Expert Systems With Applications 235 (2024) 121220

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Artificial Intelligence risk measurement

Paolo Giudici [a,*], Mattia Centurelli [b], Stefano Turchetta [c]

[a] *University of Pavia, Italy*
[b] *Credito Emiliano, Italy*
[c] *Independent, Italy*

ARTICLE INFO

Keywords:
Machine learning
Sustainability
Accuracy
Fairness
Explainability
Financial risk management

ABSTRACT

Financial institutions are increasingly leveraging on advanced technologies, facilitated by the availability of Machine Learning methods that are being integrated into several applications, such as credit scoring, anomaly detection, internal controls and regulatory compliance. Despite their high predictive accuracy, Machine Learning models may not provide sufficient explainability, robustness and/or fairness; therefore, they may not be trustworthy for the involved stakeholders, such as business users, auditors, regulators and end-customers.

To measure the trustworthiness of AI applications, we propose the first Key AI Risk Indicators (KAIRI) framework for AI systems, considering financial services as a reference industry. To this aim, we map the recently proposed regulatory requirements proposed for Artificial Intelligence Act into a set of four measurable principles (Sustainability, Accuracy, Fairness, Explainability) and, for each of them, we propose a set of interrelated statistical metrics that can be employed to measure, manage and mitigate the risks that arise from artificial intelligence.

We apply the proposed framework to a collection of case studies, that have been indicated as highly relevant by the European financial institutions we interviewed during our research activities. The results from data analysis indicate that the proposed framework can be employed to effectively measure AI risks, thereby promoting a safe and trustworthy AI in finance.

1. Background and contribution

The research in this paper is motivated by the current widespread use of Artificial Intelligence (AI), which requires to develop advanced statistical methods that can measure its risks, in line with the recently proposed regulations, such as the European Artificial Intelligence Act (EU, 2022) and the American Artificial Intelligence Risk Management framework (United States National Institute of Standards and Technologies, 2022).

Indeed, machine learning models are boosting Artificial Intelligence applications in many domains, such as finance, health care and manufacturing. This is mainly due to their advantage, in terms of predictive accuracy, with respect to “classic” statistical learning models. However, although machine learning models could reach high predictive performance, they have an intrinsic non transparent (“black-box”) nature. This is a problem in regulated industries, as authorities aimed at monitoring the risks arising from the application of AI may not validate them (see e.g. Bracke et al. (2019)).

Accuracy and explainability are not the only desirable characteristics of a machine learning model. The recently proposed regulations introduce further requirements, such as robustness, cybersecurity, fairness and sustainability, within a risk-based approach to AI applications. In the regulatory context, several applications of AI have high risks and, therefore, require an appropriate risk management model. To develop such a model, we need to express the regulatory requirements in terms of statistical variables, to be measured by appropriate statistical metrics.

The main contribution of our paper is to provide an integrated risk management model for artificial intelligence applications. Recently, several papers have investigated the application AI to measure specific risks. Among them, Naim (2022) and Sundra et al. (2023) consider the application of AI in financial risk management; Bücker et al. (2022), Bussmann et al. (2020), Liu et al. (2022), Moscato et al. (2021) and Sachan et al. (2020) employ AI to measure credit risk; Giudici and Abu-Hashish (2019) and Giudici and Polinesi (2021) employ AI to measure contagion risks in algorithmic trading and crypto exchanges; Ganesh and Kalpana (2022) reviews AI methodologies for supply chain risk management, whereas Frost et al. (2019) and Kuiper et al. (2021) do so for financial intermediation; Aldasoro et al. (2022), Giudici and Raffinetti (2021), McCall et al. (2017) and Melancon et al. (2021)

[*] Corresponding author.
E-mail address: giudici@unipv.it (P. Giudici).

https://doi.org/10.1016/j.eswa.2023.121220
Received 2 May 2023; Received in revised form 20 July 2023; Accepted 14 August 2023
Available online 19 August 2023

employ AI to measure IT risk and cyber risks; Achakzai and Juan (2022) and Chen et al. (2018) employ AI to detect financial frauds and money laundering.

The previously mentioned papers propose AI methodologies to measure different types of risks, often from a financial perspective. We instead propose AI methodologies to measure the "model" risks the AI itself generates. In this sense, our contribution is quite distinctive and unique with respect to the extant literature. We also remark that our proposal for AI risk management is an integrated approach, that considers jointly all main model risks of AI. To exemplify this concept, we make a parallel between AI regulation and financial regulation. When the Basel II capital framework was released, market, credit and operational risk were identified as key statistical variables to assess the capital adequacy of a financial institution and, later, with the Basel III revision (Bank for International Settlements, 2011) it was the turn of systemic risk. Meanwhile, statistical metrics such as the Value at Risk and the Expected Shortfall (Artzner et al., 2001) and, later, the CoVaR (Adrian & Brunnermeier, 2016) have been proposed by researchers and, later, integrated by banks and regulators into an integrated measure aimed at monitoring financial risks and their coverage by banks' internal capital. Similarly to what occurred after the Basel regulations, in this paper we have deduced from the proposed AI regulations four main statistical variables to measure: Sustainability, Accuracy, Fairness and Explainability, which require the development of appropriate statistical metrics, eventually leading to an integrated measure of trustworthiness for a specific AI application, similarly to the integrated financial risk of a financial institution in the Basel regulations. The development of such metrics allows to establish not only whether an AI application is trustworthy, but also to monitor the condition of trustworthiness over time, by means of a risk management model.

The important aspect of our proposal is that we propose an integrated AI risk management model, consisting of a set of four interrelated statistical measures. The four measures can be summarised with the acronym S.A.F.E., which derives from the four considered variables: Sustainability, which refers to the resilience of the AI outputs under anomalous extreme events and/or cyber attacks; Accuracy, which refers to the predictive accuracy of the model outputs; Fairness, which refers to the absence of biases towards population groups, induced by the AI output; Explainability, which refers to the capability of the model output to be understood and oversight by humans, particularly in its driving causes. While the former two requirements are more technical, and "internal" to the AI process, the latter two are more ethical, and "external" to the AI process, involving the stakeholders of an AI system. Similarly to what occurred after the Basel regulations, in this paper we have deduced from the proposed AI regulations four main statistical variables to measure: Sustainability, Accuracy, Fairness and Explainability, which require the development of appropriate statistical metrics, eventually leading to an integrated measure of trustworthiness for a specific AI application, similarly to the integrated financial risk of a financial institution in the Basel regulations. The development of such metrics allows to establish not only whether an AI application is trustworthy, but also to monitor the condition of trustworthiness over time, by means of a risk management model. The proposed metrics consist of "agnostic" statistical tools, able to post-process the predictive output of a machine learning model in a general way, independently on the underlying data structure and statistical model.

2. KAIRI: Key Artificial Intelligence Risk Indicators

In this Section we present our main proposal: a set of Key Risk Indicators for AI Risk Management (KAIRI).

The aim of KAIRI is to help organisations to measure, manage and mitigate risks, throughout the whole AI lifecycle, from use case design to continuous monitoring in production. KAIRI is built upon the S.A.F.E. principles. For each principle, we propose metrics that can be employed in practical use cases for AI risk management, as follows.

2.1. Accuracy

The measurement of predictive accuracy draws on the comparison between the predicted and the observed evidence. Typically, simpler models (e.g. regression models) are less accurate than more complex models (e.g. random forest models).

To measure accuracy, we should distinguish the case of continuous response variables from that of categorical response variables. When the target response variable is continuous, measures such as the Root Mean Square Error (RMSE) of the predictions could be employed. To use RMSE as a Key Risk Indicator, we suggest to accompany it with a statistical test, such as Diebold–Mariano's (Diebold & Mariano, 1995), which can be employed to decide whether the predictions from two competing models differ significantly. The lower the RMSE, the better the model and, when the corresponding p-value is lower than a set threshold (such as 5%), we reject equality of the predictions: the models are significantly different.

When the target response is categorical, measures based on the false positives and false negatives, such as the Area Under the ROC Curve (AUROC), could be employed. To use AUROC as a KRI we suggest a statistical test, such as DeLong's (DeLong et al., 1988), with which we can decide whether the classifications from two competing models differ significantly. The ROC curve represents, for a given set of cut-off points, the True Positive Rate against the False Positive Rate. The greater the AUROC, the better the model and, for a small p-value, we reject equality of the predictions.

2.2. Sustainability

The predictions from a machine learning model, especially when a large number of explanatory variables is considered, may be altered by the presence of "extreme" data points, deriving from anomalous events, or from cyber data manipulation. To improve the robustness and the cybersecurity of AI applications (their sustainability), we propose to extend variable selection methods available for probabilistic models, such as linear regression, to non-probabilistic models, such as random forests and neural network models. Doing so, we obtain a more parsimonious model which, while not significantly losing predictive accuracy, will be simpler and, therefore, more stable against extreme variations in the data.

For probabilistic models, the difference in likelihood between two nested models can be calculated, and a statistical test, such as the F-test for a continuous response, or the Chi-square test for a binary response, can be implemented, to assess whether the difference is significant. For example, if models are compared in a backward perspective, removing one variable at a time, and starting from the fully parameterised model (with the highest likelihood), the test can provide a stopping rule, based on the p-values of the test. The procedure can lead to a model that, while still accurate, is more parsimonious and, therefore, more sustainable than the full model.

For non probabilistic models, we propose a similar stepwise procedure, but ordering variables not in terms of their likelihood, which cannot be calculated, but in terms of an explainability criterion, such as their Shapley values. We then propose to employ a stopping rule based on the comparison of predictive accuracy, either in terms of Diebold and Mariano's test (when the response variable is continuous) or in terms of the DeLong's test (when the response variable is binary. The stopping rule will imply to add variables as long as predictions are different, that is, until the difference in predictive accuracy between two consecutive models is large, and the p-value smaller than a set threshold.

This provides, to our knowledge, the first model selection criteria for (non-probabilistic) ML models.

2.3. Explainability

Explainability has a positive impact on all the stakeholders involved with an AI system. A data scientist can better manage technical model evolutions if the system is interpretable. Auditors and regulators can better validate AI models. Also, end-customers would like to be informed about the reasons of the predictions that involve them (e.g. in credit lending applications) and on how their data is processed.

While some AI models are explainable by design, in terms of their parameters (e.g. regression models), others are black boxes (e.g. neural networks and random forests). For black-box models, we propose to apply Shapley values, introduced by Shapley (1953), and adapted to the explainable AI context by Bracke et al. (2019) and Lundberg and Lee (2017).

Shapley values were introduced in the game theory context. Assume there is a game for each observation to be predicted. For each game, the players are the model predictors and the total gain is equal to the predicted value, obtained as the sum of each predictor's contribution. The (local) effect of each variable $k\,(k = 1, \dots, K)$, on each observation i, $(i = 1, \dots, N)$, can then be calculated as follows:

$$\phi(f(\hat{X}_i)) = \sum_{X' \subseteq C(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [\hat{f}(X' \cup X_k)_i - \hat{f}(X')_i], \qquad (1)$$

where K represents the number of predictors, X' is a subset of $|X'|$ predictors, $\hat{f}(X' \cup X_k)_i$ and $\hat{f}(X')_i$ are the predictions of the ith observation obtained with all possible subset configurations, respectively including variable X_k and not including variable X_k.

Once Shapley values are calculated, for each observation, the global contribution of each explanatory variable to the predictions is obtained summing or averaging them over all observations. For our KAIRI model, we propose as Key Risk Indicator the percentage of Shapley values of the variables contained in the selected machine learning model. If the model is fully parameterised, the KRI is equal to 100%; but, in the realistic setting of a more parsimonious model, easier to explain, the KRI will assume values between 0% and 100%.

We recall that, for a white box model, to evaluate whether the explainability of a variable is statistically significant, we can apply the T-test to the corresponding estimated parameters. A similar procedure can be applied to a black-box model, replacing the explanatory variables with the corresponding Shapley values. In both cases, a small p-value of the test will indicate a significant explainability.

2.4. Fairness

Fairness is a property that essentially requires that AI applications do not present biases among different population groups. Fairness should be verified at both the data (input) level and the prediction (output) level.

To measure fairness we propose to extend the Gini coefficient, originally developed to measure the concentration of income in a population, to the measurement of the concentration of the explanatory variables which may be affected by bias, in terms of their estimated parameters (for an explainable model), or in terms of their Shapley values (for a black-box model.)

Our proposal can be illustrated as follows, in terms of Shapley values (and similarly for the estimated parameters). Let p_m, $m = 1, \dots, M$ be the considered population groups, with X_k, $k = 1, \dots, K$ an explanatory variable, and let v_{mk} be the mean Shapley values associated with the kth variable and the mth population group. The values v_{mk} can be organised in a tabular format, as in Table 1.

From Table 1 we can calculate the ratios $q_{mk} = v_{mk}/v_{\cdot k}$, where $v_{\cdot k} = \sum_{m=1}^{M} v_{mk}$, each of which represents the quota of the kth variable "owned" by the mth population group. The Gini coefficient can then be applied to the obtained ratios, obtaining a measure of concentration of a variable's importance among different population groups.

Table 1
Mean Shapley values for K explanatory variables and M population groups.

	X_1	…	X_k	…	X_K
p_1	v_{11}	…	v_{1k}	…	v_{1K}
$\vdots$	…	…	…	…	…
p_m	v_{m1}	…	v_{mk}	…	v_{mK}
$\vdots$	…	…	…	…	…
p_M	v_{M1}	…	v_{Mk}	…	v_{MK}
	$v_{\cdot 1}$	…	$v_{\cdot k}$	…	$v_{\cdot K}$

For a given explanatory variable, similar ratios lead to a Gini coefficient close to 1, indicating that the effect of that variable is fair across different population groups. On the other hand, a Gini coefficient close to 0 indicates that a variable's effect largely depend on some groups, indicating bias.

In terms of the proposed KAIRI risk management framework, the Gini coefficient for a variable is a value in $[0, 1]$, that can be used as a Key Risk Indicator that indicate its degree of fairness, with higher values indicating higher fairness.

Our proposed KRI for fairness measures how concentrated is the explainability of a variable across different population groups. To evaluate whether the concentration is significantly different from the uniform distribution, indicating fairness, we can accompany the KRI with a non parametric statistical test, such as Kolmogorov–Smirnov test. A small p-value of the test will indicate a significant departure from the uniform distribution, which indicates fairness.

Fig. 1 summarises the proposed KAIRI metrics, for each of the four considered SAFE requirements, and the corresponding statistical tests. For completeness we consider both white box and black box models.

3. Measuring SAFEty of machine learning methods

In this section we show how to use the proposed framework for the most important classes of ML models, to understand whether and how our proposed measures can be implemented.

3.1. Regression

The oldest and still highly employed learning models are regression methods, which are estimated by maximising the likelihood of the available data. They specialise in linear regression models - when the response variable is continuous - and in logistic regression models - when the response is categorical.

Regression models are "white-box" statistical learning methods, that find application in many studies. While a linear regression models specifies the response as a linear function of the explanatory variables, in a logistic regression model we have that:

$$ln((p_i)/(1 - p_i)) = \alpha + \sum_{k=1}^{K} \beta_k x_{i_k}, \qquad (2)$$

where p_i is the probability of an event (such as default) for the ith case; $x_i = (x_{i1}, \dots, x_{iK})$ is the K-dimensional vector of explanatory variables, the parameter α is the model intercept and β_k is the kth regression coefficient.

Both linear and logistic regression are explainable-by-design, being respectively linear and linear in the logarithm of the odds. However, their predictive accuracy, measured by the RMSE and the AUROC values, may be limited, due to their linear nature, especially in the presence of a large and complex database. Fairness of a regression model can be easily checked calculating the Gini variability measure for the variable coefficients estimated in different population groups. In terms of sustainability, regression models are typically based on a probabilistic model and, therefore, likelihood based statistical tests can be used, to obtain a model that, while accurate, is also parsimonious.

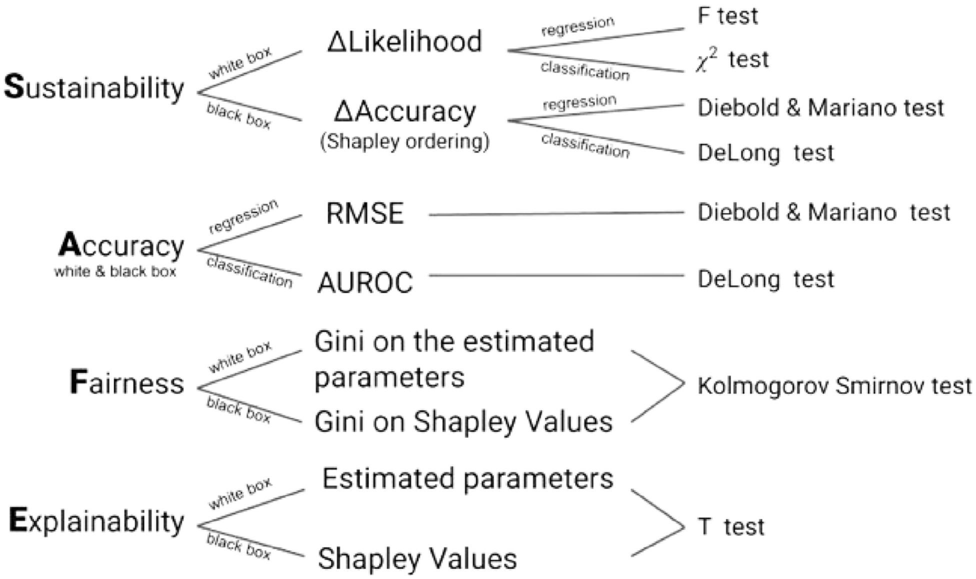

Fig. 1. The proposed KAIRI-SAFE risk measurement model.

3.2. Random forests

Ensembles of tree models, such as random forests and gradient boosting, have shown good performance in financial applications. A random forest model averages the classifications (or the predictions) obtained from a set of tree models, each of which based on a sample of training data and of explanatory variables. In a tree, the explanatory variables determine the splitting rules, with stopping decisions determined by the need to significantly reduce within group variability (as measured by the Gini coefficient, by the classification error, or by the variance).

By averaging the results from many trees, a random forest model increases predictive accuracy but loses interpretability, becoming a black box. To overcome this weakness, explainable Artificial Intelligence models, such as Shapley values, need to be employed. Fairness can be checked calculating the Gini measure for the Shapley values of the explanatory variables in different population groups. In terms of sustainability, ensemble tree models are not based on a probabilistic model and, therefore, likelihood based tests cannot be used. Sustainability can be addressed by means of the proposed model selection procedure based on the comparison of the predictive accuracy, which can lead to a parsimonious model.

3.3. Neural networks

Neural networks can be described as a structure organised according to different levels: the input, the hidden and the output layers. While the input layers receive information from the external environment and each neuron in it usually corresponds to a predictor, the output layers provides the final result to be sent outside of the system. The hidden layers define the complexity of the neural network; a neural network with many hidden layers is a deep learning model.

More formally, a generic neuron j receives n input signals $x = [x_1, x_2, \dots, x_n]$ from the neurons it is connected to in the previous layer. Each signal has an importance weight: $w_j = [w_{1j}, w_{2j}, \dots, w_{nj}]$. The same neuron elaborates the input signals through a combination function which gives rise to a value called potential, defined by $P_j = \sum_{j=1}^{n}(x_i w_{ij} - \theta_j)$, in which θ_j is a threshold which is activated only above a certain value. The output of the jth neuron, denoted with y_j, derives from the application of a function, called activation function, to the potential P_j:

$$y_j = f(P_j) = f\left(\sum_{j=1}^{n} x_i w_{ij} - \theta_j\right). \tag{3}$$

A neural network model may be highly accurate, as it can capture the non linearities present in the data. It is however a black-box model, which requires explainable Artificial Intelligence models to be explained. As for random forest models, fairness can be checked calculating the Gini coefficient for the variable Shapley values calculated in different population groups.

In terms of sustainability, neural network models are typically non probabilistic and, therefore, sustainability should be addressed choosing a parsimonious model by means of a model selection procedure, such as that provided by the proposed comparison of predictive accuracies.

4. Case studies

We have conducted several interviews and workshops with relevant European financial institutions, to identify the main companies' needs and use cases. This has led to the selection of four case studies, which are in line with those suggested in the literature, as described for example in United States National Institute of Standards and Technologies (2022).

4.1. Credit scoring

The evaluation of credit risk largely depends on the estimated probability of default (PD). This problem is usually tackled by estimating a credit scoring model, and setting a threshold to predictively classify each individual or company into one of two main classes: non-default and default, see e.g. Finlay (2011) and Tripathi et al. (2022).

In a credit scoring model, we aim to find a model that can well describe the relationship between the available explanatory variables and the default response variable.

We have applied the proposed KAIRI framework to the credit scoring data of a fintech company, that provides credit scoring of small and medium enterprises to financial institutions. The considered data sample contains eight explanatory variables, derived from the 2020 balance sheet data for over 100,000 SMEs and, as a response variable, their status (default/bonis) at the end of the year 2021. As a machine learning model, We follow Babaei et al. (2023) and apply a random forest model, after partitioning the sample in a 70% training data and a 30% test data.

Sustainability. For sustainability we have run a forward stepwise model selection procedure, inserting variables one a at a time, following

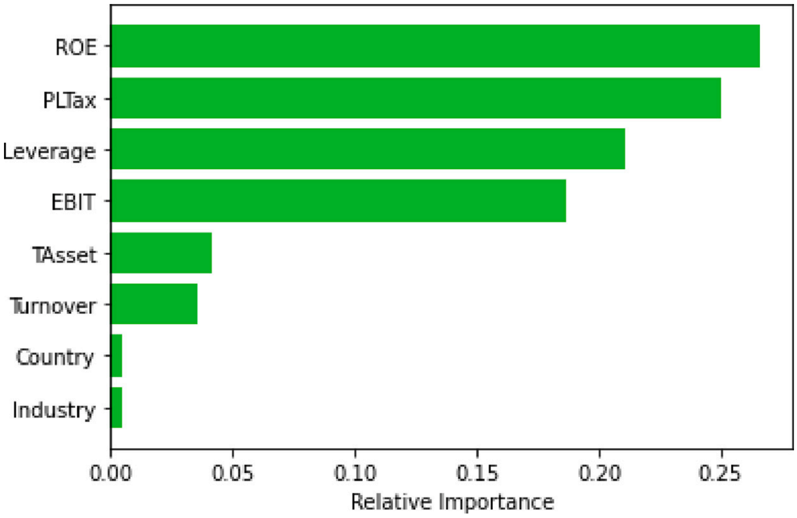

Fig. 2. Random forest feature importance for credit scoring.

Table 2
Gini coefficient and Kolmogorov–Smirnov test for variable Country Shapley values — Year 2020.

	Leverage	P/L	EBIT	ROE	Turnover	TA
$\mathcal{F}_k$	0.0362	0.0591	0.0506	0.0378	0.0566	0.1095
p-value	>0.10	>0.10	>0.10	>0.10	>0.10	>0.10

the order in Fig. 2: from “ROE” to “Industry”. Our results indicate that the p-values of the Delong tests for pairwise model comparison is less than 5% until the insertion of “EBIT”, and it raises above such threshold when “TAsset” is inserted. This leads to a model with only the first four variables selected. It is clearly more parsimonious than the full model. It also indicates that Country and Industry Sector are not significant explanations: a preliminary (unconditional) indication of fairness.

Accuracy. As the response variable to be predicted is binary, the AUROC measure, along with the DeLong test, can be employed as the accuracy KRI. The AUROC measure for the full random forest model, with all the available ten variables, turns out to be equal to 0.97, against 0.89 for a logistic regression model with the same variables. The application of DeLong’s test leads to a very small p-value, indicating that the random forest model is significantly superior with respect to logistic regression.

Fairness. Fairness can be evaluated comparing the Shapley values of the explanatory variables in the four countries in which the companies are based: France, Germany, Italy and Spain, and also in the industrial sectors in which they operate. Table 2 reports the values of the Gini coefficient and of the related Kolmogorov–Smirnov test, to evaluate the fairness of the model, across different countries.

Table 2 shows that the model is fair: there are no significant differences between countries, conditionally on each of the six considered balance sheet variables: “ROE”, “PLTax”, “Leverage”, “EBIT”, “Tasset” and “Turnover”. Similar results can be obtained when fairness across industrial sectors is considered.

Explainability. As the random forest model is black box, we can use global Shapley values as a KRI. Fig. 2 presents the mean Shapley values of all the eight considered variables.

Fig. 2 indicates that the variable “ROE” is the most explainable variable for default, followed by “PLTax”, “Leverage” and “EBIT”. In terms of our proposed KRI, these four variables explain about 90% of the Shapley values, and are all significant, applying a T-test to the Shapley regression coefficients.

4.2. AML transaction monitoring

A bank’s Anti Money Laundering (AML) function has the goal of effectively managing risks related to money laundering and terrorism financing, in compliance with the international and national regulations. Sanctions are severe and range from high fines to the stop of business. For a review see, for example, Chen et al. (2018).

Most AML functions use IT tools for monitoring unexpected transaction activity from their clients. These tools are often based on static rules which sometimes refer to outdated crime schemes, that are continuously evolving through time. This leads to producing a very high number of false positive alerts. As the AML function is obliged to process every alert and the majority of alerts are false positives, it is evident that optimising the false positive ratio is a key business issue that can significantly increase the efficiency of the AML function.

Artificial Intelligence is being adopted by an increasing number of banks to address the above mentioned challenge, and to predict whether a Unique Transaction Reference number has to be filed into a Suspicious Transaction Report (STR). The STR response can be modelled as a binary target variable.

We have applied the proposed KAIRI framework to the AML dataset of a large banking group, trained on more than 1.6 million transactions over a period of more than three years. The data includes more than 400 feature variables, including: Personal data: such as gender, nationality, date of birth; Transaction details: such as cash, control code; Financial data: overseas transfers, withdrawals, deposits; Negative events: protests, prejudicial acts, court proceedings; Income data: total assets, revenues, etc. Other risk profiles: current risk level.

The machine learning model chosen to analyse the data has been a Random Forest model, well suited for contexts like AML, characterised by several local distinct behaviours. The application of our proposed KAIRI model to the model output has led to the following conclusions.

Sustainability. To achieve a sustainable model, we have implemented a forward stepwise model selection procedure. Variables have been inserted in the model one after the other, in the order described by the feature importance plot of the random forest. This because the application of Shapley values to 400 features was computationally prohibitive. Variables have been inserted in the model as long as the p-value for the DeLong test was smaller than 5%. Doing so, the model has selected 100 variables.

To evaluate Accuracy, we have considered as a main KRI the AUROC (with the related DeLong’s p-value) along with the precision (to check on false positives performance) and the recall (to check on false negatives performance), embedded into the F1-score (which is a weighted

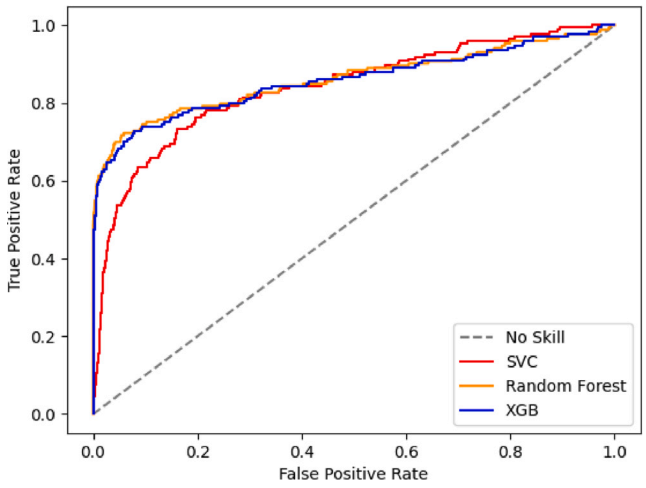

Fig. 3. Accuracy for anti-money laundering models.

average of the previous two metrics). After consultation with a business domain expert, we have agreed on a minimum threshold AUROC value of 0.7. The selected model result in an AUROC equal to 0.86, along with an F1 score of 0.79, and is therefore acceptable, having both a high predictive accuracy and a good balance between the two types of error rates.

To improve the robustness of the analysis we performed, Fig. 3 compares the accuracy of our proposed random forest model to detect money laundering against a Gradient Boosting and a Support Vector Classifier, on the same training and validation datasets.

From Fig. 3 the Random Forest model has the best performance (AUROC = 86,1%), followed by Gradient Boosting (AUROC = 85,7%) and by the Support Vector Classifier (AUROC = 83,9%). While the DeLong's test does not detect a significant difference between Random Forest and XGBoost, it does so between any of the other two models and the Support Vector Classifier.

Fairness. We analysed fairness with respect to the gender variable. Our dataset is unbalanced towards men, as it has a 58.8% of transactions by men a 41.2% by women. We applied the Gini measure, and the related Kolmogorov Smirnov test, to compare the Shapley values of each feature variable in the two population groups and the related p-value, to the distributions of the explanatory variables, and to those of their Shapley values. We obtained that, although the transactions are significantly unbalanced towards men, the selected variables are fair, with p-values of the test greater than 5%. The balance is reflected in the group specific AUROC values of the considered random forest model, equal to 87% for men and 0.85% for women.

Explainability. Explainability can be assessed on the selected model, with 100 variables. The global Shapley values indicate that the most important five variables are, in order: 1. Customer's risk profile; 2. Number of active relationships of the customer; 3. Average amount of money transfers of the customer in the previous six months; Amount of the transaction; Total amount of money transfers in the previous six months, which explain about 70% of the explainability of the selected random forest model with 100 variables. The AUROC of the model with five variables is equal to 68% against 86% of the selected model: a loss in accuracy compensated by a gain in explainability. We additionally remark that, to ease the interpretation of the model, a methodological documentation has been added to the model, to explain the model to internal users and to financial supervisors. In this way, overrides and claims by the AML functions can be used as further indicators of whether the model output is correct.

4.3. IT systems monitoring

The service industry largely depends on IT systems and procedures. Operational Risk Management best practices suggest to monitor the IT System Service Level to prevent system failures that may impact the internal processes, leading to deterioration of the customer experience.

This problem can be tackled by means of a machine learning model that estimates the risk of failure each IT procedure has and that sets thresholds to trigger alerts that will lead IT System managers to take action beforehand. Data for machine learning models can be real time data, such as system logs (Zhang et al., 2016), or batch data (Melancon et al., 2021).

A possible approach is to estimate the probability of a failure taking place for a specific IT system classifying the IT procedures into two classes: Close-to-fail and Stable. The available information for a set of N IT procedures related to T variables can be divided into two categories: "static information" (such as information related with the type of system(/s) used, the software/hardware characteristics, the implementation date) and "dynamic (or history-based) information" (such as information related with the issues and failures already managed, the tickets under management, the previous and next releases that will impact on it; the logs of the system for real time use cases). The response variable indicates whether the IT procedure will have a system/service failure in the near future and can be modelled as a binary target variable.

We have applied the proposed KAIRI framework during the development of the AI solution for IT failure detection of a large italian banking group.

The considered data consists of the historical information for 137 IT procedures along 153 weeks, for a total of 20 961 observations). As many as 258 possible explanatory variables have been retrieved during the data collection and feature engineering process. The dataset has been split into a training and testing sample: the last 16 weeks have been considered as a test set. During an exploratory data analysis, the number of variables was reduced to 35 eliminating variables with data quality issues or highly correlated with others. As a machine learning model we have implemented LightGBM, a gradient boosting tree model.

Sustainability. To achieve a Sustainable model, we have proceeded calculating variable importance, by means of Shapley values. We have then inserted variables one at a time, following the Shapley value ordering, until a significant difference in AUROC was found. The procedure has led to choosing seventeen explanatory variables.

Accuracy. The accuracy of the selected model, using AUROC as a KRI, is equal to 77.2%, a reasonably high variable for the considered problem. Indeed, for the same model, Precision is about 45%, indicating the presence of many false positives; and Recall is about 82%, indicating that most incidents are correctly predicted.

Fairness. Fairness has here an "internal" rather than an "external" impact. It can indicate, for example, whether the model works better for some IT systems, and worse for others.

Of particular interest is to assess fairness with respect to the type of hosting. To this aim, we have defined two classes of systems: those managed internally (Hosting: Facility Management) and those in outsourcing (Hosting: Outsourcing).

Fig. 4 illustrates the behaviour in time of the predictions, against the actual values, for two classes of systems: those managed internally (left panel) and those in outsourcing (right panel).

From Fig. 4, note that the predictions for systems managed internally are more accurate. Indeed, while the difference between the observed distribution of the errors in the two classes are not significantly different, using the Kolmogorov–Smirnov test, the same test applied to the Shapley values indicates that they are statistically different. This indicates a bias that comes from the model, rather than from the data.

Explainability. To evaluate the "Explainability" principle of SAFE the Shapley values of the selected model with 17 variables have been computed. Fig. 5 presents the results of the elaboration.

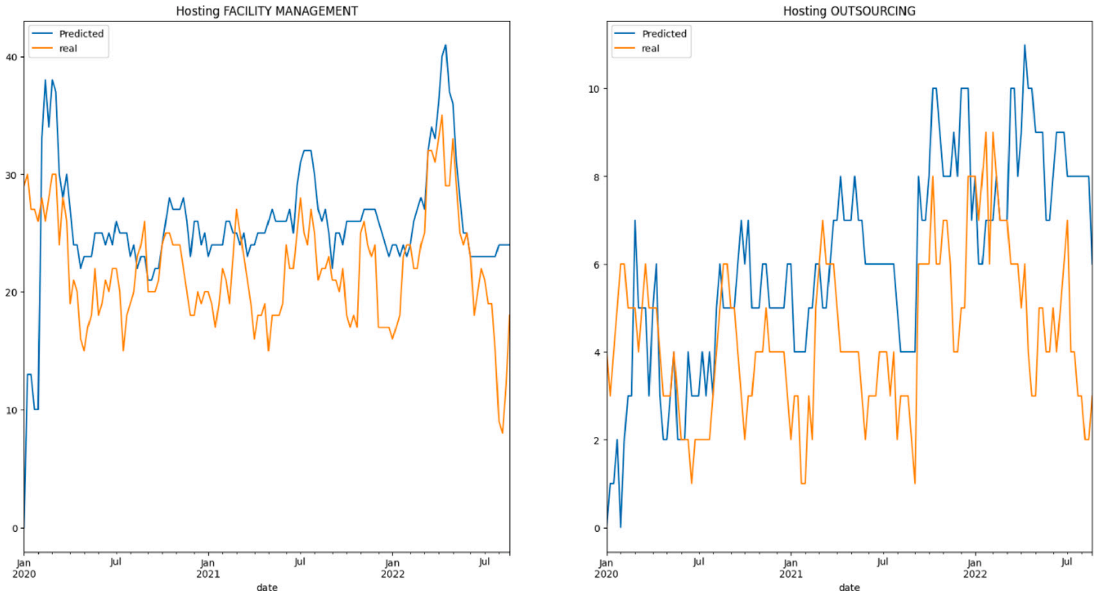

Fig. 4. Fairness of an IT system monitoring model.

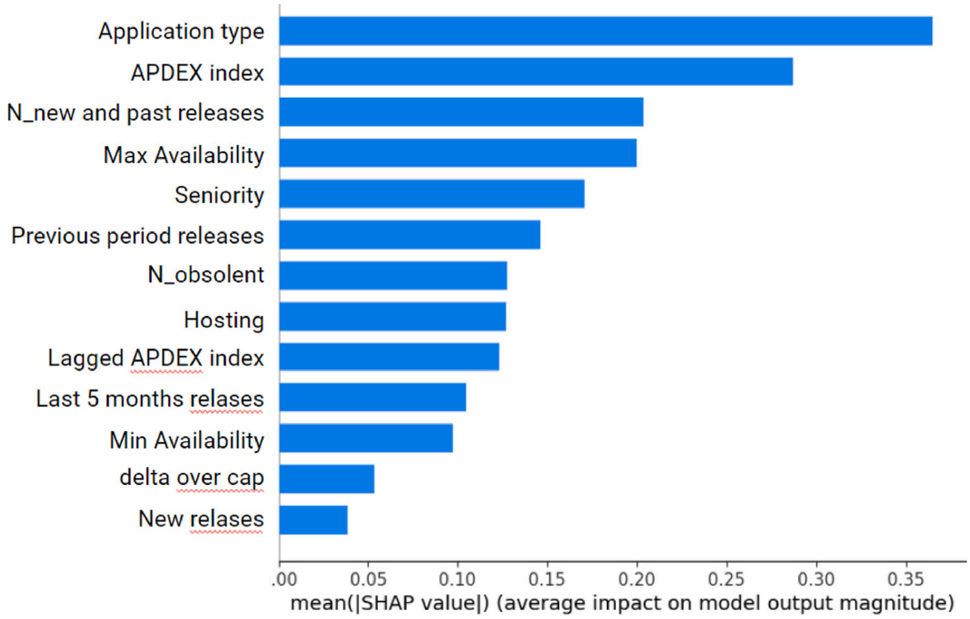

Fig. 5. XGboost feature importance for IT failure detection.

Fig. 5 indicates that the most important feature of the model is the application type, followed by the procedure performance metric APDEX and then, with almost the same impact, the number of new and past releases and the max time availability. In terms of our proposed KRI, these four variables explain about 90% of the Shapley values, and are all significant, applying a T-test to the Shapley regression coefficients.

4.4. Parmesan Cheese anomaly detection

The production of Parmesan Cheese requires the fulfilment of several quality requirements. It is a cheese that can be seasoned for a long time, even more than 10 years. Not all the producers have a warehouse and the ability to store the cheese wheels. As a result, they often outsource the seasoning process to specialised banks, which monitor the status of each cheese wheel over time, to detect anomalies.

The development of a Parmesan cheese anomaly detection model is an example of AI applied to images, as in Deeckel et al. (2018), which, in our considered bank, consists of three activities for each cheese wheel. First, the acquisition of cheese images, through the usage of photo cameras installed on cheese cleaning machines, from which the RGB values for each pixel (or section) in the taken photos are collected. Second, the acquisition of further information, such as the year and month of production of the cheese, and information on its manufacturers. Third, the application of a neural network model that, using the previous sources of information as input data, classifies the cheese wheel, predicting whether it has anomalies and the anomaly type.

Considering that the seasoning process requires that humidity is higher than 80%, the most common anomaly that can arise is related with mold, that can appear on the crust of the wheel. A less common issue takes place when the warehouse humidity is too low, possibly causing fractures on the cheese wheel crust.

To develop a solution for anomaly detection in Parmesan Cheese a Convolutional Neural Network (CNN) model has been employed in

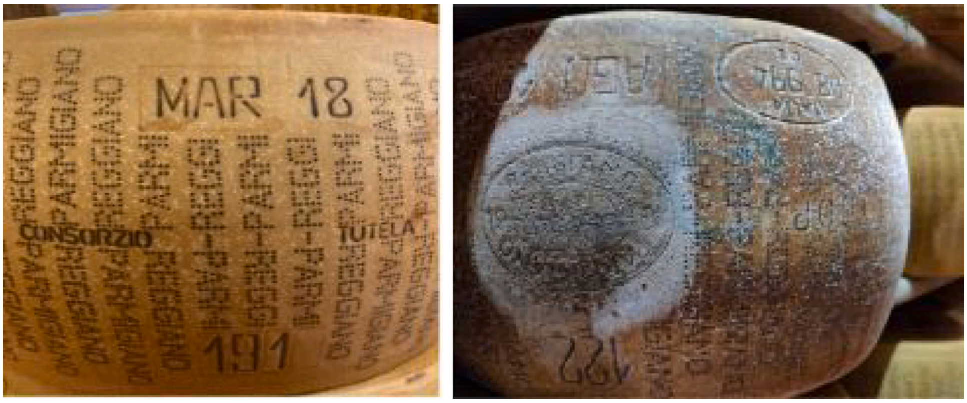

Fig. 6. Sustainability of an anomaly detection model for Parmesan cheese.

a banking group that stores parmesan cheese as an additional line of business. The input variables of the model are the RGB values for each pixel of the photos. For each cheese wheel 4 photos have been taken by the cameras and 20 000 cheese wheels have been analysed over the 500 000 available cheese wheels in the warehouse, selecting all the wheels with anomalies identified manually (about 4% of the sample considered). The response variable analysed for the model was binary: anomalous/normal cheese wheel. Data has been partitioned in a training sample (80%) and a testing sample (20%) using a stratified selection.

The application of our proposed KAIRI framework to parmesan cheese anomaly detection has led to the following conclusions.

Sustainability. Sustainability has been addressed considering an increasing numbers of hidden layers and kernel size, and calculating their AUROC, at each iteration. Due to the large number of parameters involved, Shapley values have not been calculated to determine the order of the variables but, rather, the DeLong test has been employed to compare the AUROC of the most accurate model with that of progressively simpler models. The selected model is a convolutional neural network with 87 650 parameters.

To better illustrate the actual data analysis we performed, Fig. 6 illustrates two examples of Parmesan cheese wheels.

Fig. 6 compares a “normal” cheese wheel (left figure) with an apparent “anomalous” cheese wheel (right figure), taken from a photo that was by mistake rotated upside down. Using a neural network with all pixel data, this wheel has been classified as anomalous (and as having a mold), when in reality it is not anomalous. Instead, using the more parsimonious neural network model (with 87 650 parameters), based on photo sections rather than pixels, the wheel is correctly classified as non anomalous.

Accuracy. AS the target response variable is binary with two classes: anomalous and not anomalous, model accuracy can be addressed calculating the AUROC KRI on the test set. The AUROC results to be equal to 85.3%, indicating a highly accurate model.

Fairness. Fairness has been evaluated with respect to the age of the parmesan cheese. To this end, the cheese wheels have been partitioned in samples with different ages, in months: (0; 18]; (18; 36]; (36; 60]; (60;240]. Given the large number of explanatory variables considered, fairness has been assessed directly comparing model accuracies (AUROC) rather than the conditional Shapley values. The AUROCs result to be, respectively: 87.4%, 86.1%, 85.4%, 76.8%. They are similar to each other, with the exception of the class of oldest cheese wheels. We then applied DeLong test for all pairwise comparisons: the test rejects the null hypotheses of fairness only when comparing the sample with cheese older than sixty months with the others. This highlights that the performance on that sample is the worst, indicating that the model is unfair in predicting anomalies for the old-aged cheese wheel.

Explainability. To evaluate explainability, Shapley values can be used, in principle, to evaluate which photo sections bring the most information in explaining the target response variable. However, the large number of variables imply a too high computational overhead. For this reason, explainability is carried out “manually”: when an alert is generated for a cheese wheel, the operators check the images causing the alert and the cheese status and report the outcome of their verification, not only in terms of the anomaly, but also in terms of its driving causes.

5. Conclusions

From a methodological viewpoint, we contribute to the research on Artificial Intelligence in two main ways. First, we propose an integrated AI risk management framework that can assess, in compliance with the emerging AI regulations, the risks of artificial intelligence applications, using four main statistical principles of SAFEty: Sustainability, Accuracy, Fairness, Explainability. Second, for each principle we contribute with the proposal of a set of integrated statistical metrics: the Key Artificial Intelligence Risk Indicators, that can be used to measure AI SAFEty and implement an effective AI risk management system for any AI applications.

From an applied viewpoint, we contribute to the research on the application of Artificial Intelligence by implementing the proposed SAFEty risk management system to four use cases, that have been indicated by the financial industry among the most relevant and necessary applications of AI: credit scoring, anti money laundering, IT systems monitoring and anomaly detection.

Our methodological results and applications show that our proposed AI risk management framework can support AI developers and users, but also supervisors and regulators, in the development and supervision of machine learning models which, in compliance with the regulatory requirements, can assess the safety and the trustworthiness of AI applications, and measure their risks.

What proposed has main advantages but also overheads, which should be indicated. The measurement of explainability and fairness rely on Shapley values, a highly computational intensive procedure, which may be very challenging in high dimensional situations, as it has occurred in both the AML and the Parmesan cheese anomaly detection case. To overcome this problem, mathematical research should be dedicated to the connections between sustainability and explainability, leveraging the parsimony principle to ease the computational burden required to assess explainability and, through it, fairness.

Further research could apply the proposed framework to other contexts, different from the financial industry, for which a SAFE AI can leverage the value of data, while controlling the risks it generates.

Further research is required from a mathematical and statistical viewpoint, to connect the different concepts presented in Fig. 1 and propose alternative ways to measure the SAFEty requirements and test their statistical significance, thereby providing tools that can improve the robustness of the conclusions, in real use cases. A recent related paper is Giudici and Raffinetti (2023), who proposes Lorenz Zonoids as unifying measurement tool.

We finally mention that research is needed to measure the SAFEty of machine learning models that are not supervised. What developed in this paper could be a good basis: unsupervised and/or generative models can be assessed indirectly, inserting their outputs in supervised models.

CRediT authorship contribution statement

Paolo Giudici: Ideated, Supervision, Elaborated, Writting. **Mattia Centurelli:** Elaborated, Writting. **Stefano Turchetta:** Elaborated, Writting.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The authors acknowledge financial support and research exchanges from the European Union H2020 PERISCOPE Research Grant, contract number 101016233.

References

Achakzai, M. A. K., & Juan, P. (2022). Using machine learning meta-classifiers to detect financial frauds. *Finance Research Letters*, *48*, Article 102915.

Adrian, T., & Brunnermeier, M. (2016). CoVaR. *American Economic Review*, *166*(7), 1705–1741.

Aldasoro, I., Gambacorta, L., Giudici, P., & Leach, T. (2022). The drivers of cyber risk. *Journal of financial stability*, *60*, Article 100989.

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (2001). Coherent measures of risk. *Mathematical Finance*, *9*(3), 203–228.

Babaei, G., Giudici, P., & Raffinetti, E. (2023). Explainable fintech lending. *Journal of Economics and Business*, Article 106126.

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). *Machine learning explainability in finance: an application to default risk analysis*: *Staff working paper n. 816*, London: Bank of England.

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society*, *73*(1), 70–90.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable machine learning in credit risk management. *Computational Economics*, *57*, 2013–2016.

Chen, Z., Van Khoa, L. D., & Teoh, E. N. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowldege in Information Systems*, *57*, 245–285.

Deeckel, L., Vandermeulen, R., Ruff, L., & Mandt, M. (2018). Image anomaly detection with generative adversarial networks. In *lecture notes in computer science*: *Vol. 11051*, *Machine learning and knowledge discovery in databases*. Springer.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–845.

Diebold, F., & Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

European Commission (2022). Artificial intelligence act. URL: https://artificialintelligenceact.eu.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, *210*(2), 368–378.

Frost, J., Gambacorta, L., Huang, Y., & Zbindnen, P. (2019). BigTech and the changing structure of financial intermediation. *Economic Policy*, *34*(100), 761–799.

Ganesh, A. D., & Kalpana, P. (2022). Future of artificial intelligence and its influence on supply chain risk management – A systematic review. *Computers & Industrial Engineering*, 169.

Giudici, P., & Abu-Hashish, I. (2019). What determines bitcoin exchange prices? A network VAR approach. *Finance Research Letters*, *28*, 309–318.

Giudici, P., & Polinesi, G. (2021). Crypto price discovery through correlation networks. *Annals of Operations Research*, *299*(1–2), 443–457.

Giudici, P., & Raffinetti, E. (2021). Explainable AI in cyber risk management. *Quality and Reliability Engineering International*, *38*(3), 1318–1326.

Giudici, P., & Raffinetti, E. (2023). S.A.F.E. artificial intelligence in finance. *Finance Research Letter*, *56*(104088).

Bank for International Settlements (2011). *Basel III: A global regulatory framework for more resilient banks and banking systems*. Basel Committee on Banking Supervision, 1-6-2011.

Kuiper, O., Berg, M. Van. den., Burgt, J. Van. der., & Leijnen, S. (2021). Exploring explainable AI in the financial sector: Perspectives of banks and supervisory authorities. In *Artificial intelligence and machine learning*. Springer.

Liu, W., Fan, H., & Xi, M. (2022). Credit scoring based on tree-enhanced gradient boosting decision trees. *Expert Systems with Applications*, *189*, Article 116034.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In *Proceedings of neural information processing systems 2017* (pp. 4768–4777).

McCall, J. Ainslie., Shakya, S., & Owusu, G. (2017). Predicting service levels using neural networks. In *International conference on innovative techniques and applications of artificial intelligence* (pp. 411–416). Springer.

Melancon, G., Grangier, P., Prescott-Gagnon, E., Sabourin, E., & Roussseau, L. M. (2021). A machine learning-based system for predicting service-level failures in supply chains. *INFORMS Journal on Applied Analytics*, *51*(3), 200–212, 2021.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, *165*, Article 113986.

Naim, A. (2022). Role of artificial intelligence in business risk management. *American Journal of Business Management, Economics and Banking*, *1*, 55–66.

Sachan, S., Yang, Y., Xu, E., & Li, D. (2020). An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications*, *144*, Article 113100.

Shapley, L. (1953). A value for n-person games. *Contribution to Theory of Games*, *2*, 307–317.

Sundra, B. M., Sathiyamurthi, K., & Subramanian, G. (2023). Critical evaluation of applying machine learning approaches for better handling bank risk management in the post-modern era. *Scandinavian Journal of Information Systems*, *35*(1), 1228–1231.

Tripathi, D., Shukla, A. K., Reddy, B. R., Bopche, G. S., & Chandramohan, D. (2022). Credit scoring models using ensemble learning and classification approaches: A comprehensive survey. *Wireless Personal Communications*, *123*, 785–812.

United States National Institute of Standards and Technologies (2022). AI risk management framework. URL: https://www.nist.gov/itl/ai-risk-management-framework.

Zhang, K., Xu, J., Renqiang, M., Jiang, G., & Pelechrinis, K. (2016). Automated IT system failure prediction: A deep learning approach. In *IEEE international conference on big data(2016)* (pp. 1291–1300).