a dataset or use case, preprocessing, processing, and postprocessing techniques must be consistent. The research on predictive analytics for employee turnover suffers from different approaches chosen in established manuscripts. Notably, we find that most papers are not explicit about their experimental setup.

Performance metrics: We find no common ground in metrics used for evaluation in predictive analytics for employee turnover. While most studies use at least one threshold-based metric, the use of more comprehensive threshold-independent metrics, such as Area Under the Receiver Operating Characteristic curve (AUC-ROC) and Area Under the Precision-Recall curve (AUC-PR) metrics remains limited. As different metrics might suggest different rankings of methods, agreement is needed on a set of metrics that should be reported (next to possible use-case-specific reporting).

3.4 Experimental design

In this section, we present an experimental design for evaluating multiple methods across multiple data sets for turnover prediction. Our experiments comprehensively address five identified limitations in the scoping review. We

ensure generalizability by testing our methods on 9 diverse datasets and assess the effectiveness of 14 binary classification methods. We employ proper hyperparameter tuning, along with testing the effects of class rebalancing and feature selection methods. A comprehensive evaluation is conducted using a wide set of performance metrics. Additionally, we conduct proper statistical tests to analyze and compare the performance of different classification methods. By overcoming these five limitations, our study provides a more thorough evaluation of the methods and can inform future research in this field.

3.4.1 Classification methods

Formally, turnover prediction can be described as a binary classification task. Each dataset $\mathcal{D}$ consists of N observed predictor-response pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i \in \mathbb{R}^m$ is an m-dimensional vector describing employee characteristics and $y_i \in \{0, 1\}$ is a binary feature that distinguishes between *stay* ($y = 0$) and *leave* ($y = 1$). $\mathcal{D}$ is used to train a binary classification model $s(.) : \mathbb{R}^m \rightarrow [0, 1]$. This binary classification model predicts a propensity score $s_i \in [0, 1]$ for each instance i based on the features $\boldsymbol{x}_i$. Depending on the classification threshold t_i^*, s_i is converted to a predicted class $\widehat{y}_i \in \{0, 1\}$. In our experiments, the threshold is set to 0.5. This study compares 14 different classifiers, both individual and ensemble methods, as summarized in Table 3.3. For a complete overview of which classifiers are used in which study, we refer to Table A.2 in the Appendix.

In Section 3.5.2, this study conducts two ablation studies: one assessing the effects of class rebalancing methods and another examining the impact of feature selection techniques. Since some classifiers may handle class imbalance better than others, we will explore how combining specific classifiers with class rebalancing methods or feature selection techniques might improve performance. These investigations aim to shed light on the sensitivity of various classifiers to class imbalance and feature selection, thereby guiding practitioners and researchers toward optimal model selection strategies for turnover prediction tasks.

3.4.2 Datasets

Table 3.4 describes the 9 datasets used in the experiments. The Real1, Real2, and Real3 datasets were obtained from three large Belgian organizations and are not publicly available. The remaining six datasets are publicly accessible. The IBM and Kaggle1 datasets are used in 24 and 10 of the studies included in the scoping review presented in Table 3.2, respectively.

All the datasets exhibit some degree of class imbalance (see Perc. Turnover), with the positive class representing the minority. The extent of class imbal-

Table 3.3: Overview of established classifiers

	Classifier	Abbr.	Num. Studies
Indiv. classifiers	Artificial Neural Networks	ann	12
	Decision Tree	dt	41
	K-Nearest Neighbors	knn	22
	Linear Discriminant Analysis	lda	4
	Logistic Regression	lr	33
	Naïve Bayes (Bayesian/Gaussian)	bnb /gnb	25
	Quadratic Discriminant Analysis	qda	1
	Support Vector Machine	svm	30
Ensembles	AdaBoost	ab	7
	Gradient Boosting	gb	9
	LightGBM	lgbm	3
	Random Forest	rf	36
	Extreme Gradient Boosting	xgb	12

Table 3.4: Dataset overview. S: synthetic, R: real, pub: publicly available, priv: private

Name	Avail-ability	Origin	Type	Num. Obs.	Num. Var.	Perc. Turnover	Num. Studies
Real1	priv	HR services	R	16,394	25	7.16%	0
Real2	priv	IT services	R	19,992	46	8.38%	0
Real3	priv	High-tech	R	19,413	64	6.11%	0
DS	pub	Analytics Vidhya	R	19,158	12	24.93%	0
IBM	pub	IBM	S	1,470	32	16.12%	24
Kaggle1	pub	Kaggle	S	14,249	10	23.81%	10
Kaggle2	pub	Kaggle	S	6,284	8	23.63%	0
Kaggle3	pub	Kaggle	S	19,104	14	8.46%	0
Kaggle4	pub	Kaggle	S	3,310	22	32.90%	0

ance varies across the datasets, ranging from 6.11% to 32.90%. Each dataset comprises a range of features covering demographics, employment details, organizational information, education, and talent. While certain features, such as age, gender, and organizational tenure, are common across datasets, the specific features included can vary. For further information on dataset contents, please refer to the GitHub repository where the public datasets are available.

The time-flattening of the data, i.e., the handling of the temporal aspect of turnover, slightly varies across the datasets, depending on the source. In the case of the Real1, Real2, and Real3 datasets, the continuous time aspect is divided into snapshots. In line with Rombaut and Guerry [48] and in line with industry practice, we use yearly snapshots. For Kaggle2 and Kaggle3, we adopted the snapshots as originally available in the publicly accessible datasets, which are provided on a yearly and monthly basis, respectively. Each periodic snapshot captures information about every employee in the organization and indicates whether or not the employee left the organization. In datasets with a time component, we have also included features that track changes in feature values compared to the previous period. The remaining four datasets do not include a timestamp, although some time-dependent features are incorporated (e.g., years in the current role and seniority).

3.4.3 Data preprocessing and partitioning

For each dataset, we employ a 5×2-fold cross-validation approach [131], resulting in ten distinct evaluations of out-of-sample classification performance on which we report the performance average. In each iteration, the data is divided into non-overlapping training (50%) and test (50%) sets. Additionally, to fine-tune hyperparameters and select the best models, we conduct an internal five-fold cross-validation on each training set within the 5×2-fold cross-validation loop, i.e., we adopt a so-called nested cross-validation setup.

After splitting the data, the numeric features in the training set are normalized to have a mean of zero and a variance of one. The normalization process is then extended to the corresponding features in the validation and test sets, using the mean and variance estimates obtained from the training set. Categorical features are encoded using the weights-of-evidence approach [132]. We avoid any form of data leakage by performing these operations based on the training data.

We use a full grid-search approach to optimize hyperparameters, systematically evaluating all possible combinations within predefined ranges on a separate validation set. This exhaustive exploration ensures no potential configurations are overlooked, providing a comprehensive view of the performance landscape. This approach is especially beneficial for configuring hyperparameters in benchmarking experiments, where a thorough assess-

ment of model performance is crucial. The details of the hyperparameter search space can be found in Table E.2 in the Appendix.

3.4.4 Performance metrics

The evaluation of binary classification algorithms typically involves labeling one class as positive (i.e. *turnover*) and the other class as negative (i.e. *stay*), which enables the construction of a confusion matrix. In this context, positive classes typically refer to the minority class, while negative classes describe the majority class. The classification threshold is set to 0.5. Based on the threshold-dependent confusion matrix, we calculate Accuracy, Sensitivity (also known as Recall), Specificity, Precision, and the F1-measure [126].

The Receiver Operating Characteristic (ROC) represents the trade-off between Sensitivity and the 1 - Specificity at various classification thresholds for the positive class. The area under the ROC (AUC-ROC) metric summarizes the model's ability to distinguish between positive and negative instances across different classification thresholds. However, its effectiveness in comparing methods has limitations due to its dependence on the classifier's score distribution, which can vary [133]. Given our study's objective of method comparison, we have chosen also to include the H-measure. This is the standardized measure of the expected minimum loss where Hand [133] proposes employing a symmetric $beta(x; 2, 2)$ distribution to characterize the costs of misclassification. A precise definition can be found in [133].

An alternative to AUC-ROC is AUC-PR, which calculates the area under the precision-recall (PR) curve and hence quantifies the trade-off between precision and recall at different classification thresholds. AUC-PR is the preferred metric in imbalanced learning settings like employee turnover prediction since it takes into account the prevalence of turnover by being more sensitive to the absolute number of false alarms [134]. For both threshold-independent metrics, AUC-ROC and AUC-PR, a higher value is better.

The Brier score assesses the calibration of a model's predicted probabilities by measuring the mean squared difference between the predicted probability (s_i) and the observed outcome (y_i). A lower Brier score indicates better calibration.

3.4.5 Statistical tests

To evaluate the statistical significance of the benchmarking experiment results, we employ a methodology described by Demšar [131] that involves utilizing a Friedman test in combination with a post-hoc test with Hochberg p-adjustments for multiple comparisons [135].

In this study, we compare the performance of 14 (k) different methods across 9 (N) datasets. Let r_i^j represent the rank of algorithm j on dataset i according to a specific metric. The Friedman test examines the average ranks of the algorithms, calculated as $R_j = \frac{1}{N}\sum_i r_i^j$, across the datasets. Under the null hypothesis, which assumes that all algorithms are equivalent and hence their average ranks R_j should be equal, the Friedman statistic follows a χ_F^2 distribution with $k-1$ degrees of freedom. If the null is rejected, indicating the presence of significant differences among the algorithms, we conduct a post-hoc test with Hochberg p-adjustments, which carries out a pairwise comparison where all classifiers are compared with the best-performing classifier for the chosen metric.

3.5 Empirical results

This section presents the empirical results. We divide our analysis of our results into two distinct parts. Section 3.5.1 presents our core results, where we report the results on all dataset-method combinations. In Section 3.5.2, we investigate the effect of three class balancing methods and the effect of feature selection.

3.5.1 Results

The results of the benchmarking study are presented in Table 3.5, which shows the average ranking (AR) of the classifiers across datasets per metric. The rightmost column presents the average performance and standard deviation across metrics for each classifier. The bottom row shows the Friedman χ^2 statistic to check whether at least two classifiers have significantly different performances. The p-values of the post-hoc test with Hochberg p-adjustments that compare each classifier to the best-performing one are added between brackets. Underscored values indicate a significance at the 5% level that the null hypothesis of a classifier performing equally well as the best classifier was not rejected.

We draw several insights from Table 3.5. Across all performance metrics, ensemble classifiers largely exhibit the best performance among the evaluated methods, except the artificial neural network (ann). Among the ensemble methods, there are variations in rankings, but these differences are often not statistically significant, as indicated by the underlined p-values. This suggests that the various ensemble methods generally perform on par with each other. However, to provide a more comprehensive view, it is important to note a distinction within the group of ensemble methods. Specifically, ab, one of the oldest ensemble methods [136], appears to differ from the rest of the ensemble techniques, including the more recent xgb [137]. Two widely

recognized industry standards, lr and dt, generally exhibit lower performance in terms of threshold-independent AUC-PR and AUC-ROC when compared to ensemble methods and ann. However, it is essential to highlight that their performance is significantly influenced by the dataset's characteristics. For example, when considering the IBM dataset (see Table A.8 in the Appendix), lr emerges as the top performer, while dt demonstrates weak performance.

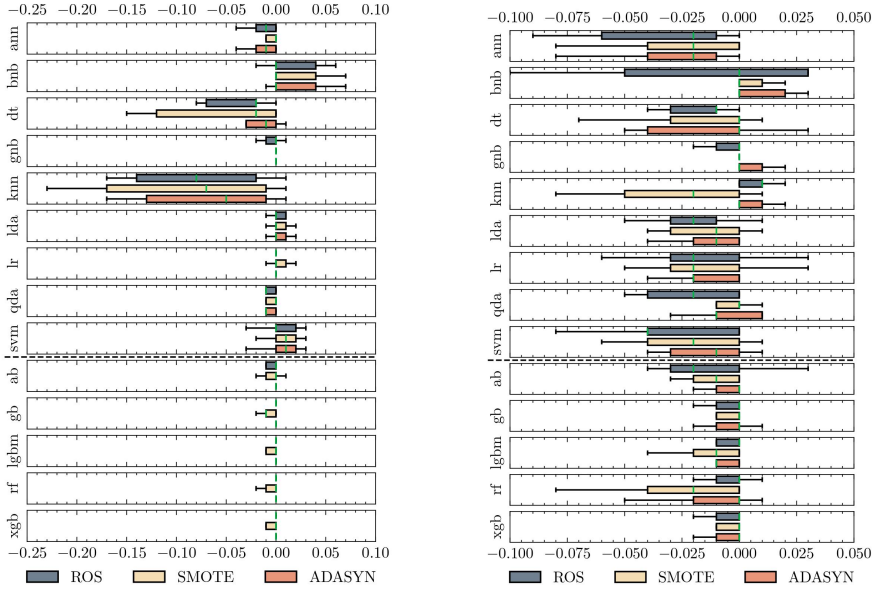
The complete results for each dataset, presented as relative rankings of classifiers per metric, are provided in Tables A.4-A.12 in the Appendix. Next to the differences in characteristics between the datasets as summarized in Table 3.4, another important difference is the unknown generation process of the synthetic datasets, resulting in considerable uncertainty regarding how accurately they reflect real-world scenarios. Consequently, caution is warranted when attempting to generalize insights derived from these synthetic datasets to real-world applications. Illustrating with the IBM dataset, lr outperforms other classifiers in terms of AUC-PR, AUC-ROC, Accuracy, and Brier Score and ranks second for F1-Score and H-Measure after lda (another linear method). These outcomes hint at a linear process in the synthetic dataset generation. Therefore, caution is crucial in generalizing these results to diverse scenarios despite the dataset's common usage (Table 3.2). It is vital to recognize that these concerns extend beyond predictive performance, encompassing potential disparities in feature predictive power and correlations that may not mirror reality. These findings underscore the importance of selecting the right classifier for a specific task, taking into account not only the classifier's inherent properties but also a match with the unique properties and complexity of the dataset under consideration.

3.5.2 Effect of class balancing and feature selection

Class balancing

Class imbalance is present in data when there is an unequal distribution of classes. This may present a significant challenge to learning algorithms [138]. An algorithm simply trained to minimize classification error might overfit the majority class, resulting in potentially high accuracy, yet low sensitivity. This is a significant threat as the minority class is typically of greatest interest, as it is in predicting employee turnover.

We show in Table 3.4 that all datasets used in the literature exhibit levels of class imbalance. In this subsection, we will investigate the severeness of class imbalance in employee turnover datasets, and answer whether such imbalance has a detrimental effect on performance. We will deploy three different resampling methods, i.e., random over-sampling (ROS), the Synthetic Minority Oversampling Technique (SMOTE) [139], and the Adaptive Synthetic (ADASYN) sampling method [140], and evaluate whether they can



(a) AUC-ROC

(b) AUC-PR

Figure 3.1: The effect of various class balancing techniques per classifier in terms of area-under-the-curve metrics, relative to the case without class balancing.

improve the predictive performance of each of the models in our main set of experiments. ROS is the simplest resampling method. It tackles imbalance by creating new samples in the minority class through random duplication of existing minority samples. With SMOTE, the minority class is oversampled by generating synthetic examples along line segments connecting the k nearest neighbors from the minority class. This technique aims to augment the representation of the minority class. ADASYN takes yet a different approach by adaptively generating minority data samples based on their distributions. It produces more synthetic data for minority class samples that are more challenging to learn while generating fewer synthetic samples for those that are easier to learn. ADASYN not only reduces the learning bias arising from the original imbalanced data distribution but also aims to adjust the decision boundary by focusing on the more difficult-to-learn samples.

In our analysis, we employed each resampling method on all nine employee turnover datasets and evaluated the resulting difference in AUC-PR and AUC-ROC for all learning algorithms. Each dataset was resampled to a 1:1 distribution of majority to minority class samples. Full results can be

found in Figure 3.1.

The effectiveness of resampling is both method-dependent and varying across datasets. The performance of ensemble methods is seemingly unaffected by resampling, in terms of AUC-ROC, suggesting robustness against class imbalance. Conversely, decision trees and k-nearest neighbors show an unexpected detrimental effect of resampling the training data. Only support vector machines, discriminant analyses, and naive Bayes approaches appear to benefit, although only slightly, from resampling. In terms of AUC-PR, we see a generally negative effect of resampling, compared to using the training data as-is. However, the impact is evidently limited with the most affected methods losing less than 10% in AUC-PR when compared to training on unsampled data.

We note that we have only evaluated the impact of resampling to a 1:1 ratio of minority to majority class observations. This ratio could be tuned as a hyperparameter in method training, potentially leading to different results. We omit this step for the sake of brevity.

We do not claim holistic results, yet our experiments suggest the use of resampling when training learning algorithms might be limited. This might stem from several factors, such as the robustness of learning algorithms against class imbalance, or the level of imbalance in the data. This insight confirms the findings of other studies on real-world classification problems with class imbalance [141], [142].

Feature selection

In this subsection, we assess the impact of feature selection on the results of the study. From a technical perspective, the objective of feature selection is to identify and retain the most informative features from a given dataset, to mitigate the curse of dimensionality and the risk of overfitting, and to reduce computational complexity, while not compromising on predictive performance. From a business perspective, especially in the context of turnover prediction, fewer, highly relevant features are preferred to obtain data-driven managerial insights, as it may be important and an objective of the analysis to understand why employees leave a company [143].

We investigate the effect of feature selection by assessing performance while incrementally increasing the number of selected features, denoted as k. We select the k highest-ranking features based on the ANOVA F-statistic which tests the relationship between that feature and the target variable. However, it is important to note two key limitations to this approach. First, this method is most effective with linear feature-target associations and may struggle with complex, nonlinear relationships. Second, if features are not independent, the F-scores can be influenced by confounding factors, potentially distorting feature importance.

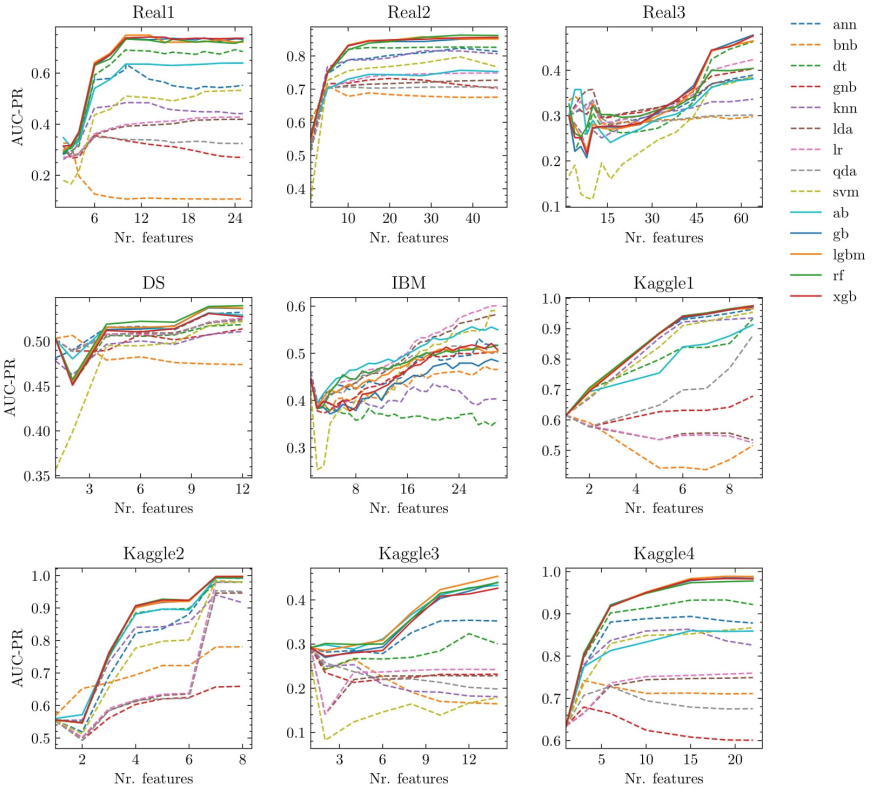


Figure 3.2: The effect of feature selection per dataset in terms of AUC-PR. The performance of single classifiers is indicated with a dashed line and of ensemble methods with a full line.

Figure 3.2 displays one dataset per panel, where performance in terms of AUC-PR is expressed in function of the number of features selected. Single classifiers are depicted using dashed lines, while ensemble methods are represented by solid lines. The plots summarize the results of a 5×2-fold cross-validation procedure where the AUC-PR is averaged over ten runs.

For the dataset Real2, performance in terms of the number of included features tends to increase for most of the classifiers and reaches a plateau after incorporating more than 10 features. The *height* of this plateau depends on the classification method. This largely aligns with prior research [128]. For the Real3, DS, IBM, and Kaggle2 datasets, we also predominantly observe upward-trending curves, but with no clear signs of reaching plateaus. For the Real3 and IBM datasets, for instance, for most of the classifiers, per-

formance continues to improve even after the last feature is added. Finally, for the Real1, Kaggle1, Kaggle3, and Kaggle4 datasets, performance deteriorates for some single classifiers when giving the models access to more features. However, as for the other datasets, we do observe performance improvements with an increase in the number of features for all ensemble models.

This last observation, in combination with the fact that the best-performing classifiers are the ensemble methods (vs. the single classifiers) for all datasets, except the IBM dataset, allows us to conclude that feature selection is generally not advisable from a predictive performance point of view. Even when considering the business need for explainability, the recommended approach is to prioritize predictive accuracy initially and subsequently address explainability (when the goal is not to test a theory) since higher predictive accuracy suggests a more profound understanding of the underlying structure in the data [69]. To make (the predictions of) an ensemble model with a lot of features explainable, various strategies can be employed. Among these, options include building a single decision tree to explain ensemble predictions, see e.g., [144], [145], or using permutation-based feature importance estimates, see e.g., [69], [146].

3.6 Conclusion

Employee turnover significantly affects organizations, resulting in the loss of valuable expertise and decreased productivity. Traditional turnover theories rely on a theoretical framework and often apply general principles to entire populations. With the abundance of available data, analytical tools can predict and explain employee turnover at a higher granularity, emphasizing either organization-specific, department-specific, or, in the most extreme case, on an individual employee level.

Utilizing data-driven insights improves decision-making processes and ultimately reduces costs through effective employee retention strategies. Nevertheless, the current body of literature on predictive analytics for employee turnover is extensive and lacks cohesion. This paper aims to focus on the intersection of data science and human resource management, targeting both professionals and academics. It achieves this by presenting a comprehensive study consisting of several components focused on traditional HR literature, conducting an extensive survey on data-driven methodologies in turnover prediction, and performing a rigorous benchmarking experiment. These components synergize to establish a clear focal point in the landscape of predictive analytics for employee turnover, aiming to illuminate the current state-of-the-art and provide a valuable reference for researchers and practitioners alike.

In the structured scoping review, we reveal the following insights that underscore the necessity for a standardized benchmarking experiment. Notably, existing predictive analytics studies on employee turnover often (i) do not disclose the datasets used or lack dataset variability and often focus on the same publicly available datasets, (ii) are ambiguous regarding data preprocessing, processing, and postprocessing techniques, (iii) lack a common set of performance metrics.

We address these insights and limitations through a benchmarking experiment. We evaluate 14 classification methods across 9 datasets, revealing ensemble classifiers' superior performance over single classifiers (except for ann) across most metrics. However, specific classifiers' effectiveness varies depending on dataset size and complexity. Additionally, our ablation study investigates class rebalancing and feature selection effects. Results show their impact varies with methodology and dataset, with ensemble methods demonstrating no benefit from these techniques despite their superior predictive performance.

In light of our findings, we offer key insights for industry practitioners striving for precise employee turnover prediction. It is recommended to adopt a use-case-driven strategy, tailoring machine learning models to organizational dynamics. Beginning with off-the-shelf methods and simple pipelines lays a foundational understanding, paving the way for refinement based on specific needs, including class rebalancing. While our results caution against prioritizing resampling, a mid-to-long-term evaluation of its necessity remains vital, contingent upon factors such as the chosen method and data imbalance. Exercise caution when generalizing insights, especially with the IBM dataset, where linear models excel but require validation for broader applicability. The effectiveness of feature selection methods hinges on various factors, including the number of features, classification method, and data structure. Some datasets show optimal predictive power with a subset of features, while others benefit from a broader set. Additionally, the chosen feature selection method significantly impacts model performance, with methods like ANOVA F-statistic tests assessing linear relationships but potentially struggling with non-linear ones.

This study has some limitations. Firstly, it is crucial to acknowledge the absence of data from external organizations, particularly in assessing perceived job alternatives. This limitation may impact the comprehensiveness of the analysis, as external factors influencing turnover decisions remain unaccounted for. Additionally, there's a trade-off between interpretability and performance in classification methods. While *black box* models enhance predictive power, they pose challenges in explaining outcomes. Although ensemble methods perform best, the experiments lack a metric for explainability, crucial for model deployment and user adoption to address skepticism

towards these models [147], [148].

Future work could explore several interesting avenues to extend the current research. Firstly, integrating unstructured data like textual performance evaluations and CVs could enhance classification algorithm performance. Secondly, acknowledging the nuanced nature of turnover, researchers could investigate scenarios where employee departure is not necessarily undesirable. A nuanced cost-sensitive approach to turnover involves understanding the varying financial implications of departures and the diverse costs of retention programs per employee. This tailored perspective informs strategic workforce management, facilitating informed decisions that account for the unique financial dynamics associated with both employee departures and retention initiatives, as the relationship between turnover and firm performance can be complex [149]. Finally, future studies could further explore the *why* behind employee turnover, unraveling the underlying motivations for departures so that HR professionals can respond to them. This corresponds with a transition from predictive to prescriptive methods, where insights are translated into actionable strategies [68], enabling the design of effective, personalized retention campaigns.

Table 3.5: Average ranking of classifiers across datasets per performance measure. The best-performing methods are emphasized in bold and second best in italic. *P*-values are added between brackets.

Classifier	AUC-PR		AUC-ROC		Accuracy		Brier Score		F1-Score		H-Measure		Precision		Recall		Specificity		Average	
ann	5.6	(0.297)	5.7	(0.207)	7.1	(0.075)	5.9	(0.090)	5.8	(0.380)	6.0	(0.419)	8.1	(0.029)	5.9	(0.524)	8.1	(0.087)	6.5	±1.0
bnb	13.4	(0.000)	12.7	(0.000)	11.6	(0.000)	10.6	(0.000)	10.6	(0.004)	10.3	(0.014)	12.4	(0.000)	9.6	(0.022)	10.8	(0.003)	11.3	±1.3
dt	7.6	(0.048)	7.8	(0.026)	5.3	(0.362)	5.2	(0.169)	5.6	(0.414)	5.7	(0.498)	7.0	(0.086)	5.7	(0.582)	7.4	(0.158)	6.4	±1.1
gnb	11.3	(0.000)	12.0	(0.000)	13.9	(0.000)	12.3	(0.000)	9.8	(0.006)	8.7	(0.061)	13.2	(0.000)	5.1	(0.723)	13.3	(0.000)	11.1	±2.8
knn	10.4	(0.001)	12.4	(0.000)	9.2	(0.007)	8.3	(0.004)	10.6	(0.004)	10.8	(0.014)	7.7	(0.044)	10.6	(0.010)	5.8	(0.474)	9.5	±2.0
lda	9.7	(0.003)	9.2	(0.003)	9.1	(0.008)	8.1	(0.005)	8.8	(0.025)	8.4	(0.065)	9.6	(0.003)	9.1	(0.039)	9.4	(0.017)	9.0	±0.5
lr	7.9	(0.038)	7.6	(0.031)	8.1	(0.031)	6.4	(0.048)	9.9	(0.006)	10.0	(0.015)	8.0	(0.030)	10.9	(0.007)	7.1	(0.193)	8.4	±1.5
qda	11.1	(0.000)	10.4	(0.000)	11.3	(0.000)	9.4	(0.001)	8.4	(0.031)	7.6	(0.124)	11.4	(0.000)	7.0	(0.270)	11.2	(0.001)	9.8	±1.7
svm	7.9	(0.038)	8.9	(0.005)	7.1	(0.075)	14.0	(0.000)	9.9	(0.006)	9.9	(0.015)	3.9	(0.787)	11.4	(0.007)	5.2	(0.608)	8.7	±3.1
ab	6.8	(0.107)	5.9	(0.176)	7.4	(0.061)	12.7	(0.000)	8.6	(0.030)	8.6	(0.064)	7.0	(0.086)	9.0	(0.041)	7.2	(0.188)	8.1	±2.0
gb	3.4	(0.933)	3.2	(0.900)	*3.3*	(0.943)	3.1	(0.660)	3.7	(0.980)	**3.8**	(1.000)	4.7	(0.547)	**4.1**	(1.000)	6.0	(0.429)	3.9	±0.9
lgbm	3.4	(0.933)	3.3	(0.869)	**3.1**	(1.000)	**2.1**	(1.000)	**3.4**	(1.000)	*4.2*	(0.897)	*3.8*	(0.787)	4.7	(0.848)	*4.2*	(0.891)	**3.6**	±0.8
rf	*3.2*	(0.975)	*3.0*	(0.920)	4.2	(0.665)	*2.9*	(0.734)	6.0	(0.322)	6.1	(0.395)	**3.1**	(1.000)	6.8	(0.312)	**3.9**	(1.000)	4.4	±1.5
xgb	**3.1**	(1.000)	**2.8**	(1.000)	3.6	(0.884)	3.8	(0.464)	**3.4**	(1.000)	4.3	(0.888)	4.4	(0.621)	*4.6*	(0.863)	4.9	(0.700)	*3.9*	±0.7
Fried. χ^2	77.6	(0.000)	86.5	(0.000)	75.1	(0.000)	101.8	(0.000)	50.9	(0.000)	40.9	(0.000)	73.8	(0.000)	46.3	(0.000)	53.9	(0.000)		

4

Data-driven internal mobility: Similarity regularization gets the job done

This paper presents a novel approach to support career management by recommending job-employee matches within an organization through data-driven insights. We build a recommender system to propose matches. The presented approach extends upon a conventional collaborative filtering recommender system, which suggests matches based on the historic performance similarities of employees. To address the prevalent challenge of the *cold start* issue in internal placement, we incorporate personal employee data with a similarity-based regularization term. This regularization term finds latent representations that are closer to each other when employees share similar personal features. This approach is evaluated using three real-life datasets and demonstrates a highly competitive performance compared to state-of-the-art benchmark methods. Overall, we make three contributions to the field of HR analytics: (i) we present a comprehensive survey of job-employee recommender systems in the context of internal mobility, (ii) we implement a similarity regularization method, and (iii) we release a first-of-its-kind HR dataset on internal mobility. All code and data used in this study are publicly available on GitHub.

4.1 Introduction

With a vast amount of available data, companies in almost every sector are devoted to exploiting this data to gain a competitive advantage [150]. Recently, many organizations have been relying on data-driven insights to support decision-making in their daily activities [151], [152]. The human resources (HR) field sees great potential in using big data, analytics, and machine learning (ML) [153]. As a result, HR departments within organizations are looking for ways to use analytical techniques to complement current evidence-based techniques with objective and data-driven insights [30]. Correspondingly, HR analytics is becoming increasingly important as a discipline and research field. There has been a growing amount of literature on the topic [154], but, despite the notable rising enthusiasm for HR analytics, the ability to apply insights from this literature in business practices remains limited [155].

As organizations are moving away from traditional organizational structures and delayering hierarchies, career ladders that guide the mobility of employees within an organization become less clear [156], [157]. This challenges traditional internal market organization theories, which mostly focus on economic and institutional factors [158], and leads to the adoption of data-driven approaches with the help of electronic human resource management (e-HRM) regimes [159], [160]. Organizations are increasingly looking into data-driven systems to support decisions on how to mobilize employees within the organization, either for career changes or for advancement in their current careers [161]. This is crucial, as internal job transitions are beneficial for employees' employability [162] and can impact, amongst others, motivation and retention behavior of employees [163]. Using a data-driven, algorithmic approach to recommend career steps has several benefits, including facilitating the critical search for solutions to hard-to-fill positions, guiding users to relevant opportunities, discovering *passive* internal candidates by better managing the internal talent pool, and streamlining manual candidate searches [164]–[166].

To support the management of internal mobility with data-driven insights, we view an employee's career within an organization from a longitudinal perspective as a sequence of activities, as is done in the field of process analytics [167]. This means that HR data can be transformed into an event log format, where events represent the execution of activities, resulting in start, completion, and/or cancellation recordings. This dynamic process perspective provides a comprehensive end-to-end view, which is suitable for the complex nature of internal mobility in modern organizations. B.1 shows an employee journey map (EJM) in the form of a directly-follows graph derived from an event log.

The literature on data-driven decision support on internal mobility is, as far as we are aware, limited to managerial concerns. No concrete algorithmic approach has been proposed yet, except for using simple distance-based or similarity-based insights for descriptive purposes [165]. In response to the lack of analytical approaches for supporting the management of internal mobility in the current literature, we propose an internal recommender system for organization-specific employee mobility.

Our proposed method focuses on predicting the performance of employees within an organization concerning specific job positions, where matches are rated with a score $y \in [0, 1]$ (see Section 4.4.1). Recommender systems (RSs) can assist in decision-making by suggesting job-employee matches and narrowing down the large search space to a personalized subspace [168], which then aids in proposing fitting jobs to employees and vice versa. When considering internal mobility as a process, visualized by an EJM, the recommender system suggests the next activity, i.e., the next job, in the trace of an employee. We start with a collaborative filtering (CF) approach that recommends job-employee matches based on the similarity with other employees. However, limited availability of data on new hires and employees with short tenure may affect the performance of RSs as a result of the cold start problem [169]. In the case of new employees or employees with short tenure, the system does not have information about their past job performances in order to make new recommendations. Consequently, the cold start problem poses a challenge for using recommender systems in internal mobility management.

Therefore, we extend the loss function of this baseline method with a similarity regularization term to incorporate additional information captured by personal employee data such as education and field of study. The proposed similarity regularization technique effectively addresses the cold start problem by incorporating additional personal information beyond past experience, in order to decrease the distance between latent representations of employees when they share similar personal features. The performance of the proposed methods is assessed using three real-life datasets containing thousands of matches from the past ten years.

Our contribution consists of three main parts. First, we examine the use of recommender systems in the setting of data-driven internal mobility. Second, we extend existing collaborative filtering methods by implementing a similarity regularization term to include personal information on employees to address the prevalent cold start problem. Third, we also publish a first-of-its-kind HR-career dataset on internal mobility in the format of an event log. This dataset considers internal mobility in event log format, spans a period of 10 years, and is made available on GitHub. As such, we ensure the reproducibility of the reported results in this paper and facilitate further research on this topic.

The rest of this paper is organized as follows. In the next section, we provide an overview of related work. In Section 4.3, we first motivate our approach with a running example, which we retake throughout this paper. Next, we explain the longitudinal approach with an event log as starting point. Then, we introduce the collaborative filtering (CF) methodology based on Matrix Factorization (MF) and the proposed similarity regularization (SR) extension. In Section 4.4, we discuss the used datasets and describe the setup of our experiments. Next, Section 4.5 presents the results and discusses the advantages and drawbacks of our approach. Finally, Section 4.6 concludes the paper.

4.2 Related work

Human Resource Analytics (HRA) as a term is relatively new, first appearing in HR literature in 2003-2004 [154]. We define HR analytics in this paper as the application of descriptive, predictive, and prescriptive analytics for data-driven human resources management. [154] and [170] provide a taxonomy of research topics in this field.

In this research, we investigate data-driven decision support for internal employee mobility. More specifically, we focus on predictive job-employee matching in a post-hire setting, considering careers from a process perspective. In our HR application, we emphasize the importance of avoiding decision automation. Instead, we focus on describing potential scenarios without prescribing specific actions. HR professionals, in collaboration with employees, select the scenario, with a human always involved in assessing the options. Thus, our emphasis lies on decision support rather than automation, aligning with predictive analytics. In contrast, descriptive analytics depict past occurrences. However, our application focuses on unseen job-employee matches rather than past events. Meanwhile, prescriptive analytics recommends optimal actions or decisions based on predictive analytics outcomes, aiming to optimize results or achieve certain goals according to a specific policy [63]. While a recommender system can directly influence decision-making, its primary function for our application is to predict the outcomes of certain job-employee matches rather than prescribe specific actions.

Table 4.1 summarizes key related work in the field of Predictive HR Analytics based on three criteria: (i) whether the dataset is publicly available or not, (ii) whether the problem statement is approached from a matching perspective or not (as discussed in Subsection 4.2.2), and (iii) whether the problem statement occurs in a post-hire setting (as discussed in Subsection 4.2.3). The bottom row highlights the positioning of our work, which addresses the research gap of providing data-driven decision support on internal mobility, which is characterized as a matching problem in a post-hire setting.

4.2.1 Predictive HR analytics

Unlike descriptive analytics, predictive analytics are forward-looking and envision the future based on observed historical data. Within the field of data-driven HRM, predictive analytics are frequently used for predicting employee performance and employee turnover, for retention strategy design, and job-employee matching. We highlight some of the most important work for each of these applications.

Performance prediction can be done for both current and future employees, i.e., candidates applying for a position. This helps HR professionals efficiently prioritize their resources by narrowing their focus to a shortlist of people [171]–[173]. Research on performance prediction in the context of recruitment and selection has been conducted by [174]–[180].

Employee turnover is a pressing issue for many organizations. Depending on the position and the difficulty of finding a replacement, the costs of turnover can range from 1.5 to 5 times an employee's annual salary [181]. While prior HRM literature has heavily relied on qualitative survey data, this approach may not always yield organization-specific insights. However, the increasing availability of longitudinal data through Human Resources Information Systems (HRISs) has enabled research on predictive analytics for employee turnover. Several studies have proposed data mining and ML techniques for turnover prediction [172], [182]–[185], and [178], [186] have explored turnover prediction in specific domains. [63] combine the prediction of performance and turnover simultaneously. As an extension, research on turnover can be taken to the next step by also developing data-driven retention policies [171].

Another field of predictive HR analytics is management-oriented and examines how businesses can and should adopt predictive HRA in their daily operations. [29], [30], [187] describe the difficulties and barriers of adopting HR analytics on the most conceptual level. To tackle those barriers, [188] propose a framework to organize HR analytics. [189] look at HR analytics from a meta-perspective by assessing and comparing different frameworks. [190]–[192] focus on practical implementation tools to successfully adopt HR analytics.

4.2.2 Matching

Matching is the process of linking potential candidates to jobs based on some criterion. There have been numerous approaches to matching, and a comprehensive overview can be found in reference works such as [63], [200]–[202]. However, this literature predominantly focuses on pre-hire matching in the context of recruitment and selection or employment services, while the focus of this research is on post-hire matching. Job recommender sys-

Table 4.1: **Predictive HR analytics overview.** In this overview, we list various applications of predictive HR analytics related to matching or a post-hire scenario, which are the two categories our research jointly encompasses. We provide details on the main problem it addresses (matching or other) and the dataset, including its availability (public or private). We also specify whether the paper discusses a post-hire setting and the types of techniques used.
We handle the following abbreviations. AR: association rules, CNN: convolutional neural networks, CS: cosine similarity, DT: decision trees, $(X)GB$: (extreme) gradient boosting, GD: graph databases, KNN: k-nearest neighbors, LDA: linear discriminant analysis, LR: logistic regression, $(M)NB$: (multinomial) naïve bayes, $MF(SR)$: matrix factorization (with similarity regularization) NLP: natural language processing, NN: neural network, RF: random forest, SVM: support vector machines, $VOBN$: variable-order bayesian network.

Ref.	Main focus	Source	#Obs.	Public	Matching	Post-hire	Techniques
[173]	selection, retention	High-tech	3,825			✓	AR, DT
[183]	turnover	unknown	1,575			✓	DT, LR, NB, SVM, RF
[76]	performance	Call center	1,037			✓	DT, NB
[193]	matching	LinkedIn	11M		✓	✓	Proportional hazards model
[194]	matching	LinkedIn	2,410		✓		CF, CS
[165]	matching	IBM	n.a.		✓	✓	KNN
[182]	turnover	HR company	13,484			✓	LR
[195]	matching, repr. learning	High-tech	>2M		✓		CNN, DT, GB, LDA, LR, NB, QDA, RF, SVM
[184]	turnover	(1) USA bank (2) IBM	14,322 1,470			✓	DT, GB, KNN, LDA, LR, NB, NN, RF, SVM, XGB
[196]	matching, repr. learning	CareerBuilder	$300K$		✓		Tripartite information graphs
[63]	recruitment	Nonprofit	$\pm 700K$		✓		DT, GB, RF, LR, NB, SVM, VOBN
[197]	matching	BOSS Zhipin	$\pm 150K$		✓		Graph NN
[198]	recruitment, matching	Baidu	>1M		✓		NN
[199]	recruitment, matching	Online vacancies	>2.5M		✓		GD, NLP
[124]	turnover	IBM	1,470	✓		✓	ANN, LR, RF, SVM, XGB
This work	matching, internal mobility	(1) High-tech (2) IT services (3) HR services	5,062 11,327 4,249	✓	✓	✓	KNN, MF, MFSR, SlopeOne, SVD

tems, which use candidate profiles and preferences to suggest job-employee matches, have seen a significant increase in popularity over the past two decades [201]. Both content-based and collaborative filtering methods have been used in the design of such systems. Research on job-employee matching using social network information and resumes can be found in the work of [179], [198], [203]–[205]. Studies on job recommender systems can be found in [166], [199]–[202], [206].

4.2.3 Post-hire setting

In this work, we approach job-employee matching from the perspective of mobilizing already existing employees within an organization rather than recruiting new candidates. The employees may be in their starting position or have prior experience within the organization. We consider an employee's past experience within an organization to be crucial for predicting the success of future job-employee matches. The field of HRA in the context of employee performance has been extensively studied by [171], [172], [174]–[178], although their research has been limited to a static view of job positions. In contrast, our research focuses on the dynamic modeling of jobs throughout an employee's career within the organization.

4.3 Methodology

This section elaborates on the motivation of this research, how an event log is taken as a starting point, and the functioning of the RS.

4.3.1 Problem definition

Figure 4.1 shows a simplified employee journey map where three employees can visit eight possible jobs. A more realistic employee journey map, based on dataset 1, is presented in B.1. Each career path, expressed as a trace in a process, is characterized by the jobs an employee has held and currently holds within an organization. This trace is supplemented with individual, employee-specific data like degree and branch of study. The process begins with employees joining an organization. Next, each employee can sequentially visit one or multiple jobs. For each job-employee combination, we observe a performance score y. The details of this performance score are discussed in Section 4.4. For example, employee 1 covers the trace of jobs 1, 3, and 5 with respective observed performance scores of 0.4, 0.8, and 0.5. To provide decision support, a recommender system can be used to suggest sensible next steps in this employee's career path by predicting a score $\hat{y}$ for each possible job-employee match and then ranking these potential matches according to this predicted outcome score.

Based on a comprehensive review of the relevant literature [29], [30] and by working closely with industry partners, we identify three key challenges that hinder the design and adoption of data-driven decision support in HR compared to other fields. These specific challenges are requirements to consider during the development of a data-driven method in the context of internal mobility.

A first challenge is the inferior quality and limited availability of data,

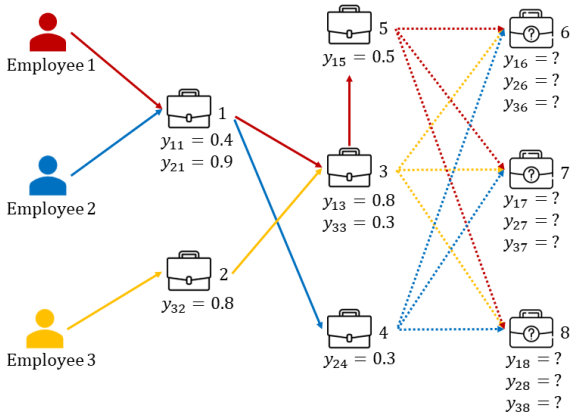


Figure 4.1: a simplified example of three internal career paths from a process perspective. This figure shows three employees, each with their own characteristics, transitioning between eight possible jobs over time. Each observed job-employee combination is characterized by a performance score y. To recommend potential next jobs, we infer future performance scores $\hat{y}$ for each possible job-employee match. The data for this example can be found in Tables 4.2 and 4.3.

making it often difficult to apply advanced methods, e.g., machine learning, that require large amounts of data. [203] describe how the performance of their models strongly hit the limits because of lacking data availability. Whereas their method is capable of handling more data to further finetune and enhance their proposed models. High-standard data accessibility is often unfeasible for many organizations without entailing issues with privacy, data protection, or depending on third-party providers. Also in our case, the available real-life data is limited. In response, in this research our objective is to develop a method that can maximally gain insights based on a limited, realistically available amount of data.

A second challenge relates to the authenticity of data that comes from external sources like social media to further enhance employee profiles. While such data sources can be used to supplement already available data, there is a risk that this might lead to misleading decisions. For example, there have been cases where social media information negatively affects the ratings of potential future employees regardless of their true qualifications [207], [208]. Therefore, in this paper, we restrict the scope to data that is typically available within the HRIS of an organization, where we only have data available on employees and historic job-employee matches without having

any information on job properties.

A third challenge concerns the accountability, fairness, and explainability of often-used opaque ML methods. An infamous example is the Amazon case where a secret AI recruiting tool was removed after showing bias to the detriment of women [209]. Because of the growing need for interpretable algorithms [210]–[212], the objective is to develop a method that takes into account the importance of explainability and transparency, which is further discussed in Section 4.5.2.

4.3.2 Event log as starting point

Our analysis represents job-employee matches and their corresponding outcomes using an event log format [213]. An event log is a dataset containing events that are recordings of the execution of an activity within a business process linked to a particular case [214]. The case or process instance is the entity being handled by the process that is analyzed. Events refer to process instances. A trace is a sequence of events. E.g., typical examples in process analytics include the traversal of a customer applying for a loan, or a patient going through a healthcare process where the customer and patient are the case identifiers respectively. In this representation, each case corresponds to an employee, each activity represents a job, and each trace represents an employee's career within the organization. Formally, we have access to a dataset $\mathcal{D} = \{(u_k, t_k^s, t_k^e, v_k, \mathbf{x}_k, y_k,) : k = 1, \ldots, K\}$. An example of $\mathcal{D}$ can be found in Table 4.2.

Each tuple $(u_k, t_k^s, t_k^e, v_k, \mathbf{x}_k, y_k)$ represents a job-employee match with employee u_k holding job v_k from time t_k^s to time t_k^e with an observed outcome y_k. This outcome y_k is defined by the end-user of the system and can be a function of any feature of interest or the combination of multiple features, based on the preferences and the intent of the end user. This is further discussed in Section 4.4.1. Employee u_k has features $\mathbf{x}_k$, e.g., the branch of study, degree, age, and gender. In total, we observe K job-employee matches where multiple tuples can refer to the same employee.

Personalized RSs provide suggestions to a user based on their profile. In the setting of this research, this means that an RS provides a relevant next step in a career to an employee, based on their profile. An employee profile $\mathcal{D}_i \subseteq \mathcal{D}$ of person u_i consists of the combination of $|\mathcal{D}_i|$ tuples where $|\mathcal{D}_i|$ is the number of jobs this employee has occupied within this organization, i.e., the length of the trace. Hence, $|\mathcal{D}_i|$ tuples contribute to one employee profile. The employee identifier u_i is unique for each employee and the same for all tuples that belong to this employee. Additionally, we assume that each unique job v can be executed at most once by each employee u. Consequently, an employee profile $\mathcal{D}_i$ consists of (i) the visited jobs with their corresponding performance score y and (ii) personal information $\mathbf{x}_i$.

Table 4.2: Synthetic example data $\mathcal{D}$ in the format of an event log. Figure 4.1 in Section 4.3.1 visualizes this event log as a directly-follows graph.

u	t^s	t^e	v	x_1	x_2	x_3	x_4	x_5	y	
1	10/2014	06/2016	Job 1	MSc	Physics	F	1975	1	0.4	$\mathcal{D}_1$
1	07/2016	02/2019	Job 3	MSc	Physics	F	1975	1	0.8	
1	03/2019	07/2022	Job 5	MSc	Physics	F	1975	1	0.5	
2	09/2009	02/2016	Job 1	BSc	Finance	M	1981	0.8	0.9	$\mathcal{D}_2$
2	03/2016	07/2022	Job 4	BSc	Finance	M	1981	0.8	0.3	
3	06/2016	03/2019	Job 2	PhD	Electronics	M	1977	1	0.8	$\mathcal{D}_3$
3	04/2019	07/2022	Job 3	PhD	Electronics	M	1977	1	0.3	
4	...	...	...	...	...	...	...	...	...	

4.3.3 Collaborative filtering

Two types of collaborative filtering (CF) algorithms are commonly used: neighborhood-based and model-based. Neighborhood-based approaches focus on the similarity between either users or items, while model-based approaches leverage RS information to construct a model that generates the recommendations [215]. The latter category encompasses the latent factor approach, a method we both apply and extend in our current work.

Extensive literature exists within the RS domain, particularly in CF, aiming to improve performance through CF enhancement [164]. Particularly in sparse datasets, where users have provided ratings for only a few items or items have been rated by only a few users, the quality of recommendations is significantly compromised [216]. To boost CF performance, many approaches involve the use of customized similarity metrics to tackle data sparsity issues. Here, only rating vectors are used as input, with the impact of data sparsity on similarity measures significantly influencing RS accuracy. A detailed overview of diverse customized similarity metrics and their performance can be found in reference work such as [217], [218].

CF starts from an observed outcome matrix $R \in \mathbb{R}^{m \times n}$ describing the outcome of m employees on n jobs which can be directly derived from the information on u, v, and y provided in $\mathcal{D}$. Tables 4.2 and 4.3 provide examples on dataset $\mathcal{D}$ and matrix $R \in \mathbb{R}^{m \times n}$. We observe K job-employee matches, which results in matrix R with a sparsity of $(1 - \frac{K}{m \times n})$. The latent factor approach that we adopt aims to factorize (MF) this matrix R by two matrices $U \in \mathbb{R}^{l \times m}$ and $V \in \mathbb{R}^{l \times n}$ with $l < \min(m, n)$.

Equation 4.1 represents the loss function $\mathcal{L}$. Following the approach of [219], [220], we use gradient descent to obtain two matrices U and V. Ultimately, with these two matrices, the matrix $\hat{R} = U^T V$ with a predicted outcome $\hat{y}_{ij}$ for each combination of u_i and v_j is calculated.

Table 4.3: Synthetic example data in the format of observed job-employee rating matrix $R^{m \times n}$, derived from $\mathcal{D}$.

v / u	Job 1	Job 2	Job 3	Job 4	Job 5	Job 6
1	0.4		0.8		0.5	...
2	0.9			0.3		...
3		0.8		0.3		...
4	...	...	...	...	...	...

$$\min_{U,V} \mathcal{L}(R, U, V) = \underbrace{\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} \left(R_{ij} - U_i^T V_j \right)^2}_{\textbf{(i)}} + \underbrace{\frac{\lambda_1}{2} \|U\|_F^2 + \frac{\lambda_2}{2} \|V\|_F^2}_{\textbf{(ii)}} \quad (4.1)$$

Term (i) expresses the difference between the observed outcomes in matrix R and the reconstructed outcomes $U^T V$. Since matrix R is sparse, I_{ij} is the indicator function that takes value 1 if the combination of the employee u_i and job v_j is observed in the training set and takes value 0 otherwise. Term (ii) consists of two regularization terms with hyperparameters λ_1 and λ_2. $\| \cdot \|_F^2$ denotes the Frobenius form, which is element-wise defined for a matrix A by Equation 4.2. The complete algorithm of model-based CF through conventional MF and its implementation in this work can be found in B.2.

$$\|\mathrm{A}\|_F \equiv \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} \left| a_{ij} \right|^2} \quad (4.2)$$

4.3.4 Matrix factorization with similarity regularization

The model-based conventional version of CF, however, is limited in that it only considers the past job positions of employees with their corresponding performance scores when determining similarity and generating latent representations of employees. Consequently, it does not take into account personal information $\mathbf{x}_i$. Employees with limited tenure or new hires may not possess enough historical data to calculate representative latent representations and consequently provide reliable recommendations. Therefore, matching them with appropriate jobs within the organization may be difficult.

To overcome this limitation, we enhance the traditional matrix factorization approach with similarity regularization SR. This concept is related to the idea of social network regularization, introduced by [221]. This approach implements the intuition that users of a recommender system value

recommendations from users close in their social network, e.g., good friends, more than from other users [222]. In this work, we introduce a resembling approach by identifying similar employees based on personal information stored in $\mathbf{x}$. Depending on the available data, personal information may include the branch of study, degree, date of birth, full-time equivalent, type of contract, location of employment, and marital status. Some features are numerical, others are categorical. The selection of features on which similarity is calculated can be done either manually by the user of the system based on domain knowledge, or can be done by testing the system's performance on a separate validation set.

We calculate the similarity between two employees using a metric Sim (Equation 4.3) to identify resembling peers. The more similar two employees u_i and u_p are, e.g., by having the same degree, the more similar their latent representations U_i and U_p should be. To enforce this, the SR term compares an employee u_i to their $m-1$ peers u_p individually and adjusts the latent representations accordingly. This is represented by term (iii) in Equation 4.4, where $\beta > 0$.

The similarity metric $Sim(x, y)$ handles a mix of numerical and categorical components by combining and weighting a numerical and categorical metric: $\mathrm{Sim}(\mathbf{x}, \mathbf{y}) = \gamma \cdot \mathrm{NumSim}(\mathbf{x}^{num}, \mathbf{y}^{num}) + (1-\gamma) \cdot \mathrm{CatSim}(\mathbf{x}^{cat}, \mathbf{y}^{cat})$ where $\mathbf{x}^{num}$ are the numerical and $\mathbf{x}^{cat}$ the categorical variables of vector $\mathbf{x}$. The parameter γ is set to the fraction of numerical features in the data. For $NumSim$ we use the cosine similarity. For $CatSim$ we use the Jaccard index to measure the overlap between two categorical vectors. This index will be directly proportional to the number of attributes in which they have an equal value [223]. The precise formulation of this similarity metric is given by Equation (4.3) where n is the number of numerical features, k is the number of categorical features, and w is set to $\frac{1}{k}$. For clarity in notation, superscripts num and cat are omitted.

A large value of $Sim(i, p)$ indicates that the distance between feature vectors U_i and U_p should be low and vice versa. By integrating the regularization term into the combined loss function, we introduce information about the relationships or similarities between distinct entities such as employees. Consequently, when regularization is employed to incorporate supplementary information, it goes beyond its conventional role of preventing overfitting. Instead, it becomes a tool for injecting employee characteristics into the learning process.

$$\text{Sim}(\mathbf{x},\mathbf{y}) = \gamma \cdot \underbrace{\frac{\sum_{i=1}^{n} \mathrm{x}_i \mathrm{y}_i}{\sqrt{\sum_{i=1}^{n} \mathrm{x}_i^2}\sqrt{\sum_{i=1}^{n} \mathrm{y}_i^2}}}_{NumSim} + (1-\gamma) \cdot \underbrace{\sum_{i=1}^{k} \Big(w \cdot s\left(x_i, y_i\right) \Big)}_{CatSim} \tag{4.3}$$

$$\text{with } s\left(x_i, y_i\right) = \begin{cases} 1 & \text{if } x_i = y_i \\ 0 & \text{otherwise} \end{cases}$$

The complete loss function of MF with SR consists of three terms (Equation 4.4). The SR term is labeled as **(iii)**.

$$\min_{U,V} \mathcal{L}(R,U,V) = \underbrace{\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(R_{ij} - U_i^T V_j\right)^2}_{\textbf{(i)}} + \underbrace{\frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2}_{\textbf{(ii)}} + \underbrace{\frac{\beta}{2}\sum_{i=1}^{m}\sum_{p=1}^{m} \text{Sim}(i,p)\left\|U_i - U_p\right\|_F^2}_{\textbf{(iii)}} \tag{4.4}$$

The full algorithm for MF with SR is represented by Algorithm 1. Specifically, the extension to regular MF (i.e., as presented in B.2) and $MFSR$ are step 6 where γ_i is calculated, step 7 where $\beta\gamma_i$ is added, and step 10 were (iii) is added to the composite loss function.

4.4 Experimental evaluation

In this section, the implementation of our proposed method is tested on three real-life datasets of which two are private and one is made publicly available. First, we describe the datasets and performance score y. Next, we outline the setup of the experiment.

4.4.1 Data

Description of observed y We use a performance score $y \in [0,1]$ to assess the quality of each job-employee match $(u, t^s, t^e, v, \mathbf{x}, y)$ in the dataset $\mathcal{D}$. Taking into account the trade-off between costs and benefits of onboarding people, the measure y that is defined exhibits a sigmoidal relationship with lead time. Specifically, lower lead times correspond to low values for y. As lead time increases, y rises until it saturates for longer lead times. This

Algorithm 1: Matrix Factorization with Similarity Regularization

Input : observed ratings R, initial matrices U and V, learning rate α, regularization parameters λ_1 and λ_2, learning steps n, stopping threshold t, similarity regularization parameter β, similarity metric Sim

Output : matrix $\hat{R} = U^T V$ with estimated ratings

for *steps* $= 1, 2, \ldots, n$ **do**

  for *each element* $R_{i,j}$ **do**

    ; > $i \in \{u_1, u_2, \ldots, u_m\}$, $j \in \{v_1, v_2, \ldots, v_n\}$

    if $R_{i,j} > 0$ **then**

      $e_{i,j} := R_{i,j} - \hat{R}_{i,j}$; > calculate error

      $\gamma_i = \sum_{p=1}^{m} \mathrm{Sim}(i,p)\left(U_i - U_p\right)$; > sim. reg.

      $U_i := U_i + \alpha(\underbrace{e_{i,j} V_j}_{\text{(i)}} + \underbrace{\lambda_1 U_i}_{\text{(ii)}} + \underbrace{\beta\gamma_i}_{\text{(iii)}})$; > update vector U_i

      $V_j := V_j + \alpha(\underbrace{e_{i,j} U_i}_{\text{(i)}} + \underbrace{\lambda_2 V_j}_{\text{(ii)}})$; > update vector V_j

  $e := \underbrace{\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(R_{ij} - U_i^T V_j\right)^2}_{\text{(i)}} + \underbrace{\frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2}_{\text{(ii)}}$

  $+ \underbrace{\frac{\beta}{2}\sum_{i=1}^{m}\sum_{p=1}^{m} \mathrm{Sim}(i,p)\left\|U_i - U_p\right\|_F^2}_{\text{(iii)}}$; > sim. reg.

  if *error* $e < t$ **then**

    break ; > stop if error falls below threshold t

Return : $\hat{R} = U^T V$

relationship is formally given by Equation 4.5, where t is expressed in years, so that $y(1) = 0.12$, $y(2) = 0.5$, $y(3) = 0.88$, and $y(4) = 0.98$.

$$y(t) = \frac{1}{1 + e^{-2(t-2)}} \tag{4.5}$$

This approach is based on discussions with the industry partner that delivered the data and perceives this to be a viable approach for using the RS in the setting of filling hard-to-fill positions. By combining insights from the literature [224] and our own experience, we have identified various factors that make it difficult to accurately and appropriately measure job performance and define the metric y. These factors include a lack of consistent monitoring of job specifications and work outcomes, constraints posed by privacy and ethical concerns, the potential influence of biases on individual performance assessments, discrepancies between intended and actual measurements, the interdependence of complex jobs in which outcomes result from the collective efforts of multiple employees, and the untraceable nature of certain employee behaviors.

Given these challenging factors, we have opted to use a performance metric based on lead times, which is always observed and a commonly-used approach in the field of labor economics [225]. Specifically, we treat a job match as a *pure experience good*, where employees with satisfactory matches are more inclined to stay, while those with lower-quality matches tend to leave early [226], [227].

However, we recognize that our definition of y serves as a proxy for true job-employee success. Other metrics and definitions may exist, depending on available data and the precise objectives of the recommender system. These variables to determine y could include salary increase, performance gain, skill enhancement, education, or training. Nonetheless, exploring these alternatives would warrant a separate research study and is beyond the scope of this manuscript.

y and ongoing cases In the context of HR and job-employee matching in specific, a match may be censored if the employee leaves the company or changes roles within the company after the end of the data capturing period [228]. This type of censoring can impact the outcome variable y, as it takes lead times into consideration. To ensure the validity and accuracy of our results, we choose to exclude ongoing observations, and withhold a random subsample to reduce the risk of bias and increase the validity of the results. The downside of leaving out these observations is the increased sparsity of matrix R, which could negatively affect the accuracy of our method [229], [230].

Dataset description We have access to three real-life event logs, one of which is fully anonymized and made publicly available along with the publication of this paper. A summary of the characteristics of these datasets is presented in Table 4.4. All three datasets are extracted from an HRIS and represent three different companies. Dataset 1 originates from a high-tech R&D company with ±3,000 employees. It contains twelve years of HR data, resulting in over 5,062 observed job-employee matches. Dataset 2 is provided by a company active in IT services with ±4,500 employees. It spans a period of ten years, resulting in 11,327 observed matches. Dataset 3 is provided by a company active in HR services with ±1,500 employees. The data spans a period of 10 years, resulting in 3,792 observed matches.

Personal features stored in x The personal employee data $\mathbf{x}$ that is available for analysis varies among the datasets. Depending on the dataset, $\mathbf{x}$ may include information on degree, the field of study, contract type, location of employment, and percentage employed. The values of these features may change over time for a small number of observations. For these observations, the features are updated to the first-observed values, respecting the out-of-time aspect for validation and testing. Other features that are available in each dataset but not considered in this analysis include nationality, marital status, zip code of residence, and gender.

Anonymous dataset Dataset 3 is publicly available on GitHub. It is obtained from an HR services provider and fully anonymized. Due to confidentiality concerns, the original features and additional background information about the data cannot be disclosed. Personal information $\mathbf{x}$ consists of eight categorical variables ($V01 - V08$) and three numerical variables ($V09 - V11$) which are normalized between 0 and 1. When an employee leaves the organization, this is indicated with the *leave* activity. The data covers the period from 2012 to 2021.

Table 4.4: Overview of datasets

Dataset	Source	Public	#Empl.	#Jobs	#Matches	Timeframe	$R^{m \times n}$ sparsity
1	High-tech		±3000	±250	5,062	2009-2021	99.3%
2	IT Services		±4500	±200	11,327	2012-2022	98.7%
3	HR Services	✓	±1500	±250	3,792	2012-2021	99.0%

4.4.2 Experimental setup

Each dataset is split into a training, validation, and test set with proportions of 0.5/0.25/0.25 in an out-of-time fashion. A job-employee rating matrix

$R^{m \times n}$ is created based on the performance scores of the observed matches. An example is shown in Table 4.3.

To be able to calculate latent representations of employees, at least an employee's first job is required in the training set. Therefore, if a certain employee u_i has no observed first job in the training set, we add one from the validation or test set. This approach is consistent with the focus of our research, which is on matching current employees with new jobs within the company rather than placing unemployed individuals in open vacancies.

To evaluate the performance of our proposed methods, we use both metrics for predicted score accuracy and metrics for assessing the relative ranking of proposed matches. The latter type of metric is used because it allows us to compare the expected performance of matches relative to other proposed matches, rather than just looking at the predicted scores themselves.

The Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), as shown in Equations 4.6 and 4.7, are used to measure the prediction accuracy of the actual performance score of a match. $\mathcal{D}_{test}$ denotes the test set of matches that are used for evaluation. Since the set of observed ground-truth matches is extremely sparse, we tweak the metrics' definition by only summing matches present in $\mathcal{D}_{test}$.

$$MAE = \frac{\sum_{(i,j) \in \mathcal{D}_{test}} \left| R_{ij} - \widehat{R}_{ij} \right|}{|\mathcal{D}_{test}|} \tag{4.6}$$

$$RMSE = \sqrt{\frac{\sum_{(i,j) \in \mathcal{D}_{test}} \left(R_{ij} - \widehat{R}_{ij} \right)^2}{|\mathcal{D}_{test}|}} \tag{4.7}$$

To assess the relative ranking of matches, we use the Spearman and Kendall rank correlation coefficients, which measure the extent to which the order of the match scores is predicted correctly. These metrics are deployed from a job point of view, i.e, we evaluate the proposed ranking of employees per job with the true ranking in the test set. Moreover, only jobs with more than two matches are considered, as a ranking is only possible with at least two matches. A higher score indicates better performance.

We evaluate the performance of matrix factorization with ($MFSR$) and without (MF) similarity regularization on three datasets. We compare the performance with six other methods: Singular Value Decomposition (SVD), *SlopeOne*, and various nearest neighbor approaches utilizing cosine similarity (KNN_c), Pearson similarity (KNN_p), difference-based similarity (KNN_{smd}), and triangle-based cosine measure (KNN_{ta}) [217].

The selection of these six methods was motivated as follows: SVD closely resembles our $MFSR$ approach and is a well-established model-based CF technique. *SlopeOne* was chosen for its simplicity, efficiency, and widespread

adoption. KNN was included due to its extensive usage in the literature [164] and its versatility in employing different similarity measures [215]. KNN_c and KNN_p were added for their simplicity and common application. Conversely, KNN_{smd} and KNN_{ta} represent more advanced techniques that are state-of-the-art in neighborhood-based CF [217].

SVD decomposes the job-employee matrix R into $U\Sigma V^T$ where U and V are left and right singular matrices and Σ is a rectangular diagonal matrix with non-negative real numbers on the diagonal, to approximate R and make predictions.

The SlopeOne algorithm predicts item ratings based on the average deviation of ratings for similar items and users [231]. It calculates pairwise differences between all items from matrix R and generates a deviation matrix storing average deviation values for each item pair to make predictions.

In recommender systems, KNN predicts item ratings based on the ratings of k nearest items or users. It constructs a user-item matrix, computes similarities using a chosen distance metric, identifies the nearest items or users, and predicts the rating of the target item by the user based on their ratings. Many similarity metrics are available, ranging from basic (such as cosine and Pearson) to advanced methods tailored to address typical RS challenges like data sparsity, demonstrating state-of-the-art performance [217].

SVD, *SlopeOne*, KNN_c, and KNN_p were implemented using the Surprise package with the Python implementation provided by [232]. KNN_{smd} and KNN_{ta} are based on the implementation of [217].

We perform a rigorous grid search on a separate out-of-time validation set for optimizing each method's hyperparameters. A full overview of the considered grid and the optimal hyperparameters for each method and dataset are summarized in B.3. These hyperparameters are tuned separately for each method and dataset.

4.5 Results and discussion

This section first presents the experimental results. Then, the discussion highlights the advantages and covers the potential drawbacks of our proposed method and of using data-driven methods in HRM in a more general context.

4.5.1 Results

Table 4.5 summarizes the experimental results. The $MFSR$ method consistently performs best across all datasets in terms of the $RMSE$ and Kendall rank correlation. Moreover, $MFSR$ always outperforms MF in all metrics over all datasets, highlighting the benefit of including the similarity regularization term. The state-of-the-art benchmark methods KNN_{smd} and

KNN_{ta} display similar performance and often perform best. The other methods have varying performances across the datasets and metrics.

Table 4.5: Summary of results. The table presents the results of evaluating three model-based CF methods (MF, $MFSR$, and SVD) and five memory-based CF methods ($SlopeOne$, KNN_c, KNN_p, KNN_{smd}, and KNN_{ta}) on three different datasets using four metrics. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are both measures of prediction error, with lower values indicating better performance. The Spearman and Kendall coefficients are both measures of rank correlation, with values closer to 1 indicating a stronger positive correlation. The best results are highlighted in bold.

Dataset	Metric	$MFSR$	MF	SVD	$SlopeOne$	KNN_c	KNN_p	KNN_{smd}	KNN_{ta}
1	MAE	0.171	0.176	0.179	0.166	0.184	0.189	**0.157**	0.159
	$RMSE$	**0.204**	0.234	0.222	0.210	0.225	0.224	0.217	0.216
	$Spearman$	0.265	0.172	0.166	0.317	0.063	-0.018	**0.339**	0.322
	$Kendall$	**0.409**	0.334	0.358	0.340	0.285	0.250	0.358	0.374
2	MAE	0.188	0.215	0.195	0.220	0.263	0.206	**0.185**	0.188
	$RMSE$	**0.237**	0.287	0.259	0.285	0.317	0.257	0.282	0.281
	$Spearman$	0.048	0.028	**0.114**	0.105	0.066	-0.029	0.092	0.076
	$Kendall$	**0.267**	0.219	0.244	0.236	0.220	0.165	0.240	0.247
3	MAE	**0.164**	0.174	0.168	0.173	0.178	0.174	0.177	0.170
	$RMSE$	**0.201**	0.219	0.211	0.217	0.217	0.208	0.221	0.214
	$Spearman$	0.194	0.216	0.172	0.226	-0.028	0.050	0.243	**0.290**
	$Kendall$	**0.386**	0.346	0.345	0.325	0.253	0.253	0.372	0.363

The experiment's findings, consolidated across the three datasets are summarized in Table 4.6, showcasing the average ranking (AR) of methods over datasets per metric. To test the experimental results for statistical significance, we employ a methodology described by Demšar [131], leveraging a combination of the Friedman test and a post-hoc Dunn test with Holm p-adjustments for multiple comparisons [233].

The bottom of Table 4.6 presents the Friedman χ^2 statistic, assessing whether there are statistically significant differences in performance among the methods. Enclosed in brackets are the p-values derived from the post-hoc test. The underscored p-values in the table denote significance at the 5% level, rejecting the null hypothesis of equal performance compared to the best method per metric.

From this table, we conclude the following. First, although table C.1 indicates that MFSR only scores best once in terms of MAE, on aggregate over the three datasets this method still is ranked first on average. Second, in terms of RMSE, the performances of MF and MFSR are significantly different, highlighting the benefit of adding the similarity regularization term. Third, in terms of Spearman, the Friedman χ^2 test statistic indicates some

significant difference but the post-hoc test fails to establish the difference to be significant. This may be a result of the lower power of the latter as described by [131]. In general, the results suggest a good performance of $MFSR$, but for establishing the difference to be statistically significant, typically more data sets are needed to increase the power of the tests.

Table 4.6: Average ranking of methods across datasets per performance measure. The best-performing methods are emphasized in bold. P-values are added between brackets. The P-values of methods that perform statistically significantly differently at the 5% level from the best-performing one are underlined.

Method	MAE		RMSE		Spearman		Kendall	
MFSR	**2.3**	(1.000)	**1.0**	(1.000)	5.0	(0.915)	**1.0**	(1.000)
MF	5.7	(0.857)	7.3	(0.025)	5.3	(0.844)	5.7	(0.307)
SVD	4.0	(0.998)	3.7	(0.959)	4.3	(0.976)	3.7	(0.941)
SlopeOne	4.7	(0.986)	4.7	(0.742)	2.7	(0.998)	5.3	(0.419)
KNN_c	7.7	(0.137)	6.7	(0.078)	6.7	(0.307)	7.0	(0.044)
KNN_p	6.0	(0.757)	3.3	(0.976)	7.7	(0.081)	7.7	(0.013)
KNN_{smd}	3.0	(0.998)	5.7	(0.317)	**2.0**	(1.000)	3.3	(0.976)
KNN_{ta}	2.7	(0.998)	3.7	(0.959)	2.3	(0.998)	2.3	(0.990)
Fried. χ^2	12.1	(0.096)	14.6	(0.042)	15.0	(0.036)	18.7	(0.009)

4.5.2 Discussion

This study explores the potential advantages of data mining and machine learning techniques in HRM. Specifically, the use of data-driven recommender systems for job matching and career path management is examined. While these methods have the potential to improve decision support in HRM, there are also risks and drawbacks that must be considered. In the following discussion, the advantages, drawbacks, and limitations of the current study are analyzed in greater depth.

To mitigate the risk of making bad decisions that is associated with the use of these data-driven methods, it is to use them as a tool for decision support, rather than for automating decision-making. HR professionals and employees obviously should have the final say in any proposed job-employee match, but recommendations that are made by such a system would also have to be examined and interpreted before acting on them.

Advantages One advantage of the proposed method is its ability to address the *cold start* problem, where little information is available about past

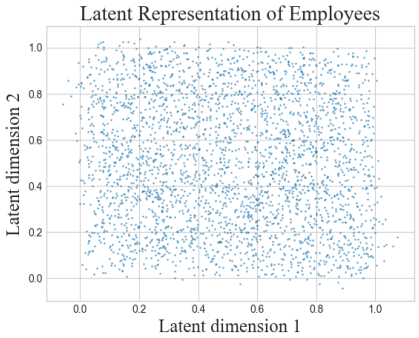
experience of employees. By utilizing similarity regularization, these methods can create meaningful embeddings based on the similarity with other employees even in the absence of extensive information about an employee in question's past experience.

Additionally, collaborative filtering, which utilizes matrix factorization techniques, maps both employees and jobs to a latent factor space in such a way that job-employee matches are the result of inner products within this space [234]. For employees, the latent factors could potentially correspond to essential dimensions such as possessing specific technical proficiency or effective management competency. Even if this were not the case, despite the fact that the latent factors themselves may not seem transparent, they still function as indicators of the closeness and resemblances between jobs and among employees. This addresses the specific challenges listed in Section 4.3.1.

Figure 4.2 displays a two-dimensional t-SNE plot of the higher-dimensional latent representation of employees [235]. The effect of similarity regularization on the latent representation can be observed by comparing the two panels, as clusters between similar employees are distinctly visible. For the purpose of simplicity, the similarity metric for similarity regularization in this example is based on *education level* only. When multiple variables are taken into account for calculating similarity, or when considering a higher-dimensional representation, the clusters become visually less distinguishable in a two-dimensional plot.

Drawbacks One of the main drawbacks of RSs is their tendency towards over-specialization, which can lead to a lack of exploration and novelty in decision-making [200]–[202]. This phenomenon, referred to as the serendipity problem, has been addressed in previous research through the incorporation of randomness to explore more novel options [236]–[238]. However, it is worth noting that this issue may be less significant in cases where routine decisions are the primary focus or where the user of the system is aware of this issue.

Additionally, RSs may prioritize short-term considerations, such as the next job an employee should take, rather than considering longer-term career development. Another limitation of RSs is their domain-specific nature. Because of current or historic HR strategies and policies, observed matches in the data do not happen at random. In other words, some of the observations on job-employee matches and transitions are missing in a non-random way (MNAR) and are related to some characteristics of the data that are not observed [239]. This can make it difficult to accurately extrapolate insights to a new domain, i.e. transitions between jobs that have never been observed before, and may also impact the validity of evaluation measures. Special-



(a) Two-dimensional latent representation for MF without similarity regularization for dataset 1.

(b) Two-dimensional latent representation for MF with similarity regularization on *education* for dataset 1.

Figure 4.2: The effect of similarity regularization. The two panels display the m two-dimensional latent representation of employees (U_i with $i = 1, 2, \ldots, m$). Panel (a) displays a much more *scattered* representation, whereas panel (b) clearly shows a more clustered representation.

ized methods such as multiple imputation, inverse probability weighting, or causal methods could mitigate or even eliminate this bias [240].

It is also worth noting that historical data, which is often used to train these systems, may not accurately reflect current HR policies or job transition patterns as a result of concept drift. To evaluate job-employee matches based on the metric y as defined in this manuscript is a strong simplification. In reality, the appointment and evaluation of employees are influenced by many factors, including more nuanced soft HR-based evaluation criteria. Furthermore, this method does not consider the capacity of positions or any other constraints or HR policies that may impact the feasibility of recommended job-employee matches. It is also highly sensitive to proper hyperparameter tuning, requiring careful validation in order to achieve reliable results. Finally, while this method has been evaluated for its ability to predict employee performance, it does not consider factors such as diversity, serendipity, novelty, and robustness in its evaluation metrics.

4.6 Conclusion

With the increasing trend of companies utilizing big data, analytics, and machine learning to gain a competitive advantage, also the HR field sees significant potential in using such tools to support decision-making. As a re-

sult, HR departments are seeking to complement traditional evidence-based techniques with data-driven insights to recommend career steps, streamline manual candidate searches, and manage the internal talent pool.

We present a data-driven approach for internally matching employees with job opportunities and guiding career path decisions using a recommender system with collaborative filtering as a baseline. This method's performance is further improved by incorporating personal information about employees with a similarity-based regularization term. Doing so addresses the cold start problem which is present due to limited data on new hires and employees with short tenure. This data limitation makes it difficult to generate meaningful, similarity-based latent employee representations. By considering both the perspective of the employee and the organization, our approach is able to provide more targeted and effective recommendations for internal mobility.

The efficacy of our approach was rigorously tested on three real-life datasets comprising thousands of job-employee matches. The results of this evaluation showed that our method outperformed the default collaborative filtering approach in terms of ranking performance. Python code of the presented method and experiments is provided, along with an anonymized HR dataset on internal mobility in the format of an event log, so as to facilitate replication of the presented results and to spur further research.

Overall, the presented results highlight the potential of using data-driven approaches to support internal mobility management in organizations and provide objective and data-driven insights that can complement the blend of HR professionals' expertise and intuition on the one hand and evidence-based techniques from more traditional scientific HR literature on the other hand. The adoption of such approaches has the potential to transform the way organizations manage the careers of their employees and make informed decisions about internal mobility.

Future work will focus on mitigating the effect of selection bias, i.e., controlling for the current HR policies that result in the non-random selection of the next jobs in a career path. We believe this issue could be addressed with causal methods.

Part II

Methodological advances

5
Robust instance-dependent cost-sensitive classification

Instance-dependent cost-sensitive (IDCS) learning methods have proven useful for binary classification tasks where individual instances are associated with variable misclassification costs. However, we demonstrate in this paper by means of a series of experiments that IDCS methods are sensitive to noise and outliers in relation to instance-dependent misclassification costs and their performance strongly depends on the cost distribution of the data sample. Therefore, we propose a generic three-step framework to make IDCS methods more robust: (i) detect outliers automatically, (ii) correct outlying cost information in a data-driven way, and (iii) construct an IDCS learning method using the adjusted cost information. We apply this framework to cslogit, a logistic regression-based IDCS method, to obtain its robust version, which we name r-cslogit. The robustness of this approach is introduced in steps (i) and (ii), where we make use of robust estimators to detect and impute outlying costs of individual instances. The newly proposed r-cslogit method is tested on synthetic and semi-synthetic data and proven to be superior in terms of savings compared to its non-robust counterpart for variable levels of noise and outliers. All our code is made available online at https://github.com/SimonDeVos/Robust-IDCS.

5.1 Introduction

Classification is a well-studied machine learning task that involves the assignment of instances to a predefined set of outcome classes. Cost-sensitive classification methods take into account asymmetric costs related to incorrectly classifying instances across various classes [9], [241]. Such misclassification costs may either be class-dependent, i.e., equal for all instances of a class, or instance-dependent, i.e., vary across instances.

Classification methods are adopted to support or automate business decision-making, e.g., for credit scoring [242] or customer churn prediction [243]. Note that in both applications, misclassified instances involve variable costs. For instance, the cost of a misclassified churner equals the future customer lifetime value, whereas a misclassified non-churner typically involves a much smaller cost, i.e., the cost of targeting the customer with the retention campaign. Either or both may be instance-dependent or class-dependent depending on the characteristics of the particular application setting.

A broad variety of cost-sensitive (CS) and instance-dependent cost-sensitive (IDCS) classification methods have been proposed in the literature as reviewed and experimentally evaluated by [244] and [245]. A prominent approach that is adopted by both CS and IDCS methods for taking misclassification costs into account is to weigh instances proportionally with the misclassification cost involved when learning a classification model.

In this article, we raise the question of whether IDCS classification methods are sensitive to outliers and noise in the data. No prior work seems to have addressed this question, which nonetheless is of significant practical importance given the broad adoption and potential monetary impact of using biased classification models for decision-making.

To address these shortcomings, we present the results of a series of experiments to evaluate the robustness of IDCS classification methods with respect to outlying costs in the data, which highlight the potential bias and vulnerability of IDCS classification methods. We propose a robust approach to IDCS classification by extending the existing cslogit approach [14]. An important benefit is the automatic and reliable detection of outliers in the data. These outliers may not only spoil the resulting analysis (as illustrated in this article) but can also contain valuable information. A robust analysis can thus provide better insight into the structure of the data.

The following section outlines related work on IDCS learning and discusses both cslogit and robustness. Next, in Section 5.3, simulations on synthetic data are presented that motivate the need for robust IDCS learning, which we develop in Section 5.4. Section 5.5 presents the results of experiments illustrating the excellent performance of the proposed robust IDCS learning method, denoted r-cslogit, in comparison with both logit and cslogit. We conclude and present directions for future research in Section 5.6.

5.2 Related work

Elkan [9] introduces a learning paradigm where different misclassification errors incur different penalties depending on the predicted and actual class, with applications to, for example, detecting transaction fraud and credit scoring. The benefits and costs of different predictions can be summarized in a two-dimensional instance-dependent cost matrix with one dimension for the predicted value and another dimension for the ground truth. Given these benefits and costs, each new instance should be assigned to the class that leads to the lowest expected cost, which is calculated by means of conditional probabilities.

5.2.1 IDCS learning, cslogit, and robustness

For certain applications, benefits and costs depend not only on the class but also on the instance itself. Therefore, instance-dependent cost-sensitive learning considers a more detailed, lower level of granularity than class-dependent costs. For these applications, using instance-dependent costs instead of class-dependent costs leads to a decreased total misclassification cost [245], [246].

Several instance-dependent cost-sensitive methodologies have been proposed in the literature, with recent overviews given by [244] and [245]. Especially relevant to our work are methodologies that adjust the learning algorithm to incorporate instance-dependent costs. Instance-dependent cost-sensitive variants have been proposed for several common machine learning classifiers, such as boosting [14], [247], [248], support vector machines [246], decision trees [249], [250], and logistic regression [14], [251].

In this work, we will build upon an instance-dependent cost-sensitive version of logistic regression. Following [14], we will refer to this method as cslogit. Logistic regression is a widely used method for binary classification tasks. To extend logistic regression to its IDCS counterpart, Bahnsen, Aouada, Stojanovic, *et al.* [252] and Höppner, Baesens, Verbeke, *et al.* [14] propose an objective function that combines both cost-sensitivity and instance-dependent learning, resulting in instance-dependent costs for optimization. The application of this objective function yields significant improvements in terms of higher savings compared to cost-insensitive or class-dependent cost-sensitive models in the context of, for example, credit scoring and transaction fraud detection.

Classical nonrobust methods for regression, such as least squares or maximum likelihood techniques, try to fit the model optimally to all the data. As a result, these methods are heavily influenced by data outliers. This implies that outliers may bias the parameter estimates and confidence intervals,

and thus, hypothesis tests may become unreliable and/or uninformative. In contrast, robust methods can resist the effect of outliers to avoid distorted results and false conclusions. As an important benefit, they allow the automatic detection of outliers as observations that deviate substantially from the robust fit. It is important to note that the detected outliers are not necessarily errors in the data. The presence of outliers may reveal that the data are more heterogeneous than has been assumed and that it can be handled by the original statistical model. Outliers can be isolated or may come in clusters, indicating that there are subgroups in the population that behave differently. Many different approaches to robust regression have been proposed, and a good overview can be found in reference works such as [253]–[255]. In the context of generalized linear models (GLMs), various robust alternatives have been presented, such as [256]–[260]. Robust logistic regression has been studied by [261]–[269].

5.2.2 Preliminaries

The dataset $\mathcal{D}$ consists of N observed predictor-response pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ and is used to train a binary classification model $s(.)$. The costs C_i correspond to the cost matrix defined in Table 5.1.

This binary classification model predicts a probability score $s_i \in [0, 1]$ for each instance i based on the features $\boldsymbol{x}_i$. Depending on the classification threshold t_i^*, s_i is converted to a predicted class $\widehat{y}_i \in \{0, 1\}$.

For models trained with AEC (Equation (5.3)), savings remain relatively stable across different thresholding strategies [245]. Therefore, we use a default threshold of 0.5.

A binary logistic regression predicts a probability score that an observation belongs to the positive class. This probability score is calculated by Equation (5.1), where β_0 is the bias term, $\beta_1 \dots \beta_d$ the learned weights and $\boldsymbol{x}_i$ are the features of a particular observation i:

$$s_i = s_{(\beta_0, \boldsymbol{\beta})}(\boldsymbol{x}_i) = \frac{1}{1 + e^{-z}} \text{ where } z = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{id}. \quad (5.1)$$

This probability score is then compared to a threshold to categorize each of these observations into classes. The objective function of a logistic regression is the likelihood that is maximized or the cross-entropy loss that is minimized. For a single sample with true label $y_i \in \{0, 1\}$ and a probability score $s_i = P(Y = 1)$, the cross-entropy loss is presented by Equation (5.2):

$$L_{\log}(y_i, s_i) = -\Big(y_i \log(s_i) + (1 - y_i) \log(1 - s_i)\Big). \quad (5.2)$$

Note that this equation does not take into account any costs. Because this objective function assigns equal weights to each misclassification, it does

not necessarily correspond to the underlying business problem where costs are to be minimized. The reason for this is twofold: misclassification costs are different per class and per instance. The real business objective is to minimize the average expected total cost of the binary classifier.

We build upon the instance-dependent cost-sensitive logistic (cslogit) model as proposed by [252] and [14]. Cslogit minimizes an instance-dependent cost-sensitive objective function corresponding to the real business objective of minimizing costs in domains such as customer churn prediction, credit scoring, and direct marketing [270], [271]. Dependent on this business objective, also other cost matrices can be considered. For example, Höppner, Baesens, Verbeke, *et al.* [14] propose the cost matrix $C_i(0 \mid 0) = 0$, $C_i(0 \mid 1) = A_i$, $C_i(1 \mid 0) = c_f$, and $C_i(1 \mid 1) = c_f$ for the detection of transfer fraud where c_f is a fixed administrative fee. Alternatively, Bahnsen, Aouada, and Ottersten [251] propose the cost matrix $C_i(0 \mid 0) = 0$, $C_i(0 \mid 1) = L_{gd}$, $C_i(1 \mid 0) = r_i + C^a_{FP}$, and $C_i(1 \mid 1) = 0$ for credit scoring where L_{gd} is the loss given default, r_i is the loss in profit by rejecting what could have been a good customer, and C^a_{FP} is the cost related to the assumption that the financial institution will not keep the amount of the declined applicant unused. However, the reason for this work is to address the need for robustness and to propose a solution to solve this potential issue in a generic way, regardless of its application. Therefore, to present an application-agnostic methodology and preferring the most simple cost matrix, this work utilizes a symmetric cost matrix.

Equation (5.3) shows the average expected cost (AEC), the cost-sensitive objective function that is used by cslogit, given a symmetric cost matrix, as shown in Table 5.1:

$$\begin{aligned} AEC(s(\mathcal{D})) &= \frac{1}{N} E[\text{Cost}(s(\mathcal{D})) \mid \boldsymbol{X}] \\ &= \frac{1}{N} \sum_{i=1}^{N} \Big(y_i [s_i C_i(1 \mid 1) + (1 - s_i) C_i(0 \mid 1)] \\ &\quad + (1 - y_i)[s_i C_i(1 \mid 0) + (1 - s_i) C_i(0 \mid 0)] \Big) \\ &= \frac{1}{N} \sum_{i=1}^{N} \Big(A_i (y_i (1 - s_i) + (1 - y_i) s_i) \Big). \end{aligned} \tag{5.3}$$

In Equation (5.3), each observation i is a pair of d features $\boldsymbol{x}_i = (x_{i1}, ..., x_{id})$ and a binary response label $y_i \in \{0, 1\}$.

Across multiple models, the total cost as a metric is not unambiguously interpretable, as datasets with high instance-dependent costs might have a higher total misclassification cost but still have a better relative score. Proceeding with the idea of normalizing the total classification costs of a model

Table 5.1: Symmetric cost matrix for cslogit

	Actual 0	Actual 1
Predicted as 0	$C_i(0 \mid 0) = 0$	$C_i(0 \mid 1) = A_i$
Predicted as 1	$C_i(1 \mid 0) = A_i$	$C_i(1 \mid 1) = 0$

presented in [272], [251] introduce a more interpretable metric: *Savings*. This metric represents the relative improvement of the cost of a newly proposed model, $Cost(s(\mathcal{D}))$, compared to the cost of using an empty model that assigns all instances to a single class, $Cost_{empty}(\mathcal{D})$. $Cost_{empty}(\mathcal{D})$ is calculated by taking the minimum of the costs incurred when classifying all instances as either belonging to the negative or positive class:

$$Cost_{empty}(\mathcal{D}) = \min\left\{Cost\left(s_0(\mathcal{D})\right), Cost\left(s_1(\mathcal{D})\right)\right\}. \tag{5.4}$$

Using the $Cost_{empty}(\mathcal{D})$ of an empty model as a factor to normalize total costs, *Savings* of the model $s(\mathcal{D})$ are calculated by Equation (5.5):

$$Savings(s(\mathcal{D})) = 1 - \frac{Cost(s(\mathcal{D}))}{Cost_{empty}(\mathcal{D})}. \tag{5.5}$$

5.3 Sensitivity analysis

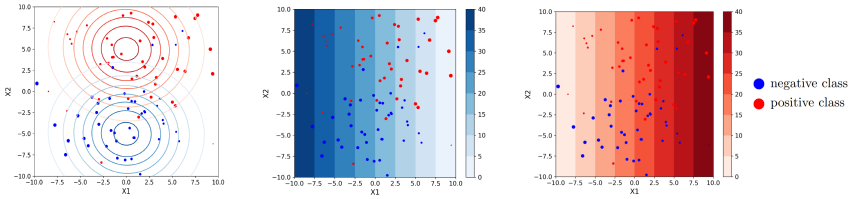
Data can contain outliers in terms of misclassification costs due to various reasons, such as missing data, invalid observations, or typos. By incorporating instance-dependent costs in the learning algorithm, outliers in these misclassification costs could potentially have a large impact on instance-dependent cost-sensitive learning methodologies such as cslogit. Therefore, we test the sensitivity of cslogit to these outliers and examine to what extent this is a shortcoming of this method.

5.3.1 Simulation setup

We analyze the sensitivity to outlying costs through a series of simulations on synthetic data. The different synthetic datasets all share the following properties. Each observation is visualized by a dot, with the size of the dot corresponding to its misclassification cost. The positive class is presented in red and the negative class in blue. Each observation has, other than its misclassification cost and label, two features: X_1 and X_2. X_1 is the feature for the misclassification cost A. For the positive class, this cost is positively related to X_1. Cases of the negative class have a negative relation between X_1 and their cost. The underlying function is given by Equation (5.6):

$$A_i = \begin{cases} 20 + 2x_{1i} & \text{for the positive class,} \\ 20 - 2x_{1i} & \text{for the negative class.} \end{cases} \tag{5.6}$$

Panel (b) and (c) in Figure 5.1 visualize this equation.



(a) Class distribution

(b) Generated costs: negative class

(c) Generated costs: positive class

Figure 5.1: The setup for synthetic data. Panel (a) displays the distribution of the negative and positive class, dependent on X_2. Panels (b) and (c) represent the misclassification costs of observations from the negative and positive classes as a linear function of X_1, generated by Equation (5.6). The three panels all show a sample of size 50 per class.

X_2 is the feature that determines the two distributions of classes 0 and 1. The two class distributions are a 2-dimensional Gaussian, sharing the same standard deviations. Observations from the negative class are sampled from $N(\mu_0, \sigma_0^2, \nu_0, \tau_0^2, \rho)$ and observations from the positive class from $N(\mu_1, \sigma_1^2, \nu_1, \tau_1^2, \rho)$.

μ_0 and μ_1 are both equal to 0, while $\nu_0 = -5$ and $\nu_1 = 5$. The variances $\sigma_0^2, \tau_0^2, \sigma_1^2$ and τ_1^2 are equal to 4. As there is no correlation between the two dimensions X_1 and X_2, ρ is equal to 0. The cases of the positive class have a higher X_2 value than the cases of the negative class. Panel (a) in Figure 5.1 displays these class distributions by which data are generated.

To generate outliers, both in the synthetic setup (Sections 5.3 and 5.5.1) and in the sensitivity analysis on real data (Section 5.5.2), the multivariate distribution of $\mathbf{X}$ remains unchanged since we only focus on outliers in the observed costs of instances. We use the *Tukey-Huber* contamination model to generate the cost distribution with outliers where the Dirac function is applied to generate costs of any size [254].

Given these settings for instance-dependent costs and class distribution, observations of the negative class with a high associated cost are expected to be located in the third quadrant and observations of the positive class with a high associated cost in the first quadrant. The symmetric cost matrix used for the examples on synthetic data is presented in Table 5.1 as introduced in Section 5.2.2.

5.3.2 Results

Within each setting, two classifiers are compared: logit and cslogit. They are both linear classifiers and propose a distinctly different decision boundary based on the training data. Since the data are only two-dimensional, these decision boundaries can be visually represented by lines. The logit and cslogit models' proposed boundaries are respectively colored in red and blue. The normal behavior of both models in the default settings of examples on synthetic data is visualized in Panel a of Figure 5.2. This figure motivates the need for a robust version of cslogit. Three examples of synthetic data where logit and cslogit are tested are shown. Panel a shows the normal behavior of cslogit and logit in the default case. Panel b displays the case where a large outlier is added. The blue decision boundary of cslogit shifts, while the red decision boundary of logit remains stable. Note that the effect of the outlier can be subtle, i.e. it *pulls* on the decision boundary, resulting in a slight rotation, without actually being classified correctly. Panel c displays the case where random noise is added to the misclassification costs. The decision boundary of cslogit shifts even further, resulting in an almost perpendicular boundary in comparison with Panel a. We further elaborate on the exact setting of the examples on synthetic data in Section 5.5.1.

5.4 Robust IDCS

To overcome the sensitivity of instance-dependent cost-sensitive classifiers to outlying costs, we introduce a three-step framework to make IDCS methods robust by detecting outliers and adapting their cost matrix. Hence, the final model will be trained using a less volatile and more rigid set of costs. The resulting robust classification model will also yield automatic outlier detection. The concrete implementation of this framework is represented by Algorithm 2.

To estimate the misclassification costs of observations in a robust manner in step 1, a regression with Huber loss is applied. Concretely, we estimate the cost A_i of observation i as a function of its features $\mathbf{x}_i$ and label y_i with a linear regression with a Huber loss function: $\hat{A}_i = f(\boldsymbol{x}_i, y_i)$. A formalization of robustness in statistics started with the work of [273]. Interestingly, his ground-breaking results and well-known loss function are still widely used today in the field of statistics and machine learning. The Huber loss function is defined by Equation (5.7). This results in a regression that is less sensitive to outliers than traditional regression methods, which often use a squared error loss.

Algorithm 2: Robust IDCS

Input : $\mathcal{D} = \{(\mathbf{x}_i, A_i, y_i) : i = 1, \ldots, N\}$ where $\mathbf{x}_i$ is a feature vector, A_i is the associated misclassification cost and $y_i \in \{0, 1\}$ is the response label of an observation i.

Output : Robust IDCS predictions $\hat{y}$ of label y

Step 1: Detect outliers.

Train a linear regression model with Huber loss so that:

$\hat{A} = f(\mathbf{X}, Y)$

Initialize set $\mathcal{S}_{outlier} := \varnothing$

for *each observation* i **do**

  if *absolute value of the standardized residuals* > 2.5 **then**

    add observation $(\mathbf{x}_i, A_i, y_i)$ to set $S_{outlier}$

    remove observation $(\mathbf{x}_i, A_i, y_i)$ from $\mathcal{D}$

Step 2: Impute instance-dependent misclassification cost.

for *each observation* $i \in \mathcal{S}_{outlier}$ **do**

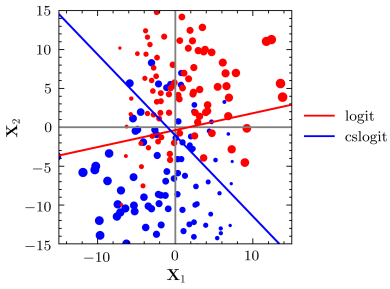
  replace A_i with $\hat{A}_i$

$\mathcal{DI} := \mathcal{D} \cup \mathcal{S}_{outlier}$

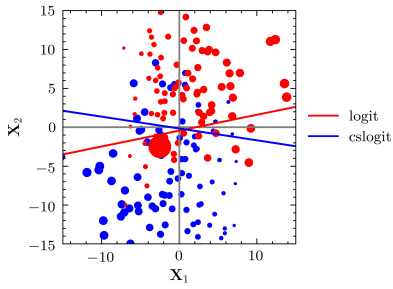
Step 3: Apply the IDCS method.

Apply cslogit to the new set $\mathcal{DI}$

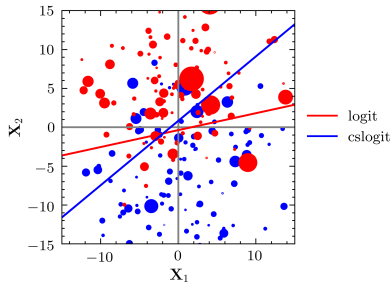
Return : $\hat{y}$



(a) Example on synthetic data: *Optimal* behavior of logit and cslogit



(b) Example on synthetic data: Influence of outliers on cslogit



(c) Example on synthetic data: Influence of noise on cslogit

Figure 5.2: The instability of cslogit's decision boundary.

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \leq \delta, \\ \delta\left(|a| - \frac{1}{2}\delta\right) & \text{otherwise.} \end{cases} \tag{5.7}$$

Next, to detect outliers, we compare the absolute value of the standardized residuals with a cutoff value of a normal distribution [274]. If this value exceeds 2.5, we consider it an outlier and add it to the initially empty set $\mathcal{S}_{outlier}$.

The observed cost A_i of observation i is operationally defined as an outlier if, for an estimator $\hat{A} = f(X, Y)$, the absolute value of the standardized residual ϵ_i is larger than 2.5.

Figure 5.3 further clarifies the concept of a conditional outlier on the setting of synthetic data where noise is added to the observed costs, as is displayed in Figure 5.2c. In this figure, the black line displays the estimated

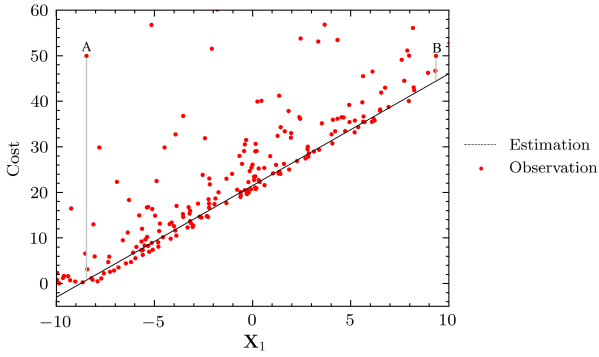


Figure 5.3: Cost estimation in function of X_1.

costs, as predicted by the linear regression model with Huber loss (Algorithm 2, line 2). The red dots represent the costs of observations in function of X_1. Consider the two observations A and B. To check whether their observed costs are outliers, we look at the standardized residuals, represented by the grey vertical lines for those two observations. Both observations A and B have an observed cost of 50. The standardized residual of A exceeds 2.5, whereas for B it is smaller than 2.5. Consequently, although both observations have the same cost, the cost of A is considered an outlier, whereas the cost of B is not.

By doing so, the costs A that are outliers, conditional on their features $\mathbf{X}$ and label Y, are detected. In step 2, the observed outlying costs A of all observations in $\mathcal{S}_{outlier}$ are imputed with their estimated counterpart $\hat{A}$. This results in a robust cost matrix (Table 5.2).

Table 5.2: Symmetric cost matrix for r-cslogit

	$y = 0$	$y = 1$
$\hat{y} = 0$	$C_i(0 \mid 0) = 0$	$C_i(0 \mid 1) = \begin{cases} \hat{A}_i = f(\boldsymbol{x}_i, y_i) & \text{if outlier,} \\ A_i & \text{otherwise} \end{cases}$
$\hat{y} = 1$	$C_i(1 \mid 0) = \begin{cases} \hat{A}_i = f(\boldsymbol{x}_i, y_i) & \text{if outlier,} \\ A_i & \text{otherwise} \end{cases}$	$C_i(1 \mid 1) = 0$

Equation (5.8) retakes the cost-sensitive objective function AEC, given by Equation (5.3), but adapts it to the new robust cost-matrix given by Table 5.2. Indicator function $\mathbb{1}_o$ takes value 1 if the cost A_i of observation i is classified as an outlier.

$$\begin{aligned}\mathrm{AEC}(s(\mathcal{D})) &= \frac{1}{N} E[\mathrm{Cost}(s(\mathcal{D})) \mid \boldsymbol{X}] \\ &= \frac{1}{N} \sum_{i=1}^{N} \Bigg[\mathbb{1}_o \Big(\hat{A}_i \left(y_i \left(1 - s_i\right) + \left(1 - y_i\right) s_i \right) \Big) \\ &\quad + \left(1 - \mathbb{1}_o\right) \Big(A_i \left(y_i \left(1 - s_i\right) + \left(1 - y_i\right) s_i \right) \Big) \Bigg]\end{aligned} \tag{5.8}$$

5.5 Results

This section discusses the performance of logit, cslogit, and the novel r-cslogit on synthetic data and tests their sensitivity on real data with additional outliers. In the reported experiments, symmetric cost matrices are taken into account. However, the use of alternative cost matrices as presented in Section 5.2.2 yields similar results concerning robustness.

The performance of binary classification algorithms is typically measured by labeling one class as positive and the other class as negative and constructing a confusion matrix. Positive classes are typically used to describe the minority class, and negative classes are used to describe the majority class. From the confusion matrix, we count the following numbers. True Negatives (TN) are the number of correctly classified negative cases. False Positives (FP) are the number of negative cases incorrectly classified as positive. False Negatives (FN) are the number of positive cases incorrectly classified as negative. True Positives (TP) are the number of correctly classified positive cases. With these numbers, we define *Sensitivity* or *Recall*, *Specificity*, and *Precision* by Equations 5.9, 5.10, and 5.11:

$$Sensitivity = Recall = \frac{TP}{TP + FN} \tag{5.9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5.10}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.11}$$

Equation 5.12 defines $F1$-measure, which is the Harmonic mean of precision and recall.

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{5.12}$$

The area under the receiver operating curve (AUC) of a classifier can be interpreted as a measure of the probability that a randomly chosen minority

case is predicted to have a higher score than a randomly chosen majority case. Therefore, a higher AUC indicates better classification performance [14]. Class distributions and misclassification costs are not taken into account in calculating the AUC.

Equation 5.13 defines the Brier score, where s_i is the predicted probability and y_i is the observed outcome. This metric measures the mean squared difference between the predicted probability and the actual outcome and is used to assess whether the model's predictions are calibrated probabilities. A lower score is better.

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(s_i - y_i)^2 \tag{5.13}$$

5.5.1 Synthetic data

In this subsection, we reuse the examples on synthetic data introduced in Section 5.3.1 to demonstrate how the possible shortcomings of cslogit can be countered by deploying the more robust r-cslogit. Figure 5.4 displays the decision boundaries of logit, cslogit, and r-cslogit in red, blue, and green, respectively.

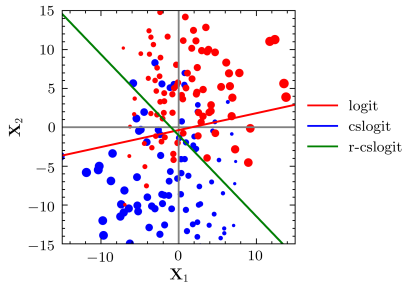
Synthetic data: Three settings

The basic setting of the examples on synthetic data in Panel a is the same as explained in Subsection 5.3.1. In Panel b, an additional outlier of the positive class is added in the third quadrant with a cost equal to 400. The robust method first estimates its cost with a linear Huber regression to be 13.75 and flags it as an outlier. Next, the cost for this instance of 400 is changed to its estimated cost of 13.75. In Panel c, we add noise to the costs. Hence, the misclassification costs are generated by Equation (5.14), where the noise ϵ_i is sampled from a lognormal distribution with parameters $\mu = 2$ and $\sigma = 1.5$.

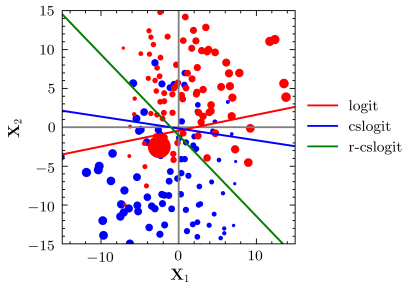
$$A_i = \begin{cases} 20 + 2x_{1i} + \epsilon_i & \text{for the positive class,} \\ 20 - 2x_{1i} + \epsilon_i & \text{for the negative class.} \end{cases} \tag{5.14}$$

Description of results

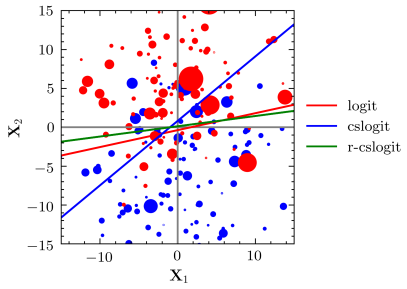
Figure 5.4 visualizes the decision boundaries of the three models. In Panel a, the decision boundaries of cslogit and r-cslogit overlap as regression with Huber Loss can perfectly predict the underlying function of associated misclassification costs as a function of X_1 in the absence of noise or outliers.



(a) Example on synthetic data: Default case



(b) Example on synthetic data: One outlier



(c) Example on synthetic data: Additional noise

Figure 5.4: Superiority of r-cslogit.

Panel b displays the case where one outlier is added. The decision boundary of the logit model is not affected by the size of misclassification costs. Hence, it is not influenced by the outlier and remains unchanged, demonstrating normal behavior as defined before. The blue decision boundary of the cslogit model is strongly influenced by outliers. The objective function takes into account the full misclassification costs of the observations in the training set, including the excessive outliers. As a consequence, the behavior of the cslogit model has been completely disrupted. This is strongly in conflict with its normal behavior, as the decision boundary is almost tilted by a quarter turn. This tilted decision boundary results in poor predictive classification power, making the cslogit model to be of inferior quality. The green decision boundary of r-cslogit remains largely unchanged, as it is robust against the single added outlier.

Performance metrics are summarized in Table 5.3. We consider the cost-

Table 5.3: Results on synthetic data (i) in the default setting, (ii) with an outlier, and (iii) with additional noise added to the amounts. The sample size is set to 300, i.e., 150 per class. The data is generated according to the setting as explained in Section 5.3.1. We apply a 2×5-fold cross-validation procedure with a train/test split ratio of 0.8/0.2. The best-performing methods are indicated in bold. A full analysis on synthetic data with different settings for class imbalance and outlier size can be found in Appendix C.1. We report the average together with the standard deviation over these 10 runs.

Setting	Method	Savings	F1	AUC	Sensitivity	Specificity	Brier
default setting	logit	0.68 ± 0.08	**0.84±0.04**	**0.85±0.04**	**0.83±0.04**	**0.88±0.05**	**0.14±0.04**
	cslogit	**0.80±0.05**	0.77 ± 0.05	0.77 ± 0.06	0.78 ± 0.03	0.77 ± 0.11	0.23 ± 0.06
	r-cslogit	**0.80±0.05**	0.77 ± 0.05	0.77 ± 0.06	0.78 ± 0.03	0.77 ± 0.11	0.23 ± 0.06
with outlier	logit	0.68 ± 0.08	**0.84±0.04**	**0.86±0.04**	**0.83±0.04**	**0.88±0.05**	**0.14±0.04**
	cslogit	0.65 ± 0.15	0.82 ± 0.05	0.83 ± 0.05	0.81 ± 0.05	0.8219 ± 0.07	0.17 ± 0.05
	r-cslogit	**0.80±0.05**	0.77 ± 0.05	0.77 ± 0.06	0.78 ± 0.03	0.77 ± 0.11	0.23 ± 0.06
with noise	logit	0.68 ± 0.08	**0.84±0.04**	**0.86±0.04**	**0.83±0.04**	**0.88±0.05**	**0.14±0.04**
	cslogit	0.57 ± 0.09	0.76 ± 0.06	0.77 ± 0.05	0.77 ± 0.06	0.83 ± 0.04	0.17 ± 0.05
	r-cslogit	**0.78±0.06**	0.82 ± 0.03	0.83 ± 0.04	0.81 ± 0.05	0.86 ± 0.05	0.15 ± 0.03

sensitive metric *Savings* introduced in Section 5.2.2 and cost-independent metrics *Sensitivity*, *Specificity*, $F1$, *AUC*, and *Brier* score.

r-cslogit outperforms logit and cslogit in terms of Savings when we add an outlier and noise. Moreover, the performance in terms of Savings remains unchanged after adding an outlier. In the default case of setting one, r-cslogit and cslogit are equivalent, as they make the exact same predictions. When considering cost-insensitive metrics, logit performs best. A full analysis on synthetic data where we experiment with different settings of class imbalance and outlier size can be found in Appendix C.1.

5.5.2 Sensitivity analysis on real data

In this subsection, we analyze the sensitivity of the three methods in an experiment with real data where we add an additional outlier, gradually increasing in size. To add outliers, we randomly select an observation and change its class label and instance-dependent misclassification cost.

This setup is similar to the second setup with synthetic data as presented in the previous subsection. The performance is measured by the cost-sensitive metric *Savings* as described before as well as the cost-independent metrics *Sensitivity*, *Specificity*, $F1$, *AUC*, and *Brier* score. The measurement of performance makes use of five-fold cross-validation with a stratified split on class distribution that is repeated twice with a different random initialization.

Description of the dataset

The dataset on which the three methods are tested is the Kaggle Credit Card Fraud Detection dataset [275]. The dataset dates from September 2013 and contains transactions made by European credit cardholders. A total of 492 out of 284,807 transactions are fraudulent, resulting in a high class imbalance. The numerical input features $V1, V2, \ldots, V28$ are the results of a PCA transformation to anonymize the dataset. *Time* and *Amount* have not been transformed. The feature *Time* is not taken into consideration in this experiment and is therefore dropped in the preprocessing phase. The feature *Amount* is the transaction amount, which is of high importance in cost-sensitive instance-dependent learning and translates into our setting as the instance-dependent misclassification cost. The feature $Class \in \{0, 1\}$ indicates whether a transaction is fraudulent or not.

Results

Table 5.4 contains the results of a 2×5-fold cross-validation procedure for the Kaggle Credit Card Fraud Detection dataset. We measure each classifier's performance averaged over the ten (2×5) test sets with the metrics *Savings*, $F1$, AUC, *Sensitivity*, $Specificity$, and *Brier* score, where instance-independent thresholds are applied.

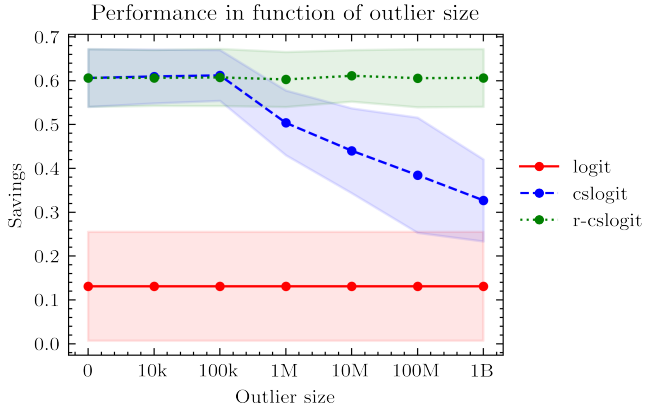


Figure 5.5: Sensitivity analysis on real data.

In terms of *Savings*, logit is always outperformed by cslogit and r-cslogit. When adding an outlier, r-cslogit outperforms cslogit. Note that the performance of r-cslogit remains stable for all considered metrics when increasing the size of the outlier. In terms of cost-insensitive metrics AUC, $Specificity$,

Table 5.4: Sensitivity analysis on real data resulting from a two times five-fold cross validation procedure on the Kaggle Credit Card Fraud Detection dataset. The size of the outlier is gradually increased. Each metric is based on 10 (2×5) out-of-sample performance estimates over the 10 test sets. We report the average together with the standard deviation over these 10 runs.

Outlier size	Method	Savings	F1	AUC	Sensitivity	Specificity	Brier
	logit	0.13 ± 0.14	0.72 ± 0.01	**0.97±0.01**	0.62 ± 0.04	**0.99±0.00**	**0.00±0.00**
0	cslogit	**0.61±0.06**	0.81 ± 0.01	0.93 ± 0.02	0.78 ± 0.04	0.99 ± 0.00	0.00 ± 0.00
	r-cslogit	**0.61±0.06**	**0.81±0.02**	0.93 ± 0.02	**0.78±0.04**	0.99 ± 0.00	0.00 ± 0.00
	logit	0.13 ± 0.14	0.72 ± 0.01	**0.97±0.01**	0.62 ± 0.04	**0.99±0.00**	**0.00±0.00**
10K	cslogit	**0.61±0.06**	**0.81±0.01**	0.93 ± 0.02	**0.78±0.04**	0.99 ± 0.00	0.00 ± 0.00
	r-cslogit	0.59 ± 0.07	0.81 ± 0.01	0.93 ± 0.02	0.78 ± 0.04	0.99 ± 0.00	0.00 ± 0.00
	logit	0.13 ± 0.14	0.72 ± 0.01	**0.97±0.01**	0.62 ± 0.04	**0.99±0.00**	**0.00±0.00**
100K	cslogit	0.61 ± 0.06	**0.81 ± 0.01**	0.93 ± 0.02	0.78 ± 0.04	0.99 ± 0.00	0.00 ± 0.00
	r-cslogit	**0.61±0.06**	0.81 ± 0.01	0.93 ± 0.02	**0.78±0.04**	0.99 ± 0.00	0.00 ± 0.00
	logit	0.13 ± 0.14	0.72 ± 0.01	**0.97±0.01**	0.62 ± 0.04	**0.99±0.00**	**0.00±0.00**
1M	cslogit	0.54 ± 0.06	0.72 ± 0.03	0.89 ± 0.04	0.78 ± 0.04	0.99 ± 0.00	0.00 ± 0.00
	r-cslogit	**0.60±0.60**	**0.81±0.01**	0.93 ± 0.02	**0.78±0.03**	0.99 ± 0.00	0.00 ± 0.00
	logit	0.13 ± 0.14	0.72 ± 0.01	**0.97±0.01**	0.62 ± 0.04	**0.99±0.00**	**0.00±0.00**
10M	cslogit	0.41 ± 0.13	0.66 ± 0.04	0.87 ± 0.05	0.76 ± 0.06	0.99 ± 0.00	0.00 ± 0.00
	r-cslogit	**0.61±0.06**	**0.81±0.01**	0.93 ± 0.02	**0.78±0.04**	0.99 ± 0.00	0.00 ± 0.00
	logit	0.13 ± 0.14	0.72 ± 0.01	**0.97±0.01**	0.62 ± 0.04	**0.99±0.00**	**0.00±0.00**
100M	cslogit	0.32 ± 0.24	0.62 ± 0.04	0.85 ± 0.05	0.73 ± 0.09	0.99 ± 0.00	0.00 ± 0.00
	r-cslogit	**0.60±0.06**	**0.81±0.01**	0.93 ± 0.02	**0.78±0.04**	0.99 ± 0.00	0.00 ± 0.00
	logit	0.13 ± 0.14	0.72 ± 0.01	**0.97±0.01**	0.62 ± 0.04	**0.99±0.00**	**0.00±0.00**
1B	cslogit	0.29 ± 0.27	0.63 ± 0.05	0.87 ± 0.04	0.76 ± 0.05	0.99 ± 0.00	0.00 ± 0.00
	r-cslogit	**0.61±0.06**	**0.81±0.01**	0.93 ± 0.02	**0.78±0.04**	0.99 ± 0.00	0.00 ± 0.00

and *Brier* score, logit performs best. In terms of $F1$ and *Sensitivity*, logit is outperformed by either cslogit or r-cslogit. This could be due to the effect of class imbalance and is in line with previous findings of Höppner, Baesens, Verbeke, *et al.* [14]. The results in terms of Savings are visualized in Figure 5.5. Since the logit model is not cost-sensitive, its performance remains constant after adding an outlier. The performance of cslogit is strongly disrupted after the cost of the outlier is set to 1 M or larger. This corresponds with the shift of the two-dimensional linear decision boundary, as shown by the findings of the examples on synthetic data. Even though the dataset contains over 280,000 instances, a single outlier, albeit a large outlier, can unhinge the cslogit method. When increasing the misclassification cost of a single outlier, the performance of r-cslogit remains stable. It is certainly more robust to this additional noise than its non-robust counterpart, as the individual outlier is detected and its cost is imputed with an estimated, expected cost. The shaded areas in Figure 5.5 represent the variability of performance over different folds in cross-validation. In contrast to the variability of cslogit, which increases drastically, the variability in the performance of r-cslogit remains stable.

5.6 Conclusion

Instance-dependent cost-sensitive (IDCS) learning methods take into account variable misclassification costs across instances in the training data in learning a classification model. This allows for optimizing the performance of the resulting classification model in terms of the misclassification costs rather than the classification accuracy.

In this article, we present the results of a series of experiments on synthetic data to demonstrate the sensitivity of IDCS methods to outliers and noise in the data. We show that the resulting classification model may be highly sensitive to outlying instance-dependent costs, in learning an instance-dependent cost-sensitive classification model. Consequently, using existing cost-sensitive models in the presence of noise or outliers can result in large misclassification costs.

To address this potential vulnerability, we propose a generic, IDCS-method-independent, three-step framework to develop robust IDCS methods with respect to the effects of random variability and noise. In the first step, instances with outlying misclassification costs are detected. In the second step, outlying costs are corrected in a data-driven way. In the third step, an IDCS learning method is applied using the adjusted instance-dependent cost information.

This generic framework is subsequently applied in combination with cslogit, which is a logistic regression-based IDCS method, to obtain its robust version named r-cslogit. The robustness of this approach is introduced in the first two steps of the generic framework by making use of robust estimators to detect and impute outlying costs of individual instances. The newly proposed r-cslogit method is tested on synthetic and semi-synthetic data. The results show that the proposed method is superior in terms of cost savings when compared to its non-robust counterpart for variable levels of noise and outliers.

6

Decision-centric fairness: Evaluation and optimization for resource allocation problems

Data-driven decision support tools play an increasingly central role in decision-making across various domains. In this work, we focus on binary classification models for predicting positive-outcome scores and deciding on resource allocation, e.g., credit scores for granting loans or churn propensity scores for targeting customers with a retention campaign. Such models may exhibit discriminatory behavior toward specific demographic groups through their predicted scores, potentially leading to unfair resource allocation. We focus on demographic parity as a fairness metric to compare the proportions of instances that are selected based on their positive outcome scores across groups. In this work, we propose a decision-centric fairness methodology that induces fairness only within the decision-making region—the range of relevant decision thresholds on the score that may be used to decide on resource allocation—as an alternative to a global fairness approach that seeks to enforce parity across the entire score distribution. By restricting the induction of fairness to the decision-making region, the proposed decision-centric approach avoids imposing overly restrictive constraints on the model, which may unnecessarily degrade the quality of the predicted scores. We empirically compare our approach to a global fairness approach on multiple (semi-synthetic) datasets to identify scenarios in which focusing on fairness where it truly matters, i.e., decision-centric fairness, proves beneficial.

6.1 Introduction

In an increasingly data-driven world, algorithms and machine learning models play a crucial role in business decision-making. In this paper, we focus on predictive models that are used to optimize resource allocation, particularly on binary classification models that predict positive-outcome scores for this purpose. Such models are widely used across various domains, including marketing, where retention incentives are offered based on churn propensity scores [276], [277]; credit risk management, where loans are granted based on default risk assessments [278]; and fraud detection, where investigative resources are allocated based on predicted fraud risk [279]. These models can exhibit discriminatory behavior toward specific demographic groups through their predicted scores—that is, the predicted scores may follow different distributions across demographic groups—potentially leading to unfair resource allocations. Such discriminatory behavior can originate from biases in the historical data that is used to train the models or from inherent differences between groups in their tendency to belong to the positive class, which—although statistically justified—may be considered unacceptable discrimination when acted upon.

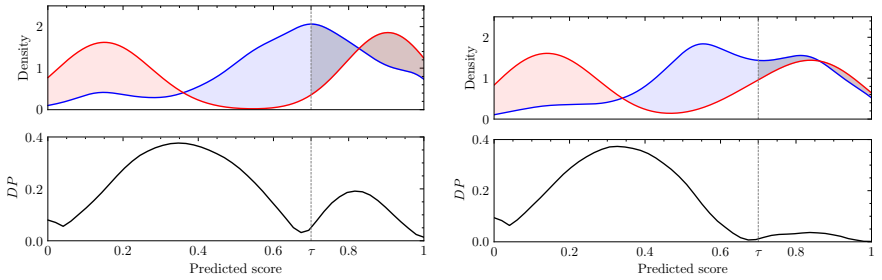
Fairness is central to the acceptability of algorithm-informed decisions, particularly in domains where these decisions can significantly affect individuals' access to resources or opportunities [280]. While much of the existing research has focused on preventing gender-based discrimination in pricing to comply with regulatory standards [281]–[283], the relevance of fairness extends well beyond pricing models to key business functions such as credit risk assessment, targeted marketing, and fraud detection. Fairness is closely tied to principles embedded in non-discrimination laws, particularly in the EU and US, which emphasize equitable outcomes across demographic groups and provide a foundation for fairness criteria like demographic parity [284]. Similar to the pricing context, discrimination in credit risk management [285], [286] and fraud detection is legally prohibited (e.g., to prevent gender-based discrimination and ethnic profiling). Additionally, in marketing, fairness considerations are seen by some companies as crucial for building and maintaining both a diverse customer base and a positive reputation.

Traditional approaches to evaluating algorithmic fairness typically rely on output-based metrics that assess disparities in average predictions or error rates between demographic groups [21]. These approaches are built on the idea that protected attributes, such as gender or race, should not impact a predicted score (e.g., a customer's default risk score). Although such metrics provide a general view, they often overlook more subtle nuances in algorithmic behavior [287]. Recent work has expanded fairness analyses by incorporating higher-order moments of the output distribution, such as the

variance [288], and by comparing entire output distributions [289]. More specifically, Han, Jiang, Jin, *et al.* [289] propose distribution-level variants of demographic parity, a fairness metric that compares the proportions of positive class predictions—the predictions on which we would potentially act in a resource allocation setting—across groups. While Han, Jiang, Jin, *et al.* [289] focus on model evaluation from a fairness perspective, Peeperkorn and De Vos [290] show that these distribution-level fairness notions can be used to develop predictive models that are intrinsically more fair. Building on these works, we propose a pragmatic *decision-centric fairness approach* to classification for resource allocation optimization. Specifically, rather than focusing on inducing demographic parity across the entire output distribution (i.e., ensuring a proportionally equal number of positive outcomes at all possible decision thresholds on the predicted scores), which we term a *global fairness approach*, we propose to concentrate on the "decision-making region" by inducing parity only within the range of relevant thresholds used for resource allocation, as visualized in Figure 6.1. For example, in a customer retention campaign, interventions are typically targeted at customers with a high predicted churn propensity score. Since resource allocation decisions affect only instances within this high-risk-score region, fairness should be enforced within this region, across all relevant application-dependent thresholds. The proposed decision-centric approach aims to ensure fairness where it matters, while achieving better predictions compared to a global fairness approach. The latter imposes overly strict constraints on the model to enforce fairness also outside the decision-making region, where it will not affect real-world decisions, potentially degrading the predictive quality of the generated scores more than necessary.

While achieving fairness in terms of demographic parity is straightforward using a group-dependent decision threshold post-hoc when decisions are made in batch, optimizing classification models to be inherently more fair without focusing on a single decision threshold is useful in common online decision-making settings (i.e., on a continuous basis). In such settings, resource constraints—such as marketing budgets, loan-granting capacity, or investigative resources—may change over time [291], causing the decision threshold to vary within a certain decision-making region. In these cases, setting group-dependent decision thresholds post hoc to achieve parity is not possible, and retraining the model each time a new threshold (within the decision-making region) is adopted due to changing resource constraints can be (too) costly.

Our main contributions are as follows: (i) we introduce and formalize the concept of decision-centric fairness for resource allocation optimization; (ii) we propose a decision-centric fairness approach to optimize classification models that are used for resource allocation; (iii) we propose a decision-



(a) No fairness induction (b) Induced decision-centric fairness

Figure 6.1: Densities of predicted scores $\tilde{y}$ for two demographic groups (with protected attributes $s = 0$ and $s = 1$, in blue and red, respectively), along with the corresponding demographic parity (DP) across all possible thresholds. By inducing decision-centric fairness, we aim to achieve demographic parity in the decision-making region, i.e., where $\tilde{y} > \tau$, to ensure a proportionally equal number of positive outcomes across the two groups at all thresholds within this region.

centric predictive performance metric for classification models; and (iv) we empirically compare our proposed decision-centric fairness methodology to a global fairness approach on multiple (semi-synthetic) datasets to identify scenarios where, from a decision-centric evaluation perspective, focusing on fairness only where it truly matters, outperforms imposing fairness globally across the output domain.

The remainder of this paper is structured as follows. In Section 6.2, we formalize the problem setting and discuss common fairness metrics and related work. Our proposed decision-centric fairness optimization methodology is presented in Section 6.3. Section 6.4 outlines the experimental design, with results presented and discussed in Section 6.5. Finally, Section 6.6 presents our conclusions and outlines possible future research directions.

6.2 Background and related work

In this section, we formalize the problem setting, provide a discussion of common fairness metrics, and motivate the use of demographic parity in business decisions.

6.2.1 Classification and resource allocation

Suppose that we have a data set $\mathcal{D} = \{(\boldsymbol{x_i}, y_i, s_i)\}_{i=1}^{N}$ with N the number of instances with each a feature vector $\boldsymbol{x}_i \in \mathbb{R}^k$, a binary response variable $y_i \in \{0, 1\}$, and a binary protected attribute $s_i \in \{0, 1\}$. Let $f : \mathbb{R}^{k+1} \to [0, 1]$ be a model that maps instances to a score $\tilde{y} \in [0, 1]$ which allows ranking instances from low to high score for belonging to the positive class. A predicted class $\hat{y} \in \{0, 1\}$ is obtained by setting a decision threshold $\tau \in [0, 1]$. Instances with a predicted score below the threshold, $\tilde{y} < \tau$, are assigned to the negative class, while those with a score above the threshold are assigned to the positive class. In resource allocation applications, resources are allocated to instances that are classified in the positive class based on the predicted score and the adopted threshold. In practice, the decision threshold is often determined post-training, as resource constraints that affect the threshold may only be known at runtime (e.g., the available investigative capacity for fraud detection), and/or because the threshold is to be optimized at runtime (e.g., to maximize the profitability of a retention campaign by taking into account the customer lifetime value of customers that are classified as churners [277]).

6.2.2 Fairness notions

We focus on situations in which algorithmic fairness issues may arise as a result of using model scores in combination with a (variable) decision threshold τ to decide whether or not resources are allocated [292]. Specifically, we consider settings where an action (e.g., offering a retention incentive) is taken for instances with a model score $\tilde{y} \geq \tau$ [293]. A standard approach to determine whether a classification model conforms to a given notion of fairness is to evaluate its outcome distribution with respect to a number of protected attributes, such as gender or race [294]. However, the current literature on algorithmic fairness offers a wide range of fairness notions and metrics [16], [21].

Direct vs. indirect discrimination Algorithmic discrimination comes in two types [16]: direct discrimination, where decisions are based on protected attributes, and indirect discrimination, where decisions hurt protected groups even without explicitly using protected attributes [295]. Direct discrimination is typically avoided by excluding protected attributes or perfectly correlated variables [296]. Therefore, this paper focuses on indirect discrimination. Even without using sensitive data directly, discriminatory behavior can occur through proxy features [297]. For example, a churn model might offer a retention incentive based on customer usage, but if that us-

age pattern is linked to a protected characteristic, the incentive may end up favoring one group over another [298].

Individual vs. group fairness Fairness metrics are typically categorized into two primary types [294]: individual fairness and group fairness. Individual fairness is based on the principle that similarly situated individuals should receive comparable outcomes [299], yet its practical application is hindered by the challenge of defining a robust, context-independent metric for similarity [300]. Alternative formulations of individual fairness are provided by causal and counterfactual fairness measures, which rely on explicitly modeling hypothetical scenarios at the individual level [301]. In contrast, group fairness evaluates statistical differences in outcomes across various demographic segments, facing the difficulty of selecting an appropriate metric—since even popular measures like demographic parity and equal opportunity are often mutually incompatible [302], [303]. In this paper, we focus exclusively on group fairness, as it is more prevalent in practice and its enforcement and measurement are comparatively more straightforward [304].

Group fairness metrics Group fairness evaluation methods typically translate philosophical or political fairness ideals into statistical parity metrics that apply to model outputs [305]. For instance, demographic parity in classification compares the proportions of positive class predictions ($\hat{y} = 1$) across groups. Alternatively, by incorporating ground truth labels (y), one can assess disparities in error rates—equality of opportunity, for example, compares true positive rates between groups [306]. These conventional output-based metrics focus on first-order statistics, evaluating average outcomes or error rates across groups [280]. Although these group fairness notions are widely used, many alternative definitions exist [21]. However, these output-based evaluations mainly rely on threshold-dependent first-order statistics and lack interpretability, potentially overlooking discriminatory behavior captured by higher-order moments or the broader prediction distribution [307], [308].

Distribution-level fairness metrics Recent work expands on traditional group fairness metric—which typically focus solely on first moments—by incorporating higher-order statistics, such as the variance [287], [288], or by comparing entire output distributions [289]. More specifically, Han, Jiang, Jin, *et al.* [289] argue that the standard demographic parity metric fails to detect unfairness due to its threshold dependency, i.e., a slight variation in the decision threshold may potentially re-introduce unfairness. Therefore, they propose two distribution-level variants of demographic parity:

- Area Between Probability density function Curves (ABPC):

$$\text{ABPC} = \int_0^1 |f_0(x) - f_1(x)| \, \mathrm{d}x, \tag{6.1}$$

  where $f_0(x)$ and $f_1(x)$ are the probability density functions (PDFs) of the predicted scores for two demographic groups, characterized by a protected attribute s, with $s = 0$ and $s = 1$, respectively.
- Area Between Cumulative density function Curves (ABCC):

$$\text{ABCC} = \int_0^1 |F_0(x) - F_1(x)| \, \mathrm{d}x, \tag{6.2}$$

  where $F_0(x)$ and $F_1(x)$ are the cumulative distribution functions (CDFs) of the predicted scores for two demographic groups, characterized by a protected attribute s, with $s = 0$ and $s = 1$, respectively.

While we agree that threshold-sensitivity is a challenging problem when implementing demographic parity in practice, we argue in Section 6.3 that the proposed ABPC and ABCC metrics are too rigid as they enforce demographic parity across the entire output distribution. This allows the use of decision thresholds across the entire output distribution, including regions where, in practice, the decision threshold would never be set due to resource constraints that limit the number of instances that can be acted upon (e.g., those that can be targeted with a retention campaign).

6.2.3 Demographic parity in resource allocation

Fairness in business operations is increasingly recognized as a fundamental concern [309]. When allocating resources based on predictive models, businesses must ensure that their decision-making processes adhere to well-defined fairness criteria [310]. However, as fairness in algorithmic decision-making is inherently multifaceted, choosing an appropriate fairness notion is non-trivial [19], [311]. Many widely-used fairness metrics, such as demographic parity and equal opportunity, are mutually incompatible except under highly constrained conditions [303]. This incompatibility is further exacerbated by deep-rooted philosophical disagreements on which notions are most appropriate in different contexts [312]. The choice between demographic parity and equal opportunity ultimately hinges on the evaluator's underlying assumptions and worldview [313], making it imperative to ground fairness considerations in regulatory and ethical frameworks that guide real-world business decisions.

In the European Union, non-discrimination legislation and fairness requirements stemming from hard and soft law sources are highly influential

in this context. For instance, the EU guidelines on discrimination in insurance explicitly require fairness in the access to and supply of goods and services [284]. Similarly, ensuring fairness in insurance pricing and preventing discriminatory effects in customer engagement strategies are also legal obligations [282]. This broader interpretation of fairness aligns closely with the principle of demographic parity, which seeks to equalize the likelihood of favorable outcomes across demographic groups, independent of underlying differences in target variable distributions and potential historical injustices. Beyond EU law, the concept of demographic parity also intersects with US anti-discrimination regulations, particularly in the context of disparate impact analysis [314]–[316]. Despite this apparent close connection between what is legally required and demographic parity, EU law does not foresee a specific implementation where fairness must be enforced across all potential model outputs [317]. In fact, if individuals fall outside the actionable range of scores regardless of whether a model has been globally or locally constrained, the legal impact remains unchanged when fairness is implemented locally. This suggests that constraints focused solely on the decision-making region could potentially fulfill legal fairness or non-discrimination objectives equally well as methods pursuing global demographic parity.

At the same time, pursuing global demographic parity often comes at the cost of predictive performance and calibration. Enforcing demographic parity usually involves imposing constraints on the model that can reduce predictive performance [318], [319]. Furthermore, demographic parity and calibration are fundamentally incompatible unless the base rates for the positive class are identical between the demographic groups [306]. These trade-offs are critical considerations for businesses that must balance fairness objectives with operational effectiveness and the reliability of predicted scores.

6.3 Decision-centric demographic parity

In this section, we first propose decision-centric variants of the distribution-level demographic parity metrics, ABPC and ABCC. We then outline how these metrics can be leveraged to induce decision-centric fairness in classification models used for resource allocation.

6.3.1 Evaluating decision-centric fairness

The two distribution-level variants of demographic parity proposed by Han, Jiang, Jin, *et al.* [289], introduced in Section 6.2.2, are threshold-independent. As discussed in Section 6.1, this is an important property, as predictive models are often deployed in online decision-making settings, in which resource

constraints may change over time [291], causing the decision threshold to vary. Additionally, the decision threshold may depend on instance-dependent costs and benefits [277]. In a resource allocation context, however, we argue that these distribution-level variants of demographic parity, which allow for the use of decision thresholds across the entire output distribution, are overly strict. Specifically, they penalize deviations from demographic parity even in regions with predictions that will never be acted upon in practice due to the aforementioned resource constraints. To address this, we propose decision-centric variants of the two distribution-level demographic parity metrics that only penalize unfairness within a relevant decision-making region:

$$\mathrm{ABPC}_\tau = \int_\tau^1 |f_0(x) - f_1(x)| \, \mathrm{d}x, \tag{6.3}$$

$$\mathrm{ABCC}_\tau = \int_\tau^1 |F_0(x) - F_1(x)| \, \mathrm{d}x. \tag{6.4}$$

That is, we adapt the previously defined ABPC (Equation (6.1)) and ABCC (Equation (6.2)) metrics by restricting their integration domains to the decision-making region $[\tau, 1]$, making ABPC and ABCC special cases of the more general proposed ABPC_τ and ABCC_τ metrics.

6.3.2 Inducing decision-centric fairness

The ABPC and ABCC distribution-level demographic parity metrics [289] allow us to *evaluate* algorithmic fairness across full score distributions. Building on this work, Peeperkorn and De Vos [290] operationalize a *global fairness approach*, which allows *inducing* algorithmic fairness across full score distributions, by training a neural classifier using a composite loss function that combines standard binary cross-entropy loss $\mathcal{L}_{\mathrm{BCE}}$ with an unfairness penalty $\mathcal{L}_{\mathrm{unfairness}}$ for deviations from demographic parity between the distributions of predicted scores for two groups defined by a protected attribute s:

$$\mathcal{L} = (1 - \lambda) \cdot \mathcal{L}_{\mathrm{BCE}} + \lambda \cdot \mathcal{L}_{\mathrm{unfairness}}, \tag{6.5}$$

where the hyperparameter λ controls the trade-off between predictive performance and fairness. They propose to use Integral Probability Metrics (IPMs) [320], notably the 1-Wasserstein distance[1], to quantify differences in predicted score distributions between demographic groups over the entire domain $[0, 1]$. However, in resource allocation applications, fairness constraints applied globally (i.e., over the entire output domain) can impose unnecessary

[1] Efficient methods are available to approximate the distance and its gradients, enabling its use in a neural network objective function, as it can be directly minimized using common frameworks for neural network training.

rigidity, aiming to enforce fairness also outside the decision-making region where it will not affect real-world decisions, potentially degrading the predictive quality of the generated scores more than necessary. Therefore, we propose a *decision-centric fairness approach* specifically tailored to induce fairness only within the decision-making region (i.e., over the domain $[\tau, 1]$), thereby focusing only on the predictions that are relevant for resource allocation.

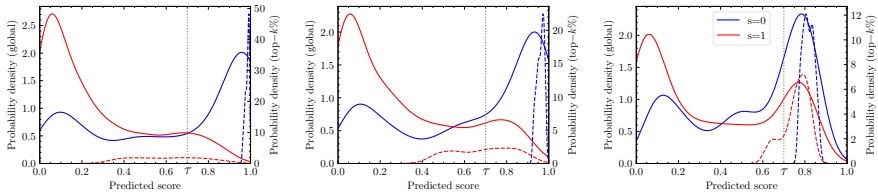
Our approach operationalizes this concept by applying the unfairness penalty exclusively to the top-$k\%$ of the predicted scores of each protected group:

$$\mathcal{L}_{\text{unfairness}} = \text{IPM}(\tilde{y}_0^{(k\%)}, \tilde{y}_1^{(k\%)}), \tag{6.6}$$

where $\tilde{y}_0^{(k\%)}$ and $\tilde{y}_1^{(k\%)}$ denote the distributions of the top-$k\%$ predicted scores for the protected groups with $s = 0$ and $s = 1$, respectively, and the IPM corresponds to the 1-Wasserstein distance [2]. To determine the percentile threshold $k\%$, we first train a baseline unconstrained model (with $\lambda = 0$) and set $k\%$ to match the proportion of instances with scores above the decision threshold τ on a validation set, irrespective of the protected attribute s. In this way, the unfairness penalty specifically focuses on actionable instances.

Conceptually, this means we alter the composition of the set of instances with $\tilde{y} \geq \tau$, as produced by a model without any penalty for unfairness, in order to induce greater fairness in the decision-centric demographic parity sense. This effect is illustrated in Figure 6.2, which shows how different values of the hyperparameter λ impact the predicted score distributions for two demographic groups ($s = 0$ and $s = 1$) on a test set after training a model for 30 epochs. Panel 6.2a shows the obtained score distributions for $\lambda = 0$, i.e., a model optimized solely for classification error. If only instances in the decision-making region (i.e., those with a predicted score $\tilde{y} \geq \tau = 0.7$) are acted upon, this model would result in clearly unfair resource allocation. Panels 6.2b and 6.2c, in contrast, induce decision-centric fairness with increasing strength, resulting in increasingly fair score distributions in the decision-making region (as reflected by the increasing overlap between the top-$k\%$ score distributions across demographic groups), potentially at the cost of predictive performance.

[2] An alternative, more directly aligned with the goal of inducing fairness in the decision-making region $[\tau, 1]$, is a quantile-based approach that characterizes the differences between the score distributions in a discretized manner using a histogram-based [321] unfairness penalty. Although this allows for the direct use of the decision threshold τ—which defines the relevant decision-making region—in optimization, training proved unstable in preliminary experiments, as even very small values of λ led to all predicted scores $\tilde{y}$ falling below τ. The percentile-based approach prevents this downward biasing of predicted scores.



(a) $\lambda = 0.0$ (b) $\lambda = 0.3$ (c) $\lambda = 0.6$

Figure 6.2: The decision-centric fairness approach for different values of λ. Score distributions (PDFs) on a test set after training each model for 30 epochs are shown, split by the protected attribute ($s = 0$ and $s = 1$). Solid lines represent the distributions of all predicted scores, while dotted lines represent the top-k% score distributions. The decision threshold τ is shown as a vertical line. An animated version of these plots is available in the GitHub repository.

6.4 Experimental design

This section provides a detailed overview of the experimental design. We first describe the datasets used and the process of creating semi-synthetic data to introduce (additional) bias into the historical data, providing a way to control the level of discriminatory behavior when left unmitigated. This allows us to test the sensitivity of the different fairness induction methods to varying levels of bias in the data used to train classification models for resource allocation. Next, we introduce and motivate the use of a decision-centric and threshold-independent measure based on the precision-recall curve to assess predictive performance, complementing the fairness evaluation using the decision-centric fairness metrics introduced in Section 6.3.1. Finally, we provide details on the problem configuration setups and the hyperparameter combinations tested. The code for reproducing the experiments is publicly available at `https://github.com/SimonDeVos/DCF`.

6.4.1 Data

We use three datasets for our experiments: *TelecomKaggle* (public), *Churn* (proprietary), and *Adult* (public). The first two datasets are directly relevant in the context of resource allocation, as they involve predicting churn propensity scores, whereas the last one is included because it is a standard

dataset in the algorithmic fairness literature[3]. An overview of dataset characteristics is provided in Table 6.1.

The need for fairness induction arises when biases are present in the historical data used to train classification models, or when there are inherent differences between groups in their tendency to belong to the positive class, which—although statistically justified—are considered unacceptable discrimination when acted upon. Specifically, if the data itself is unbiased and there are no inherent differences between groups, a classification model optimized solely for classification error will naturally produce fair predicted scores, and consequently, a fair resource allocation. To test the sensitivity of the fairness induction methods discussed in Section 6.3.2, we control for the first issue mentioned above by varying the levels of discriminatory bias in the historical data for the *TelecomKaggle* dataset. Based on the raw *TelecomKaggle* dataset, we create three semi-synthetic datasets to introduce (additional) bias into the historical data. This bias is systematically introduced through *informed label flipping*. Specifically, within one protected group, we selectively flip ground-truth outcome labels from 0 to 1. The details of this procedure are outlined in Algorithm 4 in Appendix D.1. Figures D.1–D.3 display the score distributions obtained using no fairness induction, i.e., reflecting the baseline discriminatory behavior present in the datasets. Figure D.1 illustrates the effect of introducing additional bias through informed label flipping. As the bias increases, the violation of demographic parity becomes more pronounced—both globally across the entire domain $[0, 1]$ and within the decision-making region $[\tau, 1]$.

After introducing bias through informed label flipping (where applicable), each dataset is split into training, validation, and test sets using a fixed ratio of 0.34/0.33/0.33. The training set is used to learn model parameters, the validation set is used for hyperparameter tuning (i.e., selecting the configuration with the lowest validation loss for $\lambda = 0$), and the test set is used for final model evaluation. A relatively large portion of the data is allocated to the validation and test sets to ensure sufficient resolution and stability in assessing performance and fairness metrics, particularly when analyzing decision-centric results across subgroups and thresholds.

6.4.2 Evaluation metrics

Our evaluation examines both predictive performance and fairness through a decision-centric lens, enabling the assessment of how the deployment of the models would impact resource allocation.

[3]Note that it is difficult to find public datasets in the fields of credit risk management and fraud detection, as in these domains *fairness through unawareness* is often used to comply with existing regulations; hence, no information on potential protected attributes is (publicly) available in those datasets [322], [323].

Table 6.1: Overview of dataset characteristics. By introducing additional bias in *TelecomKaggle*, we increase the ground-truth class imbalance within the protected group $s = 1$. For an overview of the score distributions obtained without fairness induction, i.e., reflecting the baseline discriminatory behavior present in the datasets, we refer to Figures D.1–D.3 in Appendix D.1.2.

Dataset	Protected attribute s	# Vars.	# Obs.	Bias rate	$s = 0$		$s = 1$		Class balance	
					$y = 0$	$y = 1$	$y = 0$	$y = 1$	$y = 0$	$y = 1$
TelecomKaggle	'Sex'	39	7,032	0.25	0.27	0.22	0.37	0.13	0.64	0.36
				0.50	0.18	0.31	0.37	0.13	0.55	0.45
				0.75	0.09	0.40	0.37	0.13	0.46	0.54
Churn	'Sex'	10	44,942	-	0.38	0.27	0.18	0.17	0.56	0.44
Adult	'Sex'	14	32,561	-	0.46	0.20	0.29	0.04	0.76	0.24

To evaluate the predictive performance of a classification model in the context of online resource allocation with a dynamic decision threshold τ, driven by dynamic resource constraints (and/or instance-dependent costs and benefits), we need a threshold-independent evaluation metric. Moreover, this metric should focus on predictions within the decision-making region, as these are the predictions we will potentially act upon. Specifically, for predictions across the relevant decision thresholds $[\tau, 1]$, we aim to maximize both precision and recall. Focusing on precision alone can be misleading, as a model with fewer predictions for which $\tilde{y} \geq \tau$ may appear more precise simply because it predicts $\tilde{y} \geq \tau$ only for the most confident cases, disregarding recall. Conversely, focusing solely on recall may reward models that capture more true positives at the cost of a higher false positive rate, leading to wasted resources. Precision-recall (PR) curves address this trade-off by considering both metrics simultaneously for all possible thresholds. To restrict the thresholds of interest to those within the decision-making region, we introduce the decision-centric performance metric AUC-PR$_\tau$. This metric summarizes precision and recall by calculating the area under a partial PR curve, which only considers thresholds above τ. This ensures that the performance assessment incorporates the deployment of the classification model in combination with all decision thresholds within the decision-making region. Figure 6.3 illustrates how AUC-PR$_\tau$ is obtained.

In addition to evaluating decision-centric predictive performance via AUC-PR$_\tau$, we also evaluate the decision-centric fairness of the classification models. To this end, we use the ABPC$_\tau$ (Equation (6.3)) and ABCC$_\tau$ (Equation (6.4)) metrics introduced in Section 6.3.1.

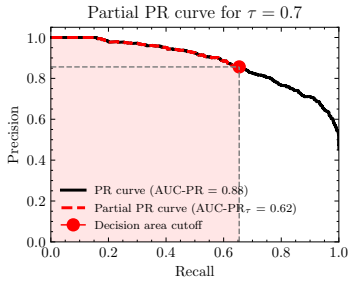


Figure 6.3: An example partial precision-recall curve and its corresponding AUC-PR$_\tau$, which summarizes precision and recall for all decision thresholds within the decision-making region $[\tau, 1]$. This region corresponds to the top-left segment of the PR curve starting at τ.

6.4.3 Problem and hyperparameter configurations

The most important hyperparameter in our experimental study is λ, as it controls the importance assigned to the unfairness penalty during training. Due to the predictive performance-fairness trade-off (see Section 6.2.3), however, objectively tuning this hyperparameter is often not possible. Therefore, we train neural classification models for $\lambda \in \{0, 0.05, 0.10, \ldots, 0.95\}$ for both the global and decision-centric fairness approaches. These models will be assessed using the concept of Pareto-optimality, as explained in more detail in Section 6.5. To determine all other hyperparameters, tuning is performed for the model with $\lambda = 0$ on each dataset. The hyperparameter configuration that results in the lowest validation loss is selected, and the same configuration is subsequently used for all values of λ. To quantify dissimilarities between score distributions for different demographic groups, i.e., for $\mathcal{L}_{\text{unfairness}}$, we use the implementation by Shalit, Johansson, and Sontag [320] to approximate the 1-Wasserstein distance IPM [324] (and its gradients) using Sinkhorn distances [325], [326]. For further details, we refer to Appendix B.1 of Shalit, Johansson, and Sontag [320]. Additional information on the model architecture, implementation, training, and hyperparameter tuning is provided in Appendix D.2.

In addition to testing the sensitivity of the different fairness induction methods to varying levels of discriminatory bias in the data (via the creation of semi-synthetic datasets with additional bias introduced; see Section 6.4.1), we also assess their sensitivity to the size of the decision-making region by varying the decision threshold τ.

6.5 Results and discussion

In this section, we present our results by comparing the Pareto fronts of the global and decision-centric fairness induction approaches. Specifically, we visualize the trade-offs achieved between decision-centric predictive performance—measured by AUC-PR$_\tau$ (higher is better)—and decision-centric fairness—measured by ABPC$_\tau$ or ABCC$_\tau$ (lower is better)—for the different values of λ; and we construct Pareto fronts for both the global and decision-centric approaches by identifying all models that represent optimal trade-offs under Pareto-optimality (i.e., models for which no objective can be improved without worsening the other). The model without fairness induction (i.e., $\lambda = 0$) is also visualized and marked with a red star.

We structure our presentation and discussion of the results around the following three key questions:

- **Q1**: What is the impact of adopting a decision-centric versus a global approach to inducing fairness in the decision-making region on predictive performance?
- **Q2**: How do the size of the decision-making region and varying levels of discriminatory bias in the historical data affect the differences in predictive performance between a decision-centric versus a global fairness induction approach?
- **Q3**: Which metric is most appropriate for evaluating decision-centric fairness, and how should a model be selected for deployment?

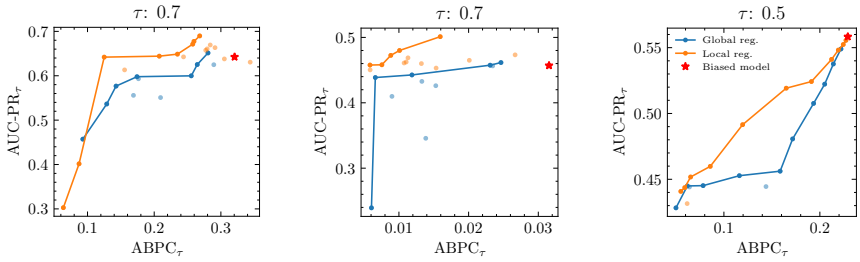
To address these questions, we selectively highlight relevant results. A comprehensive set of results is presented in Appendix D.3.

6.5.1 Q1: Impact of decision-centric versus global fairness approach on predictive performance

We first investigate how decision-centric fairness optimization compares to the global approach by evaluating models trained with varying values of λ. Figures 6.4 and 6.5 show the results for three datasets using ABPC$_\tau$ and ABCC$_\tau$ as the fairness metric, respectively.

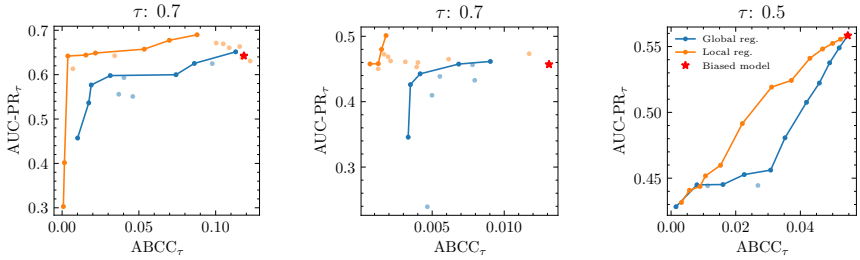
The plots clearly illustrate the advantage of our proposed decision-centric method (orange curves) over the global approach (blue curves). As λ increases, model performance moves along a trade-off curve, prioritizing fairness at the expense of predictive performance. However, the decision-centric approach consistently delivers superior trade-offs, underscoring its effectiveness in balancing fairness with predictive performance specifically within the decision-making region.

The trade-off varies by dataset. For example, in the *TelecomKaggle* and *Churn* datasets, fairness can be substantially improved with relatively minor



(a) *TelecomKaggle* (bias rate 0.5) (b) *Churn* (c) *Adult*

Figure 6.4: Pareto fronts illustrating the trade-off between predictive performance and fairness (AUC-PR$_\tau$ vs. ABPC$_\tau$) for the decision-centric (orange) and global (blue) fairness induction approaches for three datasets.



(a) *TelecomKaggle* (bias rate 0.5) (b) *Churn* (c) *Adult*

Figure 6.5: Pareto fronts illustrating the trade-off between predictive performance and fairness (AUC-PR$_\tau$ vs. ABCC$_\tau$) for the decision-centric (orange) and global (blue) fairness induction approaches for three datasets.

reductions in predictive performance. Conversely, for the *Adult* dataset, even a slight increase in λ leads to a drop in predictive performance, though the decision-centric approach still yields a more favorable trade-off than the global method.

Moreover, adding an unfairness penalty—whether decision-centric or global—can further enhance both fairness and predictive performance. This effect is visible in the plots, where certain models lie higher and further to the left than the model without unfairness penalty (i.e., with $\lambda = 0$) marked with a red star. This shows that the unfairness penalty, in some cases, also serves as a regularization mechanism.

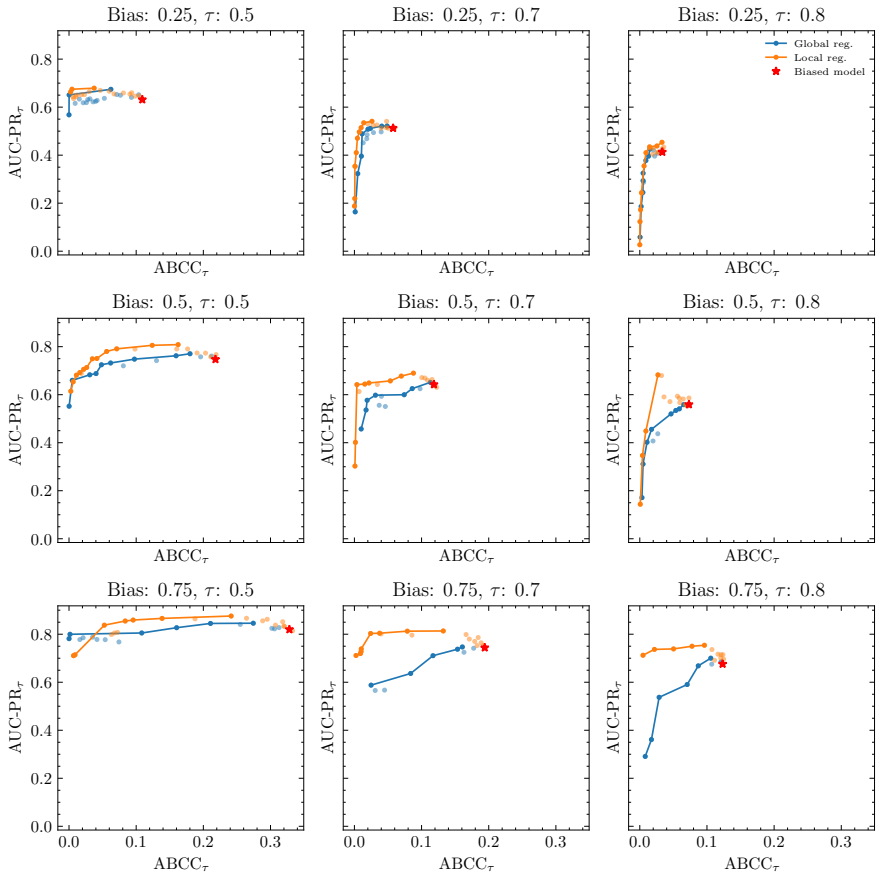


Figure 6.6: Results on the *TelecomKaggle* dataset for different bias rates, with decision-centric fairness measured by ABCC$_\tau$. The figure illustrates the effect of a varying size of the decision-making region with $\tau = 0.5, 0.7, 0.8$ (columns) on the decision-centric fairness-predictive performance trade-off. The orange and blue lines represent decision-centric and global fairness induction, respectively, while the model without unfairness penalty (i.e., with $\lambda = 0$) is marked with a red star.

6.5.2 Q2: Impact of decision-making region size and level of discriminatory bias in historical data

We analyze the influence of varying the decision threshold τ and semi-synthetic bias rates on model performance, both in terms of predictive performance and fairness, using the *TelecomKaggle* dataset. Using the informed

label flipping method to generate semi-synthetic datasets with varying bias levels allows us to systematically investigate the impact of bias on model outcomes. Results are shown in Figure 6.6, where rows represent different bias rates (0.25, 0.50, 0.75) and columns represent different decision thresholds (τ values of 0.5, 0.7, and 0.8), with fairness measured by ABCC_τ. For fairness results in terms of ABPC_τ, see Figure D.4 in the appendix.

The models without unfairness penalization (i.e., with $\lambda = 0$, indicated by red stars) illustrate how fairness deteriorates as either the bias increases or the decision-making region becomes larger. Specifically, for a fixed τ, higher semi-synthetic bias consistently reduces fairness, evidenced by a rightward shift of baseline points. Similarly, for a fixed bias rate, a larger decision-making region (lower τ) also negatively impacts fairness. This is because more predictions are included in the decision-making region and thus affect decision-centric fairness calculations. Consequently, the best baseline decision-centric fairness is observed with minimal bias and a higher τ, while the worst-case scenario arises from a combination of high bias and a large decision-making region.

In comparing global (blue) and decision-centric (orange) approaches, we observe similar performance for the low bias rate (0.25), where fairness induction has a minimal potential impact as there is little bias present to be eliminated. However, for the higher bias rates (0.50 and 0.75), the decision-centric approach increasingly outperforms the global approach, demonstrating its effectiveness. The advantage of the decision-centric approach over the global approach becomes more pronounced as the decision threshold τ increases, resulting in a smaller decision-making region. Conversely, in the extreme case where $\tau = 0$, both decision-centric and global fairness induction methods coincide due to the decision-making region covering the entire prediction domain.

For results on varying the decision threshold τ for the *Churn* and *Adult* datasets, we refer to Figures D.5–D.8 in the appendix. For the *Adult* dataset, in addition to τ values of 0.5 and 0.8, we also include $\tau = 0.4$ because of the class imbalance, which leads to fewer predictions in the decision-making region.

6.5.3 Q3: Impact of decision-centric fairness metric used for evaluation and model selection

The ABCC metric directly evaluates the 1-Wasserstein distance between the score distributions of different demographic groups [289], and as such, it aligns closely with our implementation of $\mathcal{L}_{\text{unfairness}}$ for the global approach. A similar alignment exists between the ABCC_τ metric and the decision-centric fairness approach, which may (partly) explain the larger differences

observed between the Pareto fronts of the global and decision-centric approaches for the *TelecomKaggle* and *Churn* datasets (compare the plots in Figures 6.4 and 6.5).

Beyond the optimization-evaluation alignment, while the PDF-based ABPC_τ metric provides an intuitive way to quantify decision-centric fairness (see Figure 6.1), the ABCC_τ metric is also better suited to the underlying problem of using predicted scores for resource allocation, as it is sensitive to the distance that probability mass must move to align the score distributions across groups—capturing not only the existence but also the magnitude of score shifts. In contrast, ABPC (and therefore ABPC_τ) only quantifies how much probability mass is placed differently between groups, but is invariant to how far these differences are from each other. For instance, if two distributions differ only in localized regions—say, around 0.75 and 0.9 in one scenario, versus around 0.85 and 0.9 in another (with the density being lower around one value and higher around the other)—then ABPC_τ with $\tau < 0.7$ would yield the same value for both. However, ABCC_τ would favor the latter scenario. Since the likelihood of these differences—and thus the algorithmic unfairness—being canceled out through thresholding (with the decision threshold varying between τ and 1) increases when the distances between them are smaller, ABCC_τ is preferable for evaluating decision-centric fairness in resource allocation contexts.

Hence, to select a classification model for deployment in the context of resource allocation optimization, we advise decision-makers to choose a model (which, in our experiments, corresponds to selecting a fairness induction strategy together with a value for λ) that offers a good trade-off between ABCC_τ and AUC-PR_τ. Our results show that it is often possible to substantially reduce ABCC_τ through decision-centric fairness induction (and in some cases also through global fairness induction), without compromising AUC-PR_τ compared to a model without fairness induction. As such, when fairness is not mandated by strict regulatory requirements, these models are of particular interest. Performance in terms of ABPC_τ can be used to further assess a (small) subset of candidate models in a complementary manner.

6.6 Conclusion

This paper introduces and formalizes the concept of decision-centric fairness in classification models used for online resource allocation optimization with dynamic resource constraints. This novel approach aligns algorithmic fairness considerations with their potential impact on real-world resource allocation decisions, such as the fairness of credit risk scores used for loan approvals or churn propensity scores used in targeted marketing retention campaigns, while accommodating dynamic decision thresholds. Decision-centric fair-

ness redefines fairness evaluation and induction by shifting the focus from full classification score distributions to considering only the decision-making region, which includes only the range of relevant decision thresholds for a given resource allocation problem. This ensures that fairness constraints are considered and/or enforced only where they will impact real-world decisions. We propose a method to optimize classification models directly for decision-centric fairness and demonstrate that by avoiding overly strict fairness constraints, we can minimize the degradation of the predictive quality of the generated scores. Specifically, we empirically show that our decision-centric fairness methodology often leads to better decision-centric predictive performance-fairness trade-offs compared to a global fairness approach, which applies fairness constraints more naively across the full score distributions.

For certain resource allocation problems, such as targeted retention campaigns based on churn propensity scores, a causal uplift modeling approach has been shown to outperform the predictive modeling approach adopted in this work, leading to improved profitability of retention campaigns [327]. However, the predictive modeling approach remains valuable in practice, as it allows the use of a single customer churn prediction model across various retention campaigns—something that is not possible with the treatment-dependent uplift modeling approach. Nevertheless, adapting our proposed decision-centric fairness approach to optimize uplift scores appears to be a promising direction for future research. Exploring this in the context of cost-sensitive uplift modeling [328], which leverages class-dependent or instance-dependent costs and benefits to improve decision-making, also seems worthwhile, though potentially more challenging. As a first step, incorporating decision-centric fairness into the cost-sensitive predictive modeling approach [277] could serve as a starting point. In a similar vein to Devriendt, Berrevoets, and Verbeke [327], Vanderschueren, Baesens, Verdonck, *et al.* [329] recently demonstrated that framing resource allocation problems with stochastic resource constraints as a ranking problem, and subsequently relying on learning-to-rank techniques instead of classification techniques, leads to improved decision-making outcomes. This ranking approach naturally applies to all resource allocation problems (in contrast to an uplift modeling approach). Hence, incorporating a decision-centric fairness perspective into the optimization of learning-to-rank models appears to be a promising direction for future research as well. Translating such fairness-enhanced learning-to-rank models to the uplift modeling setting, building on Devriendt, Van Belle, Guns, *et al.* [330], would help complete the picture.

Beyond these methodological extensions, our work opens several additional avenues for future research concerning the fairness perspective. First, the fairness metrics used in this study center on a single group fairness no-

tion: decision-centric demographic parity. While practical and grounded in regulatory logic [284], [314], group-based metrics have notable limitations as they mask disparities at the individual level. Moreover, there might be adverse effects of reranking individuals (across or within sensitive groups), which may undermine the perceived validity of fairness [331]. Future research could examine alternative or complementary notions of fairness, such as equal opportunity [280] or counterfactual fairness [301]. Second, we adopt a percentile-based approach to operationalize decision-centric fairness induction—aligning the top$-k\%$ of each group's score distribution. This approach, however, remains a proxy for the ideal case where fairness is enforced strictly within the decision-making region. As discussed in Section 6.3.2, a quantile-based method that directly compares and penalizes disparities in score distributions above a threshold τ would align more closely with the formalization of decision-centric fairness. However, our initial experiments with such quantile-based implementations revealed training instability. Therefore, investigating how to efficiently and robustly implement such a quantile-based approach is left for future research. Next, although not yet operationalized, decision-centric fairness could naturally extend to multiple or intersecting protected attributes, suggesting a direction for future work [16], [302]. Finally, further strengthening the formal connection between decision-centric fairness and legal interpretations of *actionability* in anti-discrimination law [314] would be a valuable step toward ensuring legal as well as practical soundness. Bridging the conceptual gap between legal standards of fairness and the integration of fairness considerations into algorithm evaluation and design remains a critical step in operationalizing fairness.

7
Uplift modeling with continuous treatments: A predict-then-optimize approach

The goal of uplift modeling is to recommend actions that optimize specific outcomes by determining which entities should receive treatment. One common approach involves two steps: first, an inference step that estimates conditional average treatment effects (CATEs), and second, an optimization step that ranks entities based on their CATE values and assigns treatment to the top k within a given budget. While uplift modeling typically focuses on binary treatments, many real-world applications are characterized by continuous-valued treatments, i.e., a treatment dose. This paper presents a predict-then-optimize framework for uplift modeling with continuous treatments. First, in the inference step, conditional average dose responses (CADRs) are estimated from data using causal machine learning techniques. Second, in the optimization step, we frame the assignment task of continuous treatments as a dose-allocation problem and solve it using integer linear programming (ILP). This approach allows decision-makers to efficiently and effectively allocate treatment doses while balancing resource availability, with the possibility of adding extra constraints like fairness considerations or adapting the objective function to take into account instance-dependent costs and benefits to maximize utility. The experiments compare several CADR estimators and illustrate the trade-offs between policy value and fairness, as well as the impact of an adapted objective function. This demonstrates the framework's advantages and flexibility across diverse applications in healthcare, lending, and human resource management. All code is available on https://github.com/SimonDeVos/UMCT.

7.1 Introduction

In many applications, decision-makers are interested in learning the causal effects of a treatment on a particular outcome [332], [333]. A crucial aspect is response heterogeneity, where entities respond differently to treatments based on their characteristics [334]. In intervention-based decision-making, precisely this heterogeneity is valuable to leverage for finding a treatment assignment policy that optimizes a specific outcome variable, based on historical data, while adhering to certain constraints such as scarce resources. Uplift modeling (UM) is a set of techniques to find such policies [67]. An often-used predict-then-optimize UM approach combines conditional average treatment effect (CATE) estimation with an optimization step, where CATE estimation can be considered an inference step as part of a larger decision-making process [335].

Many real-world applications, as illustrated by Table 7.1, contain complexities beyond binary treatments and budget as the sole constraint and are better characterized by a treatment dose, i.e., where the intervention can be applied across a continuous range of values [336], [337]. Moreover, continuous treatments have benefits for making treatment allocation possibly more effective and efficient compared to their binary counterpart because the marginal utility of treatment can vary with different dose levels, and the treatments can be allocated on fine granularity.

Table 7.1: This table displays exemplary applications of a UM setting with continuous-valued treatments and their corresponding details. *Cost-sensitivity* can either refer to outcome benefits *(o)* or treatment costs *(t)* and *Constraint* to budget *(b)* or fairness *(f)*. The examples provided are illustrative and not intended to be exhaustive or fully comprehensive.

Application	**Outcome**	**Cont. Treatment**	**Cost-sensitivity**	**Constraint**
Credit	Default rate	Interest rate	Loss given default *(o)*	Regul. compliance *(f)*
Healthcare	Sickness	Medication dose	Medication price *(t)*	Equal access *(f)*
HR	Employee retention	Training hours	Hourly opportunity cost *(t)*	Instructors *(b)*
Maintenance	Machine up-time	Maintenance freq.	Machine criticality *(o)*	Spare parts avail. *(b)*

While there is a growing body of literature focusing on conditional average dose response (CADR) estimation (i.e., the continuous-valued counterpart of CATE), this line of work focuses on causal inference and finding optimal doses, largely ignoring the role of continuous treatments in constrained decision-making contexts [338]–[340]. Constraints play a vital role in embedding business requirements into decision-making, which we accomplish using an integer linear programming (ILP) formulation. While many requirements exist, fairness stands out as a key consideration and is the main exemplary constraint throughout this work [341]. Though fairness constraints are not

new in machine learning for decision-making, they are often embedded as soft constraints during model training [342], [343]. This approach, however, is limited in flexibility and modularity. By shifting the inclusion of fairness considerations to a post-processing step, i.e., after model training, decision-making is more flexible toward dynamic business requirements.

Therefore, in this paper, we develop a UM framework with continuous treatments. Our work addresses two key gaps in the current literature. First, UM has not been formally defined, with the distinction between treatment effect estimation and allocation optimization largely overlooked. Second, as far as we are aware, there are no established methods in the literature extending UM to handle (i) continuous treatments, let alone methods that manage the combined complexities of also including (ii) extra constraints (like fairness) and (iii) objective function alteration (to, for example, account for cost-sensitivity).

By addressing these gaps, our contributions are fourfold:

1. We provide a framework that clearly defines UM, and how this differentiates from a mere causal inference problem, i.e., treatment effect estimation.
2. We extend UM methods to effectively handle continuous treatments by first leveraging state-of-the-art causal ML methods for CADR estimation and then defining the optimization part as a dose-allocation problem. The ILP formulation offers flexibility to adapt the objective function (e.g., to incorporate cost-sensitivity) and to include additional constraints (e.g., to enforce fairness considerations).
3. We are the first to include fairness considerations in a UM setting as explicit constraints in an ILP formulation. By extension, to the best of our knowledge, this is also the first application of such constraints in decision-making pipelines that utilize ML predictions as input.
4. With a series of experiments, we show the capabilities of our framework and demonstrate how incorporating fairness constraints or cost-sensitive objectives influences policy outcomes, crucial for applications such as those highlighted in Table 7.1.

The remainder of this paper is structured as follows. Section 7.2 defines UM and elaborates on related literature. Section 7.3 introduces the problem formulation and establishes the necessary notation. Section 7.4 focuses on the methodology, explaining the predict-then-optimize approach, the predictive models used to estimate CADRs, and the optimization techniques employed for constrained treatment allocation. Section 7.5 details the experimental setup, including data, evaluation metrics, and the experiments, followed by a discussion of the results. Finally, Section 7.6 presents our conclusions, discusses limitations, and outlines potential further research.

7.2 Uplift modeling

In this section, we begin by discussing the purpose of UM, offering a clear definition and a detailed breakdown of its core elements. We then explore how UM can be extended to accommodate continuous treatments. Finally, we consider adjustments to UM, such as incorporating additional constraints or modifying the objective function, to better align with specific application requirements, reflecting its relevance to the domain of prescriptive analytics.

7.2.1 Purpose and definition

UM is a well-established set of methods in the field of personalized decision-making, closely aligned with the goals of prescriptive analytics [67]. Unlike traditional predictive models that identify entities likely to yield a desired outcome, UM distinguishes between baseline responders (entities likely to show positive outcomes even without intervention) and true responders that respond because of the treatment. Consequently, UM focuses on outcome changes directly attributable to the treatment assignment, rather than simply predicting positive outcome probabilities. This helps avoid suboptimal targeting and ensures assignment policies capture their true incremental impact. Traditional UM predominantly deals with binary treatments, where the main goal is to rank individuals based on their expected response to a treatment [344]. Works such as [67] provide a comprehensive overview, focusing on a setting with binary treatment effects, various modeling techniques, and the associated challenges. However, the literature often lacks a unified definition of UM, with some studies equating it to treatment effect inference [345] while others focus primarily on ranking methods [346]. To resolve this ambiguity, we propose the following UM definition:

Definition 1. Uplift modeling refers to the collection of methods where the task at hand is optimally allocating treatments, with the objective of maximizing the total benefit generated by these treatments, determined by the cumulative uplift over entities, under given constraints.

The above definition encompasses both the one-step approach (i.e., predict-and-optimize) and the two-step approach (i.e., predict-then-optimize). Notably, this definition is method-agnostic to keep it inclusive of decision-focused methods that learn policies directly and therefore do not output an explicit treatment-effect estimate. As Table 7.2 highlights, the methodological contribution of this work focuses on the two-step approach, where treatment effect estimation is a component of the broader UM framework, which also includes treatment allocation as a critical task.

Table 7.2: A schematic overview of the positioning of our work within various UM approaches. The focus of this paper is highlighted in grey. In the Decision-Focused Learning (DFL) paradigm, the goal is to directly include optimal treatment allocation in the model learning task (i.e., predict-and-optimize). Prediction-Focused Learning (PFL) consists of (i) an inference and (ii) an optimization step (i.e., predict-then-optimize).

Treatment	**Predict-and-optimize**	**Predict-then-optimize**
Binary	Learn to rank	*(i)* CATE estimation *(ii)* Treatment-allocation problem
Continuous	Learn to allocate	*(i)* CADR estimation *(ii)* Dose-allocation problem

7.2.2 Treatment effects

A treatment refers to an intervention or action that can be applied to an entity to influence a particular outcome. Treatments can be binary, where the action is *treat* or *do not treat* (e.g., offering or withholding discount). The CATE measures the expected difference in potential outcomes between treated and untreated groups, conditioned on entities' characteristics [320]. Additionally, recent developments have extended traditional UM to multi-treatment scenarios, which generalize the setup by considering more than two treatment options [347] (e.g., offering discounts through different channels), and sequential treatments, where multiple treatments can be applied over time, each potentially influencing the outcome effects of subsequent treatments [348]. However, these are not the focus of our paper and are considered future extensions.

The estimation of causal treatment effects from data is inherently challenging due to the fundamental problem of causal inference, namely, the absence of counterfactual outcomes [332]. In the case of observational data, this problem is exacerbated by the possible non-random assignment of treatments and the resulting confounding bias [333]. A real-life example is the overestimation of workplace wellness programs' impact on employee wellbeing, where non-random incentives cause treated populations to differ from the general population [349]. A naive approach may overfit on the self-selected group, missing the true causal effect. This highlights how confounding can lead to issues like Simpson's paradox, where aggregated data misrepresents trends compared to stratified data [349]. While randomized controlled trials (RCTs) are the gold standard for mitigating such biases [350], they are often impractical in real-world business settings because of cost, ethical, or operational reasons. Moreover, in the context of continuous treatments, with

virtually infinite possible doses, setting up an RCT for dose-response estimation becomes significantly more complex [337]. In contrast, observational data is often cheap and readily accessible. Therefore, methods that aim to balance the treatment and control groups have been proposed, including techniques like propensity score matching [351] or covariate balancing [320], [352].

In recent years, the consideration of continuous treatments—and consequently, CADR estimation — has gained traction because of its relevance in applications where understanding the effects of varying dose levels is critical (see Table 7.1). Estimating CADRs is inherently more challenging than CATEs because it involves a spectrum of treatment doses, rather than values 0 and 1 [339]. Also, with the theoretically infinite number of possible dose values, data sparsity and the non-uniformity of dose assignment could be challenging, making accurate effect estimation at each dose level difficult [353]. The Conditional Average Dose Effect (CADE), which can be directly derived from the CADR, generalizes the CATE for continuous treatments. Our paper specifically focuses on UM with this type of treatment. Traditionally, dose effects have been derived from average dose-response curves using methods like the Hirano-Imbens estimator [336], which extends propensity scores to continuous interventions through generalized propensity scores. However, these approaches often overlook individual-level heterogeneity in responses. Recent advancements in ML have introduced techniques for learning individualized dose-response curves [338]–[340], [354], [355]. However, analogously to the case of CATE estimation, these methods primarily focus on the inference step and do not incorporate an optimization step where the learned dose-response relationships are used to prescribe decisions [335].

In this paper, we do not make any assumptions about the shape of the CADR. Instead, we adopt a flexible, data-driven approach. In contrast, existing work, such as [356], relies on a strong structural assumption — the law of diminishing marginal utility — which implies that the marginal benefit of increasing the dose is strictly decreasing. While this assumption may hold in some applications, it limits model flexibility and generalizability. Our framework does not rely on such assumptions: the estimated dose-response relationships in our approach are non-parametric and can capture arbitrary, potentially non-monotonic shapes.

7.2.3 Allocation task

Cumulative uplift & total benefit as driver

The allocation task aims to optimize an objective value driven by (cumulative) uplift. While uplift applies to a single entity, cumulative uplift reflects the policy's overall impact. For a policy treating K entities, cumulative up-

lift is the sum of CATEs for discrete treatments or CADEs for continuous treatments across these treated entities. In certain applications, cumulative uplift directly aligns with the target objective, as seen in contexts where maximizing cumulative uplift is the explicit goal [67]. However, cumulative uplift itself does not necessarily represent the objective. For example, in value-driven analytics, where uplift might disregard expected profit, and in cases where each entity may be associated with an individual benefit and treatment cost [357]–[361]. Hence, when cumulative uplift is not the end goal in itself, allocations with lower cumulative uplift may yield a greater increase in objective value.

Predict-and-optimize

In UM, the task of allocating treatments can follow one of two paradigms, represented by the two columns in Table 7.2. On the one hand, there is the paradigm of decision-focused learning, where the optimal treatment allocation is directly learned in an integrated way [362], [363]. On the other hand, prediction-focused learning separates two distinct steps: First, an inference task, where treatment effects are estimated, followed by an optimization task that uses these estimates as input to allocate treatments [68].

In this paper, we adopt the predict-then-optimize paradigm, which is represented in the second column of Table 7.2. With this approach, we clearly distinguish the treatment estimation task as a component of the larger UM approach, which additionally includes an optimization step that takes estimated treatment effects as input [335]. In the case of binary treatments, the inference step involves estimating CATEs, followed by an optimization step, i.e., finding a policy by solving the treatment-allocation problem to determine which entities should receive treatment. Typically, this is done using a greedy ranking heuristic: entities are ranked based on their predicted CATEs, and treatments are assigned to the top-k entities until the budget is exhausted [364]. Alternatively, under a flexible budget, the optimal value of k is determined through cost-benefit analysis [360].

In contrast, the predict-and-optimize approach, or decision-focused learning (DFL), integrates prediction and optimization into a single end-to-end system [365]. This paradigm tailors predictions specifically for downstream objectives, using regret-based loss functions to align predictions with optimization goals, where $\mathbf{x}^\star(\mathbf{c})$ represents the optimal decision with full information, and $\mathbf{x}^\star(\hat{\mathbf{c}})$ is based on estimated parameters [366]:

$$\mathcal{L}_{regret} = f\left(\mathbf{x}^\star(\mathbf{c}), \mathbf{c}\right) - f\left(\mathbf{x}^\star(\hat{\mathbf{c}}), \mathbf{c}\right) \tag{7.1}$$

Although DFL approaches for UM exist [362], [363], we adopt the predict-then-optimize framework due to its simplicity, flexibility, stability, and generalizability to other treatment types. It allows using various predictive

models without being restricted by a specific optimization setup, which is especially useful for complex or pre-trained models without access to the training process [341]. Moreover, it enables flexible adjustments to the objective function and integration of additional constraints other than budget limits, such as fairness considerations, allowing for different allocations under varying budget conditions. Our approach requires modifying only the optimization step, leaving the prediction model unchanged and does not require a complete model retraining.

Allocation problem as ILP
When considering continuous treatments, the optimization step becomes a *dose-allocation problem*, where the task is to determine the optimal treatment dose for each entity. This generalizes the binary *treatment-allocation problem*, which can often be addressed by ranking heuristics. However, since for the *dose-allocation problem* the decision space expands, a greedy ranking heuristic is no longer feasible, and a more formal optimization approach is required.

Therefore, we formulate and solve the allocation problem as an ILP. To consider binary choices per dose level, we discretize the estimated CADRs into distinct dose options. This discretization is motivated by two factors: (i) it allows the ILP formulation and (ii) many practical applications only permit pre-defined dose levels with limited granularity [367]. While this discretization introduces some loss of generalization, its impact is minimal since the number of dose bins is flexible and only affects the optimization step without altering the prediction step. The number of dose bins can be freely adjusted, although more bins increase computational complexity.

Flexible optimization constraints and objectives
During the optimization step, constraints and objective functions can be flexibly adjusted. Any constraint compatible with ILP can be added, and the objective function can be adapted, as long as it remains a linear combination of decision variables, as required by the ILP. Fairness constraints and cost-sensitive objective functions are exemplary, and further modifications, such as operational constraints, can also be incorporated

Fairness as constraint The importance of fairness in algorithmic decision-making is well-recognized, especially as AI systems are increasingly deployed in high-stakes domains like criminal justice, hiring, and lending [293], [341], [368]. Algorithmic bias can lead to unfair outcomes, as demonstrated in notable cases like the COMPAS tool for recidivism prediction [369] and issues with Amazon's hiring algorithms, which were scrapped [17]. Group fairness concepts include independence, ensuring predictions are unaffected by sensitive attributes; separation, requiring conditional independence of predictions

given the outcome; and sufficiency, mandating conditional independence of the outcome given the prediction [341], [370]. The EU's AI Act underscores the regulatory focus on fairness, aiming to prevent AI from reinforcing discrimination through a risk-based framework [18]. Our framework's flexibility in setting fairness constraints enables its use across various risk categories, ensuring compliance with these requirements and balancing fairness with utility [18].

Balancing utility — such as profit and accuracy — against fairness, like equal treatment across demographics, remains a key challenge. [371] and [10] show that enhancing fairness often trades off with utility and that multiple fairness definitions may even conflict. Traditional fairness assessments emphasize output-based metrics, such as demographic parity or equal opportunity, which measure group-level outcome disparities [341]. Recent approaches consider the entire distribution of predictions or decisions, incorporating statistical moments and exploring distributional fairness [288], [342]. In decision-making, fairness extends beyond prediction to both allocation fairness—ensuring treatment levels are independent of sensitive attributes like race or gender—and outcome fairness, which ensures equitable results across groups. Although outcome predictions may be imperfect, enforcing fairness at the level of expected outcomes has been shown to be an effective and ethically justified approach in decision-making settings [10], [372]. While not ideal, these estimates still represent the best available proxy for assessing how different groups will benefit from the allocation policy. Fairness notions, which may conflict, can be incorporated as hard constraints during optimization [10], similar to this work's approach, or as soft constraints in multi-objective learning [341]. Recent research, including [343], has largely focused on fairness in observational data with binary sensitive attributes, while [373] apply causal inference to multi-stage decision-making under fairness constraints. However, these studies do not address fairness in the context of continuous treatments, a critical gap our research aims to fill.

Value-driven objective function In managerial decision-making, profit maximization or cost reduction is often the primary goal. Cost-sensitive or value-driven methods are crucial as they incorporate asymmetric costs and benefits, aligning decisions with these business objectives [14]. Cost-sensitive approaches seek to balance costs and benefits, an aspect often ignored by standard methods [357], [360]. In this work, costs and benefits are deterministic while assuming stochastic treatment effects, setting it apart from studies where both costs and benefits are modeled as stochastic [128], [374]. Cost-sensitive methods find applications across various domains, including credit risk [375], fraud detection [25], [376], customer churn [374], business failure prediction [377], and machine maintenance [378].

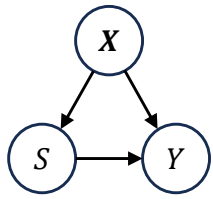


Figure 7.1: This DAG represents the assumed causal relationships between variables in the training data. $\boldsymbol{X}$: entity's pre-treatment features, S: treatment dose, Y: outcome.

7.3 Problem formulation

7.3.1 Notation

The dataset $\mathcal{D} = \{(\mathbf{x}_i, s_i, y_i)\}_{i=1}^{N}$ has N tuples of pre-treatment features $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^k$, treatment doses $S \in \mathcal{S} = [0, 1]$, and outcomes $Y \in \mathcal{Y} = [0, 1]$, where $\mathbf{x}_i$, s_i, and y_i denote the respective values for instance i. We adopt the Rubin-Neyman potential outcomes framework [334], [379], originally proposed for binary treatments $Y(0)$ and $Y(1)$, which can be extended to multi-valued or continuous treatments $Y(S)$ across $S \in [0, 1]$ [339]. For each instance, the potential outcome $y_i(s)$ reflects the response to dose s, given features $\mathbf{x}_i$. Data tuples $(\mathbf{x}, s, y)$ are generated by distributions $p(\mathbf{X})$, $p(S)$, and the observed policy Π_{obs}, which assigns treatment s_i based on $\mathbf{x}_i$, potentially introducing confounding bias. Figure 7.1 illustrates the assumed causal structure. The CADR function $\mu : \mathcal{S} \times \mathcal{X} \to [0, 1]$ is defined as $\mu(s, \mathbf{x}) = \mathbb{E}[Y(s) \mid \mathbf{X} = \mathbf{x}]$, representing the expected outcome for a given dose s and features $\mathbf{x}$. The CADE function $\tau : \mathcal{S} \times \mathcal{X} \to [-1, 1]$ is then derived from the CADR and defined as $\tau_s(\mathbf{x}) = \mathbb{E}[Y(s) - Y(0) \mid \mathbf{X} = \mathbf{x}]$, which measures the difference in expected outcomes between dose s and the baseline dose 0. Our notation is summarized in E.1.

7.3.2 Prediction step

During the prediction step, our goal is to train a model that can accurately estimate $\hat{y}(s)$, providing unbiased estimates of $\mu(s, \mathbf{x})$ and $\tau_s(\mathbf{x})$ across the entire domain of $S \in [0, 1]$. The estimated CADR for an instance with features $\mathbf{x}$ is defined as $\hat{\mu}(s, \mathbf{x}) = \mathbb{E}\left[\hat{Y}(s) \mid \mathbf{X} = \mathbf{x}\right]$. The estimated CADE for a given dose s is defined as $\hat{\tau}_s(\mathbf{x}) = \hat{\mu}(s, \mathbf{x}) - \hat{\mu}(0, \mathbf{x}) = \mathbb{E}\left[\hat{Y}(s) - \hat{Y}(0) \mid \mathbf{X} = \mathbf{x}\right]$. To use the CADE estimates as input for the optimization step, we discretize the CADRs into δ bins. For a given δ, we

define $D = \left\{ \frac{d-1}{\delta} \middle| \; d = 1, \ldots, (\delta + 1) \right\}$ so that the vector $\hat{\boldsymbol{\tau}}(\mathbf{x}) = \left(\hat{\tau}_{D_d}(\mathbf{x}) \right)_{d=1}^{\delta}$ then contains an entity's δ CADE estimates. We discuss model training and outcome estimation in Section 7.4.1.

7.3.3 Optimization step

Let $\pi : \mathcal{X} \to \{0, 1\}^{\delta}$ be an assignment policy defined for a single entity that takes its features as input and assigns a dose to this entity. In this policy output, the d^{th} element equals 1 if dose s is assigned, and 0 otherwise. Dose s corresponds to element d in D (i.e., $D_d = s$).

The cost matrix $\mathbf{C} \in \mathbb{R}^{N \times \delta}$ defines treatment costs, where $c_{i,d}$ is the treatment cost for entity i at dose s, with $c_{i,0} = 0$. For each entity i, its row in the cost matrix is denoted by $\mathbf{C}_i = (c_{i,0}, c_{i,1}, \ldots, c_{i,\delta-1})$, which lists all possible costs for that entity. For simplicity, we assume costs are directly proportional to dose levels (e.g., a dose of 0.2 has a cost of 0.2). This cost function is modular and can be easily replaced with alternative cost structures to suit different scenarios or requirements. The cost of policy π for instance i is:

$$\Psi_i(\pi) = \langle \pi(\mathbf{X}_i), \mathbf{C}_{i\cdot} \rangle \tag{7.2}$$

The uplift for an instance i (U_i) is given by the dot product $\langle \cdot, \cdot \rangle$ between the assignment policy π and the CADE vector $\boldsymbol{\tau}$. Similarly, the value gain for instance i (V_i) is the same dot product, scaled by the instance-specific benefit b_i.

Due to the fundamental problem of causal inference, historical data provides only factual outcomes, not counterfactual outcomes. Thus, in realistic scenarios, we must estimate these counterfactual outcomes. In contrast, experimental settings can leverage semi-synthetic datasets where the ground truth is known because we control the data-generating process. For a detailed discussion on the fundamental problem of causal inference, we refer readers to [333]. We define three distinct assignment policy values, each depending on the availability and usage of ground-truth versus estimated CADR values.

The expected policy value (Eq. 7.6) uses estimated CADRs for both the assignment policy and evaluation. It represents the anticipated value achievable under realistic conditions without access to counterfactual ground-truth data. The prescribed policy value V_i^{presc} (Eq. 7.7) employs estimated CADRs to determine the assignment policy but uses ground-truth CADRs for evaluation. Policies determined by maximizing V_i^{exp} yield an observable policy value V_i^{presc}. Due to differences between estimated and true causal effects,

V_i^{presc} typically diverges from V_i^{exp}. The optimal policy value V_i^{opt} (Eq. 7.8) serves as a benchmark and uses ground-truth CADRs for both policy determination and evaluation. This represents the maximum attainable policy value in a hypothetical setting where all counterfactual dose-response information is known precisely, a condition unattainable in practical applications.

Formally, these are defined as:

$$U_i^{exp} = \langle \pi(\hat{\tau}(\mathbf{X}_i)), \hat{\tau}(\mathbf{X}_i) \rangle, \tag{7.3}$$

$$U_i^{presc} = \langle \pi(\hat{\tau}(\mathbf{X}_i)), \tau(\mathbf{X}_i) \rangle, \tag{7.4}$$

$$U_i^{opt} = \langle \pi(\tau(\mathbf{X}_i)), \tau(\mathbf{X}_i) \rangle. \tag{7.5}$$

$$V_i^{exp} = U_i^{exp} b_i = \langle \pi(\hat{\tau}(\mathbf{X}_i)), \hat{\tau}(\mathbf{X}_i) \rangle b_i, \tag{7.6}$$

$$V_i^{presc} = U_i^{presc} b_i = \langle \pi(\hat{\tau}(\mathbf{X}_i)), \tau(\mathbf{X}_i) \rangle b_i, \tag{7.7}$$

$$V_i^{opt} = U_i^{opt} b_i = \langle \pi(\tau(\mathbf{X}_i)), \tau(\mathbf{X}_i) \rangle b_i. \tag{7.8}$$

For N entities with features $\mathbf{X}$ and a budget B, the policy $\overset{B}{\Pi}\colon \mathbb{R}^+ \times \mathcal{X}^N \to \{0,1\}^{(N \times \delta)}$ can be based on estimated or ground-truth CADRs. The expected optimal policy $\overset{B^{exp}}{\Pi}$ (Eq. 7.9) maximizes value within the budget using estimated CADRs. The prescribed policy assigns treatments according to $\overset{B^{exp}}{\Pi}$, but uses ground-truth CADRs. The ground-truth optimal policy $\overset{B^{opt}}{\Pi}$ (Eq. 7.10), based solely on ground-truth CADRs, is the full-information benchmark.

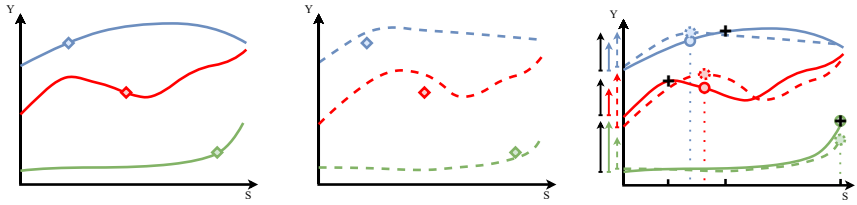
$$\overset{B^{exp}}{\Pi} = \operatorname{argmax} \left\{ \sum_{i=1}^{N} V_i^{exp}(\pi) : \sum_{i=1}^{N} \Psi_i(\pi) \leq B \right\} \tag{7.9}$$

$$\overset{B^{opt}}{\Pi} = \operatorname{argmax} \left\{ \sum_{i=1}^{N} V_i^{opt}(\pi) : \sum_{i=1}^{N} \Psi_i(\pi) \leq B \right\} \tag{7.10}$$

The implementation of this optimization formulation is discussed in Section 7.4.2.

7.4 Methodology

Figure 7.2 summarizes the predict-then-optimize approach for UM with continuous treatments under constraints.



(a) Data: Factual observations sampled from ground-truth dose-response curves

(b) Prediction step: Dose-response estimation based on factual observations

(c) Optimization step: Treatment dose allocation based on estimated dose-responses

Figure 7.2: Overview of predict-then-optimize approach for UM with continuous treatment effects for three entities. The x-axis represents the dose S, the y-axis represents the outcome Y. Key elements include true dose-response curves (solid lines), estimated dose-response curves (dashed lines), prescribed policies with their estimated and true outcomes (resp. dotted and full-lined circles), and the corresponding expected and prescribed values (resp. dotted and full arrows in color). The black ✚'s and arrows correspond to the full-information solution.

7.4.1 Predictive model for CADR estimation

To accurately estimate CADRs $\hat{\mu}$ and their associated CADE estimations $\hat{\tau}$, not addressing confounding between S and $\boldsymbol{X}$ in observational data can lead to detrimental inaccuracies or biases [333]. A variety of methods for disentangling treatment effects from confounding factors have been proposed in the literature [339], [352], [355], and this is not a prime concern in this paper. Instead, we consider the predictive methods as off-the-shelf solutions, with methods that do not apply a debiasing scheme also being applied in the experiments. One may refer to, e.g., [338], for assessments of the impact of confounding on outcomes.

Following standard practices in causal inference, we make three assumptions for estimating potential outcomes from observational data: Consistency, Ignorability, and Overlap [380], [381]. These assumptions ensure the identifiability of the CADE function $\tau_s(\mathbf{x})$, i.e., they result in Equation E.5 (see proof in E.2).

Assumption 1. *Consistency:* $\forall s \in \mathcal{S} : Y = Y(s)$. *This means that for any entity with observed treatment dose* $S = s$*, the potential outcome for this treatment dose is equal to the factual observed outcome.*

Assumption 2. *Ignorability:* $\{Y(s) \mid s \in \mathcal{S}\} \perp\!\!\!\perp S \mid \mathbf{X}$. *This means that,*

conditional on an entity's pre-treatment characteristics, the assigned treatment dose S is independent of the potential outcomes.

Assumption 3. *Overlap:* $\forall \mathbf{x} \in \mathcal{X}$ *such that* $p(\mathbf{x}) > 0, \forall s \in \mathcal{S} : 0 < p(s \mid \mathbf{x}) < 1$. *This states that all entities had a non-zero probability of being assigned any dose.*

In this work, we compare four different learning methods to estimate $\hat{\mu}$: an S-learner with random forests as the base learner (S-Learner (rf)), an S-learner with a feedforward multi-layer perceptron (MLP) without any debiasing as the base learner (S-Learner (mlp)) [382], DRNet [339], and VCNet [340]. Further details on the hyperparameters can be found in E.5.

7.4.2 ILP for the dose-allocation problem

Deciding on the optimal treatment dose for each entity, represented by policies $\overset{B}{\Pi}^{exp}$ and $\overset{B}{\Pi}^{opt}$, is equivalent to solving Equations (7.9-7.10). In the case when only binary doses can be applied, i.e., $S \in \{0, 1\}$, this problem is an instance of the knapsack problem [383]. A sensible heuristic would be to rank allocations from high to low uplift until the budget capacity is reached. This approach mirrors the calculation of a traditional Qini curve in a binary treatment setting. In our more general setting, however, any dose $S \in [0, 1]$ is permitted. To formalize this, we introduce two related functions:

$$U\colon \{0,1\}^{(N\times\delta)} \times [-1,1]^{(N\times\delta)} \to \mathbb{R}$$

$$V\colon \{0,1\}^{(N\times\delta)} \times [-1,1]^{(N\times\delta)} \times \mathbb{R}^N \to \mathbb{R}.$$

Both take as input a policy $\overset{B}{\Pi} \in \{0,1\}^{(N\times\delta)}$ (expressed as a matrix of decision variables) and a matrix $\mathbf{T} \in [-1,1]^{(N\times\delta)}$ of CADE values (estimated $\hat{\tau}$ or ground-truth τ) for the N entities. The function V additionally requires a benefit vector $\mathbf{b} \in \mathbb{R}^N$. Note that U is a special case of V in which the benefit vector $\mathbf{b}$ is set to all ones. For simplicity, we elaborate on V, with the understanding that this case also encompasses U.

To solve Equations (7.9–7.10), we propose the following ILP (Equations 7.11-7.18), where $V_i(\pi)$ in the objective is replaced by V_i^{exp} when finding $\overset{B}{\Pi}^{exp}$ and by V_i^{opt} when finding $\overset{B}{\Pi}^{opt}$. Additionally, we compare the ILP to a greedy ranking heuristic (see Section 7.5.3) Although numerous business requirements exist, we focus on fairness as the main example of constraints throughout this work. Without loss of generalization, we consider one protected binary sensitive attribute A (which is also included as an input feature to train the predictive model), where $N_{\mathrm{A=0}}$ is the index set of entities

where the protected attribute $A = 0$, and $N_{\text{A=1}}$ represents those with $A = 1$. These two sets are mutually exclusive and collectively exhaustive, so that $|N_{\text{A=0}}| + |N_{\text{A=1}}| = |N|$. To address fairness in allocation and outcomes, we introduce two slack parameters: $\epsilon_{DT} \in [0, 1]$ for allocation fairness and $\epsilon_{DO} \in [0, 1]$ for outcome fairness. A slack parameter value of $\epsilon = 0$ guarantees perfect fairness between the groups $N_{\text{A=0}}$ and $N_{\text{A=1}}$, while parameter $\epsilon = 1$ removes this fairness constraint.

Analogous to $V_i(\pi)^{exp}$, $V_i(\pi)^{presc}$, $V_i(\pi)^{opt}$ (Equations (7.6-7.8), which are defined on instance-level), and Equations (7.9-7.10) in Section 7.3, we consider three versions of V each differing in their dependence on ground-truth or estimated CADEs:

$$V^{exp} = \sum_{i=1}^{N} \sum_{d=1}^{\delta} \left(\overset{B}{\Pi}{}^{exp}_{i,d} \cdot \hat{T}_{i,d} \cdot b_i \right), \tag{7.19}$$

$$V^{presc} = \sum_{i=1}^{N} \sum_{d=1}^{\delta} \left(\overset{B}{\Pi}{}^{exp}_{i,d} \cdot T_{i,d} \cdot b_i \right), \tag{7.20}$$

$$V^{opt} = \sum_{i=1}^{N} \sum_{d=1}^{\delta} \left(\overset{B}{\Pi}{}^{opt}_{i,d} \cdot T_{i,d} \cdot b_i \right). \tag{7.21}$$

7.5 Experiments

In this section, we demonstrate our framework using semi-synthetic data, aiming to illustrate its applicability and performance in a controlled setting. The main objective is to illustrate the working of both the prediction step, i.e., the inference of continuous treatment effects, and the optimization step. We introduce relevant performance metrics and evaluate the impact of adding fairness constraints and adjusting the objective function on policy value.

7.5.1 Data

Our experiments use a semi-synthetic approach based on a real dataset about infant health and development programs (IHDP) [384]. This dataset originates from a randomized controlled trial and is frequently used as a benchmark for binary CATE estimation methods [320], [385]. Since we focus on continuous treatments, we adopt a semi-synthetic approach where both doses and outcomes are artificially generated. For this purpose, we follow the established literature [340]. Detailed information about the original dataset and the semi-synthetic data generation process is provided in E.3.

$$\max \sum_{i=1}^{N} V_i(\pi) \tag{7.11}$$

$$\text{s.t.} \sum_{i=1}^{N} \Psi_i(\pi) \leq B \qquad \text{(Budget constraint)} \tag{7.12}$$

$$\frac{1}{|N_{\text{A}=0}|} \sum_{i \in N_{\text{A}=0}} \sum_{s \in D} \pi_s(X_i) \cdot s \geq (1 - \epsilon_{DT}) \cdot \frac{1}{|N_{\text{A}=1}|} \sum_{i \in N_{\text{A}=1}} \sum_{s \in D} \pi_s(X_i) \cdot s \qquad \text{(Alloc. fairness 1)} \tag{7.13}$$

$$\frac{1}{|N_{\text{A}=0}|} \sum_{i \in N_{\text{A}=0}} \sum_{s \in D} \pi_s(X_i) \cdot s \leq (1 + \epsilon_{DT}) \cdot \frac{1}{|N_{\text{A}=1}|} \sum_{i \in N_{\text{A}=1}} \sum_{s \in D} \pi_s(X_i) \cdot s \qquad \text{(Alloc. fairness 2)} \tag{7.14}$$

$$\frac{1}{|N_{\text{A}=0}|} \sum_{i \in N_{\text{A}=0}} \sum_{s \in D} \pi_s(X_i) \cdot \tau_s \geq (1 - \epsilon_{DO}) \cdot \frac{1}{|N_{\text{A}=1}|} \sum_{i \in N_{\text{A}=1}} \sum_{s \in D} \pi_s(X_i) \cdot \tau_s \qquad \text{(Outc. fairness 1)} \tag{7.15}$$

$$\frac{1}{|N_{\text{A}=0}|} \sum_{i \in N_{\text{A}=0}} \sum_{s \in D} \pi_s(X_i) \cdot \tau_s \leq (1 + \epsilon_{DO}) \cdot \frac{1}{|N_{\text{A}=1}|} \sum_{i \in N_{\text{A}=1}} \sum_{s \in D} \pi_s(X_i) \cdot \tau_s \qquad \text{(Outc. fairness 2)} \tag{7.16}$$

$$\sum_{s \in D} \pi_s(X_i) = 1, \forall i \in \{1, \ldots, N\} \qquad \text{(Exactly one dose)} \tag{7.17}$$

$$\pi_s(X_i) \in \{0, 1\}, \forall s \in D, \forall i \in \{1, \ldots, N\} \qquad \text{(Binary decisions)} \tag{7.18}$$

7.5.2 Evaluation metrics

The predict-then-optimize approach is evaluated in two distinct steps, with separate metrics used to assess each step.

Prediction step To measure the accuracy of the predictive model in estimating individual dose-response curves, we use the *Mean Integrated Squared Error* (MISE, Equation 7.22) [339], [386]. This metric evaluates how closely the predicted dose-response curves match the true dose-response functions over the entire range $\mathcal{S}$ and hence requires information on semi-synthetic ground-truth.

$$\text{MISE} = \frac{1}{N}\sum_{i=1}^{N}\int_{s\in\mathcal{S}}\left(\mu\left(s,\mathbf{x}_i\right)-\hat{\mu}\left(s,\mathbf{x}_i\right)\right)^2 ds \tag{7.22}$$

Optimization step In the optimization step, several metrics evaluate how well the treatments are allocated, using the estimated CADEs, under the imposed constraints. The primary metric is the total Value (V), computed as the sum of individual V_i across all N entities. We distinguish V^{exp} (Eq. 7.19), V^{presc} (Eq. 7.20), and V^{opt} (Eq. 7.21) corresponding to the expected, prescribed, and optimal values, respectively. The cost-insensitive version, where costs and benefits are equal over all entities, is noted as U. *Regret* is defined as the difference between the full-information optimal value and the value achieved by the prescribed decision [366]. In other words, it measures how suboptimal the prescribed decision is compared to the optimal solution under ground-truth parameters. When CADRs are perfectly estimated, regret is zero. However, even with imperfect CADR estimations, regret can still be zero if the assignment policy remains optimal despite inaccuracies in the CADR estimates.

$$Regret = V^{opt} - V^{presc} = V\left(\overset{B}{\Pi}\left(\tau\right),\tau\right) - V\left(\Pi(\hat{\tau}),\tau\right) \tag{7.23}$$

To make *Regret* scale-independent, we also include the normalized version:

$$RegretN_B = \frac{V^{opt}-V^{presc}}{V^{opt}} = \frac{V\left(\overset{B}{\Pi}\left(\tau\right),\tau\right) - V\left(\overset{B}{\Pi}\left(\hat{\tau}\right),\tau\right)}{V\left(\overset{B}{\Pi}\left(\tau\right),\tau\right)} \tag{7.24}$$

Additionally, we assess performance across a range of budgets. The *Value Curve* plots the function $V(B) = \sum_{i=1}^{N} V_i(\pi_B)$, for $B \in [0, B_{max}]$, where B_{max} is defined as the budget required when treating all entities with dose

$s = 1$. We report the *Area Under the Value Curve* (AUVC) when costs and benefits are instance-dependent and *Area Under the Uplift Curve* (AUUC) when costs and benefits are equal for all instances. In a traditional binary treatment setting, the horizontal axis of the uplift curve typically represents the 'cumulative proportion of entities targeted' [364]. In our case, this translates to 'budget used' and differs in that the budget can target entities with finer granularity. For example, while a value of 5 on the x-axis in the traditional setting corresponds to 5 entities targeted, in our case, a budget of 5 could represent 10 entities receiving smaller doses.

7.5.3 Results and discussion

In this section, we describe the experimental evaluation of our proposed predict-then-optimize framework, focusing on its ability to handle continuous treatments, optimize dose allocation under various constraints, and balance possibly conflicting objectives such as fairness and policy value. In all experiments, the number of bins δ is fixed to 10, motivated by E.4, and the level of confounding bias is constant. All experiments are implemented using *Python 3.9* and can be reproduced with the code available on Github. All ILPs are solved using *Gurobi 11.0*. In all experimental scenarios presented, the ILP solver converged, guaranteeing optimal solutions for the given problem instances. Solver convergence was systematically verified by checking the solver status after each optimization run.

Experiment 1: Performance of dose-response estimators

The first experiment aims to compare the performance of different dose-response estimators in terms of both prediction accuracy and capacity for good dose allocation. We evaluate four dose-response estimators: S-Learner (rf) and S-Learner (mlp) [382], DRNet [339], and VCNet [340]. For each estimator, we conduct an internal 5-fold cross-validation loop, tuning hyperparameters across a grid search (details in E.5). The model with the lowest average MSE on factual outcomes over all folds is selected. In this experiment, we do not consider cost-sensitivity or fairness constraints. The budget is incrementally raised to observe the corresponding variation in U_B^{presc}. Each model's estimated CADRs are visualized in E.6.1.

We allocate treatments using two distinct methods. The first approach utilizes the ILP introduced in Section 7.4.2. The second employs a heuristic inspired by the multi-treatment framework [347]. Here, each discrete dose is treated as a separate intervention and, formally, the optimal treatment for an entity with features $\mathbf{x}_i$ is $\text{argmax}(\hat{\tau}(\mathbf{x}_i))$. Entities are then ranked according to the uplift of their optimal treatment, and doses are assigned

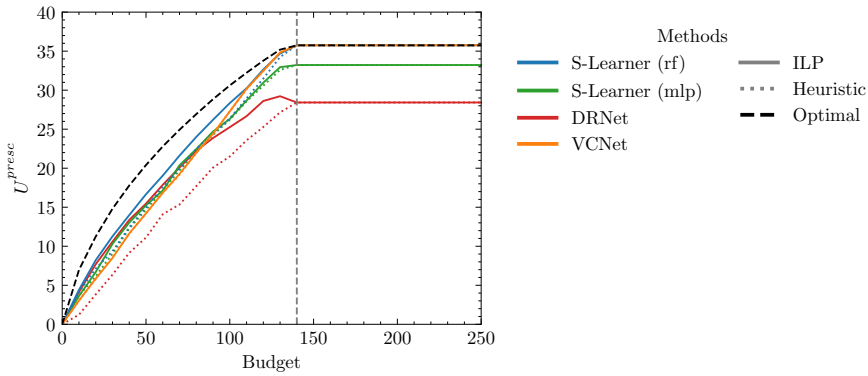


Figure 7.3: This figure shows U^{presc} for four dose-response estimators as the available budget levels increase. Doses are allocated with an ILP and a heuristic approach.

greedily in that order until the budget is depleted. The results are displayed in Figure 7.3 and Table 7.3. The analysis provides two insights.

First, compared to typical uplift curves from binary treatment scenarios (based on a ranking heuristic), both the multi-treatment heuristic and the ILP with continuous treatments as input do not fully utilize the budget. Around a budget of 140, they find a point where further treatment becomes counterproductive —i.e., it is more valuable to avoid *overtreating* certain entities— due to the non-monotonic nature of CADRs. From a budget of 140 onwards, per estimator, the heuristic and ILP approach find the same solution (i.e., selection of the dose with maximum estimated CADE) and

Table 7.3: This table shows the MISE (prediction step) and AUUC for budgets 140 and 250 (optimization step) across four dose-response estimators. The AUUC values are normalized against the optimal full-information solution. The best results are highlighted in bold and the second-best in italics.

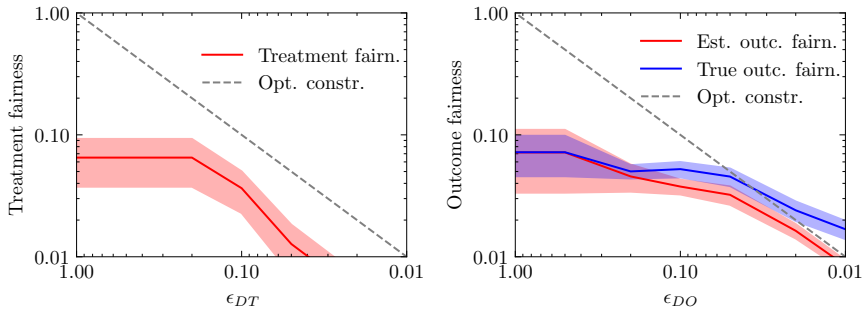
Estimator	**MISE**	**AUUC@140**		**AUUC@250**	
		Heur.	*ILP*	*Heur.*	*ILP*
S-Learner (rf)	0.060	**0.837**	**0.892**	**0.926**	**0.951**
S-Learner (mlp)	**0.044**	0.813	*0.829*	0.877	0.884
DRNet	*0.049*	0.649	0.799	0.729	0.797
VCNet	0.055	*0.827*	0.827	*0.922*	*0.922*
Optimal	0.000	1.000	1.000	1.000	1.000

they converge in terms of U^{presc}. Unlike the heuristic approach, which ranks binary treatment effects from high to low and selects the top k within the budget, continuous treatment effects have a larger search space, allowing for partial treatments. This is evident in Figure 7.3 where the Uplift curves flatten around a budget of 140. Therefore, in Table 7.3, we include AUUC not only for the full budget (i.e., @250) but also for a partial budget of 140. Around this point, for all methods considered (and the full-information optimal solution), the policy value remains constant despite increasing budget.

Second, Table 7.3 shows a misalignment between the quality of CADR estimation (measured by MISE during prediction) and the quality of treatment allocation (measured by AUUC during the optimization step). MISE does not fully capture downstream task performance, as errors in critical areas of the curve have a greater impact on treatment allocation than those in less important regions — which holds true for both heuristic and the ILP approaches.

For example, although S-Learner (mlp) has the lowest MISE, its final allocations rank second or third. DRNet shows the second-best MISE but performs worst in terms of AUUC at both budget points, irrespective of the treatment assignment method. Conversely, S-Learner (rf), despite its high MISE, outperforms all others in uplift effectiveness. This suggests that the best inference methods do not always result in optimal decision-making, a notion similarly observed in cost-sensitive learning literature, where maximizing predictive accuracy does not necessarily yield the highest profit [128]. Figure E.2a illustrates this phenomenon: the S-Learner (rf) produces low-quality estimates for $S \in [0.4, 0.6]$, resulting in a relatively high MISE. However, errors in this region have minimal impact on optimization since they correspond to low CADEs and are not selected anyway. DRNet, on the other hand, has relatively small errors overall but struggles significantly in the crucial high-impact region of $S \in [0.9, 1.0]$, where the ground-truth CADR sharply declines (see Figure E.2c). Thus, accurate dose allocation doesn't strictly require perfect predictions, provided the predictions capture the critical areas effectively. Conversely, perfect treatment allocation does not require flawless predictions; effective allocation is possible even with imperfect dose-response estimates. A further analysis of runtime for increasing problem sizes is provided in E.6.2. While the heuristic approach scales well for increasing problem sizes, it does not allow additional side constraints like fairness, which are allowed for the ILP.

Experiment 2: Fairness trade-offs in treatment allocation
The second experiment examines the trade-offs between policy value and business requirements as constraints, with a particular focus on fairness. Specifically, we examine both allocation fairness and outcome fairness, analyzing the impact of tightening or loosening these fairness constraints on



(a) Allocation fairness

(b) Outcome fairness

Figure 7.4: The effect of slack parameters ϵ_{DT} and ϵ_{DO} on, respectively, disparate treatment (allocation fairness) and disparate outcome (outcome fairness). Both panels use a logarithmic scale.

overall policy value. In this experiment, we fix the budget and disregard cost-sensitivity. We use the S-Learner (rf), which performed best in terms of AUUC, as the dose-response estimator. We vary two key fairness constraints: allocation fairness, which measures the disparity in assigned doses between groups, and outcome fairness, which reflects the difference in treatment outcomes. These constraints are regulated by two slack parameters: ϵ_{DT} for allocation fairness and ϵ_{DO} for outcome fairness.

Figure 7.4a shows the effect of parameter ϵ_{DT} on treatment fairness. Since treatment allocation is deterministic, there is no difference between the ground truth and estimates. Treatment fairness can be adjusted to any desired level by controlling ϵ_{DT}, making it straightforward to manage within the given constraints.

Figure 7.4b examines outcome fairness as regulated by the parameter ϵ_{DO}, revealing a distinct challenge compared to treatment fairness: a discrepancy between estimated and ground truth fairness. This divergence arises because optimization is based on estimated CADEs rather than true values. Although tightening ϵ_{DO} (illustrated by the red line) can restrict estimated disparities, it does not guarantee alignment with ground truth fairness (blue line). This misalignment is driven by the quality of the CADE estimates and underscores a broader challenge in predictive analytics—namely, that fairness in predictions does not always translate into fairness in actual outcomes. Nevertheless, we argue that incorporating constraints on expected outcomes across groups remains a principled and ethically sound design choice. Even when individual-level outcomes are uncertain, they still provide the most informative basis available for assessing how different groups may benefit from the allocation policy.

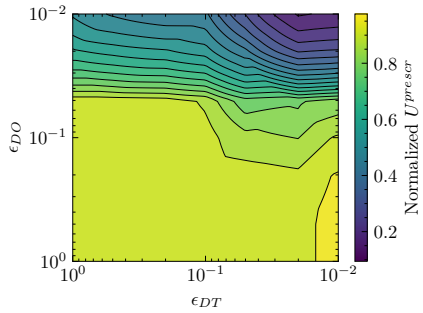
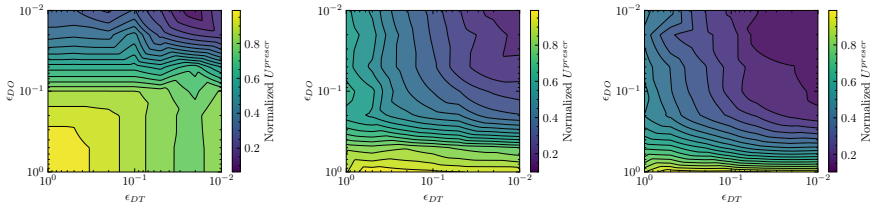


Figure 7.5: The x-axis represents the fairness constraint parameter for disparate treatment (ϵ_{DT}), and the y-axis represents the fairness constraint parameter for disparate outcome (ϵ_{DO}). Lower values of ϵ indicate stricter fairness constraints. U^{presc} is normalized between 0 and 1 and is aggregated over multiple budgets (from 25 to 250 in increments of 25). The averaged normalized U^{presc} is color-coded, where blue means lower and yellow means higher.

Figure 7.5 presents the trade-off between fairness constraints and U^{presc}. Stricter enforcement of both treatment and outcome fairness leads to lower uplift. The effects of these constraints differ; without an outcome constraint, the treatment constraint can be tightened with barely harming the policy value. Conversely, tightening the outcome constraint reduces policy value, especially when the treatment constraint is also strict. Note that finding a solution to the ILP to satisfy both fairness constraints is always possible, with the trivial available option of not allocating any treatment.

The setup of Figure 7.6 is similar to the one in Figure 7.5, but also examines varying data-generating processes where the two groups defined by the protected feature are more different. This is controlled by the parameter γ, with higher values indicating greater differences in ground-truth average treatment effects (ATEs) between the two protected groups. Figure 7.6 plots the results for three increasing values of γ (the case for $\gamma = 0$ is displayed in Figure 7.5). As the difference in ATEs between the groups increases, the effect of slack parameters ϵ_{DT} and ϵ_{DO} on the objective value (normalized U^{presc}, averaged over different budgets) becomes more pronounced. Strict slack parameters are tolerable without significant value loss when the groups are similar ($\gamma = 0$, Figure 7.5). In contrast, even mild slack parameters significantly affect the objective value when ATEs are most different ($\gamma = 10$, Panel 7.6c).



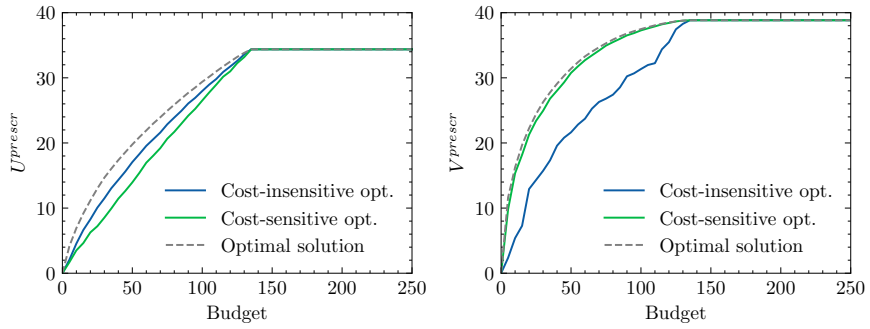
(a) $\gamma = 1$ (b) $\gamma = 5$ (c) $\gamma = 10$

Figure 7.6: Fairness constraints and varying ground-truth ATEs. This figure examines the link between the effect of the protected feature on the ground-truth ATE and the *ease* of achieving fairness. One panel is provided for each value of $\gamma \in \{1, 5, 10\}$ (the case for $\gamma = 0$ is displayed in Figure 7.5) and is aggregated over multiple budgets (from 25 to 250 in increments of 25), using the S-learner with random forests as the base learner. The averaged normalized U^{presc} is color-coded where blue means lower and yellow means higher.

Experiment 3: The impact of cost-sensitivity on utility

In many applications, the ultimate goal is profit or cost reduction, not just uplift. This experiment demonstrates that optimizing for the true objective (value) yields different policies than optimizing for uplift alone. This experiment explores the effect of incorporating cost-sensitivity into policy optimization. Building upon the previous experiments, we now account for cost-sensitivity in the adapted objective function of ILP, considering instance-dependent treatment costs, $\mathbf{C}$, and outcome benefits, $\mathbf{b}$. Figure 7.7b demonstrates that policies optimized with cost-sensitivity (green) perform better than cost-insensitive ones (blue) in terms of the cost-sensitive value (V^{presc}), particularly at lower budget levels. Conversely, Figure 7.7a shows the opposite scenario, emphasizing the need to align the optimization process with the intended objective at hand.

This experiment highlights the importance of cost-sensitivity in optimizing treatment allocation under constrained budgets. By taking into account treatment costs and outcome benefits, cost-sensitive policies achieve higher values in cost-sensitive metrics, especially when resources are limited and not all entities can receive full treatment. These policies prioritize entities with the highest benefit-to-cost ratios, improving allocation efficiency. However, they underperform on cost-insensitive utility metrics, underlining the need for alignment between optimization objectives and specific decision-making goals. The cost-sensitive policy, as expected, sacrifices some uplift if it leads to higher net value. We see differences in who gets treated: for example,



(a) Cost-insensitive U^{presc} across increasing budget.

(b) Cost-sensitive V^{presc} across increasing budget.

Figure 7.7: Comparison of cost-sensitive and cost-insensitive optimization, illustrating differences in performance based on U^{presc} and V^{presc}. The results underscore the importance of aligning the optimization objective with the downstream task.

some individuals with high uplift but low benefit may not be treated under the value-based policy, whereas they would under an uplift-only policy These findings are relevant for applications such as healthcare, lending, and human resource management, where treatment costs and benefits vary across entities, regardless of whether the goal is to maximize profitability or allocate scarce resources more efficiently. Furthermore, the modularity of the predict-then-optimize framework facilitates easy adaptation to changing objectives or constraints without requiring the retraining of predictive models. This makes it well-suited for dynamic environments where goals and resource availability evolve over time. An extension of this experiment could explore the inclusion of application-dependent and entity-dependent treatment cost functions and their interaction with the previously researched fairness constraints, adding complexity to the optimization process.

7.6 Conclusion, limitations, and further research

UM is extensively discussed in the literature, with applications such as lending, healthcare, HR, marketing, and maintenance. However, typically, no clear distinction between UM and CATE estimation is made, where CATE estimation should be viewed as a component of the broader UM field and is used as input for an optimization task. By clearly distinguishing these steps, our approach accommodates more complex treatments, as demonstrated in

our focus on continuous treatments, while allowing flexible integration of constraints like fairness considerations and cost-sensitive objectives. Our main contributions are (i) defining UM, (ii) extending it to handle continuous treatments with customizable constraints and objectives, (iii) defining fairness considerations as explicit ILP constraints, and (iv) demonstrating its capabilities through experiments. This framework's flexibility enables the use of pre-trained models and avoids challenges in decision-focused learning by allowing constraint integration without retraining, particularly advantageous when, for example, unexpected operational constraints arise. Continuous treatments allow fine-grained dose-based interventions, better suited for various applications, as shown in Table 7.1.

Our contributions are supported by the formal outlining and working of our framework in Sections 7.2 and 7.3, as well as validation by a series of experiments in Section 7.5. Experiment 1 shows the benefits of continuous treatments in an uplift setting, enabling decision-makers to maximize the total benefit of treatment allocation. Interestingly, this experiment shows that the most accurate predictive models are not always the best suited for the downstream treatment allocation. Experiment 2 explores the trade-offs between fairness constraints and policy value, revealing that stricter fairness enforcement, both in treatment allocation and outcomes, reduces overall utility, particularly when there are strong differences in ground-truth ATEs between the two groups considered. This reflects the broader challenge in algorithmic decision-making, where fairness may come at the cost of utility. Finally, Experiment 3 highlights the importance of aligning optimization with the specific objectives of the decision-making context, whether focusing on maximizing uplift or value.

While effective, our framework has limitations. The ILP approach suffers from limited scalability. In contrast, the proposed heuristic handles larger problem sizes more efficiently. However, it may yield suboptimal solutions and does not account for side constraints such as fairness. Additionally, the assumption of deterministic costs simplifies the problem but does not account for stochastic or uncertain treatment costs, which are present in other settings [361]. Another limitation is the focus on single-phase treatments, where many real-world scenarios involve sequential decisions. We also address fairness at the group level, leaving individual-level fairness for future exploration. Finally, experiments are based on a single dataset and data-generating process, which may limit generalizability.

This work is the first step in defining and providing an initial solution for UM with continuous treatments. Future research could extend this work by exploring varied allocation schemes, treatment characteristics, or sequential decision-making using dynamic policy or reinforcement learning. In cases of time dependencies, treatments are administered multiple times over

a period, or treatment effects change over time, dynamic policy learning or reinforcement learning methods could be employed to handle sequential treatments and long-term effects [387]. Extending the framework to accommodate multiple treatment types, whether mutually exclusive or not, would create new opportunities in situations where multiple intervention types are available [367]. Further research could investigate treatment effects in a network setting where spillover effects take place, making an entity's outcomes dependent on others' assigned doses and violating the stable unit treatment value assumption [388]. Additionally, while this study focuses on one-dimensional dose-based treatments, other applications may involve a high-dimensional treatment space [389]. When optimizing under uncertainty, incorporating conformal predictions, where CATE distributions or intervals around CADRs are estimated instead of point estimates, could better inform personalized decision-making [390], [391]. Finally, integrating DFL approaches, which, unlike our predict-then-optimize method, combine prediction and optimization in a single step, could potentially further improve policy outcomes by training models to directly make predictions that lead to better decisions [366]. One could also explore hybrid approaches — for example, incorporating fairness constraints directly into model training (a partial DFL approach) while still performing an optimization step for other parameters or constraints. Such methods might alleviate some of the burden on the post-prediction optimization

Epilogue

8
Industry co-creation

This dissertation was developed as part of a Baekeland mandate, a funding initiative supported by VLAIO that encourages close collaboration between academic institutions and industry partners. This research was conducted in partnership with Acerta, a leading Belgian HR services provider offering solutions in, e.g., payroll, social security, legal services, and talent management to a broad range of clients, including entrepreneurs, SMEs, and large organizations across Belgium. The structure of this dissertation aligns largely with the original project proposal. Although specific research questions evolved slightly during the course of the project, the overarching phases remain recognizable: a descriptive phase, a predictive phase, and a prescriptive phase.

Over four years, I had the opportunity to divide my efforts between KU Leuven and Acerta. At the university, my focus was on the conceptual and methodological development of advanced analytics techniques. At Acerta, I collaborated closely with BI specialists, HR analytics professionals, and various HR managers to understand their practical needs, identify pressing challenges, and iteratively transform research insights into practical tools — and conversely, to let practical challenges guide the research agenda.

The dual nature of this mandate, academic and industrial, was therefore not simply parallel, but inherently synergistic. Research questions frequently emerged directly from practical HR needs identified by Acerta and its clients, and these questions were refined through feedback from practitioners. This collaborative setup demanded research that was scientifically rigorous yet practically relevant. Moreover, it provided an opportunity to bridge the often-cited gap between academic sophistication and practical applicability.

This chapter, therefore, reflects on the iterative feedback loop between academia and industry, highlighting how practical challenges inspire novel research directions and how research outcomes inform the development of new analytical tools. Additionally, it acknowledges that the valuable collaboration with Acerta, combined with data contributions from Acerta and its clients, significantly shaped the trajectory and outcomes of this doctoral work.

In what follows, we assess how the three applied chapters (Part I) yielded insights and resulted in tools of practical relevance for Acerta and its clients. Furthermore, we discuss potential avenues for implementing and valorizing

the methodological contributions outlined in Part II. In a structured manner, we discuss operationalization, advantages, challenges, valorization, and the potential to scale to other clients.

8.1 Employee journey mapping

Chapter 2 presents a case study with Acerta where process discovery techniques are applied to HR event logs to generate EJMs, revealing the complexity of internal mobility. These maps highlight unexpected career paths, stepping stones to key roles, and hard-to-fill positions, offering a dynamic and flexible, descriptive tool for managing employee mobility.

Operationalization In addition to being implemented internally at Acerta, the tool was also deployed with two of Acerta's clients. The operationalization uses a Power BI dashboard that allows users to interactively filter and select specific data slices to display certain mobility paths. Since Power BI does not natively generate process maps, a third-party module (through a paid license) was integrated to visualize the EJMs. A screenshot of such an implementation is provided in Appendix F.1 (Figure F.1), illustrating how users can dynamically filter data and regenerate EJMs for various subsets of employees.

Advantages A key advantage of the EJM mapping exercise is its ease of deployment, stemming from relatively light data requirements. The input is an event log that captures historical job roles — covering positions, transitions, and timings — within the organization. Once the Power BI template is set up, the tool can be readily applied to provide new insights with minimal additional effort.

On one hand, it delivers directly actionable insights — for example, quantifying how many employees move from Role A to a lateral move versus how many take a vertical promotion. On the other hand, it uncovers patterns that may lead to follow-up questions. Indeed, this mapping exercise sparked interest in investigating paths leading to employee turnover, which subsequently motivated the deeper exploration of turnover prediction in Chapter 3.

Challenges Although the method itself is relatively simple, the main challenge in implementing EJMs lies in the quality of longitudinal data. Because EJMs trace job histories over time, maintaining consistent and clean records across years is non-trivial. Job titles may evolve, roles may split or merge, and inconsistent naming conventions can introduce noise. These changes must be handled carefully — for example, a modified job title should not be mistaken for a job transition. Accurate interpretation of mobility patterns depends on reliably distinguishing true role changes from data artifacts.

Valorization The value of this EJM mapping exercise is twofold. First, due to its low barrier to entry (in terms of data requirements and simplicity of modeling using a directly-follows graph miner), it can be rolled out quickly to new clients as a part of an HR analytics toolbox. This speed and ease make it an attractive "appetizer" for organizations new to data-driven HR: it provides tangible insights with minimal upfront investment. Second, the EJMs themselves deliver immediate, concrete insights. For example, organizations can see internal mobility rates, typical career pathways, and unexpected detours. These insights are valuable for HR strategy, such as identifying why certain roles struggle to retain employees or how to create more pathways for lateral career development. Additionally, positioning the EJM tool as an accessible introduction can inspire clients and stakeholders by showing the possibilities of more advanced analytics, potentially paving the way for follow-up projects (e.g., predictive or prescriptive analytics initiatives).

Potential to scale to other clients Because the EJM mapping relies on fairly standard HR data (for the three implementations, it was readily available from an ERP system with minimal preprocessing), extending it to other clients is rather straightforward once a template is in place. The main time investment is in ensuring that data is properly formatted and cleaned to avoid misleading results. With Power BI dashboards and the process mining extension, new client data can be plugged in to generate organization-specific EJMs. Thus, this tool is readily scalable, and Acerta has indeed identified it as a valuable service offering. Its low implementation complexity and high interpretability make it easy to demonstrate value to clients, either as a standalone analysis or as part of a larger HR analytics engagement.

8.2 Turnover prediction

Chapter 3 tackles employee turnover prediction, a critical issue due to the high direct and indirect costs of losing employees. It contributes a scoping review highlighting inconsistencies in the literature and a benchmarking experiment comparing 14 classification methods across 9 datasets.

Operationalization In the industrial context at Acerta, the turnover prediction work took two forms:

1. *Client-specific deep dive:* A consulting project with a particular client, focusing deeply on their internal data. This involved collecting and pooling data from multiple sources for that client, including demographic data and even assessment reports from supervisors. The goal

was to tailor a predictive model to the client's unique context and answer their specific questions about turnover risk.

2. *Generalized model across clients:* A broader turnover prediction model was trained on aggregated data from over 600 distinct organizations (for which Acerta manages payroll). This model is trained on a vast pooled dataset covering 600,000+ employee-year observations between 2019 and 2023.

In both cases, the approach followed a typical ML pipeline: data collection and cleaning, followed by predictive modeling with a post-hoc model explainer for interpretability (SHAP values in this case). Results were again reported through a Power BI dashboard, featuring interactive filters (by individual, role, department, etc.) and visualizations to examine differences across various demographic or job-related subsets. Figure F.2 in Appendix F.2 shows an example Power BI output with a beeswarm plot of SHAP values, illustrating model explainability.

Advantages The emphasis on explainability in the industrial rollout was a notable advantage. Unlike the purely predictive focus of the academic study in Chapter 3, the industry implementation highlighted why a model predicts a certain employee as high risk. Reporting SHAP values increased the validity and credibility of the model in the eyes of HR professionals and management, which is crucial for adoption – stakeholders are more likely to trust and act on the model's predictions if they understand the drivers (e.g., tenure, recent promotion, performance metrics) behind those predictions.

For the client-specific deep dive, the advantage was working with rich, context-specific data, enabling very tailored insights. The model could incorporate a wide array of features (including those unique to the client's HR systems or processes) and thus answer nuanced questions. For the generalized model across clients, the advantage lies in its scalability and immediate usefulness – a company can receive insights about turnover risk without having to provide their own data upfront. Acerta consultants could use this broad model to engage with a client by showing them patterns derived from industry-wide data ("This is what we see broadly; if you provide your data, we can refine these insights for your organization"). This approach leverages the shared structure across many companies and can serve as a conversation starter or diagnostic tool.

Challenges Several challenges emerged.

Data availability and integration (for the deep dive): Collecting and preprocessing the client's data was labor-intensive. Data resided in different systems (HR databases, performance evaluations, several isolated CSV files, etc.), and had to be carefully merged. Legacy system incompatibilities (e.g.,

changing employee ID formats over time) meant extra work to align records. Also, fields were sometimes inconsistently filled or changed definitions over the years, leading to missing or incoherent data for older records.

Data privacy and governance: Using detailed HR data triggered privacy concerns. Legal approval was needed to ensure compliance with GDPR and that employees' data could be used for analytics. This cautious approach was necessary, but it slowed down the project's start.

Data quality and consistency: Even beyond integration, ensuring clean longitudinal data was difficult. Similar to the EJM case, when working across many years, changes in how data was recorded (job titles, departments, etc.) could introduce pollution. For example, a departmental reorganization could make it appear as though some people "left" when in fact only department names changed. Vigilant cleaning and interpretation were required to avoid false signals.

Complexity of the general model: The broad model, while powerful in scale, only included features common across all clients (mostly payroll-related variables, roughly 30 features). This means it might overlook company-specific predictors of turnover that are not captured in a generalized dataset. There is a trade-off between (i) including many organizations, leading to many entries with a lower number of common variables, and (ii) analyzing just one organization, but including more detailed variables.

Valorization The value realized from these projects differs by approach.

The deep dive for this client resulted in a detailed, tailored report that addressed their key concerns. By closely analyzing their own data, we were able to highlight specific retention challenges — such as risk concentrated in certain departments or career stages — with clear implications for their HR strategy. While the insights were highly useful, producing them required significant time and effort. Replicating this level of analysis for other clients would involve a similarly heavy lift, making it valuable but not easily scalable. This kind of work is most appropriate when a client is highly motivated to understand their attrition and is prepared to share in-depth data.

The generalized model offers clear value in its broad applicability. A client can benefit from industry-wide insights with minimal effort on their part, which is a strong selling point. For Acerta, it can serve as a relatively low-effort way to engage clients: "Here are some insights from our large-scale model; if you're interested, we can dive deeper with your data to improve these insights." The broad model can deliver directly useful findings (like identifying common turnover risk factors in the sector) and doubles as a marketing or pre-sales tool to trigger further interest in advanced, tailored analytics services.

In both cases, the integration of the turnover prediction tool into Acerta's

offerings strengthens their HR analytics portfolio, either by immediately enhancing decision-making or by laying the groundwork for deeper, customized projects.

Potential to scale to other clients Scaling the turnover prediction solution to other clients depends on the approach.

Since it is already trained on a wide variety of companies, it can be applied to a new client whose payroll is managed by Acerta to give an initial risk assessment. Nearly any organization tracked in the pooled dataset can get a "turnover risk dashboard" immediately. Maintaining and updating this model with new data over time can further improve its accuracy and relevance.

The client-specific path requires more effort. To replicate the deep dive for another client, that client must commit to gathering and sharing their data. This doesn't scale seamlessly because each new deep dive will face similar hurdles (data integration, privacy checks, custom modeling). However, if a client sees the value in the generalized insights, they might be convinced to undertake this effort for a more precise analysis. Over time, as more such projects are done, Acerta could streamline the process (develop standard data request templates, cleaning scripts, etc.) to reduce the time required per client. In summary, with the combination of broad and deep modeling approaches, Acerta can both productize a turnover risk tool for the many and consult on bespoke analyses for those who are interested in a deeper analysis.

8.3 Internal mobility recommender system

Chapter 4 introduces a prescriptive analytics approach to support internal mobility by developing a data-driven recommender system. Expanding on the EJMs from Chapter 2, this recommender moves beyond simply describing existing career paths and instead proactively suggests future career moves within the organization.

Operationalization The internal mobility recommender was implemented within Acerta (for their own organizational use) rather than directly at client companies. The prototype works as follows: given a query for a specific employee, a Python script finds several "neighbor" employees (those with similar career paths or profiles) and then suggests a next job for the queried employee based on what those neighbors have done next. In essence, "employees like you went on to ..." as a recommendation.

Because precise performance scores of historical job-person matches were not readily available, the implementation uses a proxy scoring function (as

described in the paper) that estimates a match quality based on tenure in a role. The exact way this score is calculated involves modeling choices (e.g., what tenure gives the highest score), and these parameters were kept flexible so they can be adjusted as needed.

During internal pilot at Acerta, a significant adjustment was made compared to the approach of Chapter 4: changing the granularity of "jobs" into broader categories. In the original paper, each unique job title was considered a distinct item. However, for Acerta, this level of granularity (around 200 distinct job titles) proved difficult — the data became too sparse, and the recommender often suggested very common pathways (e.g., Payroll Officer I → Payroll Officer II → Payroll Officer III, which are frequent paths at Acerta). To make the system more useful, we mapped each job title to a competency profile (an internal standard categorization), reducing about 200 job titles to roughly 35 profiles. The idea was that multiple jobs correspond to the same competency profile, which (i) makes the rating matrix denser and the results more stable, and (ii) generalizes the recommendations beyond the very specific job titles to slightly broader career moves. This way, the recommender system can suggest moves that are not just the most common steps, but also lateral or unconventional moves that share underlying competency requirements.

In practice, the implementation at Acerta thus evolved towards recommendations based on competency profile transitions rather than overly granular job sequences. Each competency profile was additionally characterized by specific competencies (e.g., leadership, information processing, and interaction), with corresponding scores, enabling business logic filters within recommendations. For example, if a transition from competency profile A to B was recommended, constraints could be defined such that profile B must score higher than A on certain competencies, such as leadership. Although leadership is just one illustrative example, any business rule or constraint could be seamlessly integrated into the system.

At the time of writing, the recommender system is undergoing deployment as a *Function App*, a cloud-based solution within Microsoft Azure, with the implementation being carried out by a third-party vendor. The implementation serves as the first fully operational POC, intending to be further developed from a purely internal prototype into a robust, scalable tool.

Advantages The recommender system's prescriptive nature directly addresses a practical need: guiding employees (and HR managers) in identifying viable next career moves within the company. This has several benefits.

First, it empowers employees by making them aware of career paths they may not have previously considered. Conversely, for HR departments,

the system facilitates talent management by proactively identifying suitable internal candidates, including those who might not actively seek out new positions but possess desirable skills or experiences.

Next, we conjecture that promoting internal mobility through tailored recommendations could help organizations retain critical institutional knowledge, as internal career moves might reduce employee turnover and consequently mitigate its associated knowledge loss.

Furthermore, there is a certain data synergy with EJMs. The system leverages the same event log perspective as EJMs, meaning organizations that implemented Chapter 2's descriptive approach have already laid the groundwork for this prescriptive tool. It is a natural next step — first understand the mobility patterns, then start recommending mobility opportunities.

Challenges A limitation of the current implementation is that it is still based on relatively limited data. At this stage, only historical event logs of job-employee matches (and, by extension, competency profile-employee matches) are incorporated. To evolve the tool into a truly rich and usable instrument, it would be beneficial to supplement it with additional data from various sources. For instance, adding more granular steps within functions, or incorporating data from assessments and training programs, could enhance the system's depth. Allowing employees to input their own preferences for future roles could also make the system more personalized and flexible. However, these extensions are not yet part of the current POC and are envisioned for future iterations.

Another limitation is the absence of a direct measure of historical success in job-employee matches. Currently, we rely on the proxy of tenure duration within a role, which might be a suboptimal indicator. More nuance could be introduced by making the evaluation job-specific, recognizing that the same tenure length might signal different outcomes depending on the role. Ideally, success would be captured through direct assessments, such as structured interviews or evaluation methods, providing a more accurate and meaningful measure.

Finally, there is a critical human element: employee trust and acceptance of the tool. Career recommendations touch on sensitive personal and professional areas, and there may initially be skepticism towards algorithm-generated suggestions. Clear communication emphasizing transparency — such as providing understandable reasons for each recommendation (e.g., similarities in competencies or successful past transitions by similar colleagues) — will be crucial for fostering trust and ensuring the system's effective use.

Valorization From a practical perspective, the recommender system offers clear benefits for both employees and managers. Employees gain a powerful new tool to explore and pursue career opportunities proactively, often discovering pathways they may not have independently considered. Similarly, HR managers gain valuable assistance in filling internal vacancies more efficiently and strategically, potentially surfacing less obvious but highly qualified internal candidates. This dual-sided functionality strengthens overall talent management and succession planning within the organization.

Internally, for Acerta, deploying this recommender system also represents leadership by example. Utilizing advanced analytics within its own HR processes allows Acerta to credibly demonstrate to clients its commitment and confidence in data-driven HR solutions. Such internal use builds trust and showcases the practical benefits these tools can provide, reinforcing Acerta's reputation as an innovative and data-savvy partner.

Potential to scale to other clients In its current proof-of-concept state, the recommender system appears relatively straightforward to scale to other organizations. Its foundational components — such as historical job and competency data — are common across most enterprises. However, meaningful scalability and true commercial viability would be greatly enhanced by implementing the previously mentioned improvements. Adding richer data streams, employee-driven inputs, and a more nuanced measure of successful job matches would significantly boost the tool's effectiveness and market appeal.

When scaling to other clients, careful consideration must be given to organizational differences. The granularity of job categorization suitable for Acerta may not directly translate elsewhere, so flexibility in defining competency profiles or job groupings will be essential. Furthermore, integrating a user-friendly interface, potentially embedding the recommender within existing HR platforms or systems, would be crucial for broad adoption.

8.4 Potential implementation of other chapters

While Sections 8.1-8.3 discussed the operational details of the developed tools implemented within Acerta, several methodological advancements presented in Part II (Chapters 5-7) also hold promise for practical application. Specifically, cost-sensitive learning, decision-centric fairness, uplift modeling (with continuous treatments), and explicit fairness constraints have potential use cases relevant to Acerta's activities and client services.

Robust cost-sensitive learning (Chapter 5) addresses scenarios in which prediction errors have uneven consequences — a context encountered by

Acerta, for instance, when predicting employee turnover. Currently, Acerta's turnover prediction (as discussed in Section 8.2) relies primarily on predictive performance in combination with explainability. However, different employees and positions may incur varying degrees of replacement costs. Incorporating (robust) cost-sensitive methods would allow Acerta to help their clients prioritize retention efforts in a more nuanced manner, focusing resources on employees and hard-to-fill positions whose departure would entail the highest replacement costs. Successfully implementing this method would, however, require close collaboration with clients to accurately quantify these varying costs — an aspect not addressed in the academic part, but an initial challenge that could enhance the value of Acerta's services.

Decision-centric fairness (Chapter 6) introduces fairness evaluation specifically within subsets of model outputs relevant to actual decisions, such as, in an HR setting, shortlists for internal hiring or training and reskilling recommendations. Currently, Acerta supports numerous clients in managing internal talent mobility (as demonstrated in Sections 8.1 and 8.3), but fairness considerations have not yet been integrated. By incorporating decision-centric fairness, Acerta could directly address ethical concerns in sensitive HR processes such as selection and promotion, ensuring that recommendations or shortlists do not systematically disadvantage certain demographic groups. Practically, this approach would help Acerta and its clients comply with ethical and legal standards, improve perceptions of internal career processes, and therefore foster trust in a system, enhancing its transparency, legitimacy, and adoptability. To operationalize this successfully, Acerta would need to explicitly define fairness criteria with client organizations about protected features and relevant regulatory frameworks.

Uplift modeling with continuous treatments (Chapter 7) also presents clear opportunities, particularly in HR areas such as personalized employee retention programs or targeted training investments. Instead of merely predicting whether employees might leave, Acerta could use uplift modeling to identify the optimal amount of resources to invest in retaining specific employees. For example, rather than offering uniform retention incentives or training sessions, Acerta could guide clients toward individual-level resource allocation — such as tailored training hours — maximizing impact on a desired outcome while minimizing costs. Implementing continuous uplift modeling, however, relies on rich historical data capturing varying levels of past interventions (e.g., differentiated retention strategies previously implemented by clients). Acerta would need to facilitate such detailed data collection efforts or pilot projects before fully integrating this methodology.

Finally, both Chapters 6 and 7 address fairness but differ in their approach. The method in Chapter 6 integrates fairness directly into model training, making it suitable for scenarios where decisions occur continuously

(in time) and individually. Conversely, Chapter 7 uses fairness constraints at the batch decision-making level — ideal for scenarios involving budget allocation or the distribution of limited resources across defined groups. Acerta could leverage these complementary approaches depending on specific client needs, either embedding fairness at the individual decision level or ensuring equity across group-based interventions.

In summary, each of these advanced methods offers tangible opportunities for Acerta to expand its analytical offerings, aligning methodological advances with concrete HR challenges faced by its clients. Realizing these benefits depends on clearly defined cost and fairness criteria, enhanced data collection processes, and thoughtful alignment with legal and ethical guidelines — steps that would enable Acerta to reinforce its position as a forward-looking HR analytics provider.

9
Conclusion

Modern data collection processes, methodological advancements, and computing resources facilitate complex pattern recognition through advanced analytics [2]. Beyond descriptive analyses that reveal historical trends, predictive models enable the mapping of future scenarios. Prescriptive methods like uplift modeling guide concrete interventions under budget or fairness constraints to steer towards desired future scenarios. Building on these enabling factors, Part I and Part II work in tandem to contribute to both HR applications and methodological advances through six main chapters.

Close collaboration with Acerta Consult keeps the research grounded in real-world HR processes, ensuring that each contribution — from job-employee matching to fairness-focused interventions — is not merely novel, but remains relevant and useful in organizational settings.

The following sections summarize this dissertation's contributions, offer managerial implications, highlight limitations, and offer directions for future research.

9.1 Contributions

Chapter 2 introduced the application of process mining to internal employee mobility data. It showed how EJMs offer a dynamic way to visualize individual career paths, highlighting infrequent transitions, identifying stepping-stone positions, and revealing alternative paths beyond conventional career ladders. This approach provides HR professionals with concrete, data-driven insights on internal mobility and helps address the complexity of real-world career trajectories.

Chapter 3 focused on predicting employee turnover through a comprehensive scoping review and benchmarking of current methods. It highlighted the inconsistent methodologies in existing turnover research and then established a rigorous experiment involving multiple classification algorithms across various datasets. The findings provide a unified focal point for scholars and practitioners.

Chapter 4 tackled job–employee matching with an emphasis on addressing data scarcity problems (e.g., new hires). Integrating historical perfor-

mance with personal employee attributes through a similarity regularization term reduces cold-start issues. This similarity-driven extension proved beneficial in improving match quality for internal mobility recommendations.

Chapter 5 improved the robustness of instance-dependent, cost-sensitive classification by proposing r-cslogit, a method that detects and mitigates outliers in cost and benefit parameters. Tested extensively on (semi-)synthetic data, it showed better stability and performance gain when the outlier size increased, particularly crucial in contexts where misclassification can be costly.

Chapter 6 advances fairness in resource allocation resulting from a classification task. It enforces fair outcomes only where resource allocation occurs, i.e., a predefined actionable decision region. This approach optimizes the model's predictive capability while ensuring fairness for those receiving positive classification.

Chapter 7 extends uplift modeling to continuous treatments, moving beyond the traditional binary intervention framework. First, it estimates conditional average dose responses through causal machine learning; then, it optimizes treatment allocation via integer linear programming under budget and fairness constraints. This enables effective and efficient resource distribution, aligned with organizational goals.

9.2 Managerial implications

The enduring value of *traditional* HR theory. While analytics enhances decision-making, *traditional* HR theory remains indispensable for interpreting deeper causes of outcomes like turnover and mobility. Although not the focus of this dissertation, theory-based insights guide the design of data collection, feature selection, and interventions, ensuring that predictive models align with established frameworks. Managers should integrate these theoretical perspectives to maximize interpretability, strategic fit, and model performance.

Prioritizing data quality and compliance. Successful HR analytics initiatives depend significantly on data quality, not merely on the quantity of data collected. Managers should prioritize the systematic collection of relevant, high-quality data, informed by traditional HR theories and frameworks that identify meaningful variables. Compliance with data protection regulations such as GDPR is also crucial. HR analytics inherently involves handling sensitive employee information, and violations of privacy standards can result in legal repercussions and reputational damage. Managers must implement rigorous data governance practices, continuously evaluating their analytics frameworks to ensure compliance, ethical responsibility, and overall reliability.

Predict-and-optimize or predict-then-optimize? This dissertation highlights two distinct paradigms: *Predict-and-optimize* and *Predict-then-optimize*. Each has its advantages and disadvantages. Predict-and-optimize integrates decision-making processes directly into the predictive modeling step, enabling comprehensive management of uncertainty and potentially delivering decisions closely aligned with downstream objectives. However, this integration can introduce complexity, demanding greater computational resources and limiting flexibility. Conversely, Predict-then-optimize maintains modularity and simplicity, allowing for easier implementation and updates, though it may propagate predictive errors into suboptimal decisions. Managers must carefully consider their organization's specific context, objectives, and resource availability when selecting between these paradigms.

With a new hammer, everything looks like a nail. With the surge in availability and capabilities of analytical methods, there might be a growing temptation to apply them broadly — even where simpler approaches would suffice. Yet, not every problem requires a complex solution. Managers are advised to start with small, targeted analytics initiatives grounded in clear business objectives. For example, in turnover prediction, initiating analysis with a logistic regression model and smart feature engineering can yield actionable insights without unnecessary complexity. Avoiding solution-driven approaches — where methods dictate problems — is critical.

Importance of post-deployment analytics. Continuous monitoring of deployed analytical models is critical for sustained performance and alignment with changing organizational environments. Over time, data distributions may shift, potentially resulting in data drift and decreased model performance. Managers must establish robust monitoring protocols that detect performance degradation early, facilitating interventions such as retraining, recalibrating thresholds, or other model adjustments. Effective post-deployment monitoring protects organizations from the risks of unnoticed model failures and ensures their reliability.

Awareness of methodological assumptions. Every analytical method employed is built upon specific assumptions. Traditional statistics rest on assumptions about data distributions and sampling processes (e.g., the *I.I.D.* assumption), while causal inference methods rely on assumptions such as *consistency*, *ignorability*, and *overlap*. Managers must remain aware of these underlying assumptions and their implications in their conclusions. There should be transparency regarding these assumptions, and they should regularly be assessed against real-world conditions to prevent misinterpretations and suboptimal decisions.

Avoiding algorithmic aversion. Managers have a key responsibility in fostering an organizational culture that mitigates *algorithmic aversion* — the tendency of employees and managers to distrust or avoid algorithmic support, even when evidence shows it improves decision-making [392]. Of course, this depends on the *invasiveness* of algorithmic support. But, if desired, managers can help by creating a cultural environment that encourages continuous learning and adaptability, enabling effective integration of new technologies by framing the use of algorithms as collaborative rather than imposed [392].

Model output is not the same as probability. Interpreting model predictions as probabilities requires proper calibration. Models differ inherently in calibration quality: logistic regression typically yields naturally well-calibrated predictions, whereas boosting methods and SVMs often produce overly confident outputs near 0 or 1, or overly conservative predictions clustered around 0.5 [393]. This issue becomes even more pronounced when models are trained using multi-objective criteria such as fairness considerations (see Chapter 6). These objectives can deliberately reshape the output distribution to satisfy fairness requirements, often at the expense of calibration accuracy. Decision-makers must therefore exercise caution. Poor calibration can result in suboptimal decisions; for example, selecting employees predicted to leave with a model score above 0.7 does not imply an actual 70% probability. Similarly, when predictions are inputs for subsequent analyses, like calculating expected profit or loss, these outputs must be properly calibrated. Metrics such as the Brier score can assess calibration quality, while corrective techniques like Platt scaling can enhance the reliability of predicted probabilities [394].

9.3 Limitations and future work

Explainability and interpretability of models. One notable limitation of this dissertation is the limited focus on model explainability (except for the industry implementation as discussed in Chapter 8). While predictive accuracy and prescriptive effectiveness are central in this manuscript, a deeper exploration into opening the *black box* would enhance the practical adoption of analytics methods in HR. According to the distinction between prediction and explanation [61], HR analytics applications frequently require a balance between accurate predictions and meaningful interpretations. Future research could address this gap by developing advanced yet interpretable models or by systematically integrating post-hoc explainability techniques, facilitating greater acceptance and trust.

Incorporation of unstructured data sources. Another limitation is the exclusive reliance on structured data, neglecting the growing availability and usability of unstructured textual data. The rise of easily accessible, plug-and-play solutions leveraging large language models (LLMs) offers substantial potential for enriching HR analytics applications. Future research should explore methods for systematically incorporating textual data from sources such as employee surveys, performance reviews, and exit interviews. By using such sources of qualitative insights, analytics models could improve predictive accuracy, decision-making nuance, and contextual understanding.

Scope of fairness criteria. Chapters 6 and 7 consider one fairness notion: group-level independence. Alternative fairness criteria (e.g., separation or sufficiency), intersectionality (e.g., the combination of ethnicity and gender), or individual-level instead of group fairness, might produce different trade-offs. Extending decision-centric fairness to broader or more complex fairness notions constitutes an important next step.

Explicit modeling of uncertainty. Currently, the analytical methods presented — both predictive and prescriptive — produce outputs or recommendations based on point estimates. An important avenue for future research lies in explicitly incorporating uncertainty into these models. Advanced decision-making tools could integrate uncertainty scores or confidence intervals (through, e.g., conformal prediction), enabling more informed managerial decisions. Introducing options such as reject decisions or decision-focused learning paradigms, linking predictive and decision-making parameters explicitly, could enhance the robustness and utility of prescriptive analytics models.

Foundation models. Foundation models, pretrained on large and diverse datasets, have become central in many machine learning applications — most notably in natural language processing, where LLMs dominate. Recently, foundation models tailored to tabular data have been developed, with TabPFN [395] being a prominent example. Their performance is promising and may represent a new state of the art for certain predictive tasks, including those discussed in this manuscript. Although the current implementation of TabPFN performs best on smaller datasets, it shows great potential due to its minimal need for fine-tuning, low inference time, and its native ability to handle missing and categorical data.

References

[1] H. Chen, R. H. Chiang, and V. C. Storey, "Business intelligence and analytics: From big data to big impact," *MIS quarterly*, pp. 1165–1188, 2012.

[2] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big data*, vol. 1, no. 1, pp. 51–59, 2013.

[3] J. H. Marler and J. W. Boudreau, "An evidence-based review of hr analytics," *The International Journal of Human Resource Management*, vol. 28, no. 1, pp. 3–26, 2017.

[4] D. Ulrich and J. H. Dulebohn, "Are we there yet? what's next for hr?" *Human resource management review*, vol. 25, no. 2, pp. 188–204, 2015.

[5] D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence, and M. Stuart, "Hr and analytics: Why hr is set to fail the big data challenge," *Human resource management journal*, vol. 26, no. 1, pp. 1–11, 2016.

[6] D. Pessach, G. Singer, D. Avrahami, H. C. Ben-Gal, E. Shmueli, and I. Ben-Gal, "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming," *Decision support systems*, vol. 134, p. 113 290, 2020.

[7] European Commission, *Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation) (text with eea relevance)*, May 2016.

[8] E. Commission, *EUR-Lex - 52021PC0206 - EN - EUR-Lex*, eurlex.europa.eu, Accessed on December 2022, Apr. 2021.

[9] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, Lawrence Erlbaum Associates Ltd, vol. 17, 2001, pp. 973–978.

[10] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[11] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 469–481.

[12] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[13] W. Verbeke, D. Olaya, M.-A. Guerry, and J. Van Belle, "To do or not to do? cost-sensitive causal classification with individual treatment effect estimates," *European Journal of Operational Research*, 2022.

[14] S. Höppner, B. Baesens, W. Verbeke, and T. Verdonck, "Instance-dependent cost-sensitive learning for detecting transfer fraud," *European Journal of Operational Research*, vol. 297, no. 1, pp. 291–300, 2022.

[15] W. S. Siebert and N. Zubanov, "Searching for the optimal level of employee turnover: A study of a large uk retail organization," *Academy of management journal*, vol. 52, no. 2, pp. 294–313, 2009.

[16] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019.

[17] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," in *Ethics of data and analytics*, Auerbach Publications, 2022, pp. 296–299.

[18] European Commission, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM(2021) 206 final, 2021.

[19] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, "The measure and mismeasure of fairness," *The Journal of Machine Learning Research*, vol. 24, no. 1, pp. 14 730–14 846, 2023.

[20] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint: 1609.05807*, 2016.

[21] K. Makhlouf, S. Zhioua, and C. Palamidessi, "On the applicability of machine learning fairness notions," *ACM SIGKDD Explorations Newsletter*, vol. 23, no. 1, pp. 14–23, 2021.

[22] S. De Vos, J. De Smedt, C. Wuytens, and W. Verbeke, "Leveraging process mining to optimize internal employee mobility strategies," in *Business Process Management Cases Vol. 3: Implementation in Practice*, Springer, 2025, pp. 15–28.

[23] S. De Vos, C. Bockel-Rickermann, J. Van Belle, and W. Verbeke, "Predicting employee turnover: Scoping and benchmarking the state-of-the-art," *Business & Information Systems Engineering*, pp. 1–20, 2024.

[24] S. De Vos, J. De Smedt, M. Verbruggen, and W. Verbeke, "Data-driven internal mobility: Similarity regularization gets the job done," *Knowledge-Based Systems*, vol. 295, p. 111 824, 2024.

[25] S. De Vos, T. Vanderschueren, T. Verdonck, and W. Verbeke, "Robust instance-dependent cost-sensitive classification," *Advances in Data Analysis and Classification*, vol. 17, no. 4, pp. 1057–1079, 2023.

[26] S. De Vos, J. Van Belle, A. Algaba, W. Verbeke, and S. Verboven, "Decision-centric fairness: Evaluation and optimization for resource allocation problems," *arXiv preprint arXiv:2504.20642*, 2025.

[27] S. De Vos, C. Bockel-Rickermann, S. Lessmann, and W. Verbeke, "Uplift modeling with continuous treatments: A predict-then-optimize approach," *arXiv preprint arXiv:2412.09232*, 2024.

[28] A. Raval, "Talent wars: Why businesses have to battle to hire the best," en, *Financial Times*, Sep. 2022.

[29] J. Boudreau and W. Cascio, "Human capital analytics: Why are we not there?" *Journal of Organizational Effectiveness: People and Performance*, vol. 4, no. 2, pp. 119–126, Jan. 2017.

[30] V. Fernandez and E. Gallardo-Gallardo, "Tackling the HR digitalization challenge: Key factors and barriers to HR analytics adoption," en, *Competitiveness Review Journal*, vol. 31, no. 1, pp. 162–187, Jul. 2020.

[31] M. Dumas, M. La Rosa, J. Mendling, and H. Reijers, *Fundamentals of Business Process Management*, en. Springer Berlin Heidelberg, Mar. 2013.

[32] W. van der Aalst, "Data science in action," in *Process Mining: Data Science in Action*, W. van der Aalst, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 3–23.

[33] M. L. Van Eck, X. Lu, S. J. Leemans, and W. M. Van Der Aalst, "Pm: A process mining project methodology," in *Advanced Information Systems Engineering: 27th International Conference, CAiSE 2015, Stockholm, Sweden, June 8-12, 2015, Proceedings*, Springer, 2015, pp. 297–313.

[34] W. M. P. van der Aalst, "A practitioner's guide to process mining: Limitations of the directly-follows graph," *Procedia Computer Science*, vol. 164, pp. 321–328, Jan. 2019.

[35] P. W. Hom and R. W. Griffeth, *Employee turnover* (South-Western Series in Human Resources Management). South-Western College Publishing, 1995, ISBN: 9780538808736. [Online]. Available: `https://books.google.de/books?id=n%5C_cXAQAAMAAJ`.

[36] C. Perryer, C. Jordan, I. Firns, and A. Travaglione, "Predicting turnover intentions: The interactive effects of organizational commitment and perceived organizational support," *Management Research Review*, vol. 33, no. 9, pp. 911–923, 2010.

[37] S. M. Soltis, F. Agneessens, Z. Sasovova, and G. Labianca, "A social network perspective on turnover intentions: The role of distributive justice and social support," *Human Resource Management*, vol. 52, no. 4, pp. 561–584, 2013.

[38] R. W. Griffeth, P. W. Hom, and S. Gaertner, "A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium," *Journal of Management*, vol. 26, no. 3, pp. 463–488, 2000.

[39] M. Armstrong, *A Handbook of Human Resource Management Practice*. Kogan Page Publishers, 2006.

[40] A. Carmeli and J. Weisberg, "Exploring turnover intentions among three professional groups of employees," *Human Resource Development International*, vol. 9, no. 2, pp. 191–206, 2006.

[41] V. V. Saradhi and G. K. Palshikar, "Employee churn prediction," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1999–2006, 2011.

[42] M. J. Kavanagh and R. D. Johnson, *Human resource information systems*. SAGE Publications, Incorporated, 2020.

[43] G. DeSanctis, "Human resource information systems: A current assessment," *MIS quarterly*, pp. 15–27, 1986.

[44] I. Teinemaa, M. Dumas, M. L. Rosa, and F. M. Maggi, "Outcome-oriented predictive process monitoring: Review and benchmark," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 13, no. 2, pp. 1–57, 2019.

[45] S. Wenninger and C. Wiethe, "Benchmarking energy quantification methods to predict heating energy performance of residential buildings in germany," *Business & Information Systems Engineering*, vol. 63, pp. 223–242, 2021.

[46] S. Lessmann and S. Voß, "Customer-centric decision support: A benchmarking study of novel versus established classification models," *Business & Information Systems Engineering*, vol. 2, pp. 79–93, 2010.

[47] P. W. Hom, T. W. Lee, J. D. Shaw, and J. P. Hausknecht, "One hundred years of employee turnover theory and research.," *Journal of Applied Psychology*, vol. 102, no. 3, p. 530, 2017.

[48] E. Rombaut and M.-A. Guerry, "Predicting voluntary turnover through human resources database analysis," *Management Research Review*, vol. 41, no. 1, pp. 96–112, 2018.

[49] D. Pitts, J. Marvel, and S. Fernandez, "So hard to say goodbye? Turnover intention among US federal employees," *Public Administration Review*, vol. 71, no. 5, pp. 751–760, 2011.

[50] T. W. Ng and D. C. Feldman, "Re-examining the relationship between age and voluntary turnover," *Journal of Vocational Behavior*, vol. 74, no. 3, pp. 283–294, 2009.

[51] J. A. Grissom, S. L. Viano, and J. L. Selin, "Understanding employee turnover in the public sector: Insights from research on teacher mobility," *Public Administration Review*, vol. 76, no. 2, pp. 241–251, 2016.

[52] A. L. Kalleberg and K. A. Loscocco, "Aging, values, and rewards: Explaining age differences in job satisfaction," *American Sociological Review*, pp. 78–90, 1983.

[53] T. W. Ng and D. C. Feldman, "The relationships of age with job attitudes: A meta-analysis," *Personnel Psychology*, vol. 63, no. 3, pp. 677–718, 2010.

[54] A. Clark, A. Oswald, and P. Warr, "Is job satisfaction U-shaped in age?" *Journal of Occupational and Organizational Psychology*, vol. 69, no. 1, pp. 57–81, 1996.

[55] C. D. Crossley, R. J. Bennett, S. M. Jex, and J. L. Burnfield, "Development of a global measure of job embeddedness and integration into a traditional model of voluntary turnover.," *Journal of Applied Psychology*, vol. 92, no. 4, p. 1031, 2007.

[56] O. A. Ayodele, A. Chang-Richards, and V. González, "Factors affecting workforce turnover in the construction sector: A systematic review," *Journal of Construction Engineering and Management*, vol. 146, no. 2, p. 03 119 010, 2020.

[57] S. M. Thin, B. Chongmelaxme, S. Watcharadamrongkun, T. Kanjanarach, B. A. Sorofman, and T. Kittisopee, "A systematic review on pharmacists' turnover and turnover intention," *Research in Social and Administrative Pharmacy*, vol. 18, no. 11, pp. 3884–3894, 2022.

[58] J. W. Han, "A review of antecedents of employee turnover in the hospitality industry on individual, team and organizational levels," *International Hospitality Review*, vol. 36, no. 1, pp. 156–173, 2020.

[59] P. W. Hom, T. R. Mitchell, T. W. Lee, and R. W. Griffeth, "Reviewing employee turnover: Focusing on proximal withdrawal states and an expanded criterion.," *Psychological Bulletin*, vol. 138, no. 5, p. 831, 2012.

[60] T. W. Lee and T. R. Mitchell, "An alternative approach: The unfolding model of voluntary employee turnover," *Academy of Management Review*, vol. 19, no. 1, pp. 51–89, 1994.

[61] G. Shmueli, "To explain or to predict?" *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010. DOI: 10.1214/10-STS330. [Online]. Available: https://doi.org/10.1214/10-STS330.

[62] T. Pape, "Prioritising data items for business analytics: Framework and application to human resources," *European Journal of Operational Research*, vol. 252, no. 2, pp. 687–698, 2016.

[63] D. Pessach, G. Singer, D. Avrahami, H. Chalutz Ben-Gal, E. Shmueli, and I. Ben-Gal, "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming," en, *Decision Support Systems*, vol. 134, p. 113 290, Jul. 2020.

[64] C.-F. Chien and L.-F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," *Expert Systems with Applications*, vol. 34, no. 1, pp. 280–290, 2008.

[65] J. M. Kirimi and C. A. Moturi, "Application of data mining classification in employee performance prediction," *International Journal of Computer Applications*, vol. 146, no. 7, pp. 28–35, 2016.

[66] J.-J. Decorte, J. Van Hautte, T. Demeester, and C. Develder, "Jobbert: Understanding job titles through skills," *arXiv preprint arXiv:2109.09605*, 2021.

[67] F. Devriendt, D. Moldovan, and W. Verbeke, "A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics," *Big Data*, vol. 6, no. 1, pp. 13–41, 2018.

[68] F. Devriendt, J. Berrevoets, and W. Verbeke, "Why you should stop predicting customer churn and start using uplift models," *Information Sciences*, vol. 548, pp. 497–515, 2021.

[69] L. Breiman, "Statistical modeling: The two cultures," *Statistical science*, vol. 16, no. 3, pp. 199–231, 2001.

[70] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," *International Journal of Social Research Methodology*, vol. 8, no. 1, pp. 19–32, 2005.

[71] Z. Munn, M. D. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris, "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach," *BMC Medical Research Methodology*, vol. 18, pp. 1–7, 2018.

[72] R. Pranckutė, "Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world," *Publications*, vol. 9, no. 1, p. 12, 2021.

[73] V. Nagadevara, V. Srinivasan, and R. Valk, *Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques*, 2008.

[74] H.-Y. Chang, "Employee turnover: A novel prediction solution with effective feature selection," *WSEAS Transactions on Information Science and Applications*, vol. 6, no. 3, pp. 417–426, 2009.

[75] Q. A. Al-Radaideh and E. Al Nagi, "Using data mining techniques to build a classification model for predicting employees performance," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 2, 2012.

[76] M. A. Valle, S. Varas, and G. A. Ruz, "Job performance prediction in a call center using a naive Bayes classifier," *Expert Systems with Applications*, vol. 39, no. 11, pp. 9939–9945, 2012.

[77] M. A. Valle and G. A. Ruz, "Turnover prediction in a call center: Behavioral evidence of loss aversion using random forest and naive Bayes algorithms," *Applied Artificial Intelligence*, vol. 29, no. 9, pp. 923–942, 2015.

[78] A. M. Esmaieeli Sikaroudi, R. Ghousi, and A. Sikaroudi, "A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)," *Journal of Industrial and Systems Engineering*, vol. 8, no. 4, pp. 106–121, 2015.

[79] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, IEEE, 2017, pp. 1016–1020.

[80] İ. O. Yiğit and H. Shourabizadeh, "An approach for predicting employee churn by using data mining," in *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, IEEE, 2017, pp. 1–4.

[81] M. A. Valle, G. A. Ruz, and V. H. Masías, "Using self-organizing maps to model turnover of sales agents in a call center," *Applied Soft Computing*, vol. 60, pp. 763–774, 2017.

[82] M. M. Alam, K. Mohiuddin, M. K. Islam, M. Hassan, M. A.-U. Hoque, and S. M. Allayear, "A machine learning approach to analyze and reduce features to a significant number for employee's turn over prediction model," in *Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 2*, Springer, 2019, pp. 142–159.

[83] A. Alamsyah and N. Salma, "A comparative study of employee churn prediction model," in *2018 4th International Conference on Science and Technology (ICST)*, IEEE, 2018, pp. 1–4.

[84] S. S. Alduayj and K. Rajpoot, "Predicting employee attrition using machine learning," in *2018 International Conference on Innovations in Information Technology (IIT)*, IEEE, 2018, pp. 93–98.

[85] R. Jain and A. Nayyar, "Predicting employee attrition using XGBoost machine learning approach," in *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, IEEE, 2018, pp. 113–120.

[86] S. N. Khera and Divya, "Predictive modelling of employee turnover in Indian IT industry using machine learning techniques," *Vision*, vol. 23, no. 1, pp. 12–21, 2018.

[87] R. S. Shankar, J. Rajanikanth, V. Sivaramaraju, and K. Murthy, "Prediction of employee attrition using datamining," in *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, IEEE, 2018, pp. 1–8.

[88] S. Yadav, A. Jain, and D. Singh, "Early prediction of employee attrition using data mining techniques," in *2018 IEEE 8th International Advance Computing Conference (IACC)*, IEEE, 2018, pp. 349–354.

[89] L. Alaskar, M. Crane, and M. Alduailij, "Employee turnover prediction using machine learning," in *Advances in Data Science, Cyber Security and IT Applications: First International Conference on Computing, ICC 2019, Riyadh, Saudi Arabia, December 10–12, 2019, Proceedings, Part I 1*, Springer, 2019, pp. 301–316.

[90] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Mathematical Problems in Engineering*, vol. 2019, 2019.

[91] N. Bhartiya, S. Jannu, P. Shukla, and R. Chapaneri, "Employee attrition prediction using classification models," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, IEEE, 2019, pp. 1–6.

[92] J. Vasa and K. Masrani, "Foreseeing employee attritions using diverse data mining strategies," *International Journal of Recent Technology and Engineering*, vol. 3, pp. 620–626, 2019.

[93] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2*, Springer, 2019, pp. 737–758.

[94] Y. Sun, F. Zhuang, H. Zhu, X. Song, Q. He, and H. Xiong, "The impact of person-organization fit on talent management: A structure-aware convolutional neural network approach," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 1625–1633.

[95] X. Cai, J. Shang, Z. Jin, *et al.*, "DBGE: Employee turnover prediction based on dynamic bipartite graph embedding," *IEEE Access*, vol. 8, pp. 10 390–10 402, 2020.

[96] N. El-Rayes, M. Fang, M. Smith, and S. M. Taylor, "Predicting employee attrition using tree-based models," *International Journal of Organizational Analysis*, 2020.

[97] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. William De Luca, "Predicting employee attrition using machine learning techniques," *Computers*, vol. 9, no. 4, p. 86, 2020.

[98] P. K. Jain, M. Jain, and R. Pamula, "Explaining and predicting employees' attrition: A machine learning approach," *SN Applied Sciences*, vol. 2, pp. 1–11, 2020.

[99] Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, and B. Qiang, "RFRSF: Employee turnover prediction based on random forests and survival analysis," in *Web Information Systems Engineering–WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20–24, 2020, Proceedings, Part II 21*, Springer, 2020, pp. 503–515.

[100] L. Liu, S. Akkineni, P. Story, and C. Davis, "Using HR analytics to support managerial decisions: A case study," in *Proceedings of the 2020 ACM Southeast Conference*, 2020, pp. 168–175.

[101] A. Mhatre, A. Mahalingam, M. Narayanan, A. Nair, and S. Jaju, "Predicting employee attrition along with identifying high risk employees using big data and machine learning," in *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, IEEE, 2020, pp. 269–276.

[102] F. Ozdemir, M. Coskun, C. Gezer, and V. C. Gungor, "Assessing employee attrition using classifications algorithms," in *Proceedings of the 2020 the 4th International Conference on Information System and Data Mining*, 2020, pp. 118–122.

[103] S. Al-Darraji, D. G. Honi, F. Fallucchi, A. I. Abdulsada, R. Giuliano, and H. A. Abdulmalik, "Employee attrition prediction using deep neural networks," *Computers*, vol. 10, no. 11, p. 141, 2021.

[104] N. B. Yahia, J. Hlel, and R. Colomo-Palacios, "From big data to deep data to support people analytics for employee attrition prediction," *IEEE Access*, vol. 9, pp. 60 447–60 458, 2021.

[105] R. Chakraborty, K. Mridha, R. N. Shaw, and A. Ghosh, "Study and prediction analysis of the employee turnover using machine learning approaches," in *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, IEEE, 2021, pp. 1–6.

[106] T. Juvitayapun, "Employee turnover prediction: The impact of employee event features on interpretable machine learning methods," in *2021 13th international conference on knowledge and smart technology (kst)*, IEEE, 2021, pp. 181–185.

[107] N. Jain, A. Tomar, and P. K. Jana, "A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning," *Journal of Intelligent Information Systems*, vol. 56, pp. 279–302, 2021.

[108] N. Mansor, N. S. Sani, and M. Aliff, "Machine learning for predicting employee attrition," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11, 2021.

[109] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. Hashemkhani Zolfani, "An improved machine learning-based employees attrition prediction framework with emphasis on feature selection," *Mathematics*, vol. 9, no. 11, p. 1226, 2021.

[110] M. Pratt, M. Boudhane, and S. Cakula, "Employee attrition estimation using random forest algorithm," *Baltic Journal of Modern Computing*, vol. 9, no. 1, pp. 49–66, 2021.

[111] P. R. Srivastava and P. Eachempati, "Intelligent employee retention system for attrition rate analysis and churn prediction: An ensemble machine learning and multi-criteria decision-making approach," *Journal of Global Information Management (JGIM)*, vol. 29, no. 6, pp. 1–29, 2021.

[112] Z. Tao, C. Wu, and S. Zhao, "Research on the prediction of employee turnover behavior and its interpretability," in *Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering*, 2021, pp. 760–767.

[113] X. Wang and J. Zhi, "A machine learning-based analytical framework for employee turnover prediction," *Journal of Management Analytics*, vol. 8, no. 3, pp. 351–370, 2021.

[114] A. B. Wild Ali, "Prediction of employee turn over using random forest classifier with intensive optimized pca algorithm," *Wireless Personal Communications*, vol. 119, no. 4, pp. 3365–3382, 2021.

[115] Q. Zhu, J. Shang, X. Cai, L. Jiang, F. Liu, and B. Qiang, "CoxRF: Employee turnover prediction based on survival analysis," in *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, IEEE, 2019, pp. 1123–1130.

[116] F. K. Alsheref, I. E. Fattoh, and W. M Ead, "Automated prediction of employee attrition using ensemble model based on machine learning algorithms," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[117] S. M. Arqawi, M. A. A. Rumman, E. A. Zitawi, *et al.*, "Predicting employee attrition and performance using deep learning," *Journal of Theoretical and Applied Information Technology*, vol. 100, no. 21, 2022.

[118] S. Bhatta, I. U. Zaman, N. Raisa, S. I. Fahim, and S. Momen, "Machine learning approach to predicting attrition among employees at work," in *Artificial Intelligence Trends in Systems: Proceedings of 11th Computer Science On-line Conference 2022, Vol. 2*, Springer, 2022, pp. 285–294.

[119] K. Naz, I. F. Siddiqui, J. Koo, M. A. Khan, and N. M. F. Qureshi, "Predictive modeling of employee churn analysis for IoT-enabled software industry," *Applied Sciences*, vol. 12, no. 20, p. 10 495, 2022.

[120] E. Pekel Ozmen and T. Ozcan, "A novel deep learning model based on convolutional neural networks for employee churn prediction," *Journal of Forecasting*, vol. 41, no. 3, pp. 539–550, 2022.

[121] M. Prathilothamai, A. Sri Sakthi Maheswari, A. Chandravadhana, and R. Goutham, "Efficient approach to employee attrition prediction by handling class imbalance," in *Advances in Computing and Data Sciences: 6th International Conference, ICACDS 2022, Kurnool, India, April 22–23, 2022, Revised Selected Papers, Part II*, Springer, 2022, pp. 263–277.

[122] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," *Applied Sciences*, vol. 12, no. 13, p. 6424, 2022.

[123] S. R. Seelam, K. H. Kumar, M. S. Supritha, G. Gnaneswar, and V. V. M. Reddy, "Comparative study of predictive models to estimate employee attrition," in *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2022, pp. 1602–1607.

[124] D. Chung, J. Yun, J. Lee, and Y. Jeon, "Predictive model of employee attrition based on stacking ensemble learning," *Expert Systems with Applications*, vol. 215, p. 119 364, 2023.

[125] F. Guerranti and G. M. Dimitri, "A comparison of machine learning approaches for predicting employee attrition," *Applied Sciences*, vol. 13, no. 1, p. 267, 2023.

[126] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[127] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explorations*, vol. 6, no. 1, pp. 20–29, 2004.

[128] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

[129] C. Bockel-Rickermann, T. Verdonck, and W. Verbeke, "Fraud analytics: A decade of research organizing challenges and solutions in the field," *Expert Systems with Applications*, p. 120 605, 2023.

[130] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[131] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[132] I. Wod, "Weight of evidence: A brief survey," *Bayesian Statistics*, vol. 2, pp. 249–270, 1985.

[133] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, 2009.

[134] R. Van Belle, B. Baesens, and J. De Weerdt, "CATCHM: A novel network-based credit card fraud detection method using node representation learning," *Decision Support Systems*, vol. 164, p. 113 866, 2023.

[135] Y. Hochberg, "A sharper Bonferroni procedure for multiple tests of significance," *Biometrika*, vol. 75, no. 4, pp. 800–802, Dec. 1988, ISSN: 0006-3444. DOI: 10.1093/biomet/75.4.800. eprint: https://academic.oup.com/biomet/article-pdf/75/4/800/1170595/75-4-800.pdf. [Online]. Available: https://doi.org/10.1093/biomet/75.4.800.

[136] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[137] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[138] F. Provost, "Machine learning from imbalanced data sets 101," 2008.

[139] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[140] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, 2008, pp. 1322–1328.

[141] R. A. Bauder, T. M. Khoshgoftaar, and T. Hasanin, "Data sampling approaches with severely imbalanced big data for medicare fraud detection," in *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2018, pp. 137–142.

[142] R. Bauder and T. Khoshgoftaar, "Medicare fraud detection using random forest with class imbalanced big data," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, IEEE, 2018, pp. 80–87.

[143] D. Martens, J. Vanthienen, W. Verbeke, and B. Baesens, "Performance of classification models from a user perspective," *Decision Support Systems*, vol. 51, no. 4, pp. 782–793, 2011.

[144] M. Craven and J. Shavlik, "Extracting tree-structured representations of trained networks," *Advances in Neural Information Processing Systems*, vol. 8, 1995.

[145] H. Blockeel, L. Devos, B. Frénay, G. Nanfack, and S. Nijssen, "Decision trees: From efficient prediction to responsible AI," *Frontiers in Artificial Intelligence*, vol. 6, 2023.

[146] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research*, vol. 20, no. 177, pp. 1–81, 2019.

[147] A. Benlian, M. Wiener, W. A. Cram, *et al.*, "Algorithmic management: Bright and dark sides, practical implications, and research opportunities," *Business & Information Systems Engineering*, vol. 64, no. 6, pp. 825–839, 2022.

[148] D. Martens and F. Provost, "Explaining data-driven document classifications," *MIS Quarterly*, vol. 38, no. 1, pp. 73–100, 2014.

[149] S. De Winne, E. Marescaux, L. Sels, I. Van Beveren, and S. Vanormelingen, "The impact of employee turnover and turnover volatility on labor productivity: A flexible non-linear approach," *The International Journal of Human Resource Management*, vol. 30, no. 21, pp. 3049–3079, 2019.

[150] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," en, *Big Data*, vol. 1, no. 1, pp. 51–59, Mar. 2013.

[151] C. Holsapple, A. Lee-Post, and R. Pakath, "A unified foundation for business analytics," *Decision Support Systems*, vol. 64, pp. 130–141, Aug. 2014.

[152] A. Behl, P. Dutta, S. Lessmann, Y. K. Dwivedi, and S. Kar, "A conceptual framework for the adoption of big data analytics by e-commerce startups: A case-based approach," *Information systems and e-business management*, vol. 17, pp. 285–318, 2019.

[153] D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence, and M. Stuart, "HR and analytics: Why HR is set to fail the big data challenge," en, *Human Resource Management Journal*, vol. 26, no. 1, pp. 1–11, Jan. 2016.

[154] J. H. Marler and J. W. Boudreau, "An evidence-based review of HR analytics," en, *The International Journal of Human Resource Management*, vol. 28, no. 1, pp. 3–26, Jan. 2017.

[155] T. Rasmussen and D. Ulrich, "Learning from practice: How HR analytics avoids being a management fad," *Organizational Dynamics*, vol. 44, no. 3, pp. 236–242, Jul. 2015.

[156] D. G. Collings, K. Mellahi, and W. F. Cascio, *The Oxford Handbook of Talent Management*, en. Oxford University Press, 2017, pp. 283–300.

[157] P. Cappelli, "Talent on demand: Managing talent in an uncertain age," *Harvard Business School Press, Boston, MA*, 2008.

[158] P. Osterman, "Choice of employment systems in internal labor markets," *Industrial Relations: A Journal of Economy and Society*, vol. 26, no. 1, pp. 46–67, 1987.

[159] P. G. O'Shea and K. E. Puente, "How is technology changing talent management?" en, in *The Oxford Handbook of Talent Management*, D. G. Collings, K. Mellahi, and W. F. Cascio, Eds., Oxford University Press, Sep. 2017.

[160] P. Rogiers, S. Viaene, and J. Leysen, "The digital future of internal staffing: A vision for transformational electronic human resource management," *Intelligent Systems in Accounting, Finance and Management*, vol. 27, no. 4, pp. 182–196, 2020.

[161] M. J. Belizón and S. Kieran, "Human resources analytics: A legitimacy process," *Human Resource Management Journal*, vol. 32, no. 3, pp. 603–630, 2022.

[162] A. De Vos, S. Jacobs, and M. Verbruggen, "Career transitions and employability," *Journal of Vocational Behavior*, vol. 126, p. 103 475, 2021.

[163] M. Verbruggen, R. De Cooman, and S. Vansteenkiste, "When and why are internal job transitions successful: Transition challenges, hindrances, and resources influencing motivation and retention through basic needs satisfaction," *Group & Organization Management*, vol. 40, no. 6, pp. 744–775, 2015.

[164] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decision Support Systems*, vol. 74, pp. 12–32, 2015.

[165] D. Wei, K. R. Varshney, and M. Wagman, "Optigrow: People analytics for job transfers," in *2015 IEEE International Congress on Big Data*, IEEE, 2015, pp. 535–542.

[166] C. de Ruijt and S. Bhulai, "Job recommender systems: A review," *arXiv preprint arXiv:2111.13576*, 2021.

[167] W. Van Der Aalst, "Process mining," *Communications of the ACM*, vol. 55, no. 8, pp. 76–83, Aug. 2012.

[168] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decision Support Systems*, vol. 74, pp. 12–32, Jun. 2015.

[169] B. Lika, K. Kolomvatsos, and S. Hadjiefthymiades, "Facing the cold start problem in recommender systems," *Expert systems with applications*, vol. 41, no. 4, pp. 2065–2073, 2014.

[170] J. Garcia-Arroyo and A. Osca, "Big data contributions to human resource management: A systematic review," *The International Journal of Human Resource Management*, vol. 32, no. 20, pp. 4337–4362, Nov. 2021.

[171] E. Ribes, K. Touahri, and B. Perthame, "Employee turnover prediction and retention policies design: A case study," *arXiv preprint arXiv:1707.01377*, 2017.

[172] P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *algorithms*, vol. 4, no. 5, p. C5, 2016.

[173] C.-F. Chien and L.-F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," *Expert Systems with Applications*, vol. 34, no. 1, pp. 280–290, Jan. 2008.

[174] K.-Y. Wang and H.-Y. Shun, "Applying back propagation neural networks in the prediction of management associate work retention for small and medium enterprises," *Universal Journal of Management*, vol. 4, no. 5, pp. 223–227, 2016.

[175] X.-L. Qu, "A decision tree applied to the grass-roots staffs' turnover problem—take cr group as an example," in *2015 IEEE International Conference on Grey Systems and Intelligent Services (GSIS)*, IEEE, 2015, pp. 378–382.

[176] E. Sikaroudi, A. Mohammad, R. Ghousi, and A. Sikaroudi, "A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing)," *International Journal of Industrial and Systems Engineering*, vol. 8, no. 4, pp. 106–121, 2015.

[177] X. Gui, Z. Hu, J. Zhang, and Y. Bao, "Assessing personal performance with M-SVMs," in *2014 Seventh International Joint Conference on Computational Sciences and Optimization*, Jul. 2014, pp. 598–601.

[178] C.-Y. Fan, P.-S. Fan, T.-Y. Chan, and S.-H. Chang, "Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8844–8851, Aug. 2012.

[179] Y.-M. Li, C.-Y. Lai, and C.-P. Kao, "Incorporate personality trait with support vector machine to acquire quality matching of personnel recruitment," in *4th international conference on business and information*, 2008, pp. 1–11.

[180] S. Mehta, R. Pimplikar, A. Singh, L. R. Varshney, and K. Visweswariah, "Efficient multifaceted screening of job applicants," in *Proceedings of the 16th International Conference on Extending Database Technology*, 2013, pp. 661–671.

[181] J. C. Sesil, *Applying Advanced Analytics to HR Management Decisions: Methods for Selection, Developing Incentives, and Improving Collaboration (Paperback)*. FT Press, 2013.

[182] E. Rombaut and M.-A. Guerry, "Predicting voluntary turnover through human resources database analysis," *Management Research Review*, vol. 41, no. 1, pp. 96–112, 2018.

[183] V. V. Saradhi and G. K. Palshikar, "Employee churn prediction," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1999–2006, Mar. 2011.

[184] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in *Proceedings of SAI intelligent systems conference*, Springer, 2018, pp. 737–758.

[185] S. H. Dolatabadi and F. Keynia, "Designing of customer and employee churn prediction model based on data mining method and neural predictor," en, in *2017 2nd International Conference on Computer and Communication Systems (ICCCS)*, Krakow, Poland: IEEE, Jul. 2017.

[186] S. N. Mishra, D. R. Lama, Y. Pal, *et al.*, "Human resource predictive analytics (hrpa) for hr management in organizations," *International Journal of Scientific & Technology Research*, vol. 5, no. 5, pp. 33–35, 2016.

[187] A. Levenson and A. Fink, "Human capital analytics: Too much data and analysis, not enough models and business insights," *Journal of Organizational Effectiveness: People and Performance*, vol. 4, no. 2, pp. 145–156, Jan. 2017.

[188] T. Peeters, J. Paauwe, and K. Van De Voorde, "People analytics effectiveness: Developing a framework," en, *Journal of Organizational Effectiveness: People and Performance*, vol. 7, no. 2, pp. 203–219, Jul. 2020.

[189] W. A. Schiemann, J. H. Seibert, and M. H. Blankenship, "Putting human capital analytics to work: Predicting and driving business success," *Hum. Resour. Manage.*, vol. 57, no. 3, pp. 795–807, May 2018.

[190] C. B.-G. Hila, "An ROI-based review of HR analytics: Practical implementation tools," *Personnel Review*, vol. 48, no. 6, pp. 1429–1448, Jan. 2019.

[191] D. B. Minbaeva, "Building credible human capital analytics for organizational competitive advantage," *Human Resource Management*, vol. 57, no. 3, pp. 701–713, May 2018.

[192] M. R. Edwards and K. Edwards, *Predictive HR analytics: Mastering the HR metric*. Kogan Page Publishers, 2019.

[193] J. Wang, Y. Zhang, C. Posse, and A. Bhasin, "Is it time for a career switch?" In *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1377–1388.

[194] B. Heap, A. Krzywicki, W. Wobcke, M. Bain, and P. Compton, "Combining career progression and profile matching in a job recommender system," in *PRICAI 2014: Trends in Artificial Intelligence: 13th Pacific Rim International Conference on Artificial Intelligence, Gold Coast, QLD, Australia, December 1-5, 2014. Proceedings 13*, Springer, 2014, pp. 396–408.

[195] C. Zhu, H. Zhu, H. Xiong, *et al.*, "Person-job fit: Adapting the right talent for the right job with joint representation learning," *ACM Transactions on Management Information Systems (TMIS)*, vol. 9, no. 3, pp. 1–17, 2018.

[196] M. Liu, J. Wang, K. Abdelfatah, and M. Korayem, "Tripartite vector representations for better job recommendation," *arXiv preprint arXiv:1907.12379*, 2019.

[197] S. Bian, X. Chen, W. X. Zhao, *et al.*, "Learning to match jobs with resumes from sparse interaction data using multi-view co-teaching network," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 65–74.

[198] C. Qin, H. Zhu, T. Xu, *et al.*, "An enhanced neural network approach to person-job fit in talent recruitment," *ACM Transactions on Information Systems (TOIS)*, vol. 38, no. 2, pp. 1–33, 2020.

[199] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, and A. Seveso, "Skills2job: A recommender system that encodes job offer embeddings on graph databases," *Applied Soft Computing*, vol. 101, p. 107 049, 2021.

[200] T. A.-O. Shaha and Y. Mourad, "A survey of job recommender systems," *International Journal of Physical Sciences*, vol. 7, no. 29, pp. 5127–5142, 2012.

[201] Z. Siting, H. Wenxing, Z. Ning, and Y. Fan, "Job recommender systems: A survey," in *2012 7th International Conference on Computer Science & Education (ICCSE)*, Jul. 2012, pp. 920–924.

[202] M. N. Freire and L. N. de Castro, "E-recruitment recommender systems: A systematic review," *Knowledge and Information Systems*, vol. 63, no. 1, pp. 1–20, 2021.

[203] P. K. Roy, S. S. Chowdhary, and R. Bhatia, "A machine learning approach for automation of resume recommendation system," *Procedia Computer Science*, vol. 167, pp. 2318–2327, 2020.

[204] A. Janusz, S. Stawicki, M. Drewniak, K. Ciebiera, D. Ślęzak, and K. Stencel, "How to match jobs and candidates - a recruitment support system based on feature engineering and advanced analytics," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, Springer International Publishing, 2018, pp. 503–514.

[205] L. Zhang, D. Zhou, H. Zhu, *et al.*, "Attentive heterogeneous graph embedding for job mobility prediction," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, ser. KDD '21, Virtual Event, Singapore: Association for Computing Machinery, Aug. 2021, pp. 2192–2201.

[206] J. Bollinger, D. Hardtke, and B. Martin, "Using social data for resume job matching," in *Proceedings of the 2012 workshop on Data-driven user behavioral modelling and mining from social media*, ser. DUBMMSM '12, Maui, Hawaii, USA: Association for Computing Machinery, Oct. 2012, pp. 27–30.

[207] B. J. Bret, H. J. Walker, J. B. Gilstrap, and P. H. Schwager, "Social media snooping on job applicants: The effects of unprofessional social media information on recruiter perceptions," *Personnel Review*, vol. 48, no. 5, pp. 1261–1280, Jan. 2019.

[208] R. Slovensky and W. H. Ross, "Should human resource managers use social media to screen job applicants? managerial and legal issues in the USA," *Info*, vol. 14, no. 1, pp. 55–69, Jan. 2012.

[209] J. Dastin, "Amazon scraps secret ai recruiting tool that showed bias against women," in *Ethics of Data and Analytics*, Auerbach Publications, 2018, pp. 296–299.

[210] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., Cham: Springer International Publishing, 2019, pp. 5–22.

[211] A. Stevens, J. De Smedt, and J. Peeperkorn, "Quantifying explainability in outcome-oriented predictive process monitoring," in *Process Mining Workshops*, J. Munoz-Gama and X. Lu, Eds., Cham: Springer International Publishing, 2022, pp. 194–206, ISBN: 978-3-030-98581-3.

[212] A. V. Konstantinov and L. V. Utkin, "Interpretable machine learning with an ensemble of gradient boosting machines," *Knowledge-Based Systems*, vol. 222, p. 106 993, 2021.

[213] S. De Vos, J. De Smedt, C. Wuytens, and W. Verbeke, "Leveraging process mining to optimize internal employee mobility strategies," in *Business Process Management Cases Vol. 3*, Springer, 2024 (forthcoming).

[214] W. Van Der Aalst, A. Adriansyah, A. K. A. De Medeiros, *et al.*, "Process mining manifesto," in *Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I 9*, Springer, 2012, pp. 169–194.

[215] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[216] H. Liu, Z. Hu, A. Mian, H. Tian, and X. Zhu, "A new user similarity model to improve the accuracy of collaborative filtering," *Knowledge-based systems*, vol. 56, pp. 156–166, 2014.

[217] A. A. Amer, H. I. Abdalla, and L. Nguyen, "Enhancing recommendation systems performance using highly-effective similarity measures," *Knowledge-Based Systems*, vol. 217, p. 106 842, 2021.

[218] H. Khojamli and J. Razmara, "Survey of similarity functions on neighborhood-based collaborative filtering," *Expert Systems with Applications*, vol. 185, p. 115 482, 2021.

[219] A. Paterek, "Improving regularized singular value decomposition for collaborative filtering," in *Proceedings of KDD cup and workshop*, vol. 2007, 2007, pp. 5–8.

[220] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Major components of the gravity recommendation system," *Acm Sigkdd Explorations Newsletter*, vol. 9, no. 2, pp. 80–83, 2007.

[221] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ser. WSDM '11, Hong Kong, China: Association for Computing Machinery, Feb. 2011, pp. 287–296.

[222] M. Meire, M. Ballings, and D. Van den Poel, "The added value of social media data in b2b customer acquisition systems: A real-life experiment," *Decision Support Systems*, vol. 104, pp. 26–37, 2017.

[223] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Communications of the ACM*, vol. 29, no. 12, pp. 1213–1228, Dec. 1986.

[224] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," *California Management Review*, vol. 61, no. 4, pp. 15–42, 2019.

[225] A. J. Bowlus, "Matching workers and jobs: Cyclical fluctuations in match quality," *Journal of Labor Economics*, vol. 13, no. 2, pp. 335–350, 1995.

[226] B. Jovanovic, "Firm-specific capital and turnover," *Journal of political economy*, vol. 87, no. 6, pp. 1246–1260, 1979.

[227] G. A. Akerlof, A. K. Rose, J. L. Yellen, L. Ball, and R. E. Hall, "Job switching and job satisfaction in the us labor market," *Brookings papers on economic activity*, vol. 1988, no. 2, pp. 495–594, 1988.

[228] M. J. Van der Laan and J. M. Robins, *Unified methods for censored longitudinal data and causality*. Springer, 2003, vol. 5.

[229] S. Geuens, K. Coussement, and K. W. De Bock, "A framework for configuring collaborative filtering-based recommendations derived from purchase data," *European Journal of Operational Research*, vol. 265, no. 1, pp. 208–218, 2018.

[230] M. Bogaert, J. Lootens, D. Van den Poel, and M. Ballings, "Evaluating multi-label classifiers and recommender systems in the financial service sector," *European Journal of Operational Research*, vol. 279, no. 2, pp. 620–634, 2019.

[231] D. Lemire and A. Maclachlan, "Slope one predictors for online rating-based collaborative filtering," in *Proceedings of the 2005 SIAM International Conference on Data Mining*, SIAM, 2005, pp. 471–475.

[232] N. Hug, "Surprise: A python library for recommender systems," *Journal of Open Source Software*, vol. 5, no. 52, p. 2174, 2020. DOI: 10.21105/joss.02174. [Online]. Available: https://doi.org/10.21105/joss.02174.

[233] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, pp. 65–70, 1979.

[234] Y. Koren and R. Bell, "Advances in collaborative filtering," *Recommender Systems Handbook*, pp. 77–118, 2015.

[235] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[236] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy: Evaluating recommender systems by coverage and serendipity," in *Proceedings of the fourth ACM conference on Recommender systems*, ser. RecSys '10, Barcelona, Spain: Association for Computing Machinery, Sep. 2010, pp. 257–260.

[237] L. Iaquinta, M. de Gemmis, P. Lops, G. Semeraro, M. Filannino, and P. Molino, "Introducing serendipity in a Content-Based recommender system," in *2008 Eighth International Conference on Hybrid Intelligent Systems*, Sep. 2008, pp. 168–173.

[238] D. Kotkov, S. Wang, and J. Veijalainen, "A survey of serendipity in recommender systems," *Knowledge-Based Systems*, vol. 111, pp. 180–192, Nov. 2016.

[239] B. M. Marlin and R. S. Zemel, "Collaborative prediction and ranking with non-random missing data," in *Proceedings of the third ACM conference on Recommender systems*, ser. RecSys '09, New York, New York, USA: Association for Computing Machinery, Oct. 2009, pp. 5–12.

[240] E. Bareinboim and J. Pearl, "Controlling selection bias in causal inference," Proceedings of Machine Learning Research, vol. 22, N. D. Lawrence and M. Girolami, Eds., pp. 100–108, 2012.

[241] W. Verbeke, D. Olaya, J. Berrevoets, S. Verboven, and S. Maldonado, "The foundations of cost-sensitive causal classification," Jul. 2020. arXiv: `2007.12582 [cs.LG]`.

[242] G. Petrides, D. Moldovan, L. Coenen, T. Guns, and W. Verbeke, "Cost-sensitive learning for profit-driven credit scoring," *J. Oper. Res. Soc.*, vol. 73, no. 2, pp. 338–350, 2022.

[243] S. Lessmann, J. Haupt, K. Coussement, and K. W. De Bock, "Targeting customers for profit: An ensemble learning framework to support marketing decision-making," *Information Sciences*, vol. 557, pp. 286–301, 2021.

[244] G. Petrides and W. Verbeke, "Cost-sensitive ensemble learning: A unifying framework," *Data Mining and Knowledge Discovery*, vol. 36, no. 1, pp. 1–28, 2022.

[245] T. Vanderschueren, T. Verdonck, B. Baesens, and W. Verbeke, "Predict-then-optimize or predict-and-optimize? an empirical evaluation of cost-sensitive learning strategies," *Information Sciences*, vol. 594, pp. 400–415, 2022.

[246] U. Brefeld, P. Geibel, and F. Wysotzki, "Support vector machines with example dependent costs," in *European Conference on Machine Learning*, Springer, 2003, pp. 23–34.

[247] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "Adacost: Misclassification cost-sensitive boosting," in *Icml*, Citeseer, vol. 99, 1999, pp. 97–105.

[248] Y. Zelenkov, "Example-dependent cost-sensitive adaptive boosting," *Expert Systems with Applications*, vol. 135, pp. 71–82, 2019.

[249] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5916–5923, 2013.

[250] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive decision trees," *Expert Systems with Applications*, vol. 42, no. 19, pp. 6609–6619, 2015.

[251] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive logistic regression for credit scoring," in *2014 13th International Conference on Machine Learning and Applications*, 2014, pp. 263–269. DOI: 10.1109/ICMLA.2014.48.

[252] A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, "Feature engineering strategies for credit card fraud detection," *Expert Systems with Applications*, vol. 51, pp. 134–142, 2016.

[253] P. J. Huber and E. Ronchetti, "Robust statistics. 2nd john wiley & sons," *Hoboken, NJ*, vol. 2, 2009.

[254] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[255] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*. John Wiley & Sons, Inc., 1987, ISBN: 9780471725381.

[256] E. Cantoni and E. Ronchetti, "Robust inference for generalized linear models," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 1022–1030, 2001.

[257] A. Bergesio and V. J. Yohai, "Projection estimators for generalized linear models," *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 661–671, 2011.

[258] M. Valdora and V. J. Yohai, "Robust estimators for generalized linear models," *Journal of Statistical Planning and Inference*, vol. 146, pp. 31–48, 2014.

[259] A. Ghosh and A. Basu, "Robust estimation in generalized linear models: The density power divergence approach," *TEST*, vol. 25, no. 2, pp. 269–290, 2016.

[260] N. Štefelová, A. Alfons, J. Palarea-Albaladejo, P. Filzmoser, and K. Hron, "Robust regression with compositional covariates including cellwise outliers," *Adv. Data Anal. Classif.*, vol. 15, no. 4, pp. 869–909, Dec. 2021.

[261] H. R. Künsch, L. A. Stefanski, and R. J. Carroll, "Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 460–466, 1989.

[262] S. Morgenthaler, "Least-absolute-deviations fits for generalized linear models," *Biometrika*, vol. 79, no. 4, pp. 747–754, 1992.

[263] R. J. Carroll and S. Pederson, "On robustness in the logistic regression model," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 55, no. 3, pp. 693–706, 1993.

[264] A. M. Bianco and V. J. Yohai, "Robust estimation in the logistic regression model," in *Robust statistics, data analysis, and computer intensive methods*, Springer, 1996, pp. 17–34.

[265] C. Croux and G. Haesbroeck, "Implementing the bianco and yohai estimator for logistic regression," *Computational statistics & data analysis*, vol. 44, no. 1-2, pp. 273–295, 2003.

[266] H. D. Bondell, "Minimum distance estimation for the logistic regression model," *Biometrika*, vol. 92, no. 3, pp. 724–731, 2005.

[267] H. D. Bondell, "A characteristic function approach to the biased sampling model, with application to robust logistic regression," *Journal of Statistical Planning and Inference*, vol. 138, no. 3, pp. 742–755, 2008.

[268] G. S. Monti and P. Filzmoser, "Robust logistic zero-sum regression for microbiome compositional data," *Adv. Data Anal. Classif.*, Sep. 2021.

[269] S. Hosseinian and S. Morgenthaler, "Robust binary regression," *Journal of statistical planning and inference*, vol. 141, no. 4, pp. 1497–1509, 2011.

[270] N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–8. DOI: 10.1109/IJCNN.2010.5596486.

[271] G. I. W. Claude Sammut, *Encyclopedia of Machine Learning and Data Mining*. Springer US, 2017, ISBN: 9781489976864.

[272] C. Whitrow, D. J. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," *Data mining and knowledge discovery*, vol. 18, no. 1, pp. 30–55, 2009.

[273] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964, ISSN: 00034851. [Online]. Available: http://www.jstor.org/stable/2238020.

[274] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," en, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 1, no. 1, pp. 73–79, Jan. 2011.

[275] M. L. G. .-. ULB, *Anonymized credit card transactions labeled as fraudulent or genuine*, https://www.kaggle.com/mlg-ulb/creditcardfraud, 2018. (visited on 05/17/2021).

[276] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, 2011.

[277] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

[278] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, 2015.

[279] V. Van Vlasselaer, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "GOTCHA! Network-based fraud detection for social security fraud," *Management Science*, vol. 63, no. 9, pp. 3090–3110, 2017.

[280] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016.

[281] M. Lindholm, R. Richman, A. Tsanakas, and M. V. Wüthrich, "Discrimination-free insurance pricing," *ASTIN Bulletin: The Journal of the IAA*, vol. 52, no. 1, pp. 55–89, 2022.

[282] E. W. Frees and F. Huang, "The discriminating (pricing) actuary," *North American Actuarial Journal*, vol. 27, no. 1, pp. 2–24, 2023.

[283] X. Xin and F. Huang, "Antidiscrimination insurance pricing: Regulations, fairness criteria, and models," *North American Actuarial Journal*, vol. 28, no. 2, pp. 285–319, 2024.

[284] European Union, "Guidelines on the application of council directive 2004/113/ec to insurance, in the light of the judgment of the court of justice of the european union in case c-236/09 (test-achats)," *Official Journal of the European Union*, vol. C11, pp. 1–11, 2012. [Online]. Available: `https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:C:2012:011:0001:0011:EN:PDF`.

[285] C. Hurlin, C. Pérignon, and S. Saurin, "The fairness of credit scoring models," *Management Science*, 2024.

[286] A. C. B. Garcia, M. G. P. Garcia, and R. Rigobon, "Algorithmic discrimination in the credit domain: What do we know about it?" *AI & SOCIETY*, vol. 39, no. 4, pp. 2059–2098, 2024.

[287] F. A. Khan and J. Stoyanovich, "The unbearable weight of massive privilege: Revisiting bias-variance trade-offs in the context of fair prediction," *arXiv preprint: 2302.08704*, 2023.

[288] F. A. Khan, D. Herasymuk, and J. Stoyanovich, "On fairness and stability: Is estimator variance a friend or a foe?" *arXiv preprint: 2302.04525*, 2023.

[289] X. Han, Z. Jiang, H. Jin, *et al.*, "Retiring DeltaDP: New distribution-level metrics for demographic parity," *Transactions on Machine Learning Research*, 2023, ISSN: 2835-8856.

[290] J. Peeperkorn and S. De Vos, "Achieving group fairness through independence in predictive process monitoring," *arXiv preprint arXiv:2412.04914*, 2024.

[291] Ö. G. Ali and U. Arıtürk, "Dynamic churn prediction framework with more effective use of rare event data: The case of private banking," *Expert Systems with Applications*, vol. 41, no. 17, pp. 7889–7903, 2014.

[292] H. Weerts, L. Royakkers, and M. Pechenizkiy, "Does the end justify the means? On the moral justification of fairness-aware machine learning," *arXiv*, vol. 2022, pp. 2202–08 536, 2022.

[293] N. Kozodoi, J. Jacob, and S. Lessmann, "Fairness in credit scoring: Assessment, implementation and profit implications," *European Journal of Operational Research*, vol. 297, no. 3, pp. 1083–1094, 2022.

[294] S. Fazelpour and D. Danks, "Algorithmic bias: Senses, sources, solutions," *Philosophy Compass*, vol. 16, no. 8, e12760, 2021.

[295] A. E. Prince and D. Schwarcz, "Proxy discrimination in the age of artificial intelligence and big data," *Iowa Law Review*, vol. 105, pp. 1257–1318, 2019.

[296] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, no. 3, pp. 671–732, 2016.

[297] M. Lindholm, R. Richman, A. Tsanakas, and M. V. Wüthrich, "What is fair? Proxy discrimination vs. demographic disparities in insurance pricing," *Scandinavian Actuarial Journal*, pp. 1–36, 2024.

[298] J. Adams-Prassl, R. Binns, and A. Kelly-Lyth, "Directly discriminatory algorithms," *The Modern Law Review*, vol. 86, no. 1, pp. 144–175, 2023.

[299] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012.

[300] W. Fleisher, "What's fair about individual fairness?" In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 480–490.

[301] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017.

[302] R. Binns, "On the apparent conflict between individual and group fairness," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2020.

[303] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint:1609.05807*, 2016.

[304] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[305] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[306] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[307] A. Algaba, C. Mazijn, C. Prunkl, J. Danckaert, and V. Ginis, "LUCID-GAN: Conditional generative models to locate unfairness," in *World Conference on Explainable Artificial Intelligence*, Springer, 2023, pp. 346–367.

[308] C. Mazijn, C. Prunkl, A. Algaba, J. Danckaert, and V. Ginis, "LUCID: Exposing algorithmic bias through inverse design," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 14 391–14 399.

[309] M. De-Arteaga, S. Feuerriegel, and M. Saar-Tsechansky, "Algorithmic fairness in business analytics: Directions for research and practice," *Production and Operations Management*, vol. 31, no. 10, pp. 3749–3770, 2022.

[310] K. W. De Bock, K. Coussement, A. De Caigny, *et al.*, "Explainable ai for operational research: A defining framework, methods, applications, and a research agenda," *European Journal of Operational Research*, vol. 317, no. 2, pp. 249–272, 2024.

[311] C. Mazijn, J. Danckaert, and V. Ginis, "How do the score distributions of subpopulations influence fairness notions?" In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[312] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in *Conference on Fairness, Accountability and Transparency*, 2018.

[313] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im)possibility of fairness," *arXiv preprint: 1609.07236*, 2016.

[314] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 259–268.

[315] S. Radovanović, G. Savić, B. Delibašić, and M. Suknović, "FairDEA—removing disparate impact from efficiency scores," *European Journal of Operational Research*, vol. 301, no. 3, pp. 1088–1098, 2022.

[316] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *Journal of Machine Learning Research*, vol. 20, no. 75, pp. 1–42, 2019.

[317] M. Hanson, G. Lewkowicz, and S. Verboven, "Engineering the law-machine learning translation problem: Developing legally aligned models," *arXiv preprint arXiv:2504.16969*, 2025.

[318] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney, "Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing," in *International conference on machine learning*, PMLR, 2020, pp. 2803–2813.

[319] M. Wick, J.-B. Tristan, *et al.*, "Unlocking fairness: A trade-off revisited," *Advances in neural information processing systems*, vol. 32, 2019.

[320] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *International Conference on Machine Learning*, PMLR, 2017, pp. 3076–3085.

[321] E. Ustinova and V. Lempitsky, "Learning deep embeddings with histogram loss," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[322] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 339–348.

[323] A. Coston, K. N. Ramamurthy, D. Wei, *et al.*, "Fair transfer learning with missing protected attributes," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 91–98.

[324] C. Villani, *Optimal transport: Old and new*. Springer, 2008, vol. 338.

[325] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[326] M. Cuturi and A. Doucet, "Fast computation of wasserstein barycenters," in *International Conference on Machine Learning*, PMLR, 2014, pp. 685–693.

[327] F. Devriendt, J. Berrevoets, and W. Verbeke, "Why you should stop predicting customer churn and start using uplift models," *Information Sciences*, vol. 548, pp. 497–515, 2021.

[328] W. Verbeke, D. Olaya, M.-A. Guerry, and J. Van Belle, "To do or not to do? Cost-sensitive causal classification with individual treatment effect estimates," *European Journal of Operational Research*, vol. 305, no. 2, pp. 838–852, 2023.

[329] T. Vanderschueren, B. Baesens, T. Verdonck, and W. Verbeke, "A new perspective on classification: Optimally allocating limited resources to uncertain tasks," *Decision Support Systems*, vol. 179, p. 114 151, 2024.

[330] F. Devriendt, J. Van Belle, T. Guns, and W. Verbeke, "Learning to rank for uplift modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4888–4904, 2020.

[331] S. Goethals and T. Calders, "Reranking individuals: The effect of fair classification within-groups," *arXiv*, pp. 1–16, 2024.

[332] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.

[333] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 2009. DOI: 10.1214/09-SS057.

[334] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[335] C. Fernández-Loría and F. Provost, "Causal decision making and causal effect estimation are not the same… and why it matters," *INFORMS Journal on Data Science*, vol. 1, no. 1, pp. 4–16, 2022.

[336] K. Hirano and G. Imbens, *The propensity score with continuous treatments. applied bayesian modeling and causal inference from incomplete-data perspectives*, 2004.

[337] T. Holland-Letz and A. Kopp-Schneider, "Optimal experimental designs for dose–response studies with continuous endpoints," *Archives of toxicology*, vol. 89, pp. 2059–2068, 2015.

[338] I. Bica, J. Jordon, and M. van der Schaar, "Estimating the effects of continuous-valued interventions using generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 434–16 445, 2020.

[339] P. Schwab, L. Linhardt, S. Bauer, J. M. Buhmann, and W. Karlen, "Learning counterfactual representations for estimating individual dose-response curves," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 5612–5619.

[340] L. Nie, M. Ye, Q. Liu, and D. Nicolae, "Vcnet and functional targeted regularization for learning causal effects of continuous treatments," *arXiv preprint arXiv:2103.07861*, 2021.

[341] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[342] X. Han, Z. Jiang, H. Jin, *et al.*, "Retiring deltadp: New distribution-level metrics for demographic parity," *Transactions on Machine Learning Research*, 2023, ISSN: 2835-8856.

[343] D. Frauen, V. Melnychuk, and S. Feuerriegel, "Fair off-policy learning from observational data," *arXiv preprint arXiv:2303.08516*, 2023.

[344] W. Zhang, J. Li, and L. Liu, "A unified survey of treatment effect heterogeneity modelling and uplift modelling," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–36, 2021.

[345] P. Gutierrez and J.-Y. Gérardy, "Causal inference and uplift modelling: A review of the literature," in *International conference on predictive applications and APIs*, PMLR, 2017, pp. 1–13.

[346] F. Devriendt, J. Van Belle, T. Guns, and W. Verbeke, "Learning to rank for uplift modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4888–4904, 2020.

[347] D. Olaya, K. Coussement, and W. Verbeke, "A survey and benchmarking study of multitreatment uplift modeling," *Data Mining and Knowledge Discovery*, vol. 34, pp. 273–308, 2020.

[348] T. Hatt and S. Feuerriegel, "Sequential deconfounding for causal inference with unobserved confounders," in *Causal Learning and Reasoning*, PMLR, 2024, pp. 934–956.

[349] D. Jones, D. Molitor, and J. Reif, "What do workplace wellness programs do? evidence from the illinois workplace wellness study," *The Quarterly Journal of Economics*, vol. 134, no. 4, pp. 1747–1791, 2019.

[350] K. Verstraete, I. Gyselinck, H. Huts, *et al.*, "Estimating individual treatment effects on copd exacerbations by causal machine learning on randomised controlled trials," *Thorax*, vol. 78, no. 10, pp. 983–989, 2023, ISSN: 0040-6376. DOI: 10.1136/thorax-2022-219382. eprint: https://thorax.bmj.com/content/78/10/983.full.pdf. [Online]. Available: https://thorax.bmj.com/content/78/10/983.

[351] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[352] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *International conference on machine learning*, PMLR, 2016, pp. 3020–3029.

[353] C. Bockel-Rickermann, T. Vanderschueren, T. Verdonck, and W. Verbeke, "Sources of gain: Decomposing performance in conditional average dose response estimation," *arXiv preprint arXiv:2406.08206*, 2024.

[354] Y.-F. Zhang, H. Zhang, Z. C. Lipton, L. E. Li, and E. P. Xing, "Exploring transformer backbones for heterogeneous treatment effect estimation," *arXiv preprint arXiv:2202.01336*, 2022.

[355] C. Bockel-Rickermann, T. Vanderschueren, J. Berrevoets, T. Verdonck, and W. Verbeke, "Learning continuous-valued treatment effects through representation balancing," *arXiv preprint arXiv:2309.03731*, 2023.

[356] H. Zhou, S. Li, G. Jiang, J. Zheng, and D. Wang, "Direct heterogeneous causal learning for resource allocation problems in marketing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 5446–5454.

[357] R. M. Gubela, S. Lessmann, and S. Jaroszewicz, "Response transformation and profit decomposition for revenue uplift modeling," *European Journal of Operational Research*, vol. 283, no. 2, pp. 647–661, 2020.

[358] A. Lemmens and S. Gupta, "Managing churn to maximize profits," *Marketing Science*, vol. 39, no. 5, pp. 956–973, 2020.

[359] R. M. Gubela and S. Lessmann, "Uplift modeling with value-driven evaluation metrics," *Decision Support Systems*, vol. 150, p. 113 648, 2021.

[360] W. Verbeke, D. Olaya, M.-A. Guerry, and J. Van Belle, "To do or not to do? cost-sensitive causal classification with individual treatment effect estimates," *European Journal of Operational Research*, vol. 305, no. 2, pp. 838–852, 2023.

[361] J. Haupt and S. Lessmann, "Targeting customers under response-dependent costs," *European Journal of Operational Research*, vol. 297, no. 1, pp. 369–379, 2022.

[362] A. Betlei, E. Diemert, and M.-R. Amini, "Uplift modeling with generalization guarantees," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 55–65.

[363] T. Vanderschueren, W. Verbeke, F. Moraes, and H. M. Proença, "Metalearners for ranking treatment effects," *arXiv preprint arXiv:2405.02183*, 2024.

[364] M. Jaskowski and S. Jaroszewicz, "Uplift modeling for clinical trial data," in *ICML workshop on clinical data analysis*, vol. 46, 2012, pp. 79–95.

[365] A. N. Elmachtoub and P. Grigas, "Smart "predict, then optimize"," *Management Science*, vol. 68, no. 1, pp. 9–26, 2022.

[366] J. Mandi, J. Kotary, S. Berden, *et al.*, "Decision-focused learning: Foundations, state of the art, benchmark and future opportunities," *arXiv preprint arXiv:2307.13565*, 2023.

[367] B. Zhan, C. Liu, Y. Li, and C. Wu, "Weighted doubly robust learning: An uplift modeling technique for estimating mixed treatments' effect," *Decision Support Systems*, vol. 176, p. 114 060, 2024.

[368] A. Chouldechova and A. Roth, "A snapshot of the frontiers of fairness in machine learning," *Communications of the ACM*, vol. 63, no. 5, pp. 82–89, 2020.

[369] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, vol. 4, no. 1, eaao5580, 2018.

[370] K. Makhlouf, S. Zhioua, and C. Palamidessi, "On the applicability of machine learning fairness notions," *ACM SIGKDD Explorations Newsletter*, vol. 23, no. 1, pp. 14–23, 2021.

[371] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[372] T. Scantamburlo, J. Baumann, and C. Heitz, "On prediction-modelers and decision-makers: Why fairness requires more than a fair prediction model," *AI & SOCIETY*, pp. 1–17, 2024.

[373] R. Nabi, D. Malinsky, and I. Shpitser, "Learning optimal fair policies," in *International Conference on Machine Learning*, PMLR, 2019, pp. 4674–4682.

[374] T. Verbraken, W. Verbeke, and B. Baesens, "A novel profit maximizing metric for measuring classification performance of customer churn prediction models," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 5, pp. 961–973, 2012.

[375] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive logistic regression for credit scoring," in *2014 13th International conference on machine learning and applications*, IEEE, 2014, pp. 263–269.

[376] C. O. Vasquez, J. De Weerdt, and S. vanden Broucke, "The hidden cost of fraud: An instance-dependent cost-sensitive approach for positive and unlabeled learning," in *Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications*, PMLR, 2022, pp. 53–67.

[377] K. W. De Bock, K. Coussement, and S. Lessmann, "Cost-sensitive business failure prediction when misclassification costs are uncertain: A heterogeneous ensemble selection approach," *European Journal of Operational Research*, vol. 285, no. 2, pp. 612–630, 2020.

[378] T. Vanderschueren, R. Boute, T. Verdonck, B. Baesens, and W. Verbeke, "Optimizing the preventive maintenance frequency with causal machine learning," *International Journal of Production Economics*, vol. 258, p. 108 798, 2023.

[379] D. B. Rubin, "Direct and indirect causal effects via potential outcomes," *Scandinavian Journal of Statistics*, vol. 31, no. 2, pp. 161–170, 2004.

[380] G. W. Imbens, "The role of the propensity score in estimating dose-response functions," *Biometrika*, vol. 87, no. 3, pp. 706–710, 2000.

[381] M. Lechner, *Identification and estimation of causal effects of multiple treatments under the conditional independence assumption*. Springer, 2001.

[382] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the national academy of sciences*, vol. 116, no. 10, pp. 4156–4165, 2019.

[383] G. B. Dantzig, "Discrete-variable extremum problems," *Operations Research*, vol. 5, no. 2, pp. 266–288, 1957.

[384] J. Brooks-Gunn, F.-r. Liaw, and P. K. Klebanov, "Effects of early intervention on cognitive function of low birth weight preterm infants," *The Journal of pediatrics*, vol. 120, no. 3, pp. 350–359, 1992.

[385] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.

[386] R. Silva, "Observational-interventional priors for dose-response learning," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016.

[387] J. Berrevoets, S. Verboven, and W. Verbeke, "Treatment effect optimisation in dynamic environments," *Journal of Causal Inference*, vol. 10, no. 1, pp. 106–122, 2022.

[388] Z. Zhao, Y. Bai, R. Xiong, *et al.*, "Learning individual treatment effects under heterogeneous interference in networks," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 8, pp. 1–21, 2024.

[389] J. Berrevoets, J. Jordon, I. Bica, M. van der Schaar, *et al.*, "Organite: Optimal transplant donor organ offering using an individual treatment effect," *Advances in neural information processing systems*, vol. 33, pp. 20 037–20 050, 2020.

[390] T. Vanderschueren, J. Berrevoets, and W. Verbeke, "Noflite: Learning to predict individual treatment effect distributions," *Transactions on Machine Learning Research*, 2023.

[391] M. Schröder, D. Frauen, J. Schweisthal, K. Heß, V. Melnychuk, and S. Feuerriegel, "Conformal prediction for causal effects of continuous treatments," *arXiv preprint arXiv:2407.03094*, 2024.

[392] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management science*, vol. 64, no. 3, pp. 1155–1170, 2018.

[393] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 625–632.

[394] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[395] N. Hollmann, S. Müller, L. Purucker, *et al.*, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, no. 8045, pp. 319–326, 2025.

[396] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[397] B. Neal, "Introduction to causal inference," *Course Lecture Notes (draft)*, vol. 132, 2020.

[398] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.

Appendices

A

Predicting employee turnover: Scoping and benchmarking the state-of-the-art

A.1 Search query

Table A.1: Complete search queries for the Scopus and WoS databases

Database	Query
Scopus	TITLE-ABS-KEY ((prediction OR predicting OR forecasting OR prognosis) AND (employee OR worker OR laborer OR jobholder) AND (turnover OR attrition OR churn OR departure) AND (analytics OR data OR "machine learning")) AND PUBYEAR > 2008 AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "cp")) AND (LIMIT-TO (SUBJAREA, "COMP") OR LIMIT-TO (SUBJAREA , "BUSI") OR LIMIT-TO (SUBJAREA, "ENGI") OR LIMIT-TO (SUBJAREA, "MATH") OR LIMIT-TO (SUBJAREA , "ECON")) AND (LIMIT-TO (LANGUAGE, "English"))
WoS	TS=((prediction OR predicting OR forecasting OR prognosis) AND (employee OR worker OR laborer OR jobholder) AND (turnover OR attrition OR churn OR departure) AND (analytics OR data OR "machine learning")) and Article OR Proceeding Paper (Document Types) and Business OR Computer Science Information Systems OR Economics or Management OR Computer Science Interdisciplinary Applications OR Computer Science Software Engineering OR Engineering Industrial OR Computer Science Artificial Intelligence OR Computer Science Theory Methods OR Engineering Electrical Electronic OR Engineering Multidisciplinary (Web of Science Categories) AND English (Languages) AND PY > 2008

A.2 Established classifiers and their corresponding studies

Table A.2: Overview of established classifiers with corresponding studies.

	Classifier	Abbr.	Num. Studies	Studies
Individual classifiers	Artificial Neural Networks	ann	12	[78], [102]–[104], [111], [113], [116], [117], [119], [120], [124], [125]
	Decision Tree	dt	41	[64], [73], [75], [76], [78]–[83], [87]–[102], [104], [107]–[110], [112]–[115], [118]–[122], [125]
	K-Nearest Neighbors	knn	22	[74], [78]–[80], [82], [84], [87], [91], [93], [97], [100]–[102], [105], [109], [110], [113], [114], [118], [120], [121], [123]
	Linear Discriminant Analysis	lda	4	[73], [93], [102], [113]
	Logistic Regression	lr	33	[48], [73], [78], [80], [87]–[90], [92]–[97], [99]–[102], [104]–[107], [109], [110], [113]–[115], [119], [121]–[125]
	Naïve Bayes (Bayesian/Gaussian)	bnb /gnb	25	[41], [75]–[81], [83], [87], [89], [91], [93], [95], [97], [99], [101], [102], [105], [109], [110], [113], [115], [120], [125]
	Quadratic Discriminant Analysis	qda	1	[113]
	Support Vector Machine	svm	30	[41], [78]–[80], [82], [84], [86]–[89], [91], [93], [94], [97], [98], [101], [102], [104], [105], [107]–[110], [113], [115], [118]–[120], [122], [124]
Ensembles	AdaBoost	ab	7	[88], [89], [92], [102], [112], [113], [117]
	Gradient Boosting	gb	9	[93], [95], [100], [105], [106], [111], [113], [116], [117]
	LightGBM	lgbm	3	[96], [113], [117]
	Random Forest	rf	36	[41], [77]–[80], [82]–[84], [88], [90]–[100], [102], [105]–[107], [109], [111]–[118], [123]–[125]
	Extreme Gradient Boosting	xgb	12	[85], [93], [95], [99], [101], [104], [106], [107], [112], [113], [115], [124]

A.3 Hyperparameter search space

Table A.3: Hyperparameter search space

Method	Hyperparameter	Values	#Models
ab	n estimators	{20, 50, 100, 200}	24
	learning rate	{0.01, 0.1, 1}	
	algorithm	{'SAMME', 'SAMME.R'}	
ann	solver	{'adam', 'lbfgs'}	720
	alpha	{0.01, 0.1, 1}	
	n hidden layers	{1, 2, 3}	
	n neurons	{10, 20, 50, 100, 200}	
	activation	{'relu', 'logistic'}	
	max iter	{100, 200, 500, 1000}	
bnb	alpha	{0.01, 0.1, 1}	6
	fit prior	{True, False}	
dt	criterion	{'gini', 'entropy'}	72
	max depth	{5, 10, 20, 50}	
	min samples split	{ 2, 5, 10}	
	min samples leaf	{1, 5, 10}	
gb	n estimators	{20, 50, 100, 200}	432
	learning rate	{0.01, 0.1, 1}	
	max depth	{5, 10, 20, 50}	
	min samples split	{2, 5, 10}	
	min samples leaf	{1, 5, 10}	
gnb	var smoothing	{1e-9, 1e-8, 1e-7, 1e-6, 1e-5}	5
knn	n neighbors	{1, 5, 10, 50}	48
	weights	{'uniform', 'distance'}	
	leaf size	{10, 30, 50}	
	p	{1, 2}	
lda	shrinkage	{None, 'auto', 0.1, 0.5, 1.0}	20
	n components	{None, 2, 5, 10}	
lgbm	learning rate	{0.01, 0.1, 1}	48
	n estimators	{20, 50, 100, 200}	
	max depth	{5, 10, 20, 50}	
lr	penalty	{'none','l1', 'l2'}	48
	c	{0.001, 0.01, 0.1, 1}	
	max iter	{50, 100, 500, 1000}	
qda	reg param	{0.0, 0.01, 0.1, 0.5, 1.0}	30
	store covariance	{True, False}	
	tol	{1e-5, 1e-4, 1e-3 }	
rf	n estimators	{20, 50, 100, 200}	288
	criterion	{'gini', 'entropy'}	
	max depth	{5, 10, 20, 50}	
	min samples split	{2, 5, 10}	
	min samples leaf	{1, 5, 10}	
svm	c	{0.01, 0.1, 1, 10}	32
	kernel	{'linear', 'poly', 'rbf', 'sigmoid'}	
	gamma	{'scale', 'auto'}	
xgb	learning rate	{0.01, 0.1, 1}	48
	n estimators	{20, 50, 100, 200}	
	max depth	{5, 10, 20, 50}	

A.4 Detailed results per dataset

Table A.4: Ranking Real1

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	7	7	8	6	7	8	8	8	11
bnb	14	14	11	11	14	14	14	14	1
dt	5	5	1	2	2	3	2	3	4
gnb	13	12	14	12	11	11	13	9	14
knn	9	13	9	8	12	12	9	12	8
lda	11	8	12	9	8	7	11	7	13
lr	10	9	10	7	13	13	10	13	9
qda	12	11	13	10	10	10	12	10	12
svm	8	10	7	14	9	9	6	11	2
ab	6	6	6	13	6	6	7	6	10
gb	2	3	3	3	3	1	5	1	7
lgbm	4	4	2	1	1	2	3	2	5
rf	3	2	5	5	5	5	1	5	3
xgb	1	1	4	4	4	4	4	4	6

Table A.5: Ranking Real2

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	6	4	7	6	7	8	9	9	9
bnb	14	11	13	11	12	9	13	10	13
dt	5	7	5	5	5	2	8	2	10
gnb	13	13	14	12	14	13	14	1	14
knn	7	14	8	7	9	12	4	13	3
lda	11	12	6	8	6	5	7	5	8
lr	10	8	10	9	10	11	10	11	7
qda	12	10	11	10	11	7	12	8	12
svm	8	9	9	14	8	10	5	12	5
ab	9	5	12	13	13	14	11	14	11
gb	4	2	4	4	4	3	6	4	6
lgbm	3	6	2	2	2	4	2	6	1
rf	1	1	1	1	1	1	1	3	2
xgb	2	3	3	3	3	6	3	7	4

Table A.6: Ranking Real3

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	8	6	10	8	6	7	10	6	10
bnb	14	11	13	11	4	1	13	2	13
dt	4	9	1	6	8	9	2	9	5
gnb	7	13	14	13	10	2	14	1	14
knn	12	14	8	7	13	13	7	12	3
lda	10	8	11	9	3	4	11	4	11
lr	5	5	3	4	9	10	3	11	4
qda	13	10	12	10	7	3	12	3	12
svm	9	12	5	14	14	14	1	14	1
ab	11	7	9	12	11	11	9	10	6
gb	1	3	2	1	2	6	4	7	7
lgbm	3	2	7	3	5	8	8	8	8
rf	6	4	5	1	12	12	5	13	2
xgb	2	1	4	5	1	5	6	5	9

Table A.7: Ranking DS

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	3	4	5	5	7	6	7	7	7
bnb	14	14	14	11	11	14	14	9	12
dt	11	7	4	4	1	1	10	3	11
gnb	12	11	13	12	6	7	13	1	14
knn	13	12	12	9	12	11	11	12	5
lda	8	8	9	8	10	10	9	10	4
lr	7	9	11	7	14	13	4	14	2
qda	9	10	7	10	2	2	12	2	13
svm	10	13	8	14	9	9	1	11	3
ab	4	6	10	13	13	12	2	13	1
gb	2	2	3	2	5	5	6	6	9
lgbm	6	3	2	3	4	4	5	4	10
rf	1	1	1	1	3	3	3	5	8
xgb	5	5	6	6	8	8	8	8	6

Table A.8: Ranking IBM

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	5	9	7	6	3	3	10	3	9
bnb	12	12	11	8	7	7	11	5	10
dt	14	14	13	11	11	13	13	11	11
gnb	9	11	14	12	4	5	14	1	14
knn	13	13	10	10	14	14	4	14	2
lda	3	3	2	2	1	1	5	4	7
lr	1	1	1	1	2	2	3	6	3
qda	10	10	12	9	5	4	12	2	12
svm	2	2	6	14	13	11	1	13	13
ab	4	4	3	13	6	6	6	7	6
gb	11	8	9	7	10	10	9	10	8
lgbm	8	7	4	4	9	9	7	9	4
rf	6	5	8	3	12	12	2	12	1
xgb	7	6	4	5	8	8	8	8	5

Table A.9: Ranking Kaggle1

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	5	4	4	5	5	5	7	3	7
bnb	14	14	11	11	12	12	11	12	10
dt	8	7	5	6	6	6	6	7	6
gnb	11	11	14	12	11	11	14	11	14
knn	7	9	8	7	8	8	8	5	8
lda	12	13	13	10	14	14	13	14	12
lr	13	12	12	9	13	13	12	13	11
qda	10	10	10	8	10	10	10	9	13
svm	6	6	6	14	7	7	3	8	4
ab	9	8	9	13	9	9	9	10	9
gb	3	3	3	4	3	2	4	2	5
lgbm	4	5	7	3	4	4	5	6	3
rf	1	1	1	2	2	3	1	4	1
xgb	2	2	2	1	1	1	2	1	2

Table A.10: Ranking Kaggle2

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	4	5	5	5	5	5	10	5	10
bnb	13	13	13	11	12	12	13	12	14
dt	9	9	4	4	4	4	7	3	8
gnb	14	14	14	12	14	14	14	14	11
knn	12	12	10	9	10	10	11	9	13
lda	11	11	12	10	13	13	8	13	7
lr	7	7	7	7	7	7	9	7	9
qda	10	10	11	8	11	11	12	11	12
svm	8	8	9	14	9	9	1	10	1
ab	6	6	6	13	6	6	6	6	6
gb	3	3	1	2	1	1	4	1	5
lgbm	1	1	2	1	2	2	2	2	2
rf	4	4	8	6	8	8	5	8	4
xgb	2	1	3	3	3	3	3	4	3

Table A.11: Ranking Kaggle3

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	6	7	12	6	6	6	5	6	3
bnb	14	12	6	11	11	11	11	11	12
dt	7	6	10	4	8	8	10	8	7
gnb	9	9	14	13	4	1	9	1	11
knn	12	14	11	10	10	10	7	10	1
lda	10	10	6	8	11	11	11	11	12
lr	8	8	9	7	11	11	11	11	9
qda	11	11	13	9	9	9	8	9	2
svm	13	13	6	14	11	11	11	11	12
ab	4	3	3	12	5	5	4	5	8
gb	2	2	2	2	2	3	2	3	5
lgbm	1	1	1	1	3	4	1	4	4
rf	3	5	5	3	7	7	6	7	10
xgb	5	4	4	5	1	2	3	2	6

Table A.12: Ranking Kaggle4

Classifier	AUC-PR	AUC-ROC	Accuracy	Brier-Score	F1-Score	H-Measure	Precision	Recall	Specificity
ann	6	5	6	6	6	6	7	6	7
bnb	12	13	12	10	12	13	12	11	12
dt	5	6	5	5	5	5	5	5	5
gnb	14	14	14	13	14	14	14	7	14
knn	9	11	7	8	7	7	8	8	9
lda	11	10	11	9	13	11	11	14	11
lr	10	9	10	7	10	10	10	12	10
qda	13	12	13	11	11	12	13	9	13
svm	7	7	8	14	9	9	6	13	6
ab	8	8	9	12	8	8	9	10	8
gb	3	3	3	3	3	3	2	3	2
lgbm	1	1	1	1	1	1	1	1	1
rf	4	4	4	4	4	4	4	4	4
xgb	2	2	2	2	2	2	3	2	3

B

Data-driven internal mobility: Similarity regularization gets the job done

B.1 Employee journey map

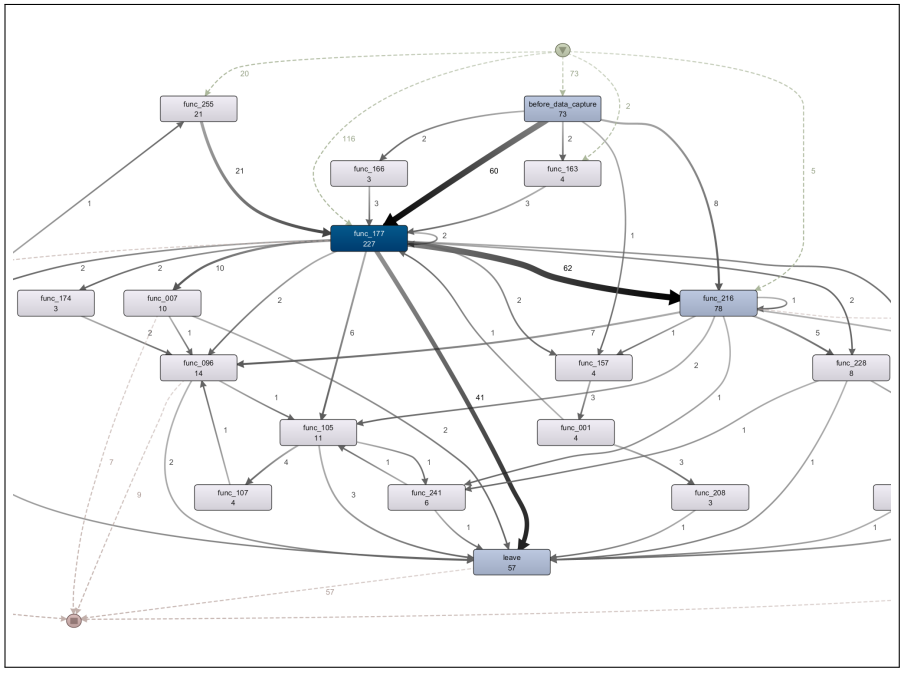


Figure B.1: An anonymized employee journey map of dataset 1 displayed as a directly-follows graph in *Disco*. Each rectangle corresponds to a job and each arc to a possible transition between these jobs. The high-quality figure is avaiable on GitHub.

B.2 Collaborative filtering through matrix factorization

Algorithm 3: Matrix Factorization

Input : observed ratings R, initial matrices U and V, learning rate α, regularization parameters λ_1 and λ_2, learning steps n, stopping threshold t

Output : matrix $\hat{R} = U^T V$ with estimated ratings

for *steps* $= 1, 2, \ldots, n$ **do**

  for *each element* $R_{i,j}$ **do**

    ; > $i \in \{u_1, u_2, \ldots, u_m\}$, $j \in \{v_1, v_2, \ldots, v_n\}$

    if $R_{i,j} > 0$ **then**

      $e_{i,j} := R_{i,j} - \hat{R}_{i,j}$; > calculate error

      $U_i := U_i + \alpha(\underbrace{e_{i,j} V_j}_{(i)} + \underbrace{\lambda_1 U_i}_{(ii)})$; > update vector U_i

      $V_j := V_j + \alpha(\underbrace{e_{i,j} U_i}_{(i)} + \underbrace{\lambda_2 V_j}_{(ii)})$; > update vector V_j

  $e := \underbrace{\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} \left(R_{ij} - U_i^T V_j\right)^2}_{(i)} + \underbrace{\frac{\lambda_1}{2} \|U\|_F^2 + \frac{\lambda_2}{2} \|V\|_F^2}_{(ii)}$;

  if *error* $e < t$ **then**

    break ; > stop if error falls below threshold t

Return : $\hat{R} = U^T V$

B.3 Hyperparameter tuning

Table B.1: This table shows the details on hyperparameter tuning for the three model-based approaches MF, $MFSR$, SVD, and four memory-based approaches with different similarity metrics. The *SlopeOne* algorithm does not have any hyperparameters to be tuned. The third column shows all the options that we consider in the grid search. The three columns on the right display the corresponding optimal hyperparameter values for each dataset.

Method	Hyperpara.	Values	Dataset 1	Dataset 2	Dataset 3
MFSR	λ	{0.01, 0.02, 0.05, 0.1}	0.02	0.05	0.02
	β	{0, 0.05, 0.1, 0.2, 0.5}	0.1	0.2	0.1
	α	{0.0001, 0.001}	0.001	0.0001	0.001
	L	{2, 3, 5, 10}	3	3	2
	steps	{1000, 2000, 3000}	3000	3000	1000
MF	λ	{0.01, 0.02, 0.05, 0.1}	0.02	0.05	0.05
	α	{0.0001, 0.001}	0.001	0.001	0.001
	L	{2, 3, 5, 10}	3	2	5
	steps	{1000, 2000, 3000}	1000	1000	3000
SVD	n_factors	{2, 3, 5, 10}	5	2	2
	n_epochs	{1000, 2000, 3000}	1000	1000	1000
	reg_pu	{0.001, 0.005, 0.01, 0.02, 0.05}	0.001	0.001	0.001
	reg_qi	{0.001, 0.005, 0.01, 0.02, 0.05}	0.001	0.001	0.001
KNN_c	k	{2, 3, 5, 10}	2	2	2
KNN_p	k	{2, 3, 5, 10}	2	2	2
KNN_{smd}	k	{2, 3, 5, 10}	3	2	2
KNN_{ta}	k	{2, 3, 5, 10}	5	2	2

C

Robust instance-dependent cost-sensitive classification

C.1 Results on synthetic data

Table C.1: This table displays the results of tests on synthetic data with no outlier and an outlier of size 100. We apply a 2×5-fold cross-validation procedure with a train/test split ratio of 0.8/0.2. In this table, r-cslogit always performs at least equally good in comparison to cslogit. In terms of the cost-sensitive metric *Savings*, logit is always outperformed by cslogit and r-cslogit. Logit performs best in terms of cost-insensitive metrics, and its performance remains stable after increasing the size of the outlier, given its cost-insensitive nature. An exception is *Specificity* with a 90/10 class imbalance and an outlier of 100. However, given the relatively high standard deviation of 0.11, these results are rather volatile because of the high class imbalance.

		No outlier			$A_{outlier}$ = 100		
Imbal.	Metric	logit	cslogit	r-cslogit	logit	cslogit	r-cslogit
50/50	*Savings*	0.68 ± 0.08	**0.80 ± 0.05**	**0.80 ± 0.05**	0.68 ± 0.08	**0.80 ± 0.05**	**0.80 ± 0.05**
	AUC	**0.86 ± 0.04**	0.77 ± 0.06	0.77 ± 0.06	**0.86 ± 0.04**	0.77 ± 0.06	0.77 ± 0.06
	F1	**0.84 ± 0.04**	0.77 ± 0.05	0.77 ± 0.05	**0.84 ± 0.04**	0.77 ± 0.05	0.77 ± 0.05
	Spec	**0.88 ± 0.05**	0.77 ± 0.11	0.77 ± 0.11	**0.88 ± 0.05**	0.77 ± 0.11	0.77 ± 0.11
	Sens	**0.83 ± 0.04**	0.78 ± 0.03	0.78 ± 0.03	**0.83 ± 0.04**	0.78 ± 0.03	0.78 ± 0.03
	Brier	**0.14 ± 0.04**	0.23 ± 0.06	0.23 ± 0.06	**0.14 ± 0.04**	0.23 ± 0.06	0.23 ± 0.06
60/40	*Savings*	0.64 ± 0.11	**0.75 ± 0.06**	**0.75 ± 0.06**	0.64 ± 0.11	**0.75 ± 0.04**	**0.75 ± 0.04**
	AUC	**0.85 ± 0.05**	0.76 ± 0.07	0.76 ± 0.07	**0.85 ± 0.05**	0.77 ± 0.06	0.76 ± 0.06
	F1	**0.88 ± 0.04**	0.79 ± 0.04	0.79 ± 0.04	**0.88 ± 0.04**	0.79 ± 0.04	0.79 ± 0.04
	Spec	**0.79 ± 0.06**	0.77 ± 0.13	0.77 ± 0.13	**0.79 ± 0.06**	0.77 ± 0.12	0.77 ± 0.12
	Sens	**0.90 ± 0.05**	0.76 ± 0.04	0.76 ± 0.04	**0.90 ± 0.05**	0.76 ± 0.03	0.76 ± 0.04
	Brier	**0.14 ± 0.05**	0.24 ± 0.06	0.24 ± 0.06	**0.14 ± 0.05**	0.24 ± 0.06	0.24 ± 0.05
70/30	*Savings*	0.52 ± 0.05	**0.70 ± 0.05**	**0.70 ± 0.05**	0.52 ± 0.05	0.70 ± 0.05	**0.70 ± 0.05**
	AUC	**0.81 ± 0.02**	0.74 ± 0.03	0.74 ± 0.03	**0.81 ± 0.02**	0.74 ± 0.03	0.74 ± 0.03
	F1	**0.89 ± 0.01**	0.84 ± 0.03	0.84 ± 0.03	**0.89 ± 0.01**	0.84 ± 0.03	0.84 ± 0.03
	Spec	**0.68 ± 0.03**	0.62 ± 0.03	0.62 ± 0.03	**0.68 ± 0.03**	0.62 ± 0.03	0.62 ± 0.03
	Sens	**0.94 ± 0.02**	0.87 ± 0.06	0.87 ± 0.06	**0.94 ± 0.02**	0.87 ± 0.05	0.87 ± 0.06
	Brier	**0.15 ± 0.02**	0.21 ± 0.04	0.21 ± 0.04	**0.15 ± 0.02**	0.22 ± 0.04	0.21 ± 0.04
80/20	*Savings*	0.37 ± 0.10	**0.58 ± 0.15**	**0.58 ± 0.15**	0.37 ± 0.10	**0.58 ± 0.15**	**0.58 ± 0.15**
	AUC	**0.80 ± 0.04**	0.74 ± 0.05	0.74 ± 0.05	**0.80 ± 0.04**	0.74 ± 0.05	0.74 ± 0.05
	F1	**0.93 ± 0.02**	0.90 ± 0.01	0.90 ± 0.01	**0.93 ± 0.02**	0.90 ± 0.01	0.90 ± 0.01
	Spec	**0.64 ± 0.06**	0.57 ± 0.13	0.57 ± 0.13	**0.64 ± 0.06**	0.57 ± 0.13	0.57 ± 0.13
	Sens	**0.95 ± 0.03**	0.91 ± 0.04	0.91 ± 0.04	**0.95 ± 0.03**	0.91 ± 0.04	0.91 ± 0.04
	Brier	**0.12 ± 0.04**	0.16 ± 0.01	0.16 ± 0.01	**0.12 ± 0.04**	0.16 ± 0.01	0.16 ± 0.01
90/10	*Savings*	0.23 ± 0.12	**0.36 ± 0.11**	**0.36 ± 0.11**	0.23 ± 0.12	**0.38 ± 0.14**	**0.38 ± 0.14**
	AUC	**0.68 ± 0.03**	0.62 ± 0.02	0.62 ± 0.02	**0.68 ± 0.03**	0.66 ± 0.05	0.64 ± 0.05
	F1	**0.95 ± 0.01**	0.89 ± 0.04	0.89 ± 0.04	**0.95 ± 0.01**	0.91 ± 0.03	0.89 ± 0.04
	Spec	**0.38 ± 0.06**	0.37 ± 0.06	0.37 ± 0.06	0.38 ± 0.06	**0.41 ± 0.11**	**0.41 ± 0.11**
	Sens	**0.97 ± 0.01**	0.86 ± 0.08	0.86 ± 0.08	**0.97 ± 0.01**	0.90 ± 0.06	0.90 ± 0.06
	Brier	**0.10 ± 0.02**	0.19 ± 0.07	0.19 ± 0.07	**0.10 ± 0.02**	0.16 ± 0.05	0.18 ± 0.06

Table C.2: This table displays the results of tests on synthetic data with an outlier of size 1,000 and an outlier of size 10,000. We apply a 2×5-fold cross-validation procedure with a train/test split ratio of 0.8/0.2. In terms of *Savings*, r-cslogit always outperforms the other two methods and remains stable after increasing the size of the outlier. Also in terms of cost-insensitive metrics, the performance of r-cslogit remains stable. After increasing the outlier size, cslogit performs worse. This is analogous to the results as displayed in Figure 5.5. Logit performs best in terms of cost-insensitive metrics and, given its cost-insensitive nature, its performance remains stable after increasing the size of the outlier. The few times that logit is outperformed by either cslogit or r-cslogit in terms of cost-insensitive metrics, the performance scores have a rather high volatility. This is predominantly the case for tests with high class imbalance.

		$A_{outlier} = 1,000$			$A_{outlier} = 10,000$		
Imbal.	Metric	logit	cslogit	r-cslogit	logit	cslogit	r-cslogit
50/50	*Savings*	0.68 ± 0.08	0.60 ± 0.15	**0.80±0.05**	0.68 ± 0.08	0.58 ± 0.11	**0.80±0.05**
	AUC	**0.86±0.04**	0.83 ± 0.05	0.77 ± 0.06	**0.86±0.04**	0.84 ± 0.03	0.77 ± 0.06
	F1	**0.84±0.04**	0.82 ± 0.05	0.77 ± 0.05	**0.84±0.04**	0.83 ± 0.03	0.77 ± 0.05
	Spec	**0.88±0.05**	0.82 ± 0.07	0.77 ± 0.11	**0.88±0.05**	0.83 ± 0.05	0.77 ± 0.11
	Sens	**0.83±0.04**	0.81 ± 0.05	0.78 ± 0.03	0.83 ± 0.04	**0.85±0.07**	0.78 ± 0.03
	Brier	**0.14±0.04**	0.17 ± 0.05	0.23 ± 0.06	**0.14±0.04**	0.16 ± 0.03	0.23 ± 0.06
60/40	*Savings*	0.64 ± 0.11	0.69 ± 0.10	**0.75±0.04**	0.64 ± 0.11	0.58 ± 0.18	**0.75±0.04**
	AUC	**0.85±0.05**	0.81 ± 0.07	0.77 ± 0.06	**0.85±0.05**	0.84 ± 0.06	0.76 ± 0.06
	F1	**0.88±0.04**	0.83 ± 0.08	0.79 ± 0.04	**0.88±0.04**	0.87 ± 0.05	0.79 ± 0.04
	Spec	0.79 ± 0.06	**0.80±0.08**	0.77 ± 0.12	**0.79±0.06**	0.78 ± 0.07	0.77 ± 0.12
	Sens	**0.90±0.05**	0.82 ± 0.13	0.76 ± 0.03	**0.90±0.05**	0.90 ± 0.06	0.76 ± 0.04
	Brier	**0.14±0.05**	0.19 ± 0.07	0.23 ± 0.06	**0.14±0.05**	0.15 ± 0.06	0.24 ± 0.05
70/30	*Savings*	0.52 ± 0.05	0.42 ± 0.16	**0.70±0.05**	0.52 ± 0.05	0.32 ± 0.20	**0.70±0.05**
	AUC	**0.80±0.02**	0.77 ± 0.05	0.74 ± 0.03	**0.80±0.02**	0.79 ± 0.06	0.74 ± 0.03
	F1	**0.89±0.01**	0.88 ± 0.02	0.84 ± 0.03	**0.89±0.01**	0.88 ± 0.03	0.84 ± 0.03
	Spec	**0.68±0.03**	0.60 ± 0.15	0.61 ± 0.03	**0.67±0.03**	0.64 ± 0.07	0.61 ± 0.03
	Sens	**0.93±0.02**	0.92 ± 0.06	0.87 ± 0.06	**0.93±0.02**	0.89 ± 0.06	0.87 ± 0.06
	Brier	**0.15±0.02**	0.17 ± 0.03	0.21 ± 0.04	**0.15±0.02**	0.16 ± 0.04	0.21 ± 0.04
80/20	*Savings*	0.37 ± 0.10	0.37 ± 0.17	**0.58±0.15**	0.37 ± 0.10	0.10 ± 0.47	**0.58±0.15**
	AUC	**0.80±0.04**	0.72 ± 0.04	0.74 ± 0.05	**0.80±0.04**	0.76 ± 0.08	0.74 ± 0.05
	F1	**0.93±0.02**	0.91 ± 0.03	0.90 ± 0.01	**0.93±0.02**	0.91 ± 0.04	0.90 ± 0.01
	Spec	**0.64±0.06**	0.47 ± 0.06	0.57 ± 0.13	**0.64±0.06**	0.60 ± 0.19	0.57 ± 0.13
	Sens	0.95 ± 0.03	**0.96±0.07**	0.91 ± 0.04	**0.95±0.03**	0.92 ± 0.08	0.91 ± 0.04
	Brier	**0.12±0.04**	0.14 ± 0.04	0.16 ± 0.01	**0.12±0.04**	0.15 ± 0.05	0.16 ± 0.01
90/10	*Savings*	0.23 ± 0.12	0.17 ± 0.13	**0.36±0.11**	0.23 ± 0.12	0.19 ± 0.16	**0.36±0.11**
	AUC	**0.68±0.03**	0.62 ± 0.05	0.64 ± 0.05	**0.68±0.03**	0.66 ± 0.06	0.64 ± 0.05
	F1	**0.95±0.01**	0.95 ± 0.00	0.89 ± 0.04	**0.95±0.01**	0.94 ± 0.01	0.89 ± 0.04
	Spec	0.38 ± 0.06	0.26 ± 0.10	**0.41±0.13**	0.38 ± 0.06	0.34 ± 0.13	**0.41±0.13**
	Sens	0.97 ± 0.01	**0.98±0.06**	0.87 ± 0.08	0.97 ± 0.01	**0.98±0.06**	0.87 ± 0.08
	Brier	**0.10±0.02**	0.10 ± 0.01	0.18 ± 0.06	**0.10±0.02**	0.11 ± 0.01	0.18 ± 0.06

D

Decision-centric fairness: Evaluation and optimization for resource allocation problems

D.1 Dataset details

D.1.1 Label flipping for inducing additional bias

Algorithm 4: Informed Label Flipping for Bias Induction

Input : $\mathcal{D} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^{N}$, protected group value $s \in \{0, 1\}$, bias rate $r \in [0, 1]$

Output: Biased dataset $\mathcal{D}'_r$

Train a classifier m on $\mathcal{D}$;

$\tilde{y} \leftarrow m(\boldsymbol{x}, s)$; `> predicted scores` $\tilde{y}_i \in [0, 1]$

$\mathcal{D}_0^s \leftarrow \{(\boldsymbol{x}_i, y_i, s_i, \tilde{y}_i) \mid s_i = s \text{ and } y_i = 0\}$; `> subset for bias`

$k \leftarrow r \cdot |\mathcal{D}_0^s|$;

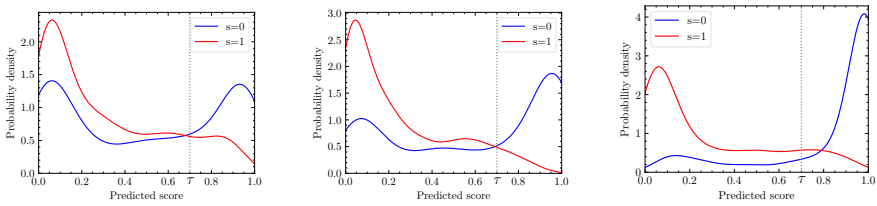
Rank $\mathcal{D}_0^s$ by $\tilde{y}$ in descending order ;

for *each* $(\boldsymbol{x}_i, y_i) \in$ *top-k of* $\mathcal{D}_0^s$ **do**

  $y_i \leftarrow 1$; `> flip labels`

Return: $\mathcal{D}'_r = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^{N}$; `> return biased dataset`

D.1.2 Baseline discriminatory behavior



(a) Bias rate 0.25 (b) Bias rate 0.50 (c) Bias rate 0.75

Figure D.1: Baseline discriminatory behavior for the *TelecomKaggle* datasets. This figure shows the score distributions (PDFs) for two groups ($s = 0$ and $s = 1$), as estimated by a model without unfairness penalty ($\lambda = 0$), under three semi-synthetic bias rates. The decision-making region is defined as $\tau = 0.7$. Higher semi-synthetic bias rates lead to increased baseline discriminatory behavior, as summarized below in Table D.1.

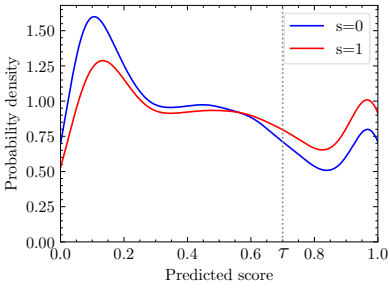


Figure D.2: Baseline discriminatory behavior for the *Churn* dataset.

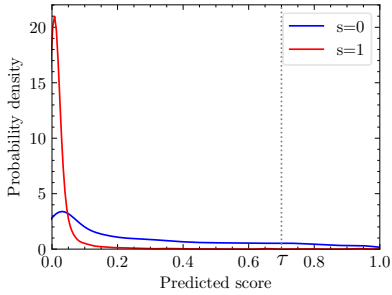


Figure D.3: Baseline discriminatory behavior for the *Adult* dataset.

Table D.1: Baseline discriminatory behavior for the various datasets for $\tau = 0.7$, corresponding to Figures D.1-D.3.

Dataset	Bias rate	$ABPC_\tau$	$ABCC_\tau$
TelecomKaggle	0.25	0.16	0.06
	0.50	0.30	0.12
	0.75	0.43	0.19
Churn	-	0.03	0.01
Adult	-	0.13	0.02

D.2 Implementation

We employ a fully-connected MLP, implemented in PyTorch, where network parameters are optimized using Adam optimizer [396] with a learning rate 0.01 for *TelecomKaggle* and *Churn*, and 0.001 for *Adult*. We use ReLu activation functions in the hidden layers. For the final layer, we use a sigmoid activation function to ensure outputs in the range $[0, 1]$. We use a standard BCE loss function for the first 15 epochs and then a composite loss function for the remainder of the training. We employ early stopping after 20 epochs without model improvement in terms of composite loss on the validation data. Per dataset, hyperparameters are selected for the case where $\lambda = 0$. To ensure enough observations remain for computing the decision-centric fairness loss—given that each training batch considers only a subset of data restricted to the top $n\%$ percentile, further divided across the two protected groups and limited to the training split—we set the batch size to 1024.

Table D.2: Hyperparameter search space.

Hyperparameter	Values	TelecomKaggle			Churn	Adult
		0.25	0.50	0.75		
Nr. of hidden layers	{2, 3, 4}	3	2	4	2	2
Hidden layer size	{16, 32, 64, 128}	64	64	32	32	64
Dropout probability	{0, 0.01, 0.1}	0	0	0	0.01	0.01
L2 regularization	{0, 0.01, 0.05}	0.01	0.01	0.01	0	0.01

D.3 Additional results

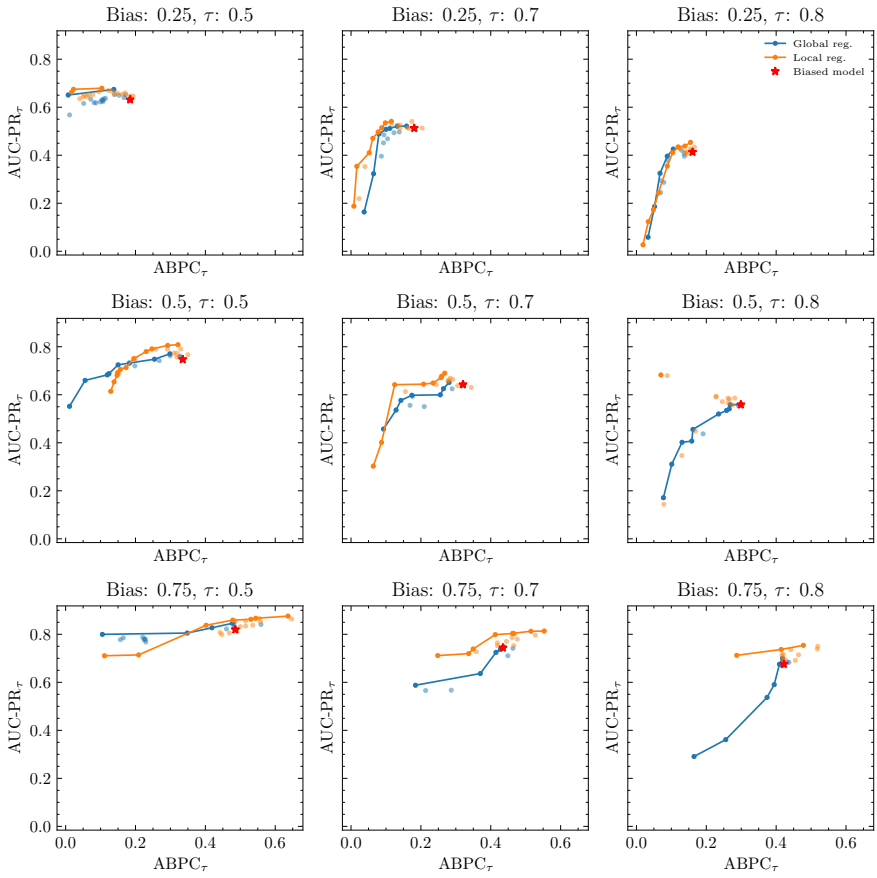


Figure D.4: Results on the *TelecomKaggle* dataset for different bias rates, with fairness measured by ABPC$_\tau$. The figure illustrates the effect of a varying size of the decision-making region with $\tau = 0.5, 0.7, 0.8$ (columns) on the decision-centric fairness-predictive performance trade-off. The orange and blue lines represent decision-centric and global fairness induction, respectively, while the model without unfairness penalty (i.e., with $\lambda = 0$) is marked with a red star.

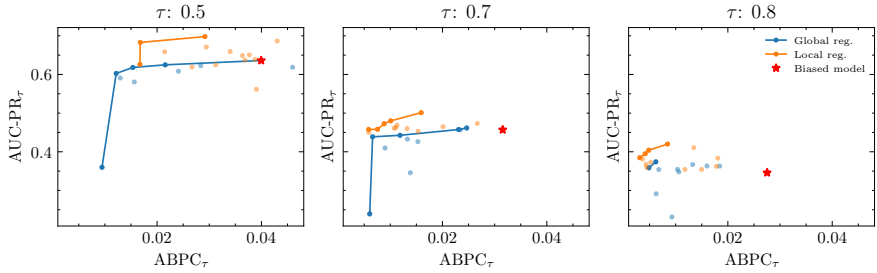


Figure D.5: Results on the *Churn* dataset with fairness measured by ABPC_τ. The figure illustrates the effect of a varying size of the decision-making region with $\tau = 0.5, 0.7, 0.8$ (columns) on the decision-centric fairness-predictive performance trade-off. The orange and blue lines represent decision-centric and global fairness induction, respectively, while the model without unfairness penalty (i.e., with $\lambda = 0$) is marked with a red star.

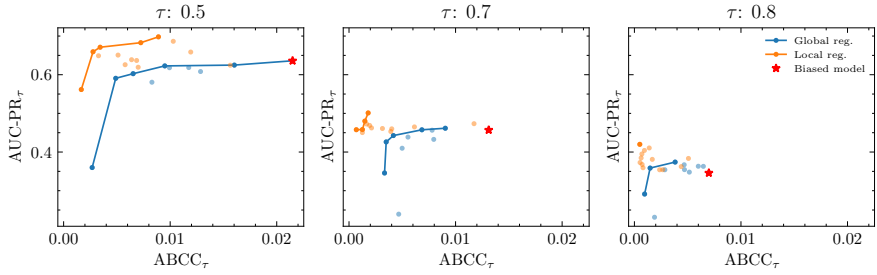


Figure D.6: Results on the *Churn* dataset with fairness measured by ABCC_τ. The figure illustrates the effect of a varying size of the decision-making region with $\tau = 0.5, 0.7, 0.8$ (columns) on the decision-centric fairness-predictive performance trade-off. The orange and blue lines represent decision-centric and global fairness induction, respectively, while the model without unfairness penalty (i.e., with $\lambda = 0$) is marked with a red star.

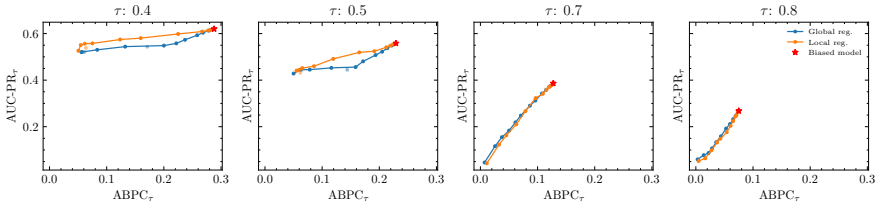


Figure D.7: Results on the *Adult* dataset with fairness measured by ABPC_τ. The figure illustrates the effect of a varying size of the decision-making region with $\tau = 0.4, 0.5, 0.7, 0.8$ (columns) on the decision-centric fairness-predictive performance trade-off. The orange and blue lines represent decision-centric and global fairness induction, respectively, while the model without unfairness penalty (i.e., with $\lambda = 0$) is marked with a red star.

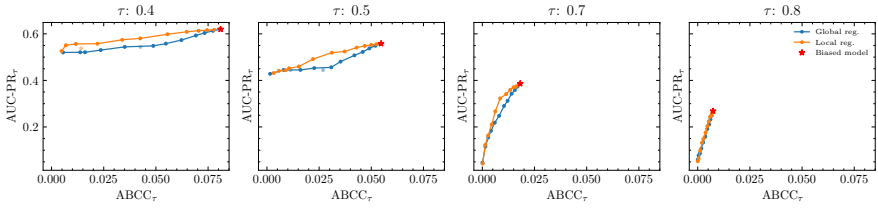


Figure D.8: Results on the *Adult* dataset with fairness measured by ABCC_τ. The figure illustrates the effect of a varying size of the decision-making region with $\tau = 0.4, 0.5, 0.7, 0.8$ (columns) on the decision-centric fairness-predictive performance trade-off. The orange and blue lines represent decision-centric and global fairness induction, respectively, while the model without unfairness penalty (i.e., with $\lambda = 0$) is marked with a red star.

E

Uplift modeling with continuous treatments: A predict-then-optimize approach

E.1 Notation overview

Table E.1: This table summarizes the notation in this paper.

$\mathcal{D} = \left\{(\mathbf{x}_i, s_i, y_i)_{i=1}^N\right\}$	Dataset containing N entities
$\mathcal{X}$	Feature space
$\mathcal{S}$	Treatment space
$\mathcal{Y}$	Outcome space
$\mathbf{X} \in \mathcal{X}$	Features
$S \in \mathcal{S}$	Treatment
$Y \in \mathcal{Y}$	Outcome
$\mu : \mathcal{S} \times \mathcal{X} \to [0,1]$	Conditional average dose response function
$\tau : \mathcal{S} \times \mathcal{X} \to [-1,1]$	Conditional average dose effect function
$\tau_s(\mathbf{x})$	Treatment effect for the features-dose pair $\{\mathbf{x}, s\}$
$\hat{\mu} : \mathcal{S} \times \mathcal{X} \to [0,1]$	Estimated conditional average dose response function
$\hat{\tau} : \mathcal{S} \times \mathcal{X} \to [-1,1]$	Estimated conditional average dose effect function
$\hat{\tau}_s(\mathbf{x})$	Estimated dose effect for the features-dose pair $\{\mathbf{x}, s\}$
δ	Number of dose bins
$\hat{\boldsymbol{\tau}}(\mathbf{x}) \in [-1,1]^\delta$	Vector with δ CADE estimates
$\pi : \mathcal{X} \to \{0,1\}^\delta$	Treatment assignment policy (unconstrained) for a single entity
$\mathbf{C} \in \mathbb{R}^{(N \times \delta)}$	Treatment cost matrix
$\Psi_i(\pi) \in \mathbb{R}_{\geq 0}$	Treatment cost for a single entity
$\mathbf{b} \in \mathbb{R}_{\geq 0}^N$	Benefit vector
$\mathbf{T} \in [-1,1]^{(N \times \delta)}$	Matrix with $(N \times \delta)$ CADE values (ground-truth)
$\hat{\mathbf{T}} \in [-1,1]^{(N \times \delta)}$	Matrix with $(N \times \delta)$ CADE values (estimated)
$U_i : \{0,1\}^\delta \times [-1,1]^\delta \to \mathbb{R}$	Policy uplift of single entity Expected (U_i^{exp}), prescribed (U_i^{presc}), and optimal (U_i^{opt})
$V_i : \{0,1\}^\delta \times [-1,1]^\delta \times \mathbb{R} \to \mathbb{R}$	Policy value of single entity Expected (V_i^{exp}), prescribed (V_i^{presc}), and optimal (V_i^{opt})
$\overset{B}{\Pi} : \mathbb{R}^+ \times \mathcal{X}^N \to \{0,1\}^{(N \times \delta)}$	Treatment assignment policy (constrained) over all entities Expected optimal ($\overset{B^{exp}}{\Pi}$), ground-truth optimal ($\overset{B^{opt}}{\Pi}$)
$U : \{0,1\}^{(N \times \delta)} \times [-1,1]^{(N \times \delta)} \to \mathbb{R}$	Policy uplift over all entities Expected (U^{exp}), prescribed (U^{presc}), and optimal (U^{opt})
$V : \{0,1\}^{(N \times \delta)} \times [-1,1]^{(N \times \delta)} \times \mathbb{R}^N \to \mathbb{R}$	Policy value over all entities Expected (V^{exp}), prescribed (V^{presc}), and optimal (V^{opt})
A	Protected sensitive attribute
$\epsilon_{DT} \in [0,1]$	Slack for allocation fairness (Disparate Treatment)
$\epsilon_{DO} \in [0,1]$	Slack for outcome fairness (Disparate Outcome)

E.2 Assumptions and mathematical justification of CADE identification

Under the Rubin-Neyman potential outcomes framework, the CADE $\tau_s(\mathbf{x}) = \mathbb{E}[Y(s) - Y(0) \mid \mathbf{X} = \mathbf{x}]$ quantifies the expected outcome of a continuous treatment dose s relative to a baseline (i.e., no treatment). Based on the proof for the binary treatment setting by Neal [397], we prove that $\tau_s(\mathbf{x})$ is identifiable as (E.5) and estimable via causal machine learning methods under three assumptions: Consistency, Ignorability, Overlap (Section 7.4.1).

Proof. From the definition of $\tau_s(\mathbf{x})$ and linearity of expectation:

$$\tau_s(\mathbf{x}) = \mathbb{E}[Y(s) - Y(0) \mid \mathbf{X} = \mathbf{x}] \tag{E.1}$$

$$= \mathbb{E}[Y(s) \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y(0) \mid \mathbf{X} = \mathbf{x}] \tag{E.2}$$

By Ignorability, potential outcomes are independent of treatment assignment given $\mathbf{X}$:

$$\mathbb{E}[Y(s) \mid \mathbf{X} = \mathbf{x}] = \mathbb{E}[Y(s) \mid \mathbf{X} = \mathbf{x}, S = s]. \tag{E.3}$$

Applying Consistency to (E.3):

$$\mathbb{E}[Y(s) \mid \mathbf{X} = \mathbf{x}, S = s] = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, S = s]. \tag{E.4}$$

Combining (E.2)–(E.4):

$$\tau_s(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, S = s] - \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}, S = 0], \tag{E.5}$$

where Overlap ensures the conditional expectations in (E.3)-(E.5) are well-defined. □

E.3 Details regarding semi-synthetic data

The original IHDP dataset contains 747 observations and 25 features. Following [340], synthetic counterfactuals are generated by Equations E.6-E.9 where $S_{\text{con}} = \{1, 2, 3, 5, 6\}$ is the index set of continuous features, $S_{\text{bin},1} = \{4, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$ and $S_{\text{bin},2} = \{16, 17, 18, 19, 20, 21, 22, 23, 24, 25\}$ are two sets of binary features, $c_1 = \mathbb{E}\left[\frac{\sum_{i \in S_{\text{dis},1}} x_i}{|S_{\text{bin},1}|}\right]$, and $c_2 = \mathbb{E}\left[\frac{\sum_{i \in S_{\text{dis},2}} x_i}{|S_{\text{bin},2}|}\right]$.

$$\tilde{t} \mid \boldsymbol{x} = \frac{x_1}{1 + x_2} + \frac{\max(x_3, x_5, x_6)}{0.2 + \min(x_3, x_5, x_6)} + \tanh\left(5 \cdot \frac{\sum_{i \in S_{\text{bin},2}} (x_i - c_2)}{|S_{\text{bin},2}|}\right) - 2 + \mathcal{N}(0, 0.25) \tag{E.6}$$

$$t = \left(1 + \exp(-2\tilde{t})\right)^{-1} \tag{E.7}$$

$$\tilde{y} \mid \boldsymbol{x}, t = \frac{\sin(3\pi t)}{1.2 - t} \cdot \left(\tanh\left(5 \cdot \frac{\sum_{i \in S_{\text{bin},1}} (x_i - c_1)}{|S_{\text{bin},1}|}\right) + \frac{\exp\left(0.2(x_1 - x_6)\right)}{0.5 + 5 \cdot \min(x_2, x_3, x_5)}\right) + \mathcal{N}(0, 0.25) \tag{E.8}$$

$$y = \frac{\tilde{y} - \min(\tilde{y})}{\max(\tilde{y}) - \min(\tilde{y})} \tag{E.9}$$

Following the literature, features are preprocessed to follow a standard normal distribution, and the generated treatments are normalized to fall within the range [0, 1] [340], [398]. Additionally, we also standardize the outcomes so that $y \in [0, 1]$.

E.4 Number of dose bins δ

Throughout the experiments, the number of available treatment bins δ is set to 10. Figure E.1 illustrates the relationship between the number of treatment bins δ, U^{presc}, and calculation time in seconds. For the semi-synthetic IHDP dataset, the increased benefit of using smaller-grained bins quickly caps out, while the required computation time continues to rise. From an application perspective, also not each granularity of doses should be considered. For instance, lending costs occur only in specific increments, and in an HR setting, training hours are typically scheduled in multiples of the session duration. Given these considerations, δ is set to 10.

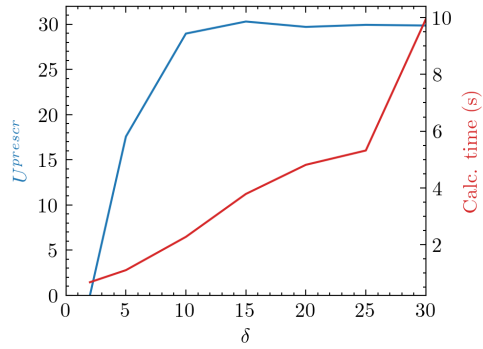


Figure E.1: Effect of available dose bins on U^{presc} and calculation time

E.5 Hyperparameter tuning

Table E.2: Hyperparameter search space. The selected hyperparameters are underlined.

Method	**Hyperparameter**	**Values**	**#Models**
S-Learner (rf)	Estimators	{10, 20, 50, 100, <u>200</u>, 500}	72
	Criterion	{*Sq. error*, <u>*Abs. error*</u>}	
	Max depth	{5, <u>15</u>, *None*}	
	Max features	{<u>*Sqrt*</u>, *Log2*}	
S-Learner (mlp)	Learning rate	{0.001, <u>0.01</u>}	128
	L2 regularization	{<u>0.0</u>, 0.1}	
	Batch size	{<u>64</u>, 128}	
	Num layers	{2, <u>3</u>}	
	Hidden size	{<u>32</u>, 64}	
	Steps	{500, 1000, 2000, <u>5000</u>}	
	Optimizer	{*Adam*}	
DRNet	Learning rate	{<u>0.001</u>, 0.01}	256
	L2 regularization	{<u>0.0</u>, 0.1}	
	Batch size	{64, <u>128</u>}	
	Num repr. layers	{2, <u>3</u>}	
	Num inf. layers	{<u>1</u>, 2}	
	Hidden size	{32, <u>64</u>}	
	Steps	{500, 1000, 2000, <u>5000</u>}	
	Num dose strata	{10}	
	Optimizer	{*Adam*}	
VCNet	Learning rate	{0.001, <u>0.01</u>}	32
	Batch size	{64, <u>128</u>}	
	Hidden size	{32, <u>64</u>}	
	Steps	{500, 1000, <u>2000</u>, 5000}	
	Optimizer	{*Adam*}	

E.6 More details on Experiment 1

E.6.1 Dose-response estimation

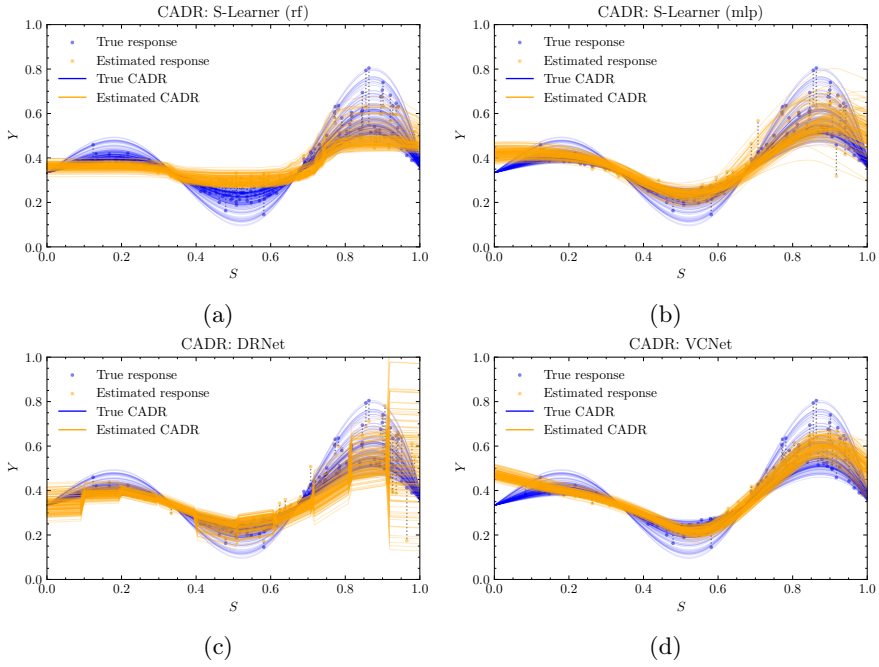


Figure E.2: This figure presents the ground-truth dose responses in blue, with the blue dots representing the factual outcomes, and their corresponding estimates in orange, for four different predictive methods on a separate test set from the IHDP dataset.

E.6.2 Scalability

We provide additional runtime analysis for increasing dataset sizes. The dataset size is semi-synthetically scaled by a factor N (with $N = 1$ corresponding to the original dataset), by randomly oversampling observations and adding slight noise to ensure that each observation remains unique. We adopt the setup of Experiment 1 (Section 7.5.3), where the available budget is also scaled by N, and compare the runtime of the ILP solution with that of the heuristic approach. Figure E.3 shows the runtime (in seconds) as a function of the problem size (measured by the scaling factor N). Both

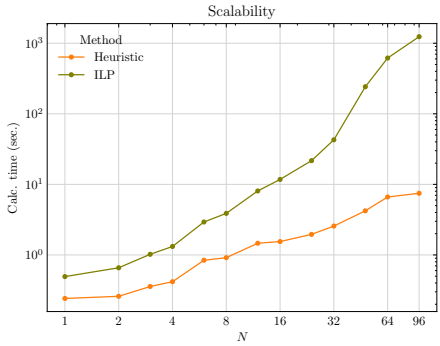


Figure E.3: Runtime of the ILP and heuristic as a function of problem size. The original dataset size corresponds with $N = 1$.

axes are shown on a logarithmic scale. The clear advantage of the heuristic method is that it scales well. However, it does not take into account any side constraints, which are allowed by the ILP. The downside of the ILP is that runtimes increase rapidly with the problem size, becoming impractical for large-scale instances. This highlights the classic trade-off: while the ILP allows for more expressive modeling and constraint handling, its computational cost means that problem size is limited. In contrast, the heuristic remains feasible in terms of computation time even as the dataset grows.

F
Industry co-creation

F.1 Implementation of EJMs

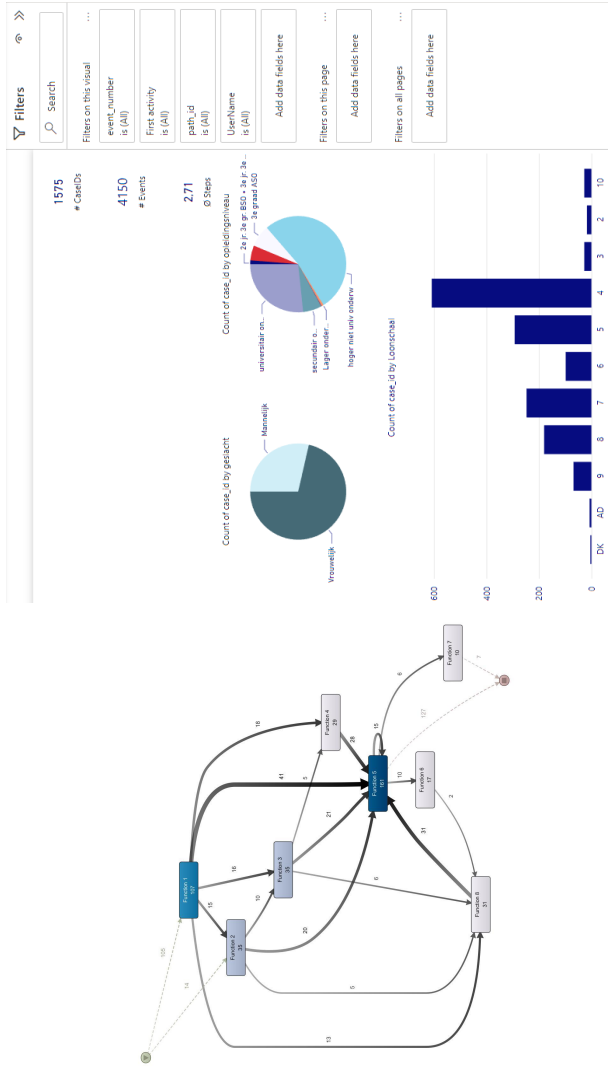


Figure F.1: An EJM as implemented in PowerBI. Filters on the right allow for any subselection of data and dynamically regenerating a new EJM.

F.2 Implementation of turnover prediction

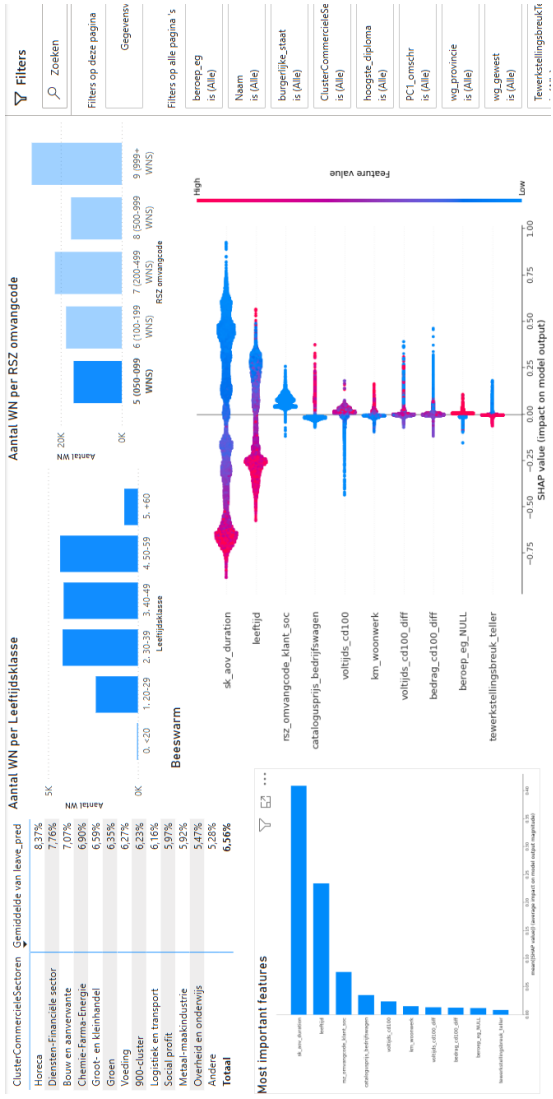


Figure F.2: Turnover prediction as implemented in PowerBI, with a beeswarm plot displaying SHAP values. Filters on the right and selecting bars in the subfigures allow for any subselection of data and dynamically regenerating a new EJM.

Publication list

Journal articles (peer reviewed)

S. De Vos, C. Bockel-Rickermann, J. Van Belle, and W. Verbeke, "Predicting Employee Turnover: Scoping and Benchmarking the State-of-the-Art," *Business & Information Systems Engineering*, 2024. doi: 10.1007/ s12599-024-00898-z.

S. De Vos, J. De Smedt, M. Verbruggen, and W. Verbeke, "Data-driven internal mobility: Similarity regularization gets the job done," *Knowledge-Based Systems*, vol. 295, Art. no. 111824, 2024. doi: 10.1016/ j.knosys.2024.111824.

S. De Vos, T. Vanderschueren, T. Verdonck, and W. Verbeke, "Robust instance-dependent cost-sensitive classification," *Advances in Data Analysis and Classification*, 2023. doi: 10.1007/ s11634-022-00533-3.

Conference proceedings (peer reviewed)

J. Peeperkorn, **S. De Vos**, *Placeholder CAISE paper: accepted, not yet published*

Book chapters (peer reviewed)

S. De Vos, J. De Smedt, C. Wuytens, and W. Verbeke, "Leveraging Process Mining to Optimize Internal Employee Mobility Strategies," in *Business Process Management Cases Vol. 3: Implementation in Practice*, Cham, Switzerland: Springer Nature, 2025, pp. 15–28. doi: 10.1007/ 978-3-031-80793-0. ISBN: 978-3-031-80792-3.

Preprints

S. De Vos, J. Van Belle, A. Algaba, W. Verbeke, S. Verboven, "Decision-centric fairness: Evaluation and optimization for resource allocation problems," *arXiv*, 2024. doi: 10.48550/ arXiv.2504.20642.

J. Peeperkorn, **S. De Vos**, "Achieving Group Fairness through Independence in Predictive Process Monitoring," *arXiv*, 2024. doi: 10.48550/

arXiv.2412.04914

S. De Vos, C. Bockel-Rickermann, S. Lessmann, and W. Verbeke, "Uplift modeling with continuous treatments: A predict-then-optimize approach," *arXiv*, 2024. doi: 10.48550/ arXiv.2412.09232.

D. Caljon, J. Vercauteren, **S. De Vos**, W. Verbeke, and J. Van Belle, "Using dynamic loss weighting to boost improvements in forecast stability," *arXiv*, 2024. doi: 10.48550/ arXiv.2409.18267.

Abstracts / Presentations / Posters

S. De Vos, J. Van Belle, A. Algaba, W. Verbeke, and S. Verboven, "Decision-Centric Fairness: Evaluation and Optimization for Classification Problems," presented at *the Joint ORBEL-NGB Conference on Operations Research*, Maastricht, Netherlands, Jan. 29–31, 2025.

S. De Vos and W. Verbeke, "A predict-then-optimize approach for uplift modeling with continuous individual treatment effects," presented at *the 33rd European Conference on Operational Research*, Copenhagen, Denmark, Jun. 30–Jul. 3, 2024.

S. De Vos, J. De Smedt, M. Verbruggen, and W. Verbeke, "A survey and benchmarking experiment of the state-of-the-art in employee turnover prediction," presented at *the 37th Annual Conference of the Belgian Operational Research Society, ORBEL 37*, Liège, May 25–26, 2023.

D. Caljon, J. Vercauteren, **S. De Vos**, and J. Van Belle, "Using adaptive loss balancing to boost improvements in forecast stability," presented at *ORBEL 37*, Liège, May 25–26, 2023.

S. De Vos, C. Wuytens, J. De Smedt, and W. Verbeke, "Process Mining-Driven Analytics of Human Resources," presented at *the 4th International Conference on Process Mining (ICPM)*, Bolzano, Italy, 2022.

S. De Vos, J. De Smedt, and W. Verbeke, "Internal Placement: Job Recommender Systems with Social Regularization," presented at *The Control Room of the Future: AI Empowered Dashboards*, Ghent, Belgium, 2022.

S. De Vos, J. De Smedt, M. Verbruggen, and W. Verbeke, "Internal Placement: Job Recommender Systems with Social Regularization," presented at *the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, Grenoble, France, Sep. 19–23, 2022.

S. De Vos, J. De Smedt, and W. Verbeke, "Human Resource Analytics: Employee Journeys from a Process Perspective," presented at *ORBEL36*, Ghent, Belgium, Sep. 12–13, 2022.

S. De Vos, T. Vanderschueren, T. Verdonck, and W. Verbeke, "Robust Instance-dependent Cost-sensitive Learning," presented at *the 32nd European Conference on Operational Research (EURO)*, Espoo, Finland, Jul. 3–6, 2022.

Code availability

GitHub repositories

The code and supplementary documentation for each chapter are available in the following GitHub repositories:

- Chapter 3: `https://github.com/SimonDeVos/turnover_prediction`
- Chapter 4: `https://github.com/SimonDeVos/RecSys_SR`
- Chapter 5: `https://github.com/SimonDeVos/Robust-IDCS`
- Chapter 6: `https://github.com/SimonDeVos/DCF`
- Chapter 7: `https://github.com/SimonDeVos/UMCT`

A list of our lab's repositories is available at: `https://github.com/VerbekeLab`.

Use of generative AI

The text, code, and images in this thesis are my own (unless otherwise specified) and generative AI has only been used in accordance with the KU Leuven guidelines and appropriate references have been added. I have reviewed and edited the content as needed and I take full responsibility for the content of the thesis. Throughout this thesis, generative AI assistance tools (ChatGPT and Grammarly) were used to assist in the writing process to check the writing's grammar, spelling, and readability.

A full list of the doctoral dissertations from the Faculty of Economics and Business can be found at:

https://www.kuleuven.be/english/research/doctoraldefences/archive