10-9-2023

How Explainable AI Methods Support Data-Driven Decision Making

Dominik Stoffels
Universität Passau, Germany, stoffe11@ads.uni-passau.de

Susanne Grabl
Universität Passau, Germany, susanne.grabl@uni-passau.de

Thomas Fischer
Universität Passau, Germany, thomas.fischer@fh-steyr.at

Marina Fiedler
Universität Passau, Germany, marina.fiedler@uni-passau.de

How Explainable AI Methods Support Data-Driven Decision-Making

Research Paper

Dominik Stoffels[1], Susanne Grabl[1], Thomas Fischer[1,2], and Marina Fiedler[1]

[1] University of Passau, Passau, Germany
{stoffe11}@ads.uni-passau.de
{susanne.grabl, marina.fiedler}@uni-passau.de
[2] University of Applied Sciences Upper Austria, Steyr, Austria
thomas.fischer@fh-steyr.at

Abstract. Explainable AI (XAI) holds great potential to reveal the patterns in black-box AI models and to support data-driven decision-making. We apply four post-hoc explanatory methods to demonstrate the explanatory capabilities of these methods for data-driven decision-making using the illustrative example of unwanted job turnover and human resource management (HRM) support. We show that XAI can be a useful aid in data-driven decision-making, but also highlight potential drawbacks and limitations of which users in research and practice should be aware.

Keywords: Explainable AI, Machine Learning, Data-Driven Decision-Making.

1 Introduction

Data-driven decision-making is becoming increasingly important in information systems (IS) research and in practice. Artificial intelligence (AI) uses data and predictive models to make decisions (Fernández-Loría et al., 2022) that will, for example, increase efficiency in organizations. Machine learning (ML) is a subdivision of AI which uses algorithms to discover patterns in complex datasets to make predictions based on the learned patterns (Hong et al., 2007; Jordan & Mitchell, 2015; Brynjolfsson & Mitchell, 2017; Sturm et al., 2021; Yang et al., 2022). For example, ML is used to predict cancer in patients based on x-ray scans (Kourou et al., 2015), to assess user engagement on social media (Shin et al., 2020), or to predict employee performance or turnover for human resource management (HRM) in an organizational context (Yuan, et al., 2021).

One category of ML algorithms on which we want to focus is supervised ML algorithms (as opposed to unsupervised ML and reinforcement learning, Benbya et al., 2021). In supervised ML algorithms the available data consists of labeled examples (i.e., each data point contains features and an associated class – a label). The goal of supervised learning algorithms is for them to learn a function that maps feature vectors (inputs) to labels (outputs) as accurately as possible (Hastie et al., 2009). Using various

18th International Conference on Wirtschaftsinformatik,
September 2023, Paderborn, Germany

features, ML models can learn and generalize patterns in the data to predict the label without knowing other correlations in this context. With knowledge of the labels, ML models can monitor their learning and reduce prediction error. In doing so, an AI system can use the ML model to make decisions based on the model's prediction. The trade-off for the high predictive accuracy of the models is limited interpretability. Therefore, a more complex model can lead to higher predictive accuracy, but also reduce interpretability (Shrestha et al., 2021). This particularly applies to black-box ML models. From a theoretical and practical perspective, these models are superior in minimizing prediction error (Goldstein et al., 2015), but the more complex the model, the less interpretable it is. For example, simple ML models such as decision trees provide a high degree of transparency; human users can understand every decision because they can directly visualize and understand decision trees. A random forest, however, which can basically consist of up to ten thousand decision trees, cannot be visualized or represented in such a way that human users can interpret it (Stoffels et al., 2022).

Since AI systems are integrated into all types of decision-making processes, an ever more urgent debate is taking place in academia about the extent to which people who develop AI or are affected by an AI-enabled decision can understand how the resulting decision mechanism works and why it arrives at a particular decision (Yang et al., 2022). Consequently, we need more interpretable approaches (Mitchell et al., 2001; Avrahami et al., 2022), referred to as explainable AI (XAI) methods (Guidotti et al., 2018; Barredo Arrieta et al., 2020; Hamm et al., 2021). Importantly, these methods not only aim to explain one particular decision (Fernández-Loría et al., 2022); they also attempt to derive actions.

XAI has become increasingly important in IS research, having already been used methodologically (Choudhury et al., 2021; Fernández-Loría et al., 2022) and empirically (Senoner et al., 2022). However, current IS research does not address the various capabilities of different XAI methods for data-driven decision-making. To address this gap, we trained seven ML models and, building on these models, we compare four different XAI methods, local model-agnostic methods, global model-agnostic methods, and counterfactual explanations. In doing so, we examine the following research questions:

Which methods are applicable to which type of decision-making? Are the explanations between the methods explicit, or can different explanations emerge?

To demonstrate the different capabilities of XAI for data-driven decision-making in an understandable way, we illustratively use organizational data to predict and explain unwanted job turnover and analyze how XAI can benefit data-driven decision-making in human resource management (HRM).

2 Research Background

Data-driven Decision-making. One of the most popular application areas for AI in organizations is in decision-making where it offers many potential benefits on a strategic level (e.g., product quality improvement, improved understanding of customer

needs, or more accurate organizational resources planning, Borges et al., 2021). Through their ability to learn from additional datapoints over time, AI systems can create knowledge that assists humans in their decision-making or they could even be applied to automate tasks and act autonomously (Shrestha et al., 2021; van den Broek et al., 2021; Shollo et al., 2022). These capabilities have been advocated as an important addition to IS theorizing (Berente et al., 2019; Miranda et al., 2022). Further, AI proponents have argued that how AI is used in practice to create working theories could be an important future avenue for creating grand theories in IS (Tremblay et al., 2021).

A particular obstacle to human interaction with AI, in practice as well as in theorizing, is the opacity of certain ML algorithms (Müller et al., 2016; Faraj et al., 2018; Berente et al., 2021). It can be difficult for professionals to make use of the knowledge created by AI systems if they cannot follow the logic that was applied to arrive at a certain result (e.g., Lebovitz et al., 2022). Various approaches, such as envelopment (Asatiani et al., 2021), have been presented to circumvent this issue. XAI could be a more direct approach to deal with the opacity of algorithmic decision-making (Gunning et al., 2019).

A promising application for XAI and data-driven decision-making could be in HRM and employee retention, given the scarcity of high-skilled employees in these contexts (Oswald et al., 2020). Unwanted employee turnover affects organizations on multiple levels, as it is expensive, ties up resources, and causes low business performance (Chowdhury et al., 2022). Hence, practitioners and scholars are trying to predict and understand unwanted employee turnover as accurately as possible (Farrell & Rusbult, 1981; Zhao et al., 2018; Oswald et al., 2020; Choudhury et al., 2021; Wang & Zhi, 2021; Yuan et al., 2021; Avrahami et al., 2022). ML and XAI have great potential to predict and reveal the reasons for unwanted job turnover in organizations and promise powerful support for data-driven decision-making in HRM to keep valued employees in the organization (Oswald et al., 2020; Chowdhury et al., 2022).

Explainable AI (XAI) and black box models. Previous research has shown that XAI can promote trust among users of black-box AI systems (Barredo Arrieta et al., 2020; Hamm et al., 2021). Black-box AI refers to the comprehensibility of AI models and systems. As the underlying models become more complex, AI systems seem to be a black-box (Meske et al., 2022). Thus, with growing complexity, a trade-off between explainability and model performance emerges, which significantly influences individuals and organizations (Alt, 2018). Therefore, to promote the adoption of AI systems, Hamm et al. (2021) assume that XAI methods should be a part of AI implementations. In addition, they emphasize the importance of XAI for organizations and developers to meet regulatory requirements that improve the AI system.

Overview of XAI Methods. There are multiple XAI methods which can provide powerful support in explaining AI systems' decisions, and which enable us to grasp the underlying patterns (Molnar, 2020). So-called *model-agnostic* XAI methods have the advantage that they can be used independently on the ML model and, therefore, count as the counterpart of *model-specific* XAI methods. Consequently, these methods have attracted a great deal of interest in the IS literature (Choudhury et al., 2021; Chowdhury et al., 2022; Fernández-Loría et al., 2022; Senoner et al., 2022). XAI methods can also be categorized based on the range of cases they aim to explain, into *local* (i.e., single

case) and *global* (i.e., all cases in a dataset) methods (Barredo Arrieta et al., 2020). Local methods can provide explanations for single predictions (i.e., for a specific employee) (Molnar, 2020). In cases of employee retention, these methods show the characteristics, such as salary, which have the greatest influence on the model decision (i.e., they indicate the main characteristics that determine employees' resignation). Yet, this only gives the reasons for the turnover of individual employees, which can vary greatly. In comparison, global methods show the influence of the characteristics in general (i.e., which affect many or all employees) (Molnar, 2020; Stoffels et al., 2022). Some XAI methods also show how a single feature influences the model outcome, so that researchers achieve fine-grained insight into the interactive pattern between the specific feature and the model outcome (Goldstein et al., 2015; Choudhury et al., 2021). While these XAI methods only show a trend of the pattern between the features and the actual prediction, counterfactual explanations can provide a quantitative explanation (Fernández-Loría et al., 2022).

Counterfactual explanations provide a practical extension to local and global methods, as they offer a means for deciding on a course of action that will change an outcome variable's value. While local and global methods commonly provide us with an overview of the importance of given features in a prediction, counterfactuals offer us a combination of features and the extent of change that needs to be implemented for the outcome to change. This can be important in practice, as high feature importance for a prediction does not necessarily affect the model's decision (Fernández-Loría et al., 2022).

Implementations of XAI methods. Current IS research mainly uses post-hoc XAI methods, such as shapley additive explanation values (SHAP) (Fernández-Loría et al., 2022; Senoner et al., 2022), local interpretable model-agnostic explanations (LIME) (Chowdhury et al., 2022), partial dependence plots (PDP) (Mehdiyev & Fettke, 2020; Choudhury et al., 2021), or counterfactual explanations (Fernández-Loría et al., 2022) implemented using e.g., diverse counterfactual explanations (DiCE) (Mothilal et al., 2020).

SHAP is technically a local model-agnostic method based on a game-theoretic approach (Lundberg & Lee, 2017), but it can also provide global explanations (Molnar, 2020). Senoner et al. (2022) use SHAP to reveal learned patterns between complex manufacturing process data and achieved quality. In contrast, Fernández-Loría et al. (2022) are very critical of this use of SHAP. They show that SHAP has limited informative value and also that features with high feature importance according to SHAP can actually have no influence on the models' decisions. Thereby, they emphasize the use of counterfactual explanations. Counterfactual explanations offer the advantage of not showing features‘ relative influence on the prediction; instead, they show how much a feature value must vary to change the model's actual prediction (Mothilal et al., 2020).

Another XAI method is partial dependence plots (PDP), which can show the impact a feature has on the model output (Mehdiyev & Fettke, 2020; Kamath & Liu, 2021). For example, the PDP shows how a single feature (e.g., employee salary) is related to the impact of termination. Thus, PDP is one of the global model-agnostic methods (Molnar, 2020). Choudhury et al. (2021) emphasize the value of PDP in gaining deep insight into models. They use it in a methodological context to explain the difference

in predictive accuracy between a simpler model (decision tree) and a more complex model (random forest). In this context, PDP show that the complex model learned more fine-grained patterns than the simpler model.

In contrast, LIME (Ribeiro et al., 2016) can show the influence of each feature on a single prediction, for example a single employee. Chowdhury et al. (2022) use this local model-agnostic method to show the potential of XAI in the context of turnover. They imply the use of XAI to gain HRM users' trust in AI systems and to improve data-driven decision-making. In contrast, John-Mathews (2021) views these types of post-hoc explanations more critically. Using LIME illustratively, he concludes that post-hoc explanations lead to misleading or partial information about the learned pattern. In addition, Stoffels et al. (2022) reveal empirically, that different patterns can arise using models of the same quality.

DiCe are an implementation that give counterfactual explanations, which answer the "what if" questions. Thus, DiCe can provide examples that are impactful enough to change the model's prediction (Mothilal et al., 2020). Besides DiCe there is a range of other implementations that can generate counterfactual explanations (Guidotti, 2022). Mothilal et al. (2021) show that the explanations of DiCe do not match those of SHAP and LIME. Further, Fernández-Loría et al. (2022) found that counterfactual explanations provide more appropriate explanations than the other methods, although the potential disadvantage of the Rashomon effect is stated.

While the use of ML models to leverage XAI is already increasingly common in IS research, and thus in the data-driven decision-making literature (Bertsimas & Kallus, 2020; de Bruijn et al., 2022; Elgendy et al., 2022), the learned models remain underinvestigated, thus requiring a deeper investigation. Also, there are no studies that comprehensively consider and critically reflect on the use of different XAI methods and their implications for data-driven decision-making in the context of unwanted job turnover and HRM support.

3 Data and Methods

We base our illustration of XAI in HRM on the publicly available dataset "IBM HR Analytics Employee Attrition & Performance" (Subhash, 2017), a synthetic dataset created by IBM that contains 35 metric and nominal features with administrative data, performance data, job satisfaction data, and data on individual characteristics (e.g., age and gender) of 1470 fictitious employees (Subhash, 2017). Our goal in illustratively using different XAI methods has been to predict turnover (0 = no turnover, 1 = turnover) based on these characteristics and to make the prediction interpretable to define measures for HRM. In the chosen dataset, 16% of the employees had resigned. The width of the features' skewness ranged from -0.55 ("*YearsSinceLastPromotion*") to 1.98 ("*WorkLifeBalance*"). We translated the ordinal variable "education" into years of education ("*YearsEducation*"). As this was a synthetic data set, there were no data quality issues or missing values.

Since it was not clear a priori which ML model would provide the best performance, a sample of models had to be selected (Choudhury et al., 2021). We applied seven different ML models: linear regression (LR) (Hosmer et al., 2013), k-nearest neighbors (KNN) (Cover & Hart, 1967), random forest (Breiman, 2001a), c-support vector classification (SVM) (Cortes & Vapnik, 1995), decision tree (DT) (Breiman et al., 1984), gradient boosting classifier (GB) (Friedman, 2002), and adaboost classifier (AdaBoost) (Freund & Schapire, 1997). Notably, we also included ML models that are not considered black-boxes (e.g., LR, DT and KNN), to showcase their performance in comparison to black-box models. For all models we used the default hyperparameters, which are set in the "scikit-learn" library (scikit-learn, 2023).

We used min-max scaling for numeric features to ensure the same range of values (Nayak et al., 2014). In addition, we used one-hot coding for nominal features (Hancock & Khoshgoftaar, 2020). To evaluate performance, we used a training-test split of 80/20, thus we used 80% as training data and 20% as test data (Hastie et al., 2009; Berrar, 2018) and the F1 score. The F1 score is the harmonic mean of precision (how many of the positive predictions are actually positive) and recall (the ratio of all positive cases that the model was able to identify correctly) (Choudhury et al., 2021). The F1 score's advantage over accuracy is that unbalanced classes are considered. Usually more employees stay in the organization than leave it, so the dataset is unbalanced. Accuracy only takes into account how many predictions are correct. Thus, if we assume that 99% of employees do not quit and only 1% of employees resign, assigning all employees to stay in the organization would already yield a prediction accuracy of 99%, even though the model learned nothing (Choudhury et al., 2021). Therefore, we chose the ML model with the highest F1 score for the further analysis with XAI. If other models had a non-significantly worse F1 score, we included these models due to their comparable predictive ability.

We applied LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), PDP (Kamath & Liu, 2021), and DiCE (Mothilal et al. 2020) to analyze the learned pattern of the best performing ML model. For SHAP we used the Kernel Explainer, because it is usable for all ML models. The more accurate Tree Explainer is only usable for tree models (Lundberg et al., 2020). Additionally, for the global explanations, we created a subsample containing only instances with employees who had resigned. Further, we checked whether the results had been influenced from seen to unseen instances. If multiple ML models achieved a comparable F1 score, we analyzed all models using these XAI methods to determine whether the different models had learned the same patterns (Breiman, 2001b; Stoffels et al., 2022).

4 Results

As usual in ML approaches, we first attempted to find an accurate predictive model for our dataset. We analyzed the F1 score the selected ML models reached, as summarized in Table 1. AdaBoost performed slightly better than GB and RF in reaching an F1 score of 85.2%. In contrast, the other models (i.e., LR, KNN, SVM, DT) performed significantly worse. SVM achieved an F1 score of only 59.8%. Subsequently, we chose the

models with the highest F1 scores to analyze the models more deeply using the selected XAI methods. Since the GB (84.8%) and RF (83.4%) scores were comparable to AdaBoost's, we also checked these models using the XAI methods.

Table 1. Average F1 scores of the trained ML models

Model	LR	KNN	RF	SVM	DT	GB	AdaBoost
F1	74.4%	66.7%	83.4%	59.8%	75.4%	84.8%	85.2%

In the following two sections, we highlight contributions that can be derived for data-driven decision-making from the use of XAI - based on the example of HRM.

4.1 Individualized Decision-Making (local methods)

Local model-agnostic methods can explain a single prediction, in this context the model's prediction of an individual employee's turnover. As Chowdhury et al. (2022) mentioned, this method can be used in different ways, including consideration of trust and understanding of the model, diagnosing the model and its performance, and determining why certain employees are likely to leave the organization. In particular, these methods can be used to define countermeasures for the individual employee. We chose three employees and analyzed their possible reasons for leaving the organization using SHAP.

Figure 1. SHAP force plots of single instances (i.e., employees)

Figure 1 shows the SHAP force plots of three employees. The features colored red indicate that they contribute to the model's prediction of turnover, while the blue features contribute to the model's prediction of no turnover. If the model predicts turnover (model output = 1), the red colored features have a greater impact on the model prediction and lead to the final model decision for turnover. Therefore, HRM can use this information to determine which features and with which relative feature values lead to turnover.

While for all three employees the lack of stock options (*StockOptionLevel*) was a reason to leave the organization, the other reasons for each differed considerably. For example, for the second and third employee, overtime (*OverTime_1*) was the main reason, while for the first employee it was the short time in the current role (*YearsInCurrentRole* = 0).

4.2 Global Decision-Making (global methods)

In contrast to local model-agnostic methods, global model-agnostic methods can indicate multiple predictions or, alternatively, the particular influence of single features on the models' predictions (Molnar, 2020). In Figure 2, the SHAP summary plot ranks the influence of the features of all employees, as well as of all employees who resigned, for the Adaboost model.

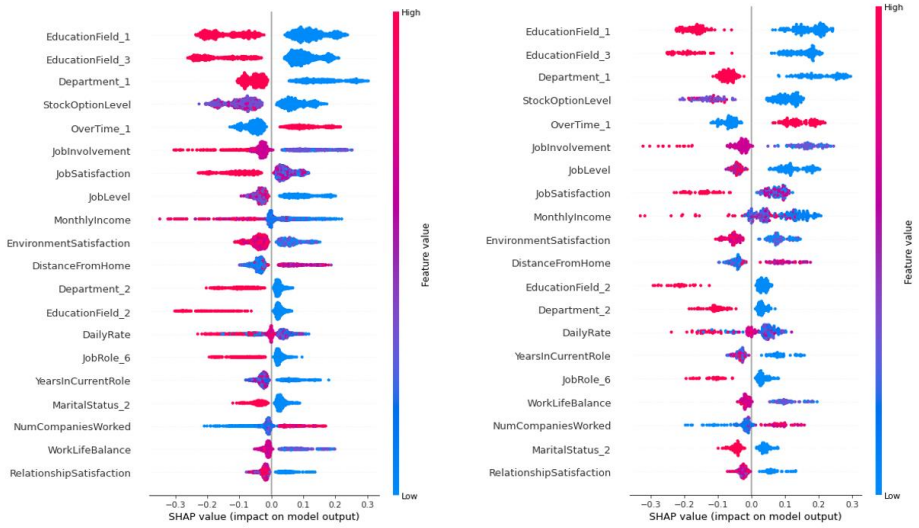

Figure 2. SHAP summary plot of all employees (left) and only employees who had resigned (right)

The features that had the greatest influence on the model decision are shown in descending order. We see that in some cases the important features are different compared to those Figure 1 shows for single instances. Still, we retain a high level of information, as we are provided with not only a ranking of features or an overall effect score, but also with an overview of the effects for each included instance, summarized in a distribution plot (Lundberg et al., 2020). The main influencing factor is found to be the employee's education (*EducationField_1*, *EducationField_3*), while the already discussed features of *OverTime_1* and *StockOptionLevel* are also amongst the most important features. Besides, SHAP shows that the most important features (up to the 6th) have not changed in the subgroup of resigned employees. Colors are used to indicate the current value of the feature. Thus, high feature values are shown with red, while low

feature values are shown with blue. For example, a high monthly income (*MonthlyIncome*) in comparison to other employees is red, while a low income is marked in blue. This can be used by HRM to identify the main reasons for unwanted turnover in the organization and could serve as a basis for defining data-driven global HRM strategies. Thus, in our example a rise in the stock options and measures to reduce overtime should be in focus, as these are identified among the most important characteristics.

SHAP, as well as other global model-agnostic methods such as PDP, can additionally visualize a single feature's influence on model output across (all) employees (Molnar, 2020; Choudhury et al., 2021). This is illustrated in Figure 3. As expected, the probability of leaving the organization decreases as work-life balance (*WorkLifeBalance*) increases. The curve could help HRM to determine an optimal value for work-life balance. As the plot shows, the impact on the model result decreases significantly if the *WorkLifeBalance* level, which is marked in red, is between 2 to 3. Thus, further increasing the *WorkLifeBalance* has little impact on the model decision, and probably a similar impact on the employee. With this help, HRM can identify on a more fine-grained basis, the extent to which certain features (i.e., employee circumstances) should be changed to reduce turnover in the organization. We see this as an important complement to local explanations and simple feature importance rankings (Choudhury et al., 2021), since the progression can highlight broader relationships.

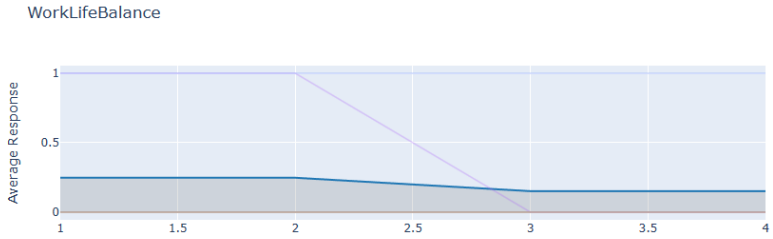

Figure 3. Influence of *WorkLifeBalance* on the model output (i.e., across all employees)

4.3 Counterfactual Explanations

SHAP and LIME both indicate only the features that have the greatest impact on the model outcome, but they do not explain how to change the features so that they will actually change the model's or employee's decision. For example, these methods do not show how much more the salary should increase to change the employee's decision about leaving the organization (Fernández-Loría et al., 2022). Further, PDP only show the average influence of a single feature on the model outcome, but the method does not provide an answer to whether an increase in a detected threshold is sufficient to change the prediction as well.

Therefore, the use of counterfactual explanations is suggested (Fernández-Loría et al., 2022), which we adopt in using DiCE (Mothilal et al., 2020). We use the functionality of the DiCE implementation to consider only characteristics that HRM can change directly or indirectly, such as raising the salary (*MonthlyIncome*) or addressing the distance from home issue (*DistanceFromHome*) (e.g., by providing a second home).

In the first row, Figure 4 shows each characteristic's values for the selected employee who left the organization. We then determined five different counterfactual explanations that actually changed the model decision to remain in the organization. We added the change in the feature value for each explanation in the next five rows. DiCE gives HRM five different options for changing the model's decision, and probably also that of the employee in question. For example, a change in JobLevel (e.g., level 1 to 2) would not change the decision in this the model (i.e., of the employee) from turnover to no turnover. With the exception of the fourth explanation, a significant increase in salary (between 848 to 2317) would be required to change the decision.

We see several advantages of using counterfactual explanations: First, HRM can decide which characteristics they can easily target to avoid unwanted turnover. Second, counterfactual explanations provide different courses of action as there is (in most cases) not just one way of changing the prediction. Finally, the model gives a precise value of the characteristics required to change the prediction, thus directing HRM toward a very specific action. Compared to the other presented XAI methods, no further interpretation is required to define actions, because SHAP and LIME give only the relative importance of each characteristic. Hence, a further HRM analysis is necessary to define a specific action such as offering a concrete salary increase. If, however, the focus is not on the formulation of measures but on the analysis of the reasons for turnover, counterfactual explanations require more investigation, as they do not directly provide a visualization of the most important influencing characteristics.

	Features							
	Monthly Income	Job Satisfication	StockOption Level	WorkLife Balance	OverTime	DistanceFrom Home	Job Involvement	JobLevel
Turnover	1261	2	0	4	0	8	2	1
DiCE - No Turnover (New feature value)	2109	2	0	3	0	8	4	1
	3578	1	0	4	0	8	4	1
	2572	2	1	3	0	8	3	1
	1009	4	0	4	0	1	4	1
	2996	1	0	4	0	1	3	1
DiCE - No Turnover (Feature value difference)	848	0	0	-1	0	0	2	0
	2317	-1	0	0	0	0	2	0
	1311	0	1	-1	0	0	1	0
	-252	2	0	0	0	-7	2	0
	1735	-1	0	0	0	-7	1	0

Figure 4. Counterfactual explanations for a single instance (i.e., an employee)

5 Discussion

We have demonstrated the use of different XAI methods to analyze organizational data and to develop explanations of turnover and related actions for HRM. Thereby, we show that data-driven actions require different XAI methods for different objectives. Global model-agnostic methods provide a global view of the model decision and the overall main influencing factors of the model. This insight can provide important data-driven support for general strategies in an organization. Local model-agnostic methods are useful for creating actions suited to individual instances, such as single employees. Nevertheless, we must address the limitations of these methods, thereby also offering potential for future research.

Different XAI Methods, Different Explanations. Global and local model-agnostic methods do not explain the model's prediction in the same way, or put differently, they provide significantly different explanations. Therefore, the patterns disclosed by the different methods are not robust and lead to different explanations and interpretations.

This leads to a major problem with the post-hoc explanations presented. While the ML model's performance can be evaluated using, for example, the F1 score for decisions on whether the model is sufficient to use, there is no actual test or metric for evaluating XAI methods. Therefore, we cannot accurately measure the error or evaluate which method would provide the correct explanations. If different methods provide different explanations and ultimately suggest different actions, this, of course, poses a significant problem for HRM or data-driven decision-making.

We have illustrated this issue by comparing the LIME and SHAP local model-agnostic methods. We find that the local explanations of LIME and SHAP are significantly different for the same employees. Table 2 illustrates two employees and the local explanations of LIME and SHAP. The two cases show significant differences in the effects the features have on the model decision. While we acknowledge that nominal features and one-hot coding can lead to confounding outcomes in XAI methods (as with *EducationField* features), other features, especially numeric features, should be closely ranked in their impact on model decision.

Table 2. LIME and SHAP for local explanations of two instances (i.e., employees)

Employee	Top three features (LIME vs. SHAP)	
First illustrative employee	EducationField_3	EducationField_1
	EducationField_1	YearsInCurrentRole
	EducationField_4	StockOptionLevel
Second illustrative employee	EducationField_3	OverTime_1
	EducationField_1	EducationField_1
	StockOptionLevel	StockOptionLevel

Therefore, users of XAI methods in practice as well as in research should be aware that these methods provide only approximate explanations and can lead to different explanations and interpretations depending on the chosen method. We suggest using at least two different methods to assess the extent to which we can trust the results. Again, counterfactual explanations have the advantage that these explanations at least change the model's actual decision (Fernández-Loría et al., 2022), which the other methods cannot achieve.

Different ML Models, Different Explanations. Not only can different XAI methods lead to different explanations, but different models of comparable quality analyzed with the same XAI method can also yield different explanations. We illustrate this issue in Figure 5, which shows the top five features in the SHAP summary plots of AdaBoost, GB, as well as RF. While the features shown here are partly similar and show small shifts (e.g, *StockOptionLevel*), there are also cases were completely different features are among the most important ones, such as *MaritalStatus_1* in the RF model.

Therefore, we should be aware of the possible effects of the *Rashomon Effect* (Breiman, 2001b) and its implications for data-driven decision-making. We recommend the following procedure as suggested in Stoffels et al. (2022): If several models have the same quality (e.g., F1 score), then all models should be examined with XAI methods. Thereafter, there are two possibilities: either there is no Rashomon Effect (i.e., the models have learned the same patterns), or the learned patterns significantly disagree. The first case requires no further action, because the models have learned robust patterns and increased our confidence in those patterns. However, in the second case, we should critically reflect on these models and the database. Thus, further analysis of the models to determine the possible cause of the discrepancy is required, as this would ensure meaningful decisions in the context of data-driven decision-making.

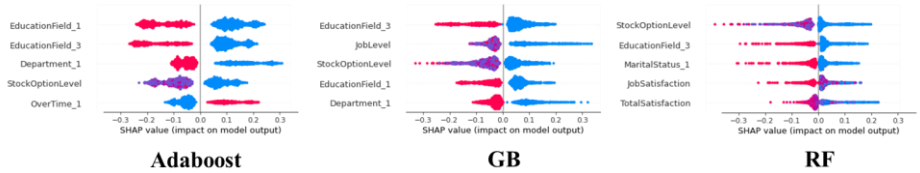

Figure 5. Comparison of SHAP summary plots (all employees)

Implications and Future Work. Given the two main shortcomings of local and global methods we have demonstrated, we see the high potential of counterfactual explanations for data-driven decision-making in practice. They are not as inconsistent as the other methods because they ensure that the explanations actually change the models' decisions. Thus, they provide not only a relative indication of which characteristics are important, but also a necessary quantitative change in those characteristics to receive a specific action toward data-driven decision-making. We have to acknowledge, however, that using counterfactual explanations requires a high degree of domain knowledge in order for them to be useful because they provide several alternatives that need to be carefully weighed against one another (Mithas et al., 2022). In addition, counterfactuals are subject to some of the drawbacks of ML and XAI, such as the Rashomon effect; therefore, those individuals choosing which specific decision alternatives to consider, need transparency and clear criteria (Artelt & Hammer, 2019).

Besides the potential limitations of XAI methods we compared in this study, we also need to mention some research limitations, since they offer potential for further investigation. First, we used synthetic data. Yet, to reveal further strengths and weaknesses of XAI methods it could be worthwhile to repeat the comparison portrayed in this study using multiple sets of real data. Second, we used only a selected set of ML algorithms and XAI method implementations, which could feasibly be extended in future research using additional methods (see e.g., Barredo Arrieta et al., 2020). The algorithms and methods could also be supplemented by further methods that are becoming increasingly popular, such as supervised clustering used to identify characteristic groups of employees (Cooper et al., 2021).

References

Alt, R. (2018), 'Electronic Markets and Current General Research', *Electronic Markets* **28**(2), 123–128.

Artelt, A. & Hammer, B. (2019), 'On the Computation of Counterfactual Explanations - A Survey', *arXiv preprin*t, http://arxiv.org/abs/1911.07749.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T. & Salovaara, A. (2021), 'Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems', *Journal of the Association for Information Systems* **22**(2), 325–352.

Avrahami, D., Pessach, D., Singer, G. & Ben-Gal, H. C. (2022), 'A Human Resources Analytics and Machine-Learning Examination of Turnover: Implications for Theory and Practice', *International Journal of Manpower* **43**(6), 1405–1424.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J. et al. (2020), 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI', *Information Fusion* **58**, 82–115.

Benbya, H., Pachidi, S. & Jarvenpaa, S. L. (2021), 'Special Issue Editorial: Artificial intelligence in Organizations: Implications for Information Systems Research', *Journal of the Association for Information Systems* **22**(2), 281–303.

Berente, N., Gu, B., Recker, J. & Santhanam, R. (2021), 'Managing Artificial Intelligence', *MIS Quarterly* **45**(3), 1433–1450.

Berente, N., Seidel, S. & Safadi, H. (2019), 'Research Commentary - Data-Driven Computationally Intensive Theory Development', *Information Systems Research* **30**(1), 50–64.

Berrar, D. (2018), 'Cross-Validation', *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* **1**(3), 542–545.

Bertsimas, D. & Kallus, N. (2020), 'From Predictive to Prescriptive Analytics', *Management Science* **66**(3), 1025–1044.

Borges, A. F. S., Laurindo, F. J. B., Spínola, M. M., Gonçalves, R. F. & Mattos, C. A. (2021), 'The Strategic Use of Artificial Intelligence in the Digital Era: Systematic Literature Review and Future Research Directions', *International Journal of Information Management* **57**, 102225.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984), *Classification and regression trees*. Taylor & Francis.

Breiman, L. (2001a), 'Random Forests', *Machine Learning* **45**, 5–32.

Breiman, L. (2001b), 'Statistical Modeling: The two Cultures', *Statistical Science* **16**(3), 199–215.

van den Broek, E., Sergeeva, A. & Huysman, M. (2021), 'When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring', *MIS Quarterly* **45**(3), 1557–1580.

de Bruijn, H., Warnier, M. & Janssen, M. (2022), 'The Perils and Pitfalls of Explainable AI: Strategies for Explaining Algorithmic Decision-Making', *Government Information Quarterly* **39**(2), 101666.

Brynjolfsson, E. & Mitchell, T. (2017), 'What can Machine Learning do? Workforce Implications', *Science* **358**(6370), 1530–1534.

Choudhury, P., Allen, R. T. & Endres, M. G. (2021), 'Machine Learning for Pattern Discovery in Management Research', *Strategic Management Journal* **42**(1), 30–57.

Chowdhury, S., Joel-Edgar, S., Kumar Dey, P., 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 Preprint*.

Cooper, A., Doyle, O. & Bourke, A. (2021), "Supervised Clustering for Subgroup Discovery: An Application to COVID-19 Symptomatology", *in* 'Machine Learning and Principles and Practice of Knowledge Discovery in Databases', Springer, pp. 408–422.

Cortes, C. & Vapnik, V. (1995), 'Support-Vector Networks', *Machine Learning* **20**(3), 273–297.

Cover, T. & Hart, P. (1967), 'Nearest Neighbor Pattern Classification', *IEEE transactions on information theory* **13**(1).

Elgendy, N., Elragal, A. & Päivärinta, T. (2022), 'DECAS: A Modern Data-Driven Decision Theory for Big Data and Analytics', *Journal of Decision Systems* **31**(4), 337–373.

Faraj, S., Pachidi, S. & Sayegh, K. (2018), 'Working and Organizing in the Age of the Learning Algorithm', *Information and Organization* **28**(1), 62–70.

Farrell, D. & Rusbult, C. E. (1981) ,'Exchange Variables as Predictors of Job Satisfaction, Job Commitment, and Turnover: The Impact of Rewards, Costs, Alternatives, and Investments', *Organizational Behavior and Human Performance* **28**(1), 78–95.

Fernández-Loría, C., Provost, F. & Han, X. (2022), 'Explaining Data-Driven Decisions made by AI Systems: The Counterfactual Approach', *MIS Quarterly* **46**(3), 1635–1660.

Freund, Y. & Schapire, R. E. (1997), 'A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting', *Journal of Computer and System Sciences* **55**(1), 119–139.

Friedman, J. H. (2002), 'Stochastic Gradient Boosting', *Computational Statistics and Data Analysis* **38**(4), 367–378.

Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. (2015), 'Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation', *Journal of Computational and Graphical Statistics* **24**(1), 44–65.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F. et al. (2018), 'A Survey of Methods for Explaining Black Box Models', *ACM Computing Surveys* **51**(5).

Guidotti, R. (2022), 'Counterfactual Explanations and How to Find Them: Literature Review and Benchmarking', *Data Mining and Knowledge Discovery.*

Gunning, D., Stefik, M., Choi, J., Miller, T. et al. (2019), 'XAI-Explainable Artificial Intelligence', *Science Robotics* **4**(37), 4–6.

Hamm, P., Wittmann, H. & Klesel, M. (2021), Explain it to Me and I will Use it: A proposal on the Impact of Explainable AI on Use Behavior, *in* 'ICIS 2021 Proceedings'.

Hancock, J. T. & Khoshgoftaar, T. M. (2020), 'Survey on categorical data for neural networks', *Journal of Big Data* **7**(1), 1–41.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The Elements of Statistical Learning*. Springer Science+Business Media.

Hong, W.-C., Wei, S.-Y. & Chen, Y.-F. (2007), 'A Comparative Test of Two Employee Turnover Prediction Models', *International Journal of Management* **24**(2), 216–229.

Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. (2013), *Applied Logistic Regression*. 3rd edn. Wiley.

John-Mathews, J.-M. (2021), Critical Empirical Study on Black-box Explanations in AI, *in* 'ICIS 2021 Proceedings'.

Jordan, M. I. & Mitchell, T. M. (2015), 'Machine Learning: Trends, Perspectives, and Prospects', *Science* **349**(6245), 255–260.

Kamath, U. & Liu, J. (2021) *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*. Springer.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. (2015), 'Machine Learning Applications in Cancer Prognosis and Prediction', *Computational and Structural Biotechnology Journal* **13**, 8–17.

Lebovitz, S., Lifshitz-Assaf, H. & Levina, N. (2022), 'To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis', *Organization Science* **33**(1), 126–148.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A. et al. (2020), 'From Local Explanations to Global Understanding with Explainable AI for Trees', *Nature Machine Intelligence* **2**(1), 56–67.

Lundberg, S. M. & Lee, S. I. (2017), A Unified Approach to Interpreting Model Predictions, *in* '31st Conference on Neural Information Processing Systems (NIPS)'.

Mehdiyev, N. & Fettke, P. (2020), Prescriptive Process Analytics with Deep Learning and Explainable Artificial Intelligence, *in* 'European Conference on Information Systems Proceedings'.

Meske, C., Bunde, E., Schneider, J. & Gersch, M. (2022), 'Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities', *Information Systems Management* **39**(1), 53–63.

Miranda, S., Berente, N., Seidel, S., Safadi, H. & Burton-Jones, A. (2022), 'Computationally Intensive Theory Construction: A Primer for Authors and Reviewers', *MIS Quarterly* **46**(2), iii–xviii.

Mitchell, T. R., Holtom, B. C., Lee, T. W. & Erez, M. (2001), 'Why people stay: Using job embeddedness to predict voluntary turnover', *Academy of Management Journal* **44**(6), 1102–1121.

Mithas, S., Xue, L., Huang, N. & Burton-Jones, A. (2022), 'Editor's comments: Causality meets diversity in information systems research', *MIS Quarterly* **46**(3), iii–xviii.

Molnar, C. (2020), *Interpretable Machine Learning.* Leanpub.

Mothilal, R. K., Sharma, A. & Tan, C. (2020), Explaining Machine Learning Classifiers through Diverse Counterfactual Explanation, *in* 'Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency', Association for Computing Machinery, pp. 607–617.

Müller, O., Junglas, I., vom Brocke, J. & Debortoli, S. (2016), 'Utilizing Big Data Analytics for Information Systems Research: Challenges, Promises and Guidelines', *European Journal of Information Systems* **25**, 289–302.

Nayak, S. C., Misra, B. & Behera, H. S. (2014), 'Impact of Data Normalization on Stock Index Forecasting', *International Journal of Computer Information Systems and Industrial Management Applications* **6**, 257–269.

Oswald, F. L., Behrend, T. S., Putka, D. J. & Sinar, E. (2020), 'Big Data in Industrial-Organizational Psychology and Human Resource Management: Forward Progress for Organizational Research and Practice', *Annual Review of Organizational Psychology and Organizational Behavior* **7**, 505–533.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016), "Why Should I Trust You?" Explaining the Predictions of Any Classifier, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining'.

scikit-learn (2023), scikit-learn. Machine learning in Python, https://scikit-learn.org/stable/index.html. Accessed: 06.06.2023.

Senoner, J., Netland, T. & Feuerriegel, S. (2022), 'Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing', *Management Science* **68**(8), 5557–6354.

Shin, D., He, S., Lee, G. M., Whinston, A. B. et al. (2020), 'Enhancing Social Media Analysis with Visual Data Analytics: A Deep Learning Approach', *MIS Quarterly* **44**(4), 1459–1492. doi:

Shollo, A., Hopf, K., Thiess, T. & Müller, O. (2022), 'Shifting ML Value Creation Mechanisms: A Process Model of ML Value Creation', *Journal of Strategic Information Systems* **31**(3), 101734.

Shrestha, Y. R., Fang He, V., Puranam, P. & von Krogh, G. (2021), 'Algorithm Supported Induction for Building Theory: How Can We Use Prediction Models to Theorize?', *Organization Science* **32**(3), 856–880.

Stoffels, D., Faltermaier, S., Strunk, K. S. & Fiedler, M. (2022), Opening the black-box of AI: Challenging Pattern Robustness and Improving Theorizing through Explainable AI Methods, *in* 'ICIS 2022 Proceedings'.

Sturm, T., Gerlach, J., Pumplun, L., Mesbah, N. et al. (2021), 'Coordinating Human and Machine Learning for Effective Organizational Learning', *MIS Quarterly* **45**(3), 1581–1602.

Subhash, P. (2017), IBM HR Analytics Employee Attrition & Performance, Kaggle - Database, https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset. Accessed: 02.11.2022.

Tremblay, M. C., Kohli, R. & Forsgren, N. (2021), 'Theories in Flux: Reimagining Theory Building in the Age of Machine Learning', *MIS Quarterly* **45**(1), 455–459.

Wang, X. & Zhi, J. (2021), 'A machine learning-based analytical framework for employee turnover prediction', *Journal of Management Analytics* **8**(3), 351–370.

Yang, G., Ye, Q. & Xia, J. (2022), 'Unbox the Black-Box for the Medical Explainable AI via Multi-Modal and Multi-Centre Data Fusion: A Mini-Review, two Showcases and Beyond', *Information Fusion* **77**, 29–52.

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*.

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B. & Zhu, X. (2018), "Employee Turnover Prediction with Machine Learning: A Reliable Approach", *in* 'Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2'. Springer International Publishing, pp. 737–758.