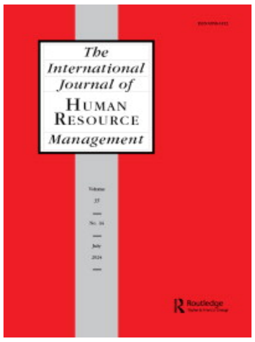


The International Journal of Human Resource Management

ISSN: 0958-5192 (Print) 1466-4399 (Online) Journal homepage: www.tandfonline.com/journals/rijh20

Machine learning with real-world HR data: mitigating the trade-off between predictive performance and transparency

Ansgar Heidemann, Svenja M. Hülter & Michael Tekieli

To cite this article: Ansgar Heidemann, Svenja M. Hülter & Michael Tekieli (2024) Machine learning with real-world HR data: mitigating the trade-off between predictive performance and transparency, The International Journal of Human Resource Management, 35:14, 2343-2366, DOI: 10.1080/09585192.2024.2335515

To link to this article: https://doi.org/10.1080/09585192.2024.2335515

Published online: 01 Apr 2024.

THE INTERNATIONAL JOURNAL OF HUMAN RESOURCE MANAGEMENT
2024, VOL. 35, NO. 14, 2343–2366
https://doi.org/10.1080/09585192.2024.2335515

Routledge
Taylor & Francis Group

Machine learning with real-world HR data: mitigating the trade-off between predictive performance and transparency

Ansgar Heidemann, Svenja M. Hülter and Michael Tekieli

Technical University Dortmund, Dortmund, Germany

ABSTRACT

Machine Learning (ML) algorithms offer a powerful tool for capturing multifaceted relationships through inductive research to gain insights and support decision-making in practice. This study contributes to understanding the dilemma whereby the more complex ML becomes, the more its value proposition can be compromised by its opacity. Using a longitudinal dataset on voluntary employee turnover from a German federal agency, we provide evidence for the underlying trade-off between predictive performance and transparency for ML, which has not been found in similar Human Resource Management (HRM) studies using artificially simulated datasets. We then propose measures to mitigate this trade-off by demonstrating the use of post-hoc explanatory methods to extract local (employee-specific) and global (organisation-wide) predictor effects. After that, we discuss their limitations, providing a nuanced perspective on the circumstances under which the use of post-hoc explanatory methods is justified. Namely, when a 'transparency-by-design' approach with traditional linear regression is not sufficient to solve HRM prediction tasks, the translation of complex ML models into human-understandable visualisations is required. As theoretical implications, this paper suggests that we can only fully understand the multi-layered HR phenomena explained to us by real-world data if we incorporate ML-based inductive methods together with traditional deductive methods.

KEYWORDS

Algorithmic HRM; human resource analytics; machine learning transparency; explainable artificial intelligence; voluntary employee turnover prediction

Introduction

ML algorithms are increasingly used in HRM research to reveal and explain multifaceted phenomena beyond linearity from data, known as an 'ML-based inductive' research method (for an example relating to employee turnover prediction, see King, 2016; Rombaut & Guerry, 2018, 2021,

CONTACT Ansgar Heidemann ansgar.heidemann@tu-dortmund.de Technical University Dortmund, Dortmund, Germany.

Choudhury et al., 2021, Erel et al., 2021, Speer, 2021, Yuan et al., 2021, Chowdhury et al., 2022). This method promises to objectively translate large amounts of raw data into new information provided by otherwise overlooked, complicated interactions between predictors with a level of efficiency not achieved by humans (van den Broek et al., 2021). Generally, this explanatory approach requires neither prior assumptions nor explicit hypotheses, which opens up various opportunities for HR research and practice (Choudhury et al., 2021). Thus, the ML-based inductive research method enables the investigation of multifaceted relationships, to gain exploratory insights and develop theory (Cheng & Hackett, 2021; Putka et al., 2018).

However, a challenge arises due to the complexity of ML algorithms, i.e. the disadvantage of not providing an easily understandable mathematical formula (Erel et al., 2021; Kellogg et al., 2020). High ML model complexity occurs when an ML algorithm capable of modelling flexible functions beyond linearity (neural networks, random forest, etc.) is applied to a large dataset with multifaceted relationships (Choudhury et al., 2021). Complex models achieve high predictive performance, which is the extent of the ML model to solve a prediction task, depending on its context and goal, with different statistical measures (e.g. root-mean-squared error for regression, accuracy for classification). Nevertheless, complex models remain opaque, in that (a) predictors are not understandable (e.g. coming from a complex system itself), (b) relationships between predictors and predictions are hidden and (c) no explanation for a specific prediction is given (Burrell, 2016; Langer & König, 2023).

Simply, there are two sides to a continuum: opacity is a lack of understanding regarding the inner workings of the ML model, while transparency is understanding the relationships between predictors and predictions (Langer & König, 2023). To ascertain the position on this continuum between transparency and opacity, ML algorithm selection is a pivotal determinant. For example, a linear regression model can be considered as inherently transparent, also known as the 'transparency-by-design' approach, because the mechanisms of the mathematical formulas and their parameters, which affect the relationships between predictors and predictions, are interpretable. Generally, it is known that the trade-off between predictive performance and transparency in ML models goes hand in hand with their complexity (Barredo Arrieta et al. 2020, p. 100), but it has not been empirically demonstrated or investigated in real-world HRM applications. Thus, we ask:

> **RQ1:** To what extent does ML in real-world HRM applications face trade-offs between predictive performance and transparency?

We select employee turnover prediction as a representative ML application in HR. This has also recently been investigated in two closely related studies (Choudhury et al., 2021; Chowdhury et al., 2022). We extend these studies by providing an empirical example of the trade-off between predictive performance and transparency in real-world data that was not found using artificially simulated datasets used in either study. Also, like these studies, we apply multiple post-hoc explanatory methods to mitigate the aforementioned trade-off. These methods aim to explain elements of a complex ML model while, such as the influence importance of predictors, maintaining its high predictive performance, which increases its transparency. However, additionally we critique the limitations and ask about the appropriate circumstances for using post-hoc explanatory methods:

> **RQ2:** Under what circumstances, may post-hoc explanatory methods be applied to understand the rationale behind complex ML model predictions?

By answering these research questions, we contribute to the literature by (1) empirically demonstrating and discussing the consequences of the trade-off between predictive performance and transparency, (2) providing a nuanced perspective when the use of post-hoc explanatory methods is justified due to lack of alternatives (3) and outlining the possibilities of revealing the multifaceted effects of complex ML model predictors combined with post-hoc explanatory methods. Thus, we respond to research calls to investigate when and how organisations can switch from opaque to transparent ML (Chowdhury et al., 2022, p. 25). As a theoretical implication, our study exemplifies the need for an inductive method based on ML, given the complexity of real-world HR phenomena that cannot be investigated to the same depth using traditional methods such as linear regression. However, caution is needed when interpreting and deriving implications, as the available post-hoc explanatory methods required for complex models can be misleading. Confirmatory studies with deductive evidence may therefore be complementarily necessary.

The paper is organised as follows. The second section reviews the literature on the multifaceted causes of employee turnover, frameworks for ML-based inductive research methods and the trade-off in ML complexity. The third section summarises the methodology used and presents the empirical dataset. The fourth section presents the results and applies three post-hoc explanatory methods to mitigate a trade-off. Finally, the fifth and sixth sections discuss post-hoc explanatory methods before concluding.

Literature review

Turnover causes are multifaceted

Employee turnover prediction is selected as a representative ML application in HRM because turnover is generally an intricate process (Yuan et al., 2021) resulting from a series of possible sequences of events or 'pathways' (Russell & Sell, 2012, p. 126). Consequently, the literature lists numerous predictors directly influencing employee turnover (e.g. Holtom et al., 2008; Rubenstein et al., 2018). This suggests that turnover is more complex than the largely linear relationships previously studied. However, turnover causes are likely to be nonlinear, with the nature of these relationships between variables changing at different points (e.g. U-shaped), or heterogeneous with different relationships for different subgroups of employees. For example, Gray and Phillips (1994) find a U-shaped relationship between age and turnover, implying that turnover is high among young employees and decreases with age; thereafter, turnover gradually increases again, due to retirement. A negative correlation between tenure and employee turnover is widespread (Rubenstein et al., 2018). Furthermore, Grissom et al. (2016) find a negative relation between salary predictors (total/increase) and turnover in public administration, albeit up to a certain level, whereas turnover at the managerial level might work differently. Lin et al. (2021) emphasise that wage increases significantly reduce voluntary turnover. In addition, they examine the moderating effect on the relationship between wage increases and turnover, finding a significant negative relationship in this regard, albeit only for workers with longer tenure.

In summary, the sophistication of relationships leading to employee turnover is not only caused by linear relationships between several predictors and turnover, but also nonlinearity and heterogeneity in employee subgroups, which lead to mathematical interactions among predictors.

Increasing algorithm complexity

Multifaceted employee turnover prediction phenomena contradict implicit assumptions of linearity in traditional ordinary least squares models, thereby strengthening the rationale for using ML for prediction and knowledge (Erel et al., 2021). Recent literature presents frameworks and proposals for inductive research methods that introduce the principles of modern predictive models (Yuan et al., 2021) and common algorithms (Putka et al., 2018), to embedded ML in scientific methodologies. For example, ML algorithms used for employee turnover prediction include random forest (Choudhury et al., 2021; Speer, 2021), extreme gradient boosting (Erel et al., 2021) and gradient boosting machines (King, 2016). These so-called 'ensemble' algorithms combine hundreds of heterogeneous trees, each of which automatically selects relevant predictors and their weights, groups data into relevant subregions and finds local dependencies (Putka et al., 2018). Similarly, increasing the number of neurons per layer or the number of layers in a neural network leads to what are known as deep learning algorithms that can achieve the same result (Vale et al., 2022).

To avoid system-based ML opacity resulting from increasing complexity, knowledge about algorithms is insufficient, as the amount of information causes information fatigue and meaninglessness (Gal et al., 2020). Complex ML algorithms learn by building their representation of a decision without considering human comprehension, thereby escaping human understanding (Burrell, 2016; Kellogg et al., 2020, p. 372). In contrast to traditional statistical and econometric approaches, non-parametric ML algorithms do not provide a representative mathematical formula that can be easily understood (Erel et al., 2021, p. 3229). However, some do provide other representations that are understandable to humans, such as the rule system of classification trees or the conditional probabilities of naive bayes classifiers (Barredo Arrieta et al. 2020). This means, the decision between a more transparent, simple ML algorithm and a more complex, opaque one is more of a continuum due to the variety of algorithm options available (Langer & König, 2023). Regardless of the specific algorithm, it is widely assumed that higher ML model complexity directly correlates with higher prediction performance, or expressed differently, solves the prediction task more effectively. Interestingly, this is not necessarily the case, as it depends on the specific prediction task and the flexibility required to approximate the underlying data, in particular the amount available and its distribution among the variables' values (Rudin, 2019).

Challenges arising from algorithm opacity

HRM is a high-stakes decision-making environment because individuals are directly affected, and data-driven decisions have various ethical and legal consequences (Gal et al., 2020). To verify fairness and nondiscrimination, input data, data analysis procedures and the links between results and conclusions must be transparent so that predictions can be validated. One reason is that correlations can be established but causal relationships are not implied. Opaque algorithms jeopardise the ability to test for adverse impacts by challenging differences between groups with existing evidence (Charlwood & Guenole, 2022, p. 7; Meijerink et al., 2021, p. 2549), thus hampering unavoidable critical thinking about predictive composites, causal inferences and subgroup differences (Putka et al., 2018, p. 711).

ML transparency is important not only for practical applications, but also for researchers who want to understand, for example, what nonlinear effects independent variables have. Somers et al. challenge the linear assumptions of HR theory regarding employee well-being and demonstrate the use of complex ML models such as neural networks to detect nonlinearities. The tools used, such as three-dimensional visualisation, help uncover the existence of nonlinear phenomena but do not provide sufficient transparency to extract explicit relationships or to include more than two predictors (Somers et al., 2021). Yuan et al. (2021) use ML to identify the key predictors of employee turnover in a sample of SMEs. However, because of the algorithm's opacity, it is vague how the predictors affect turnover. For example, perceived unfairness is the most important predictor, but it is unclear at what point perceived unfairness triggers turnover.

In summary, it is essential to the success of ML in HR applications to (1) disclose how an algorithm makes a decision, (2) ensure the right to challenge an outcome and (3) provide expertise to address these challenges (Cheng & Hackett, 2021). Ultimately, ML transparency is needed to increase trust in ML and question its responsible use (Chowdhury et al., 2022). In the absence of sufficient empirical examples, algorithm opacity is an important blind spot in algorithmic HRM research (Edwards et al., 2022, p. 5; Meijerink et al., 2021, p. 2550).

Tackling algorithm opacity with technical solutions

Research outside of HR has introduced methods to increase ML model transparency while maintaining high predictive performance, thus mitigating the aforementioned trade-off. Explainable Artificial Intelligence (XAI), also known as 'Interpretable ML', is a rapidly evolving interdisciplinary research area offering multiple technical solutions (Barredo Arrieta et al. 2020; Molnar, 2022). Moreover, XAI methods can be divided into either algorithmic models understandable to humans with appropriate knowledge (transparency-by-design), or methods using approximation to explain elements of complex—and thus opaque—models through simplified representations (post-hoc explanatory methods) (Barredo Arrieta et al. 2020; Langer & König, 2023). Interestingly, documented applications of these technical solutions remain sparse in HR (Langer & König, 2023), albeit the first examples are promising. Choudhury et al. (2021) demonstrate the use of global (enterprise-wide) post-hoc explanatory methods by applying feature importance methods and partial dependence plots to study nonlinear interactions between employee turnover and its predictors. Yakusheva et al. (2022) also use partial dependence plots to reveal an unexpected U-shaped relationship between higher staffing levels and the number of readmissions. Erel et al. (2021) provide an example of how to gain insights into opaque models of complex algorithms by quantifying the contribution of each predictor for director performance and turnover. Additionally, they demonstrate the existence of nonlinear and heterogenous relationships by breaking down the effects of predictors, using local post-hoc explanatory methods. In the HR literature, Chowdhury et al. (2022) use a local (employee-specific) post-hoc explanatory method for employee turnover prediction, called 'local agnostic model explanations' (LIME).

Knowledge gap

These examples show that post-hoc explanatory methods can be applied to complex ML models, improving our understanding of employee turnover causes. However, the proposed frameworks do not specify under what circumstances these methods should be used. One reason for this is the lack of a connection with the trade-off between predictive performance and transparency, which is not found in studies using artificially simulated datasets (Choudhury et al., 2021; Chowdhury et al., 2022). Instead, Choudhury et al. note that complex ML models (e.g. neural networks, random forests) offer only 'small performance increases over the baseline logistic regression' and explain this marginal performance gain as 'meaningful interactions and nonlinearities among variables are only relevant for a small subset of the data' (Choudhury et al., 2021, p. 48). Likewise, but independently, Chowdhury et al. come to a similar conclusion, as they find 'no significant differences' in predictive performance between transparent and more complex algorithms (Chowdhury et al., 2022, p. 13). This warrants further investigation of the trade-off between predictive performance and transparency in practical, real-world HR datasets. The findings could serve as a basis for understanding the justified use of more complex, opaque ML algorithms and subsequent post-hoc explanatory methods.

Methodology

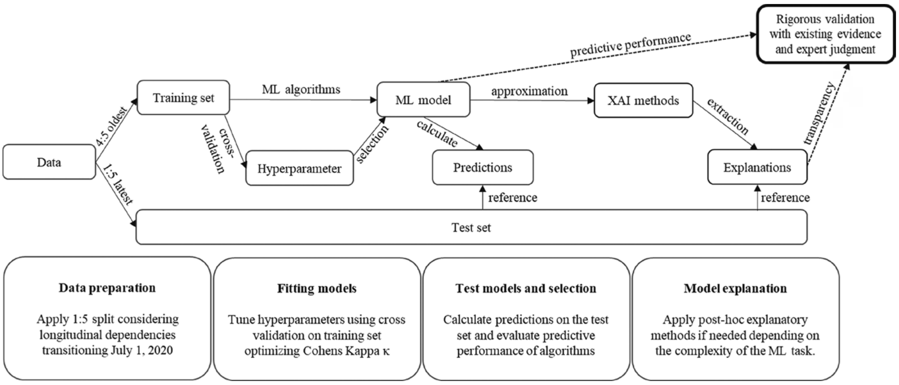
In turnover prediction, inductive research methods help analyse actual turnover in longitudinal data available through HR information systems (HRIS) (King, 2016; Rombaut & Guerry, 2018, 2021), also known as 'attrition modeling' (Speer, 2021). Building on these frameworks, we introduce a complete process for ML model development with addressing ML opacity depending on prediction task complexity. The post-hoc explanatory methods approximate a complex ML model and extract various explanations that provide the transparency needed to question the logic behind predictions. Figure 1 presents a schematic overview of an inductive research framework.

Data preparation

We adapt Rombaut and Guerry's inductive approach to predict actual voluntary employee turnover, using data from HRIS, and extend it by using complex ML algorithms and external predictors (Rombaut & Guerry, 2018). The direct effect (black arrow) is examined by bridging several implicit established constructs, such as organisational commitment and payment satisfaction, thus taking a complementary approach to survey-based methods (see Figure 2).

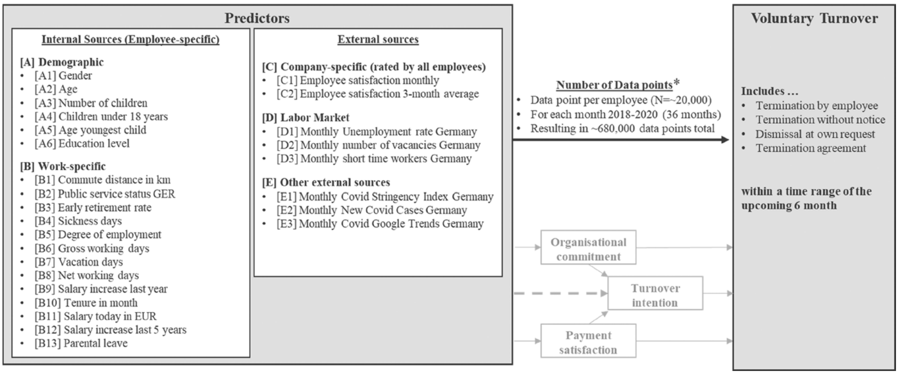
The inductive ML methodology is based on a uniquely collected dataset of 680,000 data points from a German federal agency. Each row represents an employee in combination with a specific month and contains a dichotomous variable about whether they left the company in the six months following the observed month. This time horizon was chosen based on practical response time requirements for countermeasures; it is also consistent with other research (Speer, 2021, p. 8). We used a data protection-approved process that builds on existing data agreements and anonymisation procedures during data collection.

The final dataset had 27 predictors (see Figure 2), most of which (19) originated from internal HR databases and are available for all 20,000



Rigorous validation with existing evidence and expert judgment
predictive performance
Training set
ML algorithms
ML model
approximation
XAI methods
4:5 oldest
cross-validation
selection
calculate
extraction
transparency
Data
Hyperparameter
Predictions
Explanations
1:5 latest
reference
reference
Test set
Data preparation
Apply 1:5 split considering longitudinal dependencies transitioning July 1, 2020
Fitting models
Tune hyperparameters using cross validation on training set optimizing Cohens Kappa κ
Test models and selection
Calculate predictions on the test set and evaluate predictive performance of algorithms
Model explanation
Apply post-hoc explanatory methods if needed depending on the complexity of the ML task.

Figure 1. Inductive research process using machine learning with an out-of-sample test.



Predictors
Internal Sources (Employee-specific)
[A] Demographic
[A1] Gender
[A2] Age
[A3] Number of children
[A4] Children under 18 years
[A5] Age youngest child
[A6] Education level
[B] Work-specific
[B1] Commute distance in km
[B2] Public service status GER
[B3] Early retirement rate
[B4] Sickness days
[B5] Degree of employment
[B6] Gross working days
[B7] Vacation days
[B8] Net working days
[B9] Salary increase last year
[B10] Tenure in month
[B11] Salary today in EUR
[B12] Salary increase last 5 years
[B13] Parental leave
External sources
[C] Company-specific (rated by all employees)
[C1] Employee satisfaction monthly
[C2] Employee satisfaction 3-month average
[D] Labor Market
[D1] Monthly Unemployment rate Germany
[D2] Monthly number of vacancies Germany
[D3] Monthly short time workers Germany
[E] Other external sources
[E1] Monthly Covid Stringency Index Germany
[E2] Monthly New Covid Cases Germany
[E3] Monthly Covid Google Trends Germany
Number of Data points*
Data point per employee (N=~20,000)
For each month 2018-2020 (36 months)
Resulting in ~680,000 data points total
Voluntary Turnover
Includes …
Termination by employee
Termination without notice
Dismissal at own request
Termination agreement
within a time range of the upcoming 6 month
Organisational commitment
Turnover intention
Payment satisfaction

Figure 2. Acquired longitudinal data for historical voluntary turnover. The direct impact (black) is investigated in this study.

employees over 36 months. In addition, data on global employee satisfaction and external data were integrated into the dataset. Table 1 provides a basic descriptive statistical overview of the relevant variables. Please note that the low voluntary turnover rate (below 1%) is not unusual in the public sector (e.g. Grissom et al., 2016, p. 243), thereby making prediction particularly difficult (Kuhn, 2019).

Algorithm selection

Table 2 summarises the selected algorithms chosen from common options used for employee turnover prediction (Chowdhury et al., 2022) or other inductive research (Putka et al., 2018). The advantage (transparency vs.

Table 1. Descriptive statistical overview of available variables in the acquired dataset.

Nominal variables	N (679463)	Continuous variables	Mean	Std. Dev.
A1_Gender		A2_Age	49.09	9.13
woman	71.0%	A3_Number_of_children	1.13	0.94
men	25.7%	A5_Age_youngest_children	20.46	10.09
…unknown	3.3%	B1_Commute_distance_in_km	15.28	17.62
		B3_Early_retirement_rate	6.08	17.12
A4_Children_under_18_years		B4_Sickness_days	2.07	4.73
0	72.6%	B5_Degree_of_employment	90.75	14.00
1	27.4%	B6_Gross_working_days	20.06	2.40
		B7_Vacation_days	2.33	3.55
A6_Education_level		B8_Net_working_days	15.34	5.78
lower degree university	49.5%	B9_Salary_increase_last_year	0.16	0.22
vocational training	46.7%	B10_Tenure_in_month	272.17	127.18
higher graduation university	3.8%	B11_Salary_today_EUR	4068.64	889.97
		B12_Salary_increase_last_5_years	0.27	0.26
B2_Public_service_status_GER		B13_Parental_leave	0.06	1.11
No	78.0%	C1_Overall_employee_satisfaction	3.36	0.63
Yes	22.0%	C2_Employee_satisfaction_moving_average	3.50	0.32
		D1_Monthly_unemployment_rate_Germany	5.36	0.48
Voluntary turnover		D2_Monthly_number_of_vacancies_Germany	721.10	94.60
No_Turnover	**99.2%**	D3_Monthly_short_time_workers_Germany	1134.91	1659.11
Turnover	**0.8%**	E1_Monthly_Covid_strigency_index_Germany	19.67	29.15
		E2_Monthly_New_Covid_Cases_Germany	69835.00	181549.00
		E3_Monthly_Covid_Google_trends_Germany	74.70	116.21

Table 2. Selected ML algorithms with their primary advantage according to the transparency vs. performance trade-off.

Algorithm	*R* base function/package	Advantage
Generalised Linear Model	*glm*	Transparency
Elastic Net Regression	*glmnet*	Transparency
Classification Tree	*rpart2*	Transparency
Naïve Bayes	*naivebayes*	Transparency
Random Forest	*ranger*	Performance
Extreme Gradient Boosting	*xgbDart*	Performance
Generalised Boosted Machine	*gbm*	Performance
Feed Forward Neural Network	*nnet*	Performance

performance) is drawn from the computer science literature, which order algorithms according to their comprehensible representational capabilities (Barredo Arrieta et al. 2020, p. 90).

Fitting models on training data

We use three-way partitioning by initially training several algorithms and their hyperparameters to optimise Cohen's Kappa κ on training and validation data (cross-validation on first 24 months between January 2018 and June 2020). Cohen's Kappa ” is a performance indicator that expresses the chance-adjusted proportion of correctly predicted outcomes (Cohen, 1960). It varies between zero and one, providing an interpretation similar to the traditional R-squared regression (Yakusheva et al., 2022, p. 315). By optimising Cohen's Kappa κ instead of other common evaluation methods (e.g. Accuracy, Receiver Operating Characteristic = ROC), we are able to achieve higher predictive performance across all algorithms when tuning hyperparameters, as it is better suited to address the challenge of an imbalanced dataset (Kuhn, 2019).

Evaluate model performance on test data

Evaluating the predictive performance of an ML model is critical to ensure the model's suitability for providing valuable insights. We test the predictive performance of each algorithm with out-of-sample test data (last six months between July 2020 and December 2020).

Table 3. Predictive performance measures of all used algorithms on test data.

Algorithm	Advantage	κ	ROC	Precision	Recall
Random Forest*	Performance	**0.26***	0.87	0.18	0.54
Extreme Gradient Boosting	Performance	**0.24**	0.85	0.18	0.36
Generalised Boosted Machine	Performance	**0.22**	0.87	0.18	0.34
Classification Tree	Transparency	**0.18**	0.68	0.23	0.15
Feed Forwards Neural Network	Performance	**0.16**	0.68	0.17	0.16
Generalised Linear Model	Transparency	**0.12**	0.77	0.23	0.09
Elastic Net Regression	Transparency	**0.12**	0.76	0.20	0.09
Naïve Bayes	Transparency	**0.07**	0.59	0.05	0.23

Table 4. Confusion matrices on test data of random forest (highest predictive performance overall) and classification tree (most successful alternative algorithm with advantage transparency).

Random forest				Classification tree			
		Reference				Reference	
Prediction		*No Turnover*	*Turnover*	***Prediction***		*No Turnover*	*Turnover*
	No Turnover	111,089 (96.60%)	513 (0.45%)		*No Turnover*	113,324 (98.53%)	940 (0.82%)
	Turnover	2,800 (2.43%)	**598** (0.52%)		*Turnover*	565 (0.49%)	**171** (0.15%)

Results

The predictive performance measure results are reported in Table 3.

We further compare the ML models with the highest predictive performance overall with the most successful algorithm (random forest), with the advantage of transparency (classification tree). The confusion matrices in Table 4 show the amount of cases on test data divided into positive and negative cases as well as correct and incorrect predictions.

The random forest model successfully identifies 598 (=52%) employee turnover cases. Overall, it provides a *fair* improvement over random guesses ($\kappa = 0.26$), according to common κ interpretation (Landis & Koch, 1977). Measured by the recall (the ratio of true-positive cases to all positive cases) of 0.54, the model solves over half of the prediction task while providing sufficient precision (the ratio of true-positive cases to all positive predictions) of 0.18. Thus, given the challenge of highly imbalanced classes, due to the rarity of employee turnover cases, the difficult task of prediction is solved adequately, with an acceptable number of false-positive predictions.

The classification tree model successfully identifies 171 (=15%) cases and provides a *slight* improvement over random guesses ($\kappa = 0.18$). The model successfully identifies 427 fewer cases of employee turnover than the random forest model. While the precision of 0.23 is slightly higher, the recall of 0.15 is significantly lower than the random forest. Hence, the majority of turnover cases are not detected, suggesting that the classification tree cannot predict diverse causes of employee turnover (Russell & Sell, 2012, p. 126). The same applies to the other transparent ML models with even lower overall performance ($\kappa < 0.12$). The low predictive performance of the two linear models (Generalised Linear Model, Elastic Net Regression) indicates that linear assumptions might only be valid up to a certain extent.

Two findings emerge from these results. First, while all ML models can detect instances of employee turnover better than chance ($\kappa > 0$), they differ significantly in their predictive performance. Second, a trade-off between predictive performance and transparency is revealed, with more transparent algorithms achieving lower predictive performance (see Figure 3). The only exception—classification trees achieve higher performance than feed forward neural networks—is due to tree-based methods tending to outperform neural networks for tabular data (Shwartz-Ziv and Armon 2022). Consequently, XAI's transparency-by-design approach is not sufficient for identifying various nonlinear or heterogenous causes for employee turnover as the ML model cannot predict most turnover cases.

Applying post-hoc explanatory methods to complex ML models

Thus, we apply three popular post-hoc explanatory methods as the remaining options to gain insights from the random forest model. The choice of method from among numerous alternatives is beyond the scope of this study, so we refer the reader to the technical literature (e.g. Molnar, 2022). All three methods are implemented with the R package 'IML' (Interpretable ML) (Molnar et al., 2018).

Global feature importance

Global feature importance is extracted using the random permutation strategy, which determines the factor by which the model's classification error increases when feature values are randomly shuffled, thereby breaking the relationship between the feature and the true outcome (Molnar, 2022).

Figure 4 reveals that several demographic and work-specific predictors have the highest feature importance. *B1_commute_distance* and *B4_sickness_days* are among the top three features, consistent with other studies using data from the HRIS and finding them significant (Rombaut & Guerry, 2021). External features generally have low feature importance but still contribute to predictive power. However, when comparing feature importance with other studies, it is important to note that results may vary depending on the features included in the dataset, the organisational context and changes over time (e.g. before and after Covid-19). A limitation of this high-level global feature importance method is that it is not suitable for examining whether the feature increases or decreases turnover risk.

Accumulated local effects

Accumulated Local Effects (ALE) describes how a single predictor influences prediction (strength, positive/negative contribution) on average,

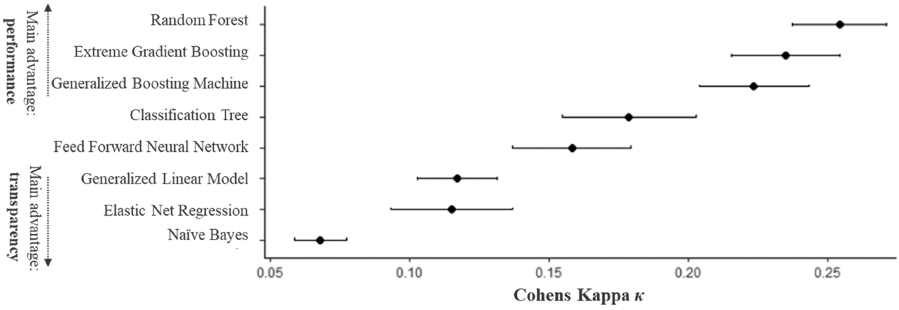


Figure 3. Predictive performance vs. transparency trade-off on test data (confidence interval 0.95).

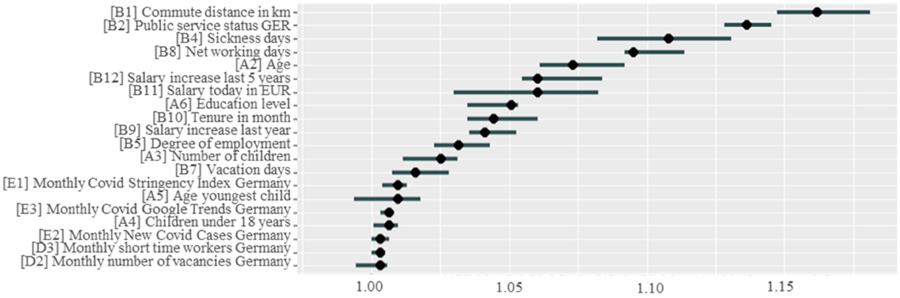


Figure 4. Permutated feature importance for the top 20 predictors (relative change in performance, confidence interval 0.95).

considering all employees in that local interval (Apley & Zhu, 2020). Compared to alternative visualisation techniques on a global level (e.g. partial dependence plots), ALE is preferred when predictors correlate (Apley & Zhu, 2020; Molnar, 2022). Figure 5 highlights the ALE plots for the top 12 predictors, sorted by descending importance.

The ALE representation helps compare results with existing empirical evidence. Consistent with research, we first find a higher turnover probability for employees with a higher *commuting distance* (Rombaut & Guerry, 2021), but there seems to be a threshold around 15 kilometres—after which turnover probability no longer increases but slowly decreases. Second, we find that high—especially complete—absenteeism is an early indicator of turnover, as reflected in the *sickness days* and *net working days* ALE plots (Rubenstein et al., 2018, p. 42).

Additionally, we find that turnover decreases for men in line with age; interestingly, turnover among women increases. Since women form the majority in the considered organisation, this also determines the direction for all employees. This chart clearly shows heterogeneity between subgroups. One HR manager cited extensive measures to retain young mothers as a particular reason for this phenomenon. Also interesting is that there is no negative correlation between *length of service* and turnover, which is widespread in the literature, indicating organisational particularity. Instead, we see a U-shaped turnover probability in relation to tenure. Overall, the *age* and *tenure* results support the finding that these relationships are not as clear-cut in practice as is often assumed (Gray & Phillips, 1994, p. 825).

Altogether, the ALE results are mostly in line with current findings; however, they also reveal organisational particularities, nonlinear relationships and interactions between predictors. However, one criticism of ALE is that it can lead to unstable and inaccurate predictions due to collinearity between predictors in intervals where instances are rare (Molnar, 2022).

Local feature effects

Finally, we apply SHapley Additive exPlanations (SHAP), a local post-hoc explanatory method computing the effect of each predictor at the employee-specific level (Lundberg and Lee 2017). SHAP's local approach is similar to its popular alternative LIME, in that it divides the complex ML model into several mathematical vicinities (Chowdhury et al., 2022). What distinguishes SHAP explanations from LIME and others is their additive nature due to the game-theoretic approach, which facilitates a simpler understanding. SHAP breaks down the probability of voluntary turnover for each employee into SHAP values that quantify increasing or decreasing effects. Ultimately, the sum of the SHAP values is the difference taken from the average predicted probability of turnover for all employees (Lundberg and Lee 2017; Erel et al., 2021). Figure 6 shows the relevant predictors with their values for two different employees, ordered by their additive explanatory contribution.

In Figure 6, a part-time employee (*Degree of employment* 27%, *Gross working days* 9) did not attend work in the last month for unknown reasons (*Net working days* 0). Together with the existing *early retirement rate* (50%) and salary-related predictors, this indicates a higher turnover risk. Interestingly, in this particular case, the *Salary today in EUR* (3,781) and *Salary increase in the last 5 years* (12%) increase turnover probability, which is not clear in the ALE plots (Figure 5). Moreover, the explanatory contribution of *Tenure*, *Sickness days*, *Age* and *Education level* is relevant in example two but not in example one.

In summary, SHAP values at the individual (local) level for both examples are mostly consistent with ALE results. Nevertheless, a closer look at the local (employee-specific) compared to the global (company-wide) level offers more opportunities to challenge the results with existing evidence. The different selection of relevant predictors, as well as the contrary explanatory contribution, argues for interactions between predictors and different reasons for turnover in employee subgroups. Thus, our results support the rationale for using ML and post-hoc methods to study employee turnover (Erel et al., 2021, p. 3247). As discussed later, it should be noted that local post-hoc explanatory methods like SHAP are often criticised because their explanations might be unstable and can therefore provide (intentionally) misleading explanations (Ghassemi et al., 2021; Vale et al., 2022).

Discussion

The increasing complexity of ML-based inductive research methods and algorithmic HRM leads to challenges, most notably ML opacity.

Consequently, stakeholders may not act optimally, based on the ML predictions and insights proposed by algorithms, or accept the results (Meijerink et al., 2021, p. 2550). Similarly, ML-based inductive research offers a useful methodology only if the ML models are sufficiently transparent to extract insights.

Predictive performance vs. transparency trade-off

Accordingly, we come back to RQ1. Our results empirically demonstrate a significant trade-off between predictive performance and transparency in real-world employee turnover prediction data that is not found in artificially simulated datasets used in similar studies (Choudhury et al., 2021; Chowdhury et al., 2022). This suggests that artificially generated datasets may not holistically capture the sophistication of the various causes of employee turnover so that transparency-by-design approaches may be sufficient. Our results suggest that real-world HR prediction tasks like employee turnover prediction face a level of complexity that cannot be adequately solved by simpler transparency-by-design algorithms. The relationships between predictors and employee turnover elude linear relationships, making nonlinearities and heterogeneous interactions essential for accurately predicting the multifaceted causes of turnover (Putka et al., 2018, p. 721). Thus, the transparency-by-design approach may not be applicable in some real-world HR applications, leaving complex ML models. This supports recent theoretical research citing that opacity is a key characteristic and default of ML due to system-based opacity, caused by ML algorithm complexity and the scale required for meaningful application (Burrell, 2016; Kellogg et al., 2020; Langer & König, 2023).

Extracting multifaceted relationships

To understand the rationale behind complex ML model predictions, such as the random forest model, we propose three post-hoc explanatory methods that support the extraction of multifaceted insights on predictions. The three proposed XAI methods support translating the patterns used by opaque ML algorithms into human-understandable results. These patterns can also be multifaceted, such as nonlinearity and heterogeneous feature interactions. To accomplish this, all three proposed methods account for the extent to which predictors influence employee turnover prediction. However, each method takes a distinct mathematical approach and focuses on different aspects of information (e.g. local vs. global). ALE helps identify nonlinear and heterogeneous causes of turnover among employee subgroups, including otherwise unnoticed insights; for example, in this empirical setting for turnover, (1) a U-shaped relationship with commuting distance (Figure 5) (2) interactions between the predictors age and gender (Figure 5), specifics of relevant predictors in diverse subgroups (Figure 6). SHAP highlights the specifics at the local level, allowing for the analysis of individuals. Such inductive findings can serve as the basis for theory development or complement deductive methods for deriving prescriptive and generalisable implications (Yuan et al., 2021, p. 3).

Criticism of post-hoc explanations

Finally, we discuss limitations of post-hoc explanations and their consequences for appropriate use. Other studies use post-hoc explanatory methods to demonstrate their capabilities for deeper investigation, although transparency-by-design approaches using linear logistic regression are sufficient (e.g. Choudhury et al., 2021; Chowdhury et al., 2022). In these cases, post-hoc explanatory methods (especially on local level) have been seriously criticised for high-stakes decision-making

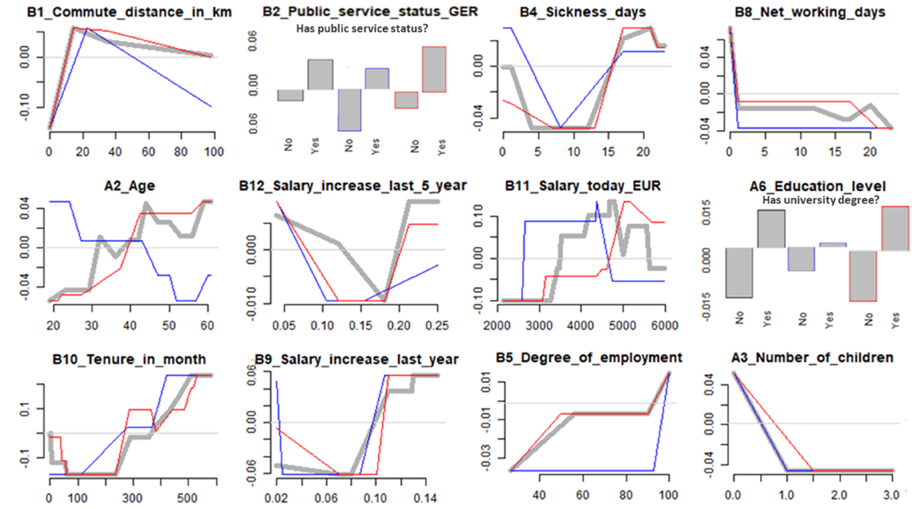


Figure 5. Accumulated local effect plots for the top 12 predictors according to permutated feature importance. Grey = all employees, blue = men, red = women.

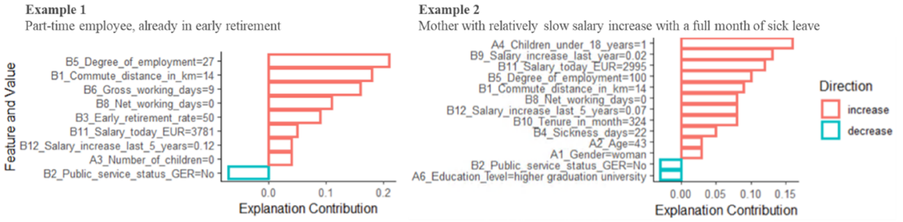


Figure 6. Top-10 SHAP values for two employees successfully predicted turnover candidates (true-positive).

like HRM by technical experts, due to several risks resulting from the simplistic inaccuracies, and instead call for transparency-by-design approaches (Rudin, 2019). Similarly, healthcare researchers refer to post-hoc explanations as a 'false hope' because stakeholders may misinterpret the ML model's capabilities, mainly due to possible explanation inaccuracy (Ghassemi et al., 2021, 745) resulting from approximating the ML model to the real-world as well as the post-hoc explanatory method to the ML model. Due to this heuristic nature, the rationale behind predictions must be carefully questioned (Cheng & Hackett, 2021). As a result of the resulting unreliability and superficial character of post-hoc explanation, Ghassemi et al. advise to not use post-hoc explanatory methods to access possible biases towards certain populations, to reassure for correct individual decisions, to increase trust or to justify acceptance of the ML model. They therefore call for using global explanatory methods only to understand how the ML model behaves at a global level and emphasise that they should be combined with rigorous prediction validation processes for different populations (Ghassemi et al., 2021, 745). Legal studies follow a similar line of reasoning, based on technical limitations such as results instability, stating that post-hoc explanatory methods cannot establish abstinence of discrimination caused by various biases embedded in ML models (Vale et al., 2022).

Justification for post-hoc explanatory methods

Concluding, in response to RQ2, post-hoc explanatory methods help examine the rationality of complex ML models to understand why they make incorrect predictions or reveal possible adverse impacts at the global level. However, given their limitations, they should be used with caution to justify individual-level personnel decisions. Post-hoc explanations should not be used as the only truth when formulating and justifying decisions where the stakes are high, but they are a helpful addition. Together, we argue for a nuanced perspective on the justified use of post-hoc explanations methods in circumstances where the transparency-by-design approach fails and managers are aware of its limitations. In these cases, despite the criticisms, post-hoc explanatory methods are the most feasible way to mitigate the trade-off between predictive performance and transparency. Furthermore, as they provide only an approximation of complex ML models, they may serve as a source of information but cannot be regarded as the definitive ground truth. Thus, the validity of ML predictions and post-hoc explanations must be critically questioned following existing evidence, theory and domain knowledge to ensure causality (Erel et al., 2021, p. 3245). As an illustration, the plausibility of the relationships presented in the ALE plot (Figure 5) is supported by employee turnover studies indicating similar results. Similarly, the consistency of patterns discovered in relation to expertise can be evaluated against other organisation-specific information, such as survey-based data and exit interviews.

Implications for research

We contribute to the ML in HRM literature in three ways. First, we empirically demonstrate the trade-off between predictive performance and transparency in real-world data. In this context, our research documents that ML opacity may be unavoidable in some real-world HR prediction tasks, as the underlying algorithms must have considerable complexity to achieve adequate performance. Therefore, transparency-by-design approaches are not always applicable, as predictive performance does not sufficiently solve the prediction task. Second, we provide a nuanced perspective on the justified use of post-hoc explanatory methods, i.e. to mitigate the trade-off between predictive performance and transparency when it occurs in the HRM application—and transparency-by-design approaches are not sufficient. Third, we demonstrate the complementary use of local and global post-hoc explanatory methods to understand nonlinearities and heterogeneity in complex ML models.

Together, these three contributions have a methodological implication and can serve as a guideline when applying the ML-based inductive research method. Based on the extent of the trade-off with different algorithms during the experimentation phase, researchers may decide to adopt a transparency-by-design approach if it is suitable to solve the prediction task adequately. Alternatively, the three proposed post-hoc explanatory methods can help extract additional nonlinear and heterogenous insights. In contrast to existing studies that use a single post-hoc explanatory method (Choudhury et al., 2021, Chowdhury et al., 2022), we provide a broader picture of available technical solutions and demonstrate the joint use of local and global post-hoc methods in a complementary manner. However, for both methods, reliance on the human ability to make ethical and moral judgments, in order to critically question and rigorously validate the results of ML algorithms should be emphasized in both research and practice. To advance HRM knowledge, we therefore advocate a hybrid methodological approach, combining ML-based inductive, exploratory findings with validation by existing evidence and theories or by subsequent deductive, confirmatory studies. Accordingly, HRM research benefits from ML-based methods by (1) testing and refining theories and (2) expanding the explanatory range of theories (Leavitt et al., 2021). Additionally, we also make a secondary contribution to the

HRM literature on employee turnover by suggesting that examining nonlinearities and heterogeneity is important in fully capturing the intricacies of the various pathways resulting in turnover.

Implications for practice

For HRM practitioners, the well-known promises of objective and accurate ML-based decisions are only valid if models are not kept opaque. Instead, causality behind predictions must be verified through a hybrid human-ML development process before they can be used for individual decision-making (van den Broek et al., 2021). In this context, applying knowledge of the revealed trade-off between predictive performance and transparency guides a more informed algorithm selection in ML development. As technically-oriented functions such as data scientists may lack an understanding of opaque ML models' impact on HRM applications (Charlwood & Guenole, 2022, p. 2), we suggest that organisations invest in educating HR staff about the consequences of ML complexity. In evaluating the potential impact on individual employees, HR managers can then weigh the predictive power or transparency of ML models on a continuum.

Educating HR decision-makers about the capabilities and limitations of post-hoc explanatory methods is also advisable; otherwise, they may not be able to 'use their tacit experience and social intelligence (based on intuitive thinking) to determine the accuracy of the model' when using complex ML models (Chowdhury et al., 2022, p. 20). Properly applied complex ML algorithms combined with post-hoc explanatory methods can mitigate the trade-off between predictive performance and transparency. This contributes to realise ML's potential to study and predict voluntary employee turnover and understand its multifaceted causes. The ALE plots' global explanations identify possible organisation-wide improvements or new retention strategies targeting influencing predictors, e.g. we find that commuting distance is an important turnover determinant, implying that a higher home-office ratio might be an effective countermeasure. ML predictions can be used in conjunction with nonlinear insights from ALE or individual-level explanations from SHAP to identify heterogeneous retention strategies for departments and demographic subgroups, or to personalise for employees, which would not be possible with linear models (Chowdhury et al., 2022, p. 15).

Ultimately, and particularly for public sector organisations, post-hoc explanatory methods might be a door-opener for ML-based methods. Public organisations must be highly transparent in their decision-making due to their high societal responsibility. Consequently, many public organisations have not yet started integrating ML-based solutions into their operations, one central reason for which is legal uncertainty related to the lack of transparency of such approaches. This was also one main reason for the examined federal agency in testing post-hoc explanatory methods and might be of similar interest for many other public sector executives seeking solutions in integrating reliable and trustworthy algorithmic HRM in their day-to-day business (Chowdhury et al., 2022, p. 24).

Limitations and directions for future research

Our first limitation is that inductive HRIS-oriented research does not provide deep insights in the same way that studies based on surveys do. Thus, pure data from HRIS and an inductive perspective should be used to complement existing theory-oriented research methods, e.g. for characterising risk groups or sub-organisational specifics (Rombaut & Guerry, 2018, p. 97). In future research, the framework presented herein with post-hoc explanatory methods can be applied not only to data available in HRIS, but also to psychological studies' survey-based data. Thus, the inductive research method can serve as a complementary framework to study multifaceted relationships between multiple predictors. Unlike deductive research, comprehensive theory is not required to specify relationships in advance, thereby helping identify nonlinear and heterogenous relationships also for psychological constructs that predictors in studies based on linear models might miss (Putka et al., 2018, p. 690). Accordingly, we endorse the management literature (Leavitt et al., 2021; Valizade et al., 2024) by recommending ML as a powerful tool for quantitative research. Our results herein support the prioritisation of algorithms in terms of 'transparency-by-design' or more complex algorithms in coordination with post-hoc explanatory methods.

Second, it is important to note that advantages and disadvantages of the three applied post-hoc explanatory methods may not be generalisable to options beyond the scope of this work. Further, HRM research should be aware of the rapid development of ML algorithms and XAI, as computer scientists attempt to resolve the trade-off by developing new transparency-by-design algorithms to increase their predictive performance, as well as new post-hoc explanatory methods (e.g. Barredo Arrieta et al. 2020).

Third, we focus on one federal agency in an in-depth examination, which means we forfeit some generalisability (Yin, 2013, p. 325). We encourage future research to investigate implications of the trade-off in diverse other (multi-)national settings and in other ML-based prediction tasks in HR besides employee turnover prediction, such as employee selection, training and management.

Conclusion

This study discloses the trade-off between predictive performance and transparency in ML empirically demonstrated in a real-world HRM application, meaning that complex—and therefore opaque—algorithms have significantly better predictive performance. For sophisticated prediction tasks such as employee turnover, the underlying algorithms must have considerable complexity and therefore cannot be adequately solved by simpler transparency-by-design ML algorithms. However, by applying three post-hoc explanatory methods to successful but opaque ML models, insights are gained that include nonlinear and heterogeneous causes of employee turnover. We argue that post-hoc explanatory methods help mitigate the trade-off if they are used properly according to their limitations and are not blindly trusted, which is why we emphasise a nuanced perspective to their justified use. We hope that this paper motivates further research regarding ML transparency as a necessity to pave the way for an ethically and a legally compliant ML-augmented decision-making process that benefits the organisation and—most importantly—all employees.

Acknowledgments

We are grateful to all participants in the 1st EIASM Workshop on People Analytics & Algorithmic Management (PAAM) 2022 in Dublin for their valuable discussions and helpful suggestions. We also sincerely thank the editor and reviewers for taking the time to review our manuscript and providing constructive feedback to improve it.

Disclosure statement

No potential conflict of interest is reported by the author(s).

Data availability statement

Data available on request due to privacy/ethical restrictions.

References

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1059–1086. https://doi.org/10.1111/rssb.12377

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Charlwood, A., & Guenole, N. (2022). Can HR adapt to the paradoxes of artificial intelligence? *Human Resource Management Journal*, *32*(4), 729–742. https://doi.org/10.1111/1748-8583.12433

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review*, *31*(1), 100698. https://doi.org/10.1016/j.hrmr.2019.100698

Choudhury, P., Allen, R. T., & Endres, M. G. (2021). Machine learning for pattern discovery in management research. *Strategic Management Journal*, *42*(1), 30–57. https://doi.org/10.1002/smj.3215

Chowdhury, S., Joel-Edgar, S., Dey, P. K., Bhattacharya, S., & Kharlamov, A. (2022). Embedding transparency in artificial intelligence machine learning models: Managerial implications on predicting and explaining employee turnover. *The International Journal of Human Resource Management*, *34*(14), 2732–2764. https://doi.org/10.1080/09585192.2022.2066981

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Edwards, M. R., Charlwood, A., Guenole, N., & Marler, J. (2022). HR analytics: An emerging field finding its place in the world alongside simmering ethical challenges. *Human Resource Management Journal*. 1-11. https://doi.org/10.1111/1748-8583.12435

Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2021). Selecting directors using machine learning. *The Review of Financial Studies*, *34*(7), 3226–3264. https://doi.org/10.1093/rfs/hhab050

Gal, U., Jensen, T. B., & Stein, M.-K. (2020). Breaking the vicious cycle of algorithmic management: A virtue ethics approach to people analytics. *Information and Organization*, *30*(2), 100301. https://doi.org/10.1016/j.infoandorg.2020.100301

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet. Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gray, A. M., & Phillips, V. L. (1994). Turnover, age and length of service: A comparison of nurses and other staff in the National Health Service. *Journal of Advanced Nursing*, *19*(4), 819–827. https://doi.org/10.1111/j.1365-2648.1994.tb01155.x

Grissom, J. A., Viano, S. L., & Selin, J. L. (2016). Understanding employee turnover in the public sector: Insights from research on teacher mobility. *Public Administration Review*, *76*(2), 241–251. https://doi.org/10.1111/puar.12435

Holtom, B. C., Mitchell, T. R., Lee, T. W., & Eberly, M. B. (2008). 5 turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future. *Academy of Management Annals*, *2*(1), 231–274. https://doi.org/10.1080/19416520802211552

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410. https://doi.org/10.5465/annals.2018.0174

King, K. G. (2016). Data analytics in human resources. *Human Resource Development Review*, *15*(4), 487–495. https://doi.org/10.1177/1534484316675818

Kuhn, M. (2019). Building predictive models in R using the caret package. Available online at https://topepo.github.io/caret/index.html, updated on 3/27/2019, checked on 12/30/2021.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical Kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*(2), 363. https://doi.org/10.2307/2529786

Langer, M., & König, C. J. (2023). Introducing a multi-stakeholder perspective on opacity, transparency and strategies to reduce opacity in algorithm-based human resource management. *Human Resource Management Review*, *33*(1), 100881. https://doi.org/10.1016/j.hrmr.2021.100881

Leavitt, K., Schabram, K., Hariharan, P., & Barnes, C. M. (2021). Ghost in the machine: On organizational theory in the age of machine learning. *Academy of Management Review*, *46*(4), 750–777. https://doi.org/10.5465/amr.2019.0247

Lin, L., Bai, Y., Mo, C., Liu, D., & Li, X. (2021). Does pay raise decrease temporary agency workers' voluntary turnover over time in China? Understanding the moderating role of demographics. *The International Journal of Human Resource Management*, *32*(7), 1537–1565. https://doi.org/10.1080/09585192.2018.1539861

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing System*. Available online at http://arxiv.org/pdf/1705.07874v2.

Meijerink, J., Boons, M., Keegan, A., & Marler, J. (2021). Algorithmic human resource management: Synthesizing developments and cross-disciplinary insights on digital HRM. *The International Journal of Human Resource Management*, *32*(12), 2545–2562. https://doi.org/10.1080/09585192.2021.1925326

Molnar, C. (2022). *Interpretable machine learning. A guide for making Black Box Models interpretable*. (2nd ed.). Lulu.

Molnar, C., Casalicchio, G., & Bischl, B. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software*, *3*(26), 786. https://doi.org/10.21105/joss.00786

Putka, D. J., Beatty, A. S., & Reeder, M. C. (2018). Modern prediction methods: New perspectives on a common problem. *Organizational Research Methods*, *21*(3), 689–732. https://doi.org/10.1177/1094428117697041

Rombaut, E., & Guerry, M.-A. (2018). Predicting voluntary turnover through human resources database analysis. *Management Research Review*, *41*(1), 96–112. https://doi.org/10.1108/MRR-04-2017-0098

Rombaut, E., & Guerry, M.-A. (2021). Determinants of voluntary turnover: A data-driven analysis for blue and white collar workers. *Work (Reading, Mass.)*, *69*(3), 1083–1101. https://doi.org/10.3233/WOR-213538

Rubenstein, A. L., Eberly, M. B., Lee, T. W., & Mitchell, T. R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. *Personnel Psychology*, *71*(1), 23–65. https://doi.org/10.1111/peps.12226

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Russell, C. J., & Sell, M. V. (2012). A closer look at decisions to quit. *Organizational Behavior and Human Decision Processes*, *117*(1), 125–137. https://doi.org/10.1016/j.obhdp.2011.09.002

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Somers, M. J., Birnbaum, D., & Casal, J. (2021). Supervisor support, control over work methods and employee well-being: New insights into nonlinearity from artificial neural networks. *International Journal of Human Resource Management*, *32*(7), 1620–1642. https://doi.org/10.1080/09585192.2018.1540442

Speer, A. B. (2021). Empirical attrition modelling and discrimination: Balancing validity and group differences. *Human Resource Management Journal*, *34*(1), 1–19. https://doi.org/10.1111/1748-8583.12355

Vale, D., El-Sharif, A., & Ali, M. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI and Ethics*, *2*(4), 815–826. https://doi.org/10.1007/s43681-022-00142-y

Valizade, D., Schulz, F., & Nicoara, C. (2024). Towards a paradigm shift: How can machine learning extend the boundaries of quantitative management scholarship? *British Journal of Management*, *35*(1), 99–114. https://doi.org/10.1111/1467-8551.12678

van den Broek, E., Sergeeva, A., & Huysman Vrije, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly*, *45*(3), 1557–1580. https://doi.org/10.25300/MISQ/2021/16559

Yakusheva, O., Bang, J. T., Hughes, R. G., Bobay, K. L., Costa, L., & Weiss, M. E. (2022). Nonlinear association of nurse staffing and readmissions uncovered in machine learning analysis. *Health Services Research*, *57*(2), 311–321. https://doi.org/10.1111/1475-6773.13695

Yin, R. K. (2013). Validity and generalization in future case study evaluations. *Evaluation*, *19*(3), 321–332. https://doi.org/10.1177/1356389013497081

Yuan, S., Kroon, B., & Kramer, A. (2021). Building prediction models with grouped data: A case study on the prediction of turnover intention. *Human Resource Management Journal*, *34*(1), 20–38. https://doi.org/10.1111/1748-8583.12396