Opening the Black-Box of AI: Challenging Pattern Robustness and Improving Theorizing through Explainable AI Methods

Dominik Stoffels
University of Passau, stoffe11@ads.uni-passau.de

Stefan Faltermaier
University of Passau, stefan.faltermaier@uni-passau.de

Kim Simon Strunk
University of Passau, kim.strunk@uni-passau.de

Marina Fiedler
University of Passau, marina.fiedler@uni-passau.de

Opening the Black-Box of AI: Challenging Pattern Robustness and Improving Theorizing through Explainable AI Methods

Completed Research Paper

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

Stefan Faltermaier
University of Passau
Passau, Germany
stefan.faltermaier@uni-passau.de

Kim Simon Strunk
University of Passau
Passau, Germany
kim.strunk@uni-passau.de

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

Abstract

Machine Learning (ML) algorithms, as approach to Artificial Intelligence (AI), show unprecedented analytical capabilities and tremendous potential for pattern detection in large data sets. Despite researchers showing great interest in these methodologies, ML remains largely underutilized, because the algorithms are a black-box, preventing the interpretation of learned models. Recent research on explainable artificial intelligence (XAI) sheds light on these models by allowing researchers to identify the main determinants of a prediction through post-hoc analyses. Thereby, XAI affords the opportunity to critically reflect on identified patterns, offering the opportunity to enhance decision making and theorizing based on these patterns. Based on two large and publicly available data sets, we show that different variables within the same data set can generate models with similar predictive accuracy. In exploring this issue, we develop guidelines and recommendations for the effective use of XAI in research and particularly for theorizing from identified patterns.

Keywords: Explainable AI, XAI, machine learning, robust patterns, data-driven theorizing

Introduction

Artificial Intelligence (AI) has the potential to impact the way we live and work. Thereby, it is important to note that AI comprises various subdomains. Machine Learning (ML) is one such subdomain. ML approaches using algorithms to discover patterns in complex data sets are increasingly used to inspire models for explaining real-world phenomena (Brynjolfsson and Mitchell, 2017; Jordan and Mitchell, 2015; Sturm et al., 2021). Thereby ML holds great potential through offering accurate predictions from vast data sets. For example, it is used to predict cancer in patients based on x-ray scans (Kourou et al., 2015), or it is applied in current Covid-19 warning systems based on video footage (Dananjayan and Raj, 2020). Further, ML is used in organizational contexts by HR management to predict employee performance or turnover (Yuan et al., 2021) or to analyze user engagement on social media (Shin et al., 2020).

Patterns identified through explorative ML application are increasingly used for data-driven theorizing and subsequent analysis (Berente et al., 2019; Choudhury et al., 2021; Lindberg, 2020). ML's ability to discover patterns in large and complex data offers tremendously powerful support to research and practice (Jacobucci and Grimm, 2020; Padmanabhan et al., 2022). However, in order to provide the promised opportunities and to inspire data-driven theorizing, patterns derived from ML application need to be robust and interpretable, which is not directly possible due to the black-box nature of ML (Fernández-Loría et al., 2022; John-Mathews, 2021). Especially, for data-driven theorizing detailed analyses of patterns are required to derive understandable, actionable, and generalizable theoretical models (Berente et al., 2019; John-Mathews, 2021; Lindberg, 2020; Shrestha et al., 2021).

To allow for evaluation and interpretability of detected patterns, recent research developed explainable artificial intelligence (XAI) methods. XAI tools offer post-hoc analyses to reveal the main determinants in a pattern are gaining attention and are increasingly used in research (e.g., Choudhury et al., 2021; Sturm et al., 2021; Senoner et al., 2021). A central reason is that XAI allows to uncover learned patterns that build the basis for black-box ML algorithms' predictions. Without XAI insights, patterns in data, received through black-box ML algorithms, cannot be understood, i.e., variables' impact on the predicted outcome remains unclear. XAI enables researchers to explore and evaluate complex patterns detected by ML (Sturm et al., 2021), and eventually to use these patterns for data-driven theorizing (Berente et al., 2019; Lindberg, 2020). To apply XAI and to evaluate the ML patterns, for instance in terms of robustness, is essential for research to benefit from ML. However, literature on the application of XAI and the implications for IS research is absent, revealing a striking void in IS research (Padmanabhan et al., 2022).

To contribute to this gap, we compare seven ML approaches to detect patterns in two publicly available and widely used data sets. We analyze the resulting patterns with two different XAI methods, showing that predictions of similar accuracy can in fact originate from different indicators within the same data set. By exploring this issue and challenging the common practice of reporting only the best model, we generate additional insights for assessing pattern robustness, interpretability, and subsequent opportunities for data-driven theorizing based on ML. Thereby, we offer two contributions for applying ML and XAI for data-driven theorizing. We provide recommendations on rigorous XAI use to analyze ML generated patterns (Padmanabhan et al., 2022) and link the insights to the ongoing discussion on using ML patterns to inspire data-driven theorizing (Berente et al., 2019; Lindberg, 2020). Thus, we contribute to understanding when, where, and how ML's analytical capabilities can benefit research effectively.

Research Background

ML is an essential subset of AI and enables discovering patterns in vast data sets, which can hardly be found without such learning algorithms, promising great benefit to research (Choudhury et al., 2021; Sturm et al., 2021). Especially, by discovering complex and often nonlinear patterns in data sets with large numbers of variables[1] and observations (Tonidandel et al., 2016), ML can help to extend our toolkit for theorizing (Shrestha et al., 2021). Thereby, ML allows for approaching the complexity of phenomena in vast data sets, promising to improve our theorizing and theoretical models. Potentially more accurate models can subsequently be analyzed by traditional means and enhance our understanding of real-world phenomena (Brynjolfsson and Mitchell, 2017; Choudhury et al., 2021; Jordan and Mitchell, 2015; Sturm et al., 2021).

For research to benefit from ML, derived patterns need to be robust and thus reliable across different ML approaches (Choudhury et al., 2021), as well as, to some extent, interpretable (John-Mathews, 2021; Lipton, 2018). Interpretability requires understandability (John-Mathews, 2021; Lipton, 2018), stressing the need for the results, if not the ML algorithm's internal structure, to be explainable (Barredo Arrieta et al., 2020). Without understandability, using black-box algorithms will prevent content-related insights. Understandable and robust ML approaches ensure that the identified patterns potentially deliver interpretable insights based on more than mere correlation and coincidence. Moving beyond coincidence is, however, vital for any ML detected pattern in research for theory building, even in explorative analyses.

Against this background, XAI gains importance in ML research, as XAI offers insights on how ML predictions and underlying patterns emerged (Hamm et al., 2021). Thereby, XAI affords the possibility to evaluate these patterns. The patterns, which black-box algorithms build their outcome predictions on, can

[1] The terms *variable* and *feature* refer to overlapping concepts. *Feature* is used in the context of ML and in the selection of predictors used, whereas *variable* is prevalent in research. Depending on context, we use *features* in the ML methodology and *variable* when referring to theorizing and interpretation.

be approximated and subsequently visualized through XAI, offering a post-hoc explanation. For many black-box ML approaches comprehending the underlying patterns learned from the data without XAI - if possible at all - requires substantial time and computational efforts. For illustration, please consider a Random Forest model. This common black-box ML approach combines a potentially large number of decision trees as basis for its outcome prediction (Breimann, 2001). Trying to precisely understand the pattern that underlies the prediction and which features in a data set impact the outcome, requires to analyze the entire number of decision trees (e.g., 1,000) manually. XAI offers an automated and interpretable approximation of the patterns the algorithm identified, making it an essential tool.

Different XAI methods provide such post-hoc explanations of which two frequently employed are Shapley Additive explanation values (SHAP) (Lundberg and Lee, 2017) and Partial Dependence Plots (PDP) (Kamath and Liu, 2021). Both, SHAP and PDP, are model-agnostic methods, i.e., they are model-independent approaches that can be used to analyze different ML models. SHAP is a local model-agnostic method (i.e., a method that aims to explain how individual predictions are made) with a game theoretic approach, which helps to explain individual predictions. Unlike other local model-agnostic methods such as the local surrogate method, the local interpretable model-agnostic explanations (LIME), SHAP offers different global interpretation methods like feature importance, feature dependence, and summary plots (Molnar, 2020), identifying important model features. As a global model-agnostic method (i.e., a method explaining how the model as a whole made predictions) PDP visualizes a feature's marginal impact on model prediction (Friedman, 2001). By providing these detailed post-hoc insights, the different XAI methods aid understanding patterns discovered by black-box ML approaches and support theorizing (Shrestha et al., 2021).

Literature on ML and XAI application grows rapidly in IS and management research (Choudhury et al., 2021; Tidhar and Eisenhardt, 2020). For example, Choudhury and colleagues (2021) identified a pattern for employee turnover by using black-box ML algorithms and analyzed the identified pattern in terms of feature importance to understand how single features impacted the outcome prediction, using XAI methods (e.g., PDP). Emphasizing the importance of interpretability and robustness of patterns is essential for harnessing ML's potential for theory building. Yet, the central role XAI is not explicitly discussed. Also Tidhar and Eisenhardt (2020) used several different ML algorithms, exploratory data analysis, and multi-case theory building. Using qualitative data from semi-structured interviews and other data sources, they built a theory for optimal revenue models in apps. While also addressing benefits of ML for theorizing and acknowledge the importance of robust patterns. However, as they report robust patterns across the different approaches, they do not extensively discuss conflicting patterns of comparable predictive power nor do they address how XAI can help to uncover this potential issue.

Besides these application-oriented studies, additional research focuses on theory building based on ML generated patterns (Berente et al., 2019; Lindberg, 2020). Berente and colleagues (2019) discuss how ML can be used in what they call Computationally Intensive Theory Development, by combining qualitative approaches and computational techniques. Lindberg (2020) develops a framework considering both, human and ML pattern recognition, as key elements for theory building. ML, as tool, can improve our theorizing by revealing patterns in data to complement human sensemaking. Thereby, both manuscripts discuss the central role the interpretation of identified patterns plays. However, due to the meta perspective these papers take in, discussing extensively how researches can identify and handle ambiguous patterns, supported by XAI, is out of their scope. Finally, there are some recent studies that explicitly address XAI in IS research (Fernández-Loría, 2022; Hamm et al., 2021; John-Mathews, 2021). Hamm and colleagues (2021) investigate XAI's impact on users' attitudes toward ML. That is, they propose a model on how XAI's abilities to make identified patterns more interpretable, impact use intention for ML (Hamm et al., 2021). Also, there are some studies with a critical perspective on XAI (Fernández-Loría et al., 2022; John-Mathews, 2021). Thereby, Fernández-Loría and colleagues (2022) show that, under certain conditions, alternative approaches are more suitable to explain data-driven decisions than common XAI methods like SHAP. Still, the authors stress the importance of making underlying patterns in data explainable. Last, building on his preliminary findings, John-Mathews (2021) stresses that transparent ML models can be more interpretable than post-hoc explanations of black-box models. Yet, given the imposed limits by using only interpretable ML approaches, XAI remain crucial for interpreting patterns in data.

To summarize extant research; ML can benefit theorizing and thereby XAI can be vital to understand ML patterns by improving their interpretability. However, there are no studies explicitly showing how XAI can uncover competing ML patterns in data and what this implies for subsequent theorizing. Thus, we address this gap by giving guidance for effective XAI use to improve and benefit the use of ML and XAI in research.

Data and Methods

For demonstrating the divergence in ML patterns, we used two real-world data sets. As the manuscript's aim is not to actually build theory empirically from the data, but to illustrate important aspects one has to consider when using ML and XAI for building theory, we will not engage in an extensive content discussion regarding the identified patterns.

First, we used data from the European Social Survey round 8 (ESS Round 8, 2016) which is publicly available[2]. We refer to this data set as ESSD (European Social Survey Data). The ESSD entails a cross-national survey across European countries using face-to-face interviews on participants' individual characteristics (e.g., age, gender, nationality), and their particular attitudes, beliefs, and behavior patterns (ESS Round 8, 2016). Overall, the ESSD comprises data from 44,387 individuals across 23 countries. We reduced the number of features on content- and context-related considerations and we conducted feature engineering by aggregating single items (i.e., questions), which were consistent in content, and calculating the mean value to construct new features (for an exhaustive list of features, see Table 3 - Appendix A). This led to a final 42 features as basis for the subsequent analysis in which we aimed to predict an individual's willingness to vote for leaving the European Union (0 = leave, 1 = remain - after recoding). Next, we aimed to gain insight on how these predictions came about. That is, we analyzed the patterns which we identified through the application of ML and XAI, i.e., we investigated which features impact the willingness to vote for leaving the European Union and how they impact this decision. Thereby, the value a feature has (simply put high or low) can increase the probability of voting for leave (0) or increase the probability of voting for remain (1). We would like to give the following illustrative example: The participants were asked how emotionally attached they are to Europe (*atcherp)*. They had to choose a value from 0-10 whereas 0 meant "not at all emotionally attached" and 10 meant "very emotionally attached". We would logically expect to see, that emotional attachment to Europe has an impact in the given context (i.e., ML/XAI should identify it as important feature), whereas high values should increase the probability of voting for remain and low values increase the probability of voting for leave.

Second, we used the adult income data set (AID) available in the UCI Machine Learning Repository (link: https://archive.ics.uci.edu/ml/datasets/adult) (Dua and Graff, 2019). The AID is a well-known, frequently used, and publicly available data set (e.g., Rosset, 2004; Agrawal et al., 2005), that covers census information collected by the United States Census Bureau, entailing individual characteristics such as age, gender, nationality, and income across 48,842 individuals (for an exhaustive list of features, see Table 4 - Appendix B). The AID comprises 14 categorical and numerical features, which we partly adapted through one-hot-encoding and min-max normalization. Based on these features and as a second model for our demonstration, we aimed to predict whether a respondent's income is above or below $50,000 (0 equals <= $50,000, 1 equals >$50,000). Again, we were interested in the prediction patterns and proceeded analogously to the analysis of the ESSD. For example, the highest level of education reached (education-num), was collected by the Census Bureau for the people in the data set. Subsequently, the different answers were put in a reasonable order, i.e., Bachelor's Degree < Master's Degree <Doctorate and so on. We would expect to see educational level impacting income, whereas high values should increase the probability of earning more than $50,000 (1) and low values increase the probability earning less than $50,000 (0).

Selection of ML and XAI Approaches

In the data analysis we compared seven well-established ML approaches (Wu et al., 2007; Hastie et al., 2009) using two different XAI methods (Kamath and Liu, 2021, Lundberg and Lee, 2017). The seven ML approaches are Artificial Neural Networks (ANN) (Hassoun, 1995), Random Forests (RF) (Breiman, 2001), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), Light Gradient Boosting Machines (LightGBM) (Ke et al., 2017), Decision Trees (DT)

[2] The complete data set as well as detailed descriptions on how the data was obtained (e.g., an exhaustive list of questions asked) can be freely accessed using the following link: https://ess-search.nsd.no/en/study/f8e11f55-0c14-4ab3-abde-96d3f14d3c76.

(Breiman et al., 1984), and Logistic Regressions (LG) (Hosmer et al., 2013), of which the first five are black-box model approaches. Due to the fact that countless ML models exist (e.g., Fernández-Delgado et al., 2014) and it is not a priori apparent, not even to experienced data scientists, which model provides the best performance for a given task, a sample of models has to be picked (Choudhury et al., 2021). We uniformly analyzed the approaches using the model-independent adaptable XAI implementations SHAP (Lundberg and Lee, 2017) and PDP to analyze and compare the learnt patterns of the models (Kamath and Liu, 2021).

Data Preprocessing, Feature Selection and Data Partitioning

Since we used real-world data, we had to check data quality and prepare it appropriately before applying the required analytic procedures (García et al., 2015; Kotsiantis et al., 2006). Entries with some missing information in the features, were removed. Categorical features were transformed using one-hot encoding (Hancock and Khoshgoftaar, 2020). In order not to lose a significant number of data points and to allow the use of an ordinal scale, we dealt with ESSD cases in which respondents abstained from answering by replacing this information with the mean value (Kotsiantis et al., 2006).

In general, there are multiple ways of selecting a final set of features (García et al., 2015; Miao and Niu, 2016). Since both of our data sets contained nominally and ordinally scaled features, we used the Mutual Information (MI) filtering method (Chakraborty et al., 2019, Shannon, 1948) to select our final 30 features for each data set. MI can be used as an effective criterion for feature selection in classification as well as regression problems (Degeest et al., 2019). MI is a quantity from information theory that indicates the strength of the statistical correlation of two random variables, or in our case, how much information a feature (respectively a set of features) has about the label. If the variables are completely independent, MI is equal zero. In contrary, a higher value means a higher dependency between the variables (Kraskov et al., 2004). Therefore, MI is a non-negative value, which measures the dependency between two variables and is based on the entropy of the variables (Cover and Thomas, 1991). Formally, this can be described by the following term (Kraskov et al., 2004): $I(X;Y) = H(X) - H(X|Y)$. Thus, $I(X;Y)$ is the mutual information of the variables X and Y, $H(X)$ is the entropy for X and $H(X|Y)$ is the conditional entropy for X at Y. However, for real data sets, MI cannot be computed directly because it is defined in probability density functions which are unknown if only a limited sample of data is available. Thereby, MI has to be estimated from the limited data set (Degeest et al., 2019), as described for example in Kraskov et al. (2004) and Ross (2014). Next, we divided data into training and testing data so that training and evaluation are based on different parts of the initial data set. For this purpose, we applied a common train/test split of 80/20% (Rácz et al., 2021). Since the features were based on different entities and differed in range, we applied a min-max normalization to prevent biased results (Nayak et al., 2014).

Since both data sets had an imbalance of classes, we employed an adaptive synthetic sampling approach (ADASYN). This approach improves the previously often used synthetic minority oversampling technique (SMOTE) approach (He et al., 2008). ADASYN ensures that the classes in both data sets are balanced evenly. Importantly, we applied this only to the training data set to prevent any bias in the test data set.

Hyperparameter Tuning, Model Training, and Evaluation

Next, we applied hyperparameter tuning to improve the models' performance (Hutter et al., 2019). Table 1 provides all parameters considered for every model. We tested 25 different sets of configurations for each model. To validate the results, we applied 10-fold cross-validation (Berrar, 2018; Hastie et al., 2009). For faster computation time, we took two approaches: First, we used a randomized grid search to find the best model parameters out of the 25 attempts (Bergstra and Bengio, 2012). Second, we used a multi-fidelity method with a subset of the set (with 4,000 data points). While neither method guarantees optimal results, they reduce computation time and likely improve model performance remarkably (Hutter et al., 2019).

After optimizing the hyperparameters, we could finally define the models and train them based on the full training data set. To select the best models, different metrics such as such as the Accuracy (Acc), a confusion matrix, recall, precision, the F1 score (which is interpreted as a harmonic mean of the precision and recall), or the receiver operating characteristics accuracy score (ROC AUC) can be considered (Choudhury et al., 2021). To provide high transparency, we report the Acc, the weighted-averaged F1 score, and the ROC AUC score. However, since the source data were unbalanced, we could not use prediction accuracy as metric for classification, since accuracy only indicates how well a certain approach assigns data points to given classes.

For example, if we assume that 99% of the data belong to class A and 1% belongs to class B, an assignment of all data points to class A would already give a prediction accuracy of 99% (Choudhury et al., 2021). In order to make a distinct model choice between ROC AUC and F1 score, we chose ROC AUC score as the appropriate metric here. The ROC AUC curve compares the model's true positive rate with the false positive rate, varying the threshold for discriminating between classes (Choudhury et al. 2021).

ML algorithm	Parameters
Artificial Neural Network	Units, hidden layers, activation function, output layer activation function, epochs, batch size, dropout rate
Random Forest	Number of estimators, maximum features, maximum depth
Support Vector Machine	Kernel, gamma, penalty of error (C)
Decision Tree	Maximum depth, criterion
Logistic Regression	Solvers, c-values
XGBoost	Gamma, subsample, colsample bytree, maximum depth, min_child_weight
LightGBM	Lambda l1, lambda l2, number of leaves, minimum number of data per leaf

Table 1. Hyperparameters used in model optimization

Finally, we checked the best models (in terms of ROC AUC scores) for each data set with the XAI approaches to gain deeper insight in relation to the learnt correlations. If the performance of the best models differed only slightly (<1%), we considered multiple models for XAI analysis. First, we used the SHAP values to analyze the features' influence in the overall context and to assess the individual predictions of the models. For this purpose, we defined a sub sample to the data set which contained 50 observations. Within these subsamples each class was represented by 25 instances. Further, to analyze the three features with the highest impact on the model prediction according to SHAP values in detail, we used PDP.

Results

Performance Evaluation and Model Selection for Comparison

As is common in ML approaches, we initially set out to find an accurate predictive model for our data set. In both cases, we analyzed the seven well-established ML approaches summarized in Table 2. In the ESSD, the ANN (74.41%) performed slightly better than the rest. However, the SVM (74.14%) and LG (74.28%) models barely performed poorer, providing comparable ROC AUC scores (<1% difference). The RF (71.37%) and DT (71.62%) performed several percent worse. In the AID, the RF (81.28%) and LG (81.69%) algorithms performed better than those of the other approaches.

Data Set	Performance metric	Machine learning algorithm						
		Artificial Neural Network	Random Forest	Support Vector Machine	Decision Tree	Logistic Regression	XG-Boost	Light-GBM
ESSD	ROC AUC	74.41	71.37	74.14	71.62	74.28	68.98	68.35
	Acc	72.09	83.74	72.02	78.93	72.86	84.08	84.28
	F1 score	74.86	83.29	74.78	79.87	75.47	82.54	82.78
AID	ROC AUC	80.66	81.28	80.09	79.78	81.69	79.86	80.09
	Acc	77.31	78.97	76.19	76.25	78.00	78.18	81.29
	F1 score	78.86	80.31	77.87	77.91	79.51	80.02	81.75

Table 2. Performance overview in percentage [%]

Subsequently, we selected the models with the highest predictive accuracy (in terms of the ROC AUC), as well as the ones deviating only slightly from the highest (ROC AUC deviating less than 1% compared to the best model) for a detailed comparison. This detailed comparison delivered the three findings explicated below. At this point, we would like to emphasize that the selected models are very similar in each performance metric and therefore have an overall comparable prediction power.

Finding 1: Model Variations within the Data Sets

For first insight into the models, we use a global interpretation method – the SHAP Summary Plot. Unlike a simple feature importance indicator, this plot shows the direction of each variable's influence on the model prediction. Also, its representation as a scatter plot, offers additional insight into a feature's impact. The colors of the individual dots indicate the initial relative value of the feature per instance. That is, if the feature has a large value, it will be colored red, if it has a comparatively small value, this dot will be colored blue. That means, it is possible to determine whether a high or low feature value is indicative of a model prediction of leaving or respectively remaining in the EU. This allows for a more fine-grained comparison across features, as the influence of the features can differ significantly depending on the other features' values for each observation.

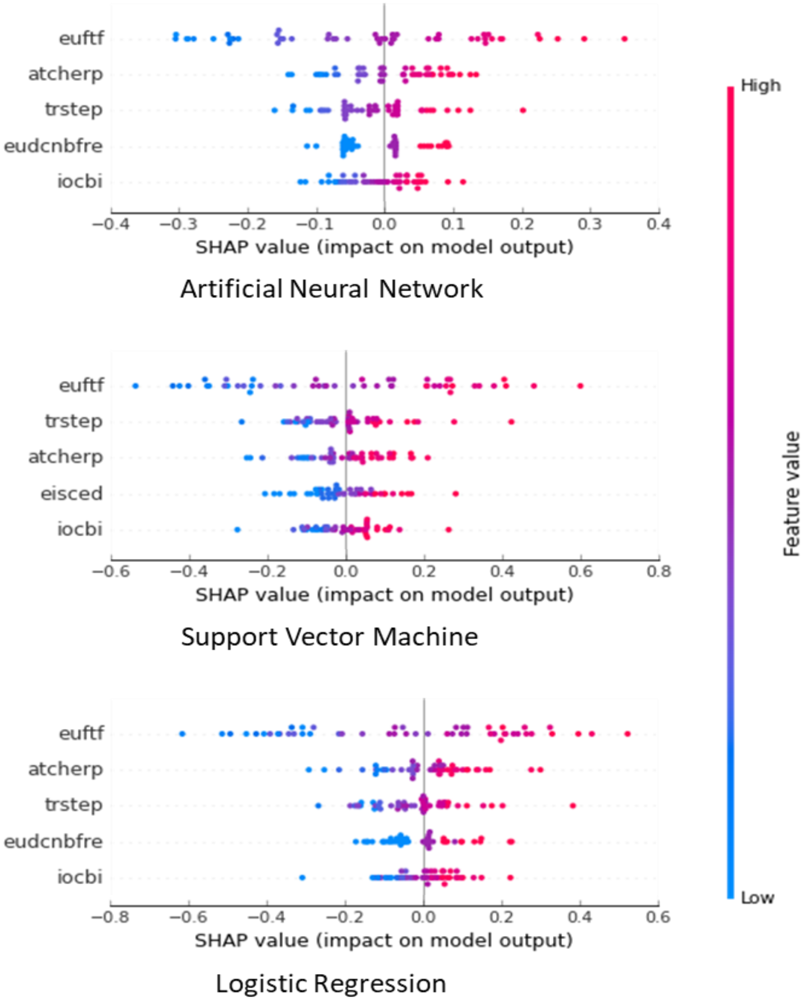
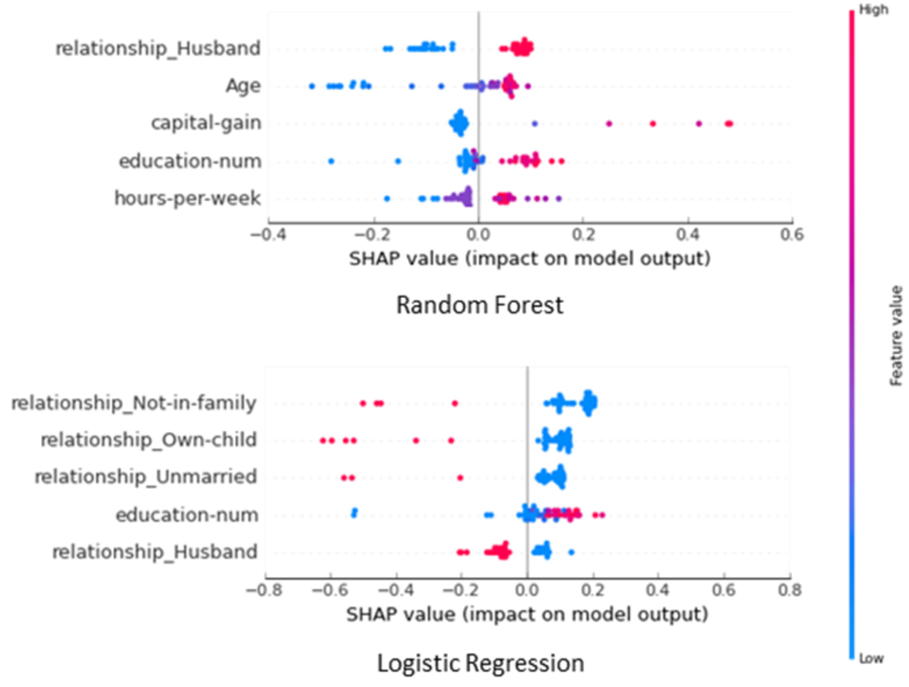


Figure 1. SHAP summary plots to analyze the ESSD features' influence and impact

Regarding the ESSD, the three models shown in Figure 1 offer comparable predictive accuracy. For all three models the three features with the largest amplitude on the model prediction are *euftf* (European Union: European unification go further or gone too far; 0 = unification already gone too far, 10 = unification go further)*, atcherp* (how emotionally attached to Europe; 0 = Not at all emotionally attached, 10 = Very emotionally attached), and *trstep* (trust in the European Parliament; 0 = No trust at all, 10 = Complete trust). In detail, the results suggest that people who think that European Unification has gone too far (*euftf*), are more likely to vote for "Leave" (and vice versa), that people who are highly emotionally attached to Europe (*atcherp*) are more likely to vote for "Remain" (and vice versa), and that people who have low trust in the European Parliament (*trstep*) are more likely to vote for "Leave" (and vice versa). The way these features influence model prediction is comparable across the different models. The only deviation we can see here, is in the order of the three most important features, as for the SVM, trstep is weighted higher than *atcherp*.

In contrast to the ESSD analysis, our analysis of the AID as represented in Figure 2, did not provide robust patterns in that across the two models with the highest accuracy in terms of ROC AUC scores, the most important features varied. Using RF, the three most impactful features for model prediction it provided, are *relationship_Husband* (where the term *relationship* represents what an individual is relative to others, and *relationship_Husband* indicates whether an individual is a husband relative to others; 0 = no husband, 1 = husband), *Age* (the age of an individual), and *capital-gain* (capital gains for an individual). That is, RF results suggest that married men (*relationship_Husband* = 1) are more likely to earn over $50,000, that younger individuals (*Age*) are more likely to earn less than $50,000, and that individuals with high capital gains (*capital-gain*) are more likely to earn over $50,000. LG, however, provided *relationship_Not-in-family* (whether an individual is in a family relationship relative to others; 0 = no family relationship, 1 = in family), *relationship_Own-child* (whether an individual is an own child relative to others; 0 = no own child, 1 = own child), and *relationship_Unmarried* (whether an individual is related by marriage; 0 = no marriage, 1 = marriage) as most influential features regarding the model prediction. That is, LG results suggest that individuals without an own family (*relationship_Not-in-Family* = 0) are more likely to earn more than $50,000, that individuals who are an own child (*relationship_Own-child* = 1) are more likely to earn less than $50,000, and that unmarried individuals (*relationship_Unmarried* = 0) are more likely to earn more than $50,000.



Random Forest

Logistic Regression

Figure 2. SHAP summary plots to analyze the AID features' influence and impact

In summary, our analyses suggest that models of similar accuracy in terms of the ROC AUC do not necessarily build on the same set of features. While the most influential features are widely consistent in the ESSD prediction, they vary substantially in the AID prediction. In the AID, only two in the top-five features overlap in the two predictions, showing that similar accuracy was achieved by building on different factors. Therefore, the patterns identified in the ESSD suggest higher consistency than in the AID.

Finding 2: Single Features differently impact Model Prediction

For further model insights, we compared PDP plots for the most important features according to our previous analyses. Analyzing PDP plots for features in the ESSD provides intriguing insight on how features' impact on predictions differs across ML approaches, as shown in Figure 3.

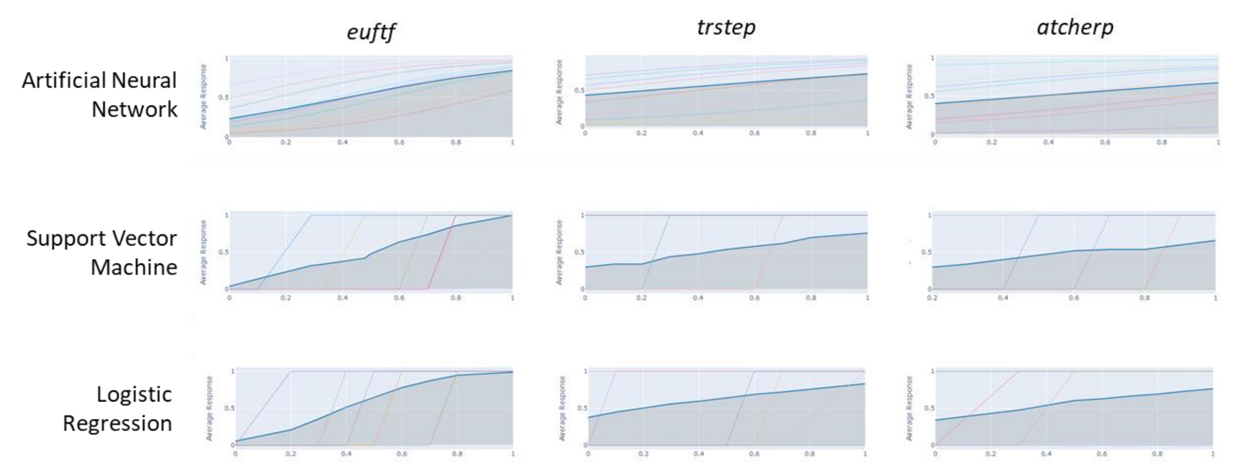


Figure 3. PDP plots of the top three features in ESSD

In detail, the plots reveal that *euftf* shows the expected tendency throughout all models (higher feature values, i.e., indicating unification should go further, lead to an increased probability of voting "Remain"). However, the course of *euftf* on the average response on the model output is quite different. The dependence graphs in the ANN run rather linearly whereas they are not linear in the SVM and LG. Despite supporting the overall tendency, the indicators suggest that patterns for the same feature differ across the different approaches.

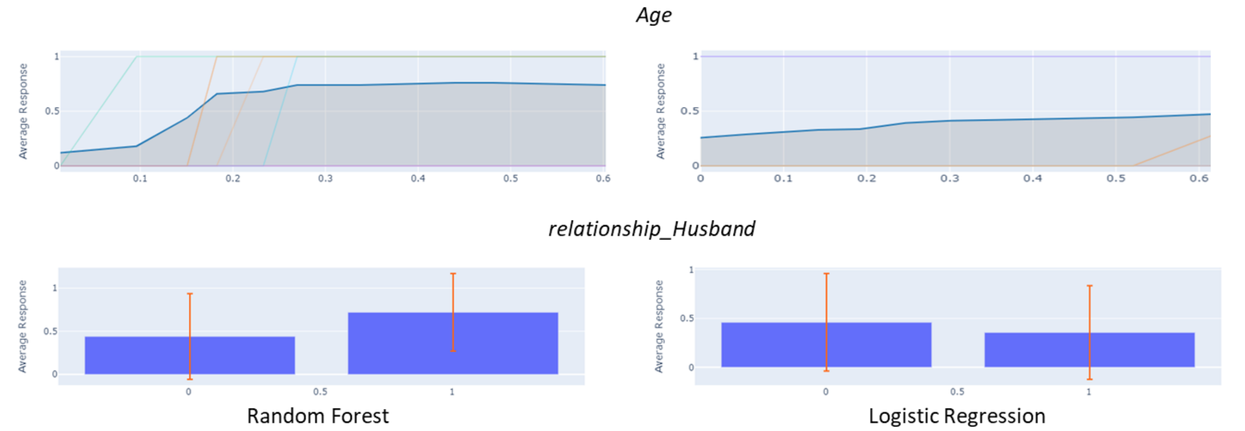


Figure 4. PDP plots for Age and relationship_Husband in the AID

In case of the AID, we chose *Age* and *relationship_Husband* for our more detailed analysis as shown in Figure 4. As these features are the most important ones in the RF model, but are far less important in the LG model, they promised to offer particularly interesting insights (of course, this also applies to the most important features of the LG model, which could have also been selected for our exemplary analysis). With the AID it is easy to recognize how the two selected features influence model prediction. As expected, an increase in *Age* positively impacts the predicted probability of earning over $50,000. However, the RF approach suggests a clearly recognizable threshold, which is not evident in the LG model.

By using XAI approaches for detailed analyses of ML outputs, we illustratively show how feature impact differs across different ML approaches. For instance, one ML approach shows a linear pattern for a certain feature, while another approach suggests a curved pattern or a threshold level for the exact same feature.

Finding 3: Opposing Feature Impact on Model Prediction between Approaches

XAI enabled us not only to reveal a shift in feature importance and the course of the relationship from feature to prediction; it also showed how different approaches suggested an opposing effect for a certain feature. This is illustrated in the SHAP summary plot and the PDP plot, which indicated that one approach (RF) suggests a feature value of 1 of *relationship_Husband* (part of the AID) to be positively related to the model outcome (average response over 0.5) and vice versa, whereas the other approach (LG) suggested the exact opposite, i.e., feature value of 1 with a negative relation (average response under 0.5) to the outcome. Thus, these different approaches suggest completely different interpretations of the features' impact.

These findings again stress the importance of critically evaluating model selection, since similar model fits do not necessarily entail uniform patterns. Our findings, therefore, illustrate that choosing an AI model based solely on best fit (i.e., accuracy), even if the relative advantage is only incremental, might have a stronger impact than one would assume. This would be particularly risky if only the single best model is considered in decision making or explored in further research, without any alternatives being reported.

Discussion

Literature on the application of ML constitutes an important stream in AI research (Padmanabhan et al., 2022), and the use of XAI to make sense of black-box ML outputs provides an especially intriguing research avenue. Our recommendations regarding XAI application, which are based on our analyses' illustrative findings, contribute to the growing and promising stream of XAI literature (e.g., Hamm et al., 2021; John-Mathews, 2021). Further, we contribute to the discussion on the benefits ML can offer for theorizing (Berente et al., 2019; Choudhury et al., 2021; Shrestha et al., 2021; Sturm et al., 2021), by demonstrating how XAI discloses potential ambiguities in identified patterns. Revealing these ambiguities affords the basis for a sophisticated understanding of patterns and thereby benefits data-driven theorizing.

Recommendations for Using XAI

To assess pattern robustness and to identify ambiguous predictions, we put forward the following recommendations for XAI use.

Compare different Models

Our findings emphasize that selecting models based merely on predictive accuracy, indicated for example by ROC AUC scores, can be problematic, as we miss information on robustness of the identified patterns. We therefore recommend that researchers always introduce different ML approaches and then compare the models the ML application delivers. Results obtained by using XAI to analyze the different ML approaches' output, demonstrated the importance of comparing different models. In the case of our first data set (ESSD) the models indicating the highest accuracy in terms of ROC AUC, provided similar patterns. In contrast, comparing the different models our second data set (AID) produced, using XAI revealed varying patterns. If various models derive their predictions from similar or identical patterns, this increases confidence in these patterns and diminishes the probability that the patterns are based merely on statistical artefacts or unknown characteristics in our data. Varying patterns, however, need further analysis to enhance our understanding of potential reasons for the disclosed regularities, and to determine the extent to which we can rely on the results for further use in research or practice.

Investigate Variations

As shown in our analyses, XAI methods further afford the opportunity to take a closer look at the models, and given that results vary across models, albeit that prediction accuracy is similar, these variations need further examination. Thereby, XAI itself allows us to detect differing patterns as well as differing impact of variables. However, XAI cannot fully explain why these patterns differ, as it only provides post-hoc explanation to black-box models, without enabling insight in how exactly the identified patterns emerged

(Barredo Arrieta et al., 2020). Thus, besides applying XAI we need to look at potential biases in data and report traditional metrics, as for example multi-collinearity, since an undetected high correlation of variables could lead to mathematically interchangeable variables[3], which in turn would distort the resulting patterns (Choudhury et al., 2021). Eventually, ambiguity in the disclosed patterns can also come from highly complex relations that characterize the phenomenon at hand. That would call for the application of measures that go beyond pattern detection through ML, i.e., for human sensemaking through research.

Be aware and critical

ML approaches are complex, allowing us to handle vast data sets and complicated phenomena, but at the same time requiring careful assessment. In the light of our findings, selecting best fit models, without critically assessing them, or selecting them based on single or too few indicators, makes little sense and seems potentially misleading. Working with ML in both research and practice, an awareness of variation detected in patterns prompts us to adapt our ML practices. A first approach is to report selective indicators, as well as to offer a wider range of metrics and particularly to consider the somewhat descriptive XAI outputs more extensively. This affords us an opportunity to critically evaluate the derived patterns. Especially, when working with openly available data sets researchers did not collect themselves, and that therefore were not guided by the researcher's preliminary theorizing, alternative patterns should be carefully considered. Considering the far-reaching consequences that can arise when we derive real-world decisions such as credit approval or when we explore options for theory development, rigorous and critical ML application is particularly important. Being aware of equally accurate analytic approaches that give divergent explanations of variables can in fact inform subsequent theorizing, while lack of such awareness might limit insight and increase the potential pitfalls of *a posteriori* theory building.

Considering these recommendations, XAI can improve ML applications in two ways. XAI aids evaluation of the robustness of identified patterns, and it can increase the understandability of the ML outputs. This benefits the theoretical models derived and theorizing inspired by the patterns gained via ML methods. Using XAI to compare multiple ML generated models with similar prediction accuracy, can deliver two results. Either, the different ML approaches learn similar models, i.e., the most important features and their impact remain consistent across different approaches, or the learned models differ across different approaches. In the first case we can confidently assume that the revealed relations stem from relations that reflect reality. Thus, even if the patterns we receive from ML are first and foremost correlation (Choudhury et al., 2021; Zhao and Hastie, 2019), reproducing patterns across approaches offers a strong basis for data-driven theorizing and improves abductive theory building that originates from these patterns. In the latter case, due to the ambiguity of non-robust patterns, they present a weaker basis for theorizing without further effort. Without XAI we risk overlooking conflicting patterns, especially when different ML approaches offer similar prediction accuracy, but diverging patterns. Revealing these differing patterns, allows us to further investigate the reasons for their ambiguity and thereby prevents us from overestimating the reliability the identified patterns provide. Second, by investigating variables' impact on predicted outcomes, we gain detailed insights on single variables' influence or detect variables that exert ambiguous influence on predictions and thus need further analysis. This lays the foundation for the understandability of the received patterns and thereby also for data driven theory development. Consequently, XAI helps to improve theorizing in IS research by offering fine-grained insights into the roles of single variables in large data.

Implications for Research

Improving Robustness for detected Patterns

Given that research increasingly uses ML algorithms due to their potential for pattern detection in complex data sets (Shrestha et al., 2021; Sturm et al., 2021), we need increasingly to consider the robustness of patterns. We agree that innovative data analysis methods alone do not promote understandability of a phenomenon (Berente et al., 2019) and that additional theorizing by researches themselves is vital (Berente et al., 2019; Sturm et al., 2021). Additionally, we strongly agree that detected patterns need researchers'

[3] As suggested, we checked this alternative explanation by analyzing the correlation matrix (see Figure 5 - Appendix C) (Choudhury et al., 2021). We found the average correlation between the different top 5 most important features of the AID was 0.14 and all features' correlation was below 0.5.

sensemaking and abductive theorizing for transferring ML into theories (Shrestha et al., 2021). However, as researchers in both our roles, i.e., in authoring manuscripts using ML algorithms or in reviewing and evaluating such manuscripts, we need to be confident that the patterns found and reported offer a stable basis for theorizing. This is a clear requirement for indicating that subsequent models are indeed properly derived and not merely *a posteriori* justification for an unstable pattern.

We agree with Sheshtra and colleagues (2021) who illustrate nicely how robust patterns are essential to theorizing, and how research which employs abductive thinking is essential to theory building. Nonetheless, being aware of the potential volatility in ML solutions, as researchers we need to ask whether we are able to detect such volatility in our own projects or in a manuscript presented to us for review. If presented with a single best fitting and unquestioned model, we will most likely not be able to find potential flaws and may thus overestimate the robustness. The XAI methods we have presented, offer additional tools for questioning the robustness of patterns derived for theorizing. Both in our own work and in reviewing others' work, with XAI methodologies we should increasingly expect additional reports in order to strengthen our confidence in the robustness of disclosed patterns.

Using XAI for Interpretability and Theorizing

Critical voices may argue that ML, essentially building on computational logical and numerical operations, appears more familiar to quantitative research (Lindberg, 2020). Considering criteria for quantitative research, black-box algorithms providing emergent solutions may fall short in rigorous theory testing. Given the limited adaptability by researchers, essentially deciding for an algorithm instead of operationalizing constructs and relationships in accordance with theory (especially when analyzing secondary data), and the emergent character of solutions, quantitative rigor may not be met. Nevertheless, interpreting this as limited usefulness of ML in research neglects the opportunities for qualitative researchers (Grover et al., 2020, Miranda et al., 2022).

Considering the rapidly growing data, accompanying the use of technology, qualitative research – traditionally relying on manual analysis of interviews and observations – struggles with excessive amounts of data (Lindberg, 2020). Early in the emergence of qualitative research in organizational and IS research, Eisenhardt (1989) acknowledged this challenge by stating that "people are notoriously poor processors of information" (p.540). This leaves two options for qualitative research: either limit qualitative research to data sets that are comprehensible by humans in a reasonable amount of time or expanding the qualitative methodological field kit (Lindberg, 2020, Rizk & Elragal, 2020). ML can contribute greatly in this endeavor (Miranda et al., 2022, Pentland et al., 2021). ML – like a microscope revealing details too tiny for the human eye – can detect patterns in data to vast for human analysis (Pentland et al. 2021) and thereby help researchers in theory building (Lindberg, 2020, Grover et al., 2020).

We argued that ML benefits theory building; but where do XAI methodologies provide additional benefits? XAI helps to increase understandability of patterns discovered through ML by providing post-hoc explanations on the importance and individual influence certain variables have on predictions. However, we need to acknowledge that this alone does not make patterns sufficiently interpretable for immediate theory building. As a complex concept, interpretability requires more than the insight on how the patterns emerged (i.e., more than understandability). In this context, research frequently discusses actionability, which provides guidelines on how to go ahead with the knowledge gained, as well as generalizability which takes us beyond a specific research case (John-Mathews, 2021; Lee and Baskerville, 2003; Lindberg, 2020). In order to generate compelling theory, these facets are vital and undisputed. Thus, we agree that next-generation theorizing is essentially based on new ways of thinking and cannot be replaced on the basis of technological advancements (e.g., big data, IoT) alone (Burton-Jones et al., 2021), that would offer some kind of automated theory building (Grover et al., 2020, Lindberg, 2020, Miranda et al., 2022). Still, we strongly argue in favor of using XAI supported ML to support and advance theorizing in IS research. Thereby, and in line with extant research (e.g., Choudhury et al., 2021; Shrestha et al., 2021), we propose using ML to exploratively detect complex patterns in empirical data, and subsequently using XAI, combined with human sensemaking, to enable interpretability of these patterns, which eventually lays the foundation for innovative IS theories based on this abductive technology supported approach. In this context, XAI plays a central role by allowing us to assess identified patterns for stable or divergent solutions. In qualitative research, triangulation with different data sets telling the same story or multi-coder approaches supporting intersubjective consistency in interpretation, are considered to benefit both rigor and theory building, the analogous argument holds for XAI methods (Miranda et al., 2022). While any pattern may inspire a researcher in formulating a theory, assessing pattern stability provides an additional criterion to theorizing (Grover et al., 2020, Weick, 1989), which should not be neglected. With qualitative research emphasizing plausibility, researchers are required to provide a plausible explanation for stable observed patterns. For unstable patterns, researchers should consider what the XAI methods suggest: potentially competing explanations. As such, researchers need to explain why an instable pattern inspiring their theorizing is superior in plausibility to other patterns and may additionally work with alternative hypotheses in their theorizing.

Limitations

Despite having conducted our research carefully, we have to acknowledge some limitations. First, given the divergent feature importance in the SHAP summary plots across models, extant research has argued that the resulting plots need to be considered with caution. Fernández-Loría et al. (2022) demonstrate the limited informative value of SHAP feature importance. They show that feature importance indicated by SHAP alone is not sufficient to explain model predictions. However, our focus was not on explaining a specific model's single predictions; rather we wanted to illustrate a descriptive comparison of different models, as well as the influence of individual features. Since we additionally used a second XAI method to match against our results and thereby identified comparable feature trends, we argue that the limitations related to SHAP are not problematic for our study's results or its contribution. However, notably, the PDP plots assumed that each respective feature is independent of the others (Molnar, 2020). Second, our analysis incorporated two open and widely used data sets. However, to understand how frequently the detected pattern ambiguity arises, future research should extend our analyses to numerous other data sets.

Conclusion

ML techniques, as approach to AI, offer great promises for analyzing large data sets to generate new insights, both for research, and decision-making in our data-driven world. Yet, we have to bear the limitations of these approaches in mind when using ML. Through XAI methods, we illustrate the problem that models of comparable fit can originate from highly different patterns. Given the potential ambiguity, we question the prevalent practices of reporting only the best fitting model without any or with only a few selective indicators in model assessment. By a) using XAI methods for investigating ML solutions, b) reporting alternative models of equal fit, and c) critically reflecting the quality and consistency ML-solutions provide, we contribute to a deeper understanding of ML and its applications in both, theory building and practice.

Appendices

Feature name; Description or question; Scale
GDP16; GDP per Capita PPP per Country 2016; metrical
nwspol; News about politics and current affairs, watching, reading or listening, in minutes; metrical
netustm; Internet use, how much time in a typical day, in minutes; metrical
polintrre; How interested in politics; ordinal 1-4 (1-not at all, 4-very interested)
iiop; Individual influence on politics; ordinal 1-5 (1-low, 5-high)
tios; Trust in own state; ordinal 0-10 (0-low, 10-high)
trstep; Trust in the European Parliament; ordinal 0-10 (0-no trust at all, 10-complete trust)
trstun; Trust in the United Nations; ordinal 0-10 (0-no trust at all, 10-complete trust)
vote; Voted last national election; Yes=1, No=2
peng; Political engagement; ordinal 0-8 (0-low, 8-high)
lr; Placement on left right ; ordinal 0-10 (left, right)
stflife; How satisfied with life as a whole; ordinal 0-10 (0-extremely dissatisfied, 10-extremely satisfied)

scoc; Satisfaction with the condition of the country; ordinal 0-10 (0-low, 10-high)
euftf; European Union: European unification go further or gone too far; ordinal 0-10 (0-unification already gone too far, 10-unification go further)
iocbi; Influence on country by immigrants; ordinal 0- 10 (0-negative influence, 10-positve influence)
happy; How happy are you; ordinal 0-10 (0-extremely unhappy, 10-extremely happy)
healthre; Subjective general health; ordinal 1-5 (1-very bad, 5-very good)
atchctr; How emotionally attached to [country]; ordinal 0-10 (0-not at all emotionally attached, 10-very emotional attached)
atcherp; How emotionally attached to Europe; ordinal 0-10 (0-not at all emotionally attached, 10-very emotionally attached)
rlgdgr; How religious are you; ordinal 0-10 (0-not at all religious,10-very religious)
ctzcntr; Citizen of country; 1=Yes, 2=No
wieq; Wealth equality important; ordinal (1-strongly disagree, 5-strongly agree)
sbi; Social benefits are important; ordinal (1-strongly disagree, 5-strongly agree)
sbni; Social benefits have negative impacts; ordinal (1-strongly disagree, 5-strongly agree)
eudcnbfre; More decisions made by EU: level of benefits in [country] become higher or lower; ordinal (1-much lower, 5-much higher)
fico; Financial concerns; ordinal (1-low concerns, 4-high concerns)
hhmmb; Number of people living regularly as member of household; metrical
gndr; Gender; 1=male, 2=female
agea; Age of respondent, calculated; metrical
domicil; Domicile, respondent's description; ordinal 1-5 (1-high level of urbanization, 5-low level of urbanization)
eisced; Highest level of education, ES – ISCED; ordinal 1-7 (1-less than lower secondary, 7-higher tertiary education)
iiwp; Influence in workplace; ordinal 0-10 (0-no influence, 10-complete control)
wkhtot; Total hours normally worked per week in main job overtime included; metrical
wrkac6m; Paid work in another country, period more than 6 months last 10 years; Yes=1, No=2
uemp5yr; Any period of unemployment and work seeking within last 5 years; Yes=1, No=2
hincsrca; Main source of household income; nominal (9 options)
hinctnta; Household's total net income, all sources; ordinal 1-10 (divided into deciles in ascending order of height)
values_1; Focus on individual life and improvement; ordinal 1-6 (1-low, 6-high)
values_2; Focus on others and the environment; ordinal 1-6 (1-low, 6-high)
values_3; Focus on traditional values like order and safety; ordinal 1-6 (1-low, 6-high)
vteurmmb; This represents the label - Would vote for [country] to remain member of European Union, or to leave; Leave = 0, Remain = 1

Table 3. Appendix A - Overview of the ESSD features

Feature name; Description; Scale
Age; The age of an individual; metrical
Workclass; A more general term denoting the employment status of an individual; nominal (8 options)
fnlwgt; This is the number of people the census thinks the entry represents; metrical (> 0)

Education; The highest level of education reached by an individual.; nominal (16 options)
education-num; The highest level of education reached in numerical form; ordinal (> 0)
marital-status; Marital status of an individual; nominal (7 options)
occupation; The overall nature of a person's employment; nominal (14 options)
relationship; Represents what this particular individual is in relation to others; nominal (6 options)
race; Descriptions of an individual's race; nominal (5 options)
sex; The biological sex of the person; nominal (Male, Female)
capital-gain; Capital gains for an individual; metrical
capital-loss; Capital loss for an individual; metrical
hours-per-week; The hours that an individual has reported to work per week; metrical
native country; An individual's country of origin; nominal (41 options)
Income; This represents the label, therefore, whether a person makes an income above or below $50,000 annually; 0 = <=50K, 1 = >50K

Table 4. Appendix B - Overview of the AID features

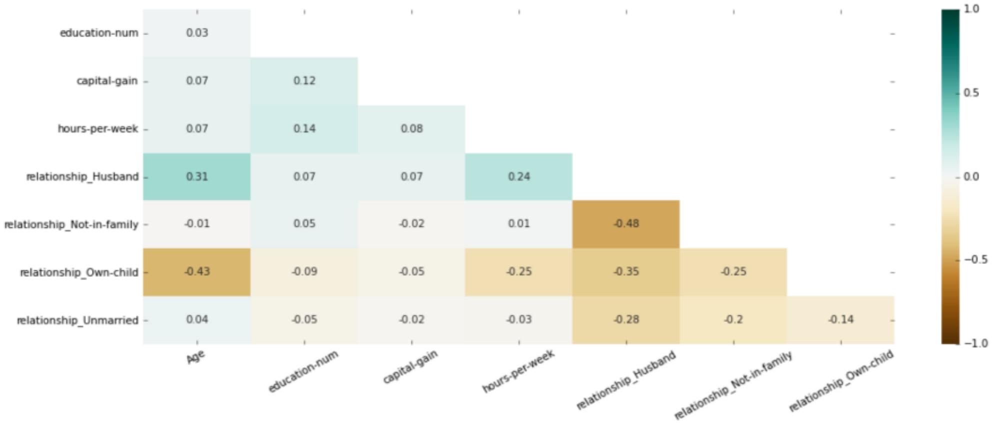


Figure 5. Appendix C - Correlation matrix for the top AID features

References

Agrawal, R., Srikant, R., and Thomas, D. 2005. "Privacy Preserving OLAP," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Baltimore, MD, pp. 251–262.

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., and Herrera, F. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82–115.

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

Bergstra, J., and Bengio, Y. 2012. "Random Search for Hyper-Parameter Optimization Yoshua Bengio," *Journal of Machine Learning Research* (13), pp. 281–305.

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

Breiman, L. 2001. "Random Forests," *Machine Learning* (45:1), pp. 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. *Classification and Regression Trees,* Belmont, CA: Wadsworth.

Brynjolfsson, E., and Mitchell, T. 2017. "What Can Machine Learning Do? Workforce Implications," *Science* (358:6370), pp. 1530–1534.

Burton-Jones, A., Butler, B., Scott, S., and Xu, S. X. 2021. "Next-Generation Information Systems Theorizing: A Call to Action," *MIS Quarterly* (45:1), pp. 301–314.

Chakraborty, A., De, R., Chatterjee, A., Schwenker, F., and Sarkar, R. 2019. "Filter Method Ensemble with Neural Networks," in *28th International Conference on Artificial Neural Networks*, Munich, Germany, pp. 755–765.

Chen, T. and Guestrin, C. 2016. "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, pp. 785-794.

Choudhury, P., Allen, R. T., and Endres, M. G. 2021. "Machine Learning for Pattern Discovery in Management Research," *Strategic Management Journal* (42:1), pp. 30–57.

Cortes, C., and Vapnik, V. 1995. "Support-Vector Networks," *Machine Learning* (20:3), pp. 273–297.

Cover, T.M. and Thomas, J.A. 1991. "Entropy, relative entropy and mutual information," *Elements of information theory* (2:1), pp. 12-13.

Dananjayan, S., and Raj, G. M. 2020. "Artificial Intelligence during a Pandemic: The COVID-19 Example," *The International Journal of Health Planning and Management* (35:5), pp. 1260–1262.

Degeest, A., Verleysen, M. and Frénay, B. 2019. "About filter criteria for feature selection in regression," in *International Work-Conference on Artificial Neural* Networks, Gran Canaria, Spain, pp. 579-590.

Dua, D., and Graff, C. 2019. "UCI Machine Learning Repository: Citation Policy," *UCI Machine Learning Repository*.

ESS Round 8. 2016. "European Social Survey Round 8 Data (2016). Data File Edition 2.2. NSD - Norwegian Centre for Research Data, Norway – Data Archive and Distributor of ESS Data for ESS ERIC."

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. 2014. "Do We Need Hundreds of Classifiers to Solve Real World Classification Problems ?," *The Journal of Machine Learning Research (15:1),* pp. 3133–3181.

Fernández-Loría, C., Provost, F., and Han, X. 2022. "Explaining Data-Driven Decisions Made by AI Systems: The Counterfactual Approach," *MIS Quarterly* (Forthcoming), pp. 1–33.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics* (29:5), pp. 1189–1232.

García, S., Luengo, J., and Herrera, F. 2015. *Data Preprocessing in Data Mining*, Cham, Switzerland: Springer.

Grover, V., Lindberg, A., Benbasat, I., and Lyytinen, K. 2020. "The perils and promises of big data research in information systems," *Journal of the Association for Information Systems* (21:2), p. 9.

Hamm, P., Wittmann, H., and 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*, Austin, TX.

Hancock, J. T., and Khoshgoftaar, T. M. 2020. "Survey on Categorical Data for Neural Networks," *Journal of Big Data* (7:1), pp. 1–41.

Hassoun, M. H. 1995. *Fundamentals of Artificial Neural Networks*, Cambridge, MA: MIT Press.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning*, New York, NY: Springer.

He, H., Bai, Y., Garcia, E. A., and Li, S. 2008. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *Proceedings of the International Joint Conference on Neural Networks*, Hong Kong, pp. 1322–1328.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. 2013. *Applied Logistic Regression*, *3rd Edition*, Chichester, UK: Wiley.

Hutter, F., Kotthoff, L., and Vanschoren, J. 2019. *Automated Machine Learning: Methods, Systems, Challenges*, Cham, Switzerland: Springer.

Jacobucci, R., and Grimm, K. J. 2020. "Machine Learning and Psychological Research: The Unexplored Effect of Measurement," *Perspectives on Psychological Science* (15:3), pp. 809–816.

John-Mathews, J.-M. 2021. "Critical Empirical Study on Black-Box Explanations in AI," in *ICIS 2021 Proceedings*, Austin, TX.

Jordan, M. I., and Mitchell, T. M. 2015. "Machine Learning: Trends, Perspectives, and Prospects," *Science* (349:6245), pp. 255–260.

Kamath, U., and Liu, J. 2021. *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*, Cham, Switzerland: Springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.Y. 2017. "Lightgbm: A highly efficient gradient boosting decision tree," in *31st Conference on Neural Information Processing Systems,* Long Beach, CA.

Kotsiantis, S. B., Kanellopoulos, D., and Pintelas, P. E. 2006. “Data Preprocessing for Supervised Leaning,” *International Journal of Computer Science* (1:2), pp. 111–117.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. 2015. “Machine Learning Applications in Cancer Prognosis and Prediction,” *Computational and Structural Biotechnology Journal* (13), pp. 8–17.

Kraskov, A., Stögbauer, H., and Grassberger, P. 2004. “Estimating Mutual Information,” *Physical Review E* (69:6), p. 066138.

Lee, A. S., and Baskerville, R. L. 2003. “Generalizing Generalizability in Information Systems Research,” *Information Systems Research* (14:3), pp. 221-243.

Lindberg, A. 2020. “Developing Theory Through Integrating Human and Machine Pattern Recognition,” *Journal of the Association for Information Systems* (21:1), pp. 90–116.

Lipton, Z. 2018. “The Mythos of Model Interpretability,” *Queue* (16:3), pp. 31–57.

Lundberg, S. M., and Lee, S.-I. 2017. “A Unified Approach to Interpreting Model Predictions,” in *31st Conference on Neural Information Processing Systems,* Long Beach, CA.

Miao, J., and Niu, L. 2016. “A Survey on Feature Selection,” *Procedia Computer Science* (91), pp. 919–926.

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

Molnar, C. 2020. *Interpretable Machine Learning*, Victoria, British Columbia, Canada: Leanpub.

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

Padmanabhan, B., Fang, X., Sahoo, N., and Burton-Jones, A. 2022. “Machine Learning in Information Systems Research,” *MIS Quarterly* (46:1).

Pentland, B., Vaast, E., and Wolf, J. R. 2021. “Theorizing process dynamics with directed graphs: A diachronic analysis of digital trace data,” *MIS Quarterly* (45:2), pp. 967-984.

Rácz, A., Bajusz, D., and Héberger, K. 2021. “Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification,” *Molecules* (26:4), pp. 1-16.

Rizk, A., and Elragal, A. 2020. “Data science: developing theoretical contributions in information systems via text analytics,” *Journal of Big Data* (7:1), pp. 1-26.

Ross, B. C. 2014. “Mutual Information between Discrete and Continuous Data Sets,” *PLoS ONE* (9:2).

Rosset, S. 2004. “Model Selection via the AUC,” in *Proceedings of the 21st international conference on machine learning*, New York, NY.

Senoner, J., Netland, T., and Feuerriegel, S. 2021a. “Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing,” *Management Science* (forthcoming).

Shannon, C. E. 1948. “A mathematical theory of communication,” *The Bell System Technical Journal* (27:3), pp. 379-423.

Shin, D., He, S., Lee, G. M., and Whinston, A. 2020. “Enhancing Social Media Analysis with Visual Data Analytics: A Deep Learning Approach,” *MIS Quarterly* (44:4), pp. 1459–1492.

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

Sturm, T., Gerlach, J., Pumplun, L., Mesbach, N., Peters, F., Tauchert, C., Nan, N., and Buxmann, P. 2021. “Coordinating Human and Machine Learning for Effective Organizational Learning,” *MIS Quarterly* (45:3), pp. 1581–1602.

Tidhar, R., and Eisenhardt, K. M. 2020. “Get Rich or Die Trying... Finding Revenue Model Fit Using Machine Learning and Multiple Cases,” *Strategic Management Journal* (41:7), pp. 1245–1273.

Tonidandel, S., King, E. B., and Cortina, J. M. 2016. “Big Data Methods: Leveraging Modern Data Analytic Techniques to Build Organizational Science,” *Organizational Research Methods* (21:3), pp. 525–547.

Weick, K. E. 1989. “Theory construction as disciplined imagination,” *Academy of Management Review* (14:4), pp. 516-531.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., and Steinberg, D. 2007. “Top 10 Algorithms in Data Mining,” *Knowledge and Information Systems* (14:1), pp. 1–37.

Yuan, S., Kroon, B., and Kramer, A. 2021. “Building Prediction Models with Grouped Data: A Case Study on the Prediction of Turnover Intention,” *Human Resource Management Journal*, pp. 1-19.

Zhao, Q., and Hastie, T. 2019. “Casual Interpretations of Black-Box Models,” *Journal of Business & Economic Statistics: A Publication of the American Statistical Association* (39:1), pp. 272–281.