Received: 28 November 2020 | Revised: 21 April 2021 | Accepted: 28 May 2021

DOI: 10.1002/qre.2939

RESEARCH ARTICLE

WILEY

Explainable AI methods in cyber risk management

Paolo Giudici | **Emanuela Raffinetti**

Department of Economics and Management, University of Pavia, Via San Felice, Pavia, Italy

Correspondence Paolo Giudici, Department of Economics and Management, University of Pavia, Via San Felice 5, Pavia, 27100, Italy.
Email: paolo.giudici@unipv.it

Funding information
European Union's Horizon 2020 training and innovation programme "FIN-TECH", Grant/Award Number: 825215

Abstract

Artificial intelligence (AI) methods are becoming widespread, especially when data are not sufficient to build classical statistical models, as is the case for cyber risk management. However, when applied to regulated industries, such as energy, finance, and health, AI methods lack explainability. Authorities aimed at validating machine learning models in regulated fields will not consider black-box models, unless they are supplemented with further methods that explain why certain predictions have been obtained, and which are the variables that mostly concur to such predictions. Recently, Shapley values have been introduced for this purpose: They are model agnostic, and powerful, but are not normalized and, therefore, cannot become a standardized procedure. In this paper, we provide an explainable AI model that embeds Shapley values with a statistical normalization, based on Lorenz Zonoids, particularly suited for ordinal measurement variables that can be obtained to assess cyber risk.

KEYWORDS
cyber risk management, Lorenz Zonoids, rank regression, Shapley values

1 | INTRODUCTION

Cyber risks can be defined as "any risk emerging from intentional attacks on information and communication technology (ICT) systems that compromises the confidentiality, availability, or the integrity of data or services" (see, e.g., Refs. 1–3). Note that, according to this definition, cyber risk does not strictly coincide with information technology (IT) operational risks, as it relates only to intentional attacks, on one hand, and it deals not only with monetary losses, but also with reputational losses, on the other.

In the last few years the number of cyber attacks on IT systems has surged: 1127 attacks occurred in 2017, against 1050 in 2016, 1012 in 2015, and 873 in 2014, with a growth of about 30% between 2014 and 2017. The trend in 2018 follows a similar behavior, with 730 cyber attacks observed only in the first half of the year.[4] Thus, the need to measure cyber risks has considerably increased.

While the scientific literature on the measurement of operational risks (see, e.g., Refs. 5, 6), based on loss data, constitutes a reasonably large body, that on cyber risk measurement is very limited. Some contributions can be found in Ruan,[7] Radanliev et al.,[8] and Shin et al.,[9] in which the focus is on the measurement of the value at risk, the maximum possible loss due to the occurrence of cyber attacks. The lack of literature on cyber risk measurement may be due to the limited availability of cyber loss data, which are typically not disclosed, to avoid reputational losses. When disclosed, they are often expressed in terms of ordered levels of severity, such as "low," "medium," or "high" severity. Unfortunately,

the ordinal classification of risks prevents the calculation of the value at risk. Although ordinal data cannot be used to calculate the value at risk, they can be used to rank risks by their "criticality," so as to prioritize interventions and, therefore, trigger mitigating actions. To our knowledge, there are very few papers that suggest how to deal with ordinal cyber data. Exceptions, that are however limited to specific issues, are Afful-Dadzie and Allen,[10] who focus on the problem of the scarcity of available data, and Hubbard and Evans,[11] Sexton et al.,[12] Hubbard and Seiersen,[13] and Facchinetti et al. (2019),[14] who introduce descriptive scoring methods.

We propose to fill this gap in the literature, providing an explainable machine learning model aimed at accurately predicting the ordinal severity levels of cyber risks. To achieve this goal, we develop a methodology that combines rank-based regression models with a rank-based Shapley value approach. We test our model on a real data set of cyber events, ordered by severity levels. The application shows that the proposed methodology is both accurate, from a predictive viewpoint, and interpretable, from an explainable viewpoint. In addition, the usage of Lorenz Zonoids to assess model performance allows to obtain results more robust with respect to data quality issues, which may lead to outlying observations. The paper is organized as follows: the next section contains our proposal; Section 3 contains the empirical findings obtained applying our model to real cyber data; finally Section 4 contains some concluding remarks.

2 | METHODOLOGY

Our proposal derives from the combination of two research streams. The first one concerns the development of models to analyze ordinal data arising in the cyber risk setting. The second one concerns the development of explainable methods to understand the results of advanced learning models. The result of the combination is a novel method for cyber risk management, which is, at the same time, predicitively accurate, interpretable, and robust.

2.1 | Rank regression models in cyber risk management

As the cyber events are typically rare and not repeatable, it is quite natural to measure them with a less demanding ordinal approach rather than using quantitative data, which are often not available. Ordinal data for cyber risk measurement can be summarized, by means of a pair of statistics for each event type: the frequency of the event, how many times it has occurred, in a given period; and the corresponding severity, the mean observed loss. In the context of ordinal data, the severity can be expressed on an ordinal scale, characterized by k distinct levels, arranged according to the corresponding magnitude. To understand the main factors impacting on cyber risks, each observed severity can be associated to a vector of explanatory variables, such as the type of attack, the technique of the attack, the victim type, and the geographical area where the event has occurred.

The statistical models typically used to explain an ordinal response variable with a set of p explanatory variables are the ordered logit or probit models (see, for instance Refs. 15 and 16). These, however, may be difficult to summarize and interpret, especially in applied contexts. We therefore resort to a linear regression model for a response variable that takes ordinal values and, in order to avoid an arbitrary assignment of the measurement scale, we resort to ranks.

Let Y be a response variable, expressed through k ordered categories. A rank $r_1 = 1$ to the smallest ordered category of Y and a rank $(r_{j-1} + n_{j-1})$ to the following ordered categories, where n_{j-1} is the absolute frequency associated with the $(j-1)$-th category and $j = 2, \dots, k$, are assigned. Based on this transformation, the phenomenon described by the Y variable can be reformulated in terms of its ranks R, where:

$$R = \left\{ \underbrace{r_1, \dots, r_1}_{n_1}, \underbrace{r_2, \dots, r_2}_{n_2}, \dots, \underbrace{r_k, \dots, r_k}_{n_k} \right\}, \tag{1}$$

with $r_1 = 1, r_2 = r_1 + n_1$ and $r_k = r_{k-1} + n_{k-1}$.

Given p explanatory variables $(X_1, \dots, X_p)$, a regression model for R can be specified as follows:

$$\hat{R} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p, \tag{2}$$

whose unknown parameters can be estimated by the classical ordinary least squares (OLS) method.

2.2 | The Shapley–Lorenz decomposition in cyber risk management

When dealing with data coming from highly regulated fields, such as energy, finance, and health, we may resort to simple or complex machine learning models. Simple machine learning models, including linear or logistic regression models, are highly interpretable but provide a limited predictive accuracy. Complex machine learning models, such as neural network models and decision tree models, fulfill the requirement of high predictive accuracy at the expense of interpretability. In order to meet both the conditions of predictive accuracy and interpretability, the idea is basically to boost accurate machine learning models with novel methodologies able to explain the predictive output. Recently, Giudici and Raffinetti17 have proposed a global explainable artificial intelligence (AI) model, named Shapley–Lorenz decomposition, which combines the interpretability power of the local Shapley value game theoretic approach (see, e.g., Ref. 18) with a more robust global approach based on the Lorenz Zonoid model accuracy tool (see, e.g., Ref. 19). The Lorenz Zonoids can be seen as a generalization of the receiver operating characteristic (ROC) curve in a multidimensional setting and, therefore, the Shapley–Lorenz decomposition has the advantage of combining predictive accuracy and explainability performance into one single diagnostics. Furthermore, the Lorenz Zonoid is intended as a measure of the mutual variability, robust to the presence of outlying observations, and can be exploited to develop partial dependence measures that allow to detect the additional contribution of a new predictor into an existing model.

Shapley values were introduced as a pay-off concept from cooperative game theory. When referring to machine learning models, the notion of pay-off corresponds to the model prediction. Thus, for any single statistical unit i $(1 = 1, \dots, n)$, the pay-offs are defined as

$$p_{off}(X_i^k) = \hat{f}(X' \cup X_k)_i - \hat{f}(X')_i, \tag{3}$$

where $\hat{f}(X')_i$ denotes the predicted values generated by the machine learning models depending only on X' predictors; $\hat{f}(X' \cup X_k)_i$ denotes the predicted values generated by the machine learning models depending both on the $|X'|$ predictors and the additional included X_k predictor. For a set of statistical units $(i = 1, \dots, n)$, the pay-off notion translated in terms of Lorenz Zonoids $(LZ(\cdot))$ is given by

$$p_{off}(X^k) = LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'}), \tag{4}$$

where $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$ are the vectors specifying the predicted values generated by the machine learning models, which include the additional explanatory variable X_k, and the predicted values generated by the machine learning models, which do not include the explanatory variable X_k, whereas $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ describe the (mutual) variability of the response variable Y explained by the models including the $X' \cup X_k$ predictors and the X' predictors, respectively.

The Shapley–Lorenz decomposition expression is the result of a combination between the Shapley value–based formula and the Lorenz Zonoid tools. Formally, the contribution of the additional variable X^k, expressed in terms of the differential contribution to the global predictive accuracy, equals to

$$LZ^{X_k}(\hat{Y}) = \sum_{X' \subseteq C(X) \setminus X_K} \frac{|X'|!(K - |X'| - 1)!}{K!} [LZ(\hat{Y}_{X' \cup X_k}) - LZ(\hat{Y}_{X'})], \tag{5}$$

where $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ measure the marginal contribution provided by the inclusion of variable X_k; K is the number of available predictors; $C(X) \setminus X_k$ is the set of all the possible model configurations that can be obtained with $K - 1$ variables, excluding variable X_k; $|X'|$ denotes the number of variables included in each possible model.

We remark that $LZ(\hat{Y}_{X' \cup X_k})$ and $LZ(\hat{Y}_{X'})$ in Equation (5) can be expressed as function of the covariance operators, that is,

$$LZ(\hat{Y}_{X' \cup X_k}) = \frac{2}{nE(\hat{Y}_{X' \cup X_k})} Cov(\hat{Y}_{X' \cup X_k}, r(\hat{Y}_{X' \cup X_k})) \quad \text{and}$$

$$LZ(\hat{Y}_{X'}) = \frac{2}{nE(\hat{Y}_{X'})} Cov(\hat{Y}_{X'}, r(\hat{Y}_{X'})),$$

TABLE 1 Frequency distribution of the cyber loss severity

Severity	Frequency
1	176
2	243
3	389

where $E(\hat{Y}_{X' \cup X_k})$ and $E(\hat{Y}_{X'})$ are the expected values of $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$, respectively; $r(\hat{Y}_{X' \cup X_k})$ and $r(\hat{Y}_{X'})$ are the rank scores associated with the $\hat{Y}_{X' \cup X_k}$ values and the $\hat{Y}_{X'}$ values.

Due to its building characteristics, the Shapley–Lorenz decomposition presents as an agnostic eXplainable AI method, which can be applied to the predictive output, regardless of which model generated it. This feature makes it suitable in all the contexts where response variables with different nature are involved. The focus on cyber risk data represents an example of application in the presence of an ordinal target variable. We remark that in such a case the response variable is transformed into ranks according to Equation (1).

3 | APPLICATION

The purpose of this application is to evaluate the performance of our cyber risk measurement proposal, based on the combination between rank regression models and the Shapley–Lorenz decomposition approach.

We employ the Clusit cyber loss database,[4] which consists of 6865 worldwide observations on serious cyber attacks, in the years 2011–2017. An attack is classified as "serious" if it has led to a significant impact, in terms of economic losses and/or damages to reputation. In this paper, we focus on a sample data, consisting of 808 cyber attacks observed in 2017, the year in which most data were observed. Severity levels are reported according to the type of attacker, technique of attacks, victims, and the corresponding continent of origin. Moreover, given the data at hand, we evaluate the model on the full sample, without splitting it into training and test sets.

We remark that the data, similarly to all cyber database available, may contain outlying observations, which can derive either from intrinsic characteristics or from measurement errors. In both cases it is necessary that the developed machine learning models are robust to outliers and data variations. This aspect suggests the use of rank-based regression models, on one hand, and of model assessment performances, which are similarly robust to data anomalies.

In terms of descriptive statistics, Tables 1 and 2 report the frequency distribution of the cyber loss severity, and of the considered four explanatory variables.

Our purpose is to detect the factors, among attacker, attack technique, victim type, and location (continent), which most affect the severity levels. To achieve this aim we first have applied our proposed rank regression model. From a descriptive viewpoint, the R^2 is equal to 0.6183, and the p-value of the associated F-test is smaller than 0.001. In Table 3 the estimated linear coefficients for the ordinal levels that correspond to attacker type (first table), continent (second table), victim type (third table), attack technique (fourth table), are presented. We break each of the four categorical variables into dummies, with the baseline cases being "Cybercrime" for type of attacker, "Africa" for continent, "Automotive" for victim, and "0-day" for attack technique. Together with the linear regression coefficients, the related p-values are also provided, showing that the geographical area, where the cyber attack occurs, has not a significant impact on its severity degree. Thus, the continent variable can be removed from the full model in favor of a more parsimonious model. In addition, the only significant effects at a significance level $\alpha = 5\%$ are: espionage/sabotage, hacktivism and information warfare for the type of attacker variable; entertainment/news, GDO/retail, online services/cloud, and research-education for the victim-type variable; phishing/social engineering and unknown for the attack technique variable.

In general, by looking at the estimated linear regression coefficients, the different cyber attack levels have the effect of decreasing the severity degree with respect to the baseline of "Cybercrime." On the contrary, the levels characterizing the attack technique and the victim type have the effect of increasing the severity degree with respect to the baselines of "0-day" and "Automative," respectively.

Although the rank regression model appears explainable by definition, it is the results of a model selection procedure whose obtained coefficients are conditional on the single chosen model. Differently, the Shapley value approach provides a measure of explainability for each single feature variable, which is based on the consideration of all possible model

TABLE 2 Frequency distributions of the explanatory variables

Continent	**Frequency**
Africa	7
America	482
Asia	112
Europe	186
Oceania	21
Type of attacker	**Frequency**
Cybercrime	600
Espionage/sabotage	82
Hacktivism	74
Information warfare	52
Victim	**Frequency**
Automotive	4
Banking/finance	65
Critical infrastructures	27
Entertainment/news	108
GDO/retail	21
Gov-Mil-LE-intelligence	159
Gov. contractors/consulting	6
Health	79
Hospitability	34
Multiple targets	71
Online services/cloud	58
Organization-ONG	6
Research-education	70
Security	10
SW/HW vendor	43
Telco	13
Others	34
Attack technique	**Frequency**
0-day	5
Account cracking	50
DDoS	33
malware	1
Malware	234
Multiple threats/APT	45
Phishing/social engineering	76
Phone hacking	2
SQLi	4
Vulnerabilities	97
Unknown	261

configurations. Finally, the Shapley–Lorenz approach further improves the measure of explainability providing a version that is normalized. We remark that the linear model coefficients can also be normalized; however, they remain conditional on the chosen model, differently from the Shapley–Lorenz measure.

We then have calculated the Shapley–Lorenz marginal contributions associated with the variables attacker, victim type, attack technique, and continent, using formula (5). When considering type of attacker (Att), victim type (Vic), attack

TABLE 3 Categorical variable reference level: cyber attack (first table): cyber crime; continent (second table): Africa; victim type (third table): automotive; attack technique (fourth table): 0-day. The intercept estimate with the related *p*-value was included in the first table

Coefficient	Estimate	*p*-value
Intercept	187.42	0.02678
Espionage/sabotage	−231.38	<0.001
Hacktivism	−39.21	0.00663
Information warfare	−222.17	<0.001
America	−11.22	0.78849
Asia	−10.20	0.81132
Europe	−18.39	0.66228
Oceania	−30.78	0.52204
Banking/finance	−21.35	0.70239
Critical infrastructures	53.35	0.36075
Entertainment/news	117.14	0.03345
GDO/retail	139.97	0.01743
Gov-Mil-LE-intelligence	−48.62	0.37632
Gov. contractors/consulting	−37.89	0.58873
Health	55.82	0.31249
Hospitability	60.33	0.28946
Multiple targets	105.20	0.06011
Online services/cloud	136.11	0.01496
Organization-ONG	66.87	0.33780
Others	60.65	0.28748
Research-education	142.26	0.01057
Security	93.28	0.14461
SW HW vendor	90.67	0.10936
Telco	73.37	0.23653
Account cracking	74.92	0.14680
DDoS	48.40	0.35773
malware	153.13	0.20044
Malware	14.19	0.77408
Multiple threats/APT	50.04	0.32898
Phishing/social engineering	120.27	0.01763
Phone hacking	103.78	0.25215
SQLi	−29.93	0.68253
Unknown	99.67	0.04516
Vulnerabilities	53.08	0.29151

technique (Tec), and continent (Con) as additional predictors, the related marginal contributions can be computed as:

$$\begin{aligned} LZ^{Att}(\widehat{Severity}) = {} & (1/4)(LZ(\hat{R}_{Att,Vic,Tec,Con}) - LZ(\hat{R}_{Vic,Tec,Con})) \\ & + (1/12)(LZ(\hat{R}_{Att,Vic,Tec}) - LZ(\hat{R}_{Vic,Tec})) \\ & + (1/12)(LZ(\hat{R}_{Att,Vic,Con}) - LZ(\hat{R}_{Vic,Con})) + (1/12)(LZ(\hat{R}_{Att,Tec,Con}) - LZ(\hat{R}_{Tec,Con})) \\ & + (1/12)(LZ(\hat{R}_{Att,Vic}) - LZ(\hat{R}_{Vic})) + (1/12)(LZ(\hat{R}_{Att,Tec}) - LZ(\hat{R}_{Tec})) \\ & + (1/12)(LZ(\hat{R}_{Att,Con}) - LZ(\hat{R}_{Con})) + (1/4)(LZ(\hat{R}_{Att})), \end{aligned}$$

TABLE 4 Marginal contribution of each explanatory variable in terms of the Shapley–Lorenz–based approach and comparison with the standard Shapley based approach

Additional covariate (X_k)	$LZ^{X_k}(\widehat{Severity})$	Global Shapley
Type of attacker	0.072	748.96
Type of victim	0.115	27.27
Technique of attack	0.058	35.06
Continent	0.032	25.67

$$\begin{aligned}LZ^{Vic}(\widehat{Severity}) = {} & (1/4)(LZ(\hat{R}_{Att,Vic,Tec,Con}) - LZ(\hat{R}_{Att,Tec,Con})) \\ & + (1/12)(LZ(\hat{R}_{Att,Vic,Tec}) - LZ(\hat{R}_{Att,Tec})) \\ & + (1/12)(LZ(\hat{R}_{Att,Vic,Con}) - LZ(\hat{R}_{Att,Con})) + (1/12)(LZ(\hat{R}_{Vic,Tec,Con}) - LZ(\hat{R}_{Tec,Con})) \\ & + (1/12)(LZ(\hat{R}_{Att,Vic}) - LZ(\hat{R}_{Att})) + (1/12)(LZ(\hat{R}_{Vic,Tec}) - LZ(\hat{R}_{Tec})) \\ & + (1/12)(LZ(\hat{R}_{Vic,Con}) - LZ(\hat{R}_{Con})) + (1/4)(LZ(\hat{R}_{Vic})),\end{aligned}$$

$$\begin{aligned}LZ^{Tec}(\widehat{Severity}) = {} & (1/4)(LZ(\hat{R}_{Att,Vic,Tec,Con}) - LZ(\hat{R}_{Att,Vic,Con})) \\ & + (1/12)(LZ(\hat{R}_{Att,Vic,Tec}) - LZ(\hat{R}_{Att,Vic})) \\ & + (1/12)(LZ(\hat{R}_{Att,Tec,Con}) - LZ(\hat{R}_{Att,Con})) + (1/12)(LZ(\hat{R}_{Vic,Tec,Con}) - LZ(\hat{R}_{Vic,Con})) \\ & + (1/12)(LZ(\hat{R}_{Att,Tec}) - LZ(\hat{R}_{Att})) + (1/12)(LZ(\hat{R}_{Vic,Tec}) - LZ(\hat{R}_{Tec})) \\ & + (1/12)(LZ(\hat{R}_{Tec,Con}) - LZ(\hat{R}_{Con})) + (1/4)(LZ(\hat{R}_{Tec})),\end{aligned}$$

$$\begin{aligned}LZ^{Con}(\widehat{Severity}) = {} & (1/4)(LZ(\hat{R}_{Att,Vic,Tec,Con}) - LZ(\hat{R}_{Att,Vic,Tec})) \\ & + (1/12)(LZ(\hat{R}_{Att,Vic,Con}) - LZ(\hat{R}_{Att,Vic})) \\ & + (1/12)(LZ(\hat{R}_{Att,Tec,Con}) - LZ(\hat{R}_{Att,Tec})) + (1/12)(LZ(\hat{R}_{Vic,Tec,Con}) - LZ(\hat{R}_{Vic,Tec})) \\ & + (1/12)(LZ(\hat{R}_{Att,Con}) - LZ(\hat{R}_{Att})) + (1/12)(LZ(\hat{R}_{Vic,Con}) - LZ(\hat{R}_{Vic})) \\ & + (1/12)(LZ(\hat{R}_{Tec,Con}) - LZ(\hat{R}_{Tec})) + (1/4)(LZ(\hat{R}_{Con})).\end{aligned}$$

On the other hand, the local Shapley values can be computed in accordance with Equation (5), by replacing the Lorenz Zonoids included in the square brackets with the pay-off in Equation (3). Note that the latter is based on a euclidean distance between predicted values under different models, differently from the Shapley–Lorenz values, based on the Gini distance, more suited to deal with ordinal variables, and more robust to outliers. We further remark that, if on the one hand, the Shapley–Lorenz values are normalized by construction (for each single feature variable) and do not require any further aggregation, on the other hand, the Shapley values are not normalized by construction but can provide a global measure of importance of each feature by summing the deviation of each variable's Shapley value from the overall mean. To avoid compensation between positive and negative deviations, the sum of the absolute Shapley values can be considered.

The results from the application of the Shapley–Lorenz values to the available data, and the comparison with the corresponding global Shapley values (specified as the sum of the absolute Shapley values) are reported in Table 4.

From Table, 4 note that, according to the Shapley–Lorenz values, the variable describing the type of victim provides the highest marginal contribution in the prediction of cyber severity, across all the possible model configurations. A further impacting variable is associated with the type of attacker, while variables with the lowest contributions are those representing the attack technique and the continent, which is intended as the geographical area where the event has occurred.

More precisely, the continent variable gives the minimum contribution to the explanation of the severity degree associated with the cyber attacks, confirming the findings derived from the application of our proposed rank regression model, according to which the continent variable is not significant.

In the Shapley–Lorenz approach perspective, the type of victim and attacker variables explain the 11.5% and the 7.2% of the mutual variability associated with the cyber attack severity degree over all the possible model configurations, respectively. The type of technique variable explains 5.8% and the continent variable only 3.2%. From an interpretational point of view, this indicates that preventive actions and mitigation measures, such as insurance coverage, should vary according to the type of victim (in our case economic activity types) rather than on the attack technique and/or on the location of the victim.

From Table 4 also note that the Shapley values, being not normalized, are more difficult to be interpreted in terms of the variables' contributions. The variable impacting more on the severity degree of the cyber attack is the type of attack, while the variable with the least effect, coherently with which emerges from the Shapley–Lorenz–based approach, is the continent variable. In addition, the technique of attack seems to explain more than the victim type, contrary to what happens when the Shapley–Lorenz values are considered. These discrepancies may be motivated by the Shapley value construction, which, as previously discussed, involves the sum of the deviation of each variable's Shapley value from the overall mean, and consequently is less robust to the presence of outlying observations. As our experiments show, this issue can be appropriately overcome by the implementation of the Shapley–Lorenz approach.

4 | CONCLUDING REMARKS

The paper proposes a new methodology to assess cyber risks, using loss data at an ordinal scale, easier to acquire with respect to continuous data.

Consistently with the ordinal nature of the data, the proposed methodology is based on a combination between rank regression model fit and Lorenz-based assessment models.

The combination of the two approaches leads to the identification of the drivers of cyber risk, which are more important to control and mitigate with insurance.

The application of the proposed method to the available data confirms that the proposed method is quite satisfactory, and provides an accurately predictive, explainable, and robust machine learning method for cyber risk management.

ORCID

Paolo Giudici https://orcid.org/0000-0002-4198-0127

NOTE

[1] It is worth noting that the rank scores $r(\hat{Y}_{X' \cup X_k})$ and $r(\hat{Y}_{X'})$ are not connected with the ranks appearing in Equation (1). Indeed, through Equation (1) the ordinal target variable is transformed into a discrete quantitative variable through the employment of ranks, which, contrary to the rank scores $r(\hat{Y}_{X' \cup X_k})$ and $r(\hat{Y}_{X'})$, which denote the positions of the $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$ values, are computed according to the procedure suggested in Section 2.1.

ACKNOWLEDGMENTS

The authors thank the European Network for Business and Industrial Statistics (ENBIS) for useful suggestions and comments during the webinars organized with the Special Interest Group on Risk Management, coordinated by the first author. The work of the authors is receiving support from the European Union's Horizon 2020 training and innovation programme "FIN-TECH," under the grant agreement number 825215 (Topic ICT-35-2018, Type of actions: CSA). The paper is the result of the joint collaboration between the two authors.

REFERENCES

1. Cebula JJ, Young LR. A taxonomy of operational cyber security risks. Technical Note, CMU/SEI-2010-TN-028, Software Engineering Institute, Carnegie Mellon University, 1-34 (2010)
2. Edgar TW, Manz DO. *Research Methods for Cyber Security*. Cambridge, MA: Elsevier; 2017.
3. Kopp E, Kaffenberger L, Wilson C. Cyber risk, market failures, and financial stability. IMF Working Paper, WP/17/185, 1-35 (2017)
4. Clusit 2018. Report on ICT Security in Italy. AIST (Italian Association for Information Security): Milan; (2018).
5. Cox LA Jr. Evaluating and improving risk formulas for allocating limited budgets to expensive risk-reduction opportunities. *Risk Anal*. 2012;32(7):1244-1252.
6. MacKenzie CA. Summarizing risk using risk measures and risk indices. *Risk Anal*. 2014;34(12):2143-2162.

7. Ruan K. Introducing cybernomics: a unifying economic framework for measuring cyber risk. *Comput Secur*. 2017;65:77-89.
8. Radanliev P, De Roure DC, Nicolescu R, et al.. Future developments in cyber risk assessment for the internet of things. *Comput Ind*. 2018;102:14-22.
9. Shin J, Son H, Heo G. Development of a cyber security risk model using Bayesian networks. *Reliab Eng Syst Saf*. 2015;134:208-217.
10. Afful-Dadzie A, Allen TT. Data-driven cyber-vulnerability maintenance policies. *J Qual Technol*. 2017;46(3):234-250.
11. Hubbard DW, Evans D. Problems with scoring methods and ordinal scales in risk assessment. *J Res Dev*. 2010;54(3):2-10.
12. Sexton J, Storlie C, Neil J. Attack chain detection, statistical analysis and data mining. *Stat Anal Data Min*. 2015;8:353-363.
13. Hubbard DW, Seiersen R. *How to Measure Anything in Cybersecurity Risk*. New York, NY: Wiley; 2016.
14. Facchinetti S, Giudici P, Osmetti SA. Cyber risk measurement with ordinal data. *Stat Methods Appl*. 2020;29:173-185.
15. McCullagh P. Regression models for ordinal data. *J Roy Statist Soc Ser B (Methodological)*. 1980;42(2):109-142.
16. Agresti A. *Analysis of Ordinal Categorical Data*, 9th ed. New York: Wiley; 2010.
17. Giudici P, Raffinetti E. Shapley-Lorenz eXplainable artificial intelligence. *Expert Syst Appl*. 2021;167.
18. Shapley LS. A value for *n*-person games. In: Kuhn H and Tucker A, eds. *Contributions to the Theory of Games*. Princeton, NJ: Princeton University Press; 1953:307-317.
19. Giudici P, Raffinetti E. Lorenz model selection. *J Classification*. 2020;37:754-768.

AUTHOR BIOGRAPHIES

Paolo Giudici is professor of statistics at the Department of Economics and Management of the University of Pavia. Giudici has been academic supervisor of about 200 master's students and 20 PhD students, currently working in the financial industry, in IT/consulting companies or as an academic researcher. The author of several scientific publications (97 in Scopus, with 1460 total citations and an h-index of 22). The main contributions to research are in the following: multivariate graphical models, network models for financial stability, correlation networks for financial technologies, operational risk management models, Bayesian data analysis, and model diagnostics, predictive accuracy, and explain ability. Giudici is coordinator of 12 funded scientific projects, among which the European Horizon2020 projects "PERISCOPE: Pan-European response to the impacts of covid-19 and future pandemics and epidemics (2020-2023)," the European Horizon 2020 project "FIN-TECH: Financial supervision and Technological compliance" (2019–2020), and the European VI programme project on "Multi industry semantic based business intelligence" (2006–2010). Giudici is chief editor of *Artificial Intelligence in Finance*, Frontiers, and associate editor of *Digital Finance*, Springer, and of *Risks*, MDPI; and is a researcher with CEPS on the monitoring of COVID-19 contagion, research fellow at the Bank for International Settlements, Basel, research fellow at the University College London center for Blockchain technologies, expert at the European Insurance and Occupational Pensions Authority (EIOPA), expert at the Italian ministry of development for the National AI strategy, member of the scientific committee of the Association of Italian Financial Risk Managers, AssoFintech, the Cryptovalues association, and is member of the following: Italian Statistical Society(SIS), Italian Econometric Society (SIDE), European Big Data Value Association (BDVA), European Network for Business and Industrial Statistics (ENBIS), and International Society for Bayesian Analysis (ISBA). Giudici is a principal investigator of research, training, and consulting projects for the following: the Italian Banking Association, Intesa SanPaolo, Unicredit, UBI, BancoBpm, MPS, BPS, Creval, Accenture, KPMG, SAS, Mediaset, and Sky.

Emanuela Raffinetti is currently research fellow in statistics at the Department of Economics and Management of the University of Pavia. She is associate editor of the *Frontiers in Artificial Intelligence* journal in the section of Artificial Intelligence in Finance. Her research activity is mainly focused on Explainable Artificial Intelligence (XAI) methods; Machine Learning model validation methods; assessment of operational and cyber risks; dependence analysis; dependence, concordance, and discordance measures; inferential issues when dealing with data set of high dimensions; subsampling methods; inequality measures in income distributions; assessment of quality and customer satisfaction; assessment of the university and educational systems.

How to cite this article: Giudici P, Raffinetti E. Explainable AI methods in cyber risk management. *Qual Reliab Eng Int*. 2021;1–9. https://doi.org/10.1002/qre.2939