Users' trust in black-box machine learning algorithms

Heitor Hoffman Nakashima and Daielly Mantovani
Department of Management, University of São Paulo, São Paulo, Brazil, and
Celso Machado Junior
Business School, Universidade Municipal de São Caetano do Sul, São Caetano do Sul, Brazil and Department of Management, Universidade Paulista, São Paulo, Brazil

Received 20 June 2022
Revised 28 August 2022
31 August 2022
13 September 2022
26 September 2022
Accepted 27 September 2022

Abstract

Purpose – This paper aims to investigate whether professional data analysts' trust of black-box systems is increased by explainability artifacts.

Design/methodology/approach – The study was developed in two phases. First a black-box prediction model was estimated using artificial neural networks, and local explainability artifacts were estimated using local interpretable model-agnostic explanations (LIME) algorithms. In the second phase, the model and explainability outcomes were presented to a sample of data analysts from the financial market and their trust of the models was measured. Finally, interviews were conducted in order to understand their perceptions regarding black-box models.

Findings – The data suggest that users' trust of black-box systems is high and explainability artifacts do not influence this behavior. The interviews reveal that the nature and complexity of the problem a black-box model addresses influences the users' perceptions, trust being reduced in situations that represent a threat (e.g. autonomous cars). Concerns about the models' ethics were also mentioned by the interviewees.

Research limitations/implications – The study considered a small sample of professional analysts from the financial market, which traditionally employs data analysis techniques for credit and risk analysis. Research with personnel in other sectors might reveal different perceptions.

Originality/value – Other studies regarding trust in black-box models and explainability artifacts have focused on ordinary users, with little or no knowledge of data analysis. The present research focuses on expert users, which provides a different perspective and shows that, for them, trust is related to the quality of data and the nature of the problem being solved, as well as the practical consequences. Explanation of the algorithm mechanics itself is not significantly relevant.

Keywords Artificial intelligence, Black-box systems, Machine learning, Trust, Explainability

Paper type Research paper

Introduction

Artificial intelligence (AI) is increasingly present in everyday life, as the fast development of machine learning (ML) techniques has made it possible to create applications such as recommendations for financial products (Farquad, Ravi, & Raju, 2012), algorithms for detecting credit card fraud, personal virtual assistants (Lu, Li, Chen, Hyoungseop, & Serikawa, 2018) and autonomous driving vehicles, among other relevant daily applications. However, how an ML algorithm achieves the result presented to the decision-maker is frequently not disclosed in the case of black-box algorithms, whose processing takes place in a closed environment. For example, it does not allow the identification of which variables

Revista de Gestão
Emerald Publishing Limited
e-ISSN: 2177-8736
p-ISSN: 1809-2276
DOI 10.1108/REGE-06-2022-0100

impacted a credit forecast or, in other words, why a certain customer had their loan application denied, while another had theirs approved.

Compared to classical statistical models, in which it is possible to clearly explain how the model's decision rules were created and for which there were discussions with stakeholders and regulators about the reasonableness of the results found, AI models do not allow any detailing of the rules created (Hoffman, Mueller, Klein, & Litman, 2018). The generation of explanations for AI systems is a key issue for both developers, who need tools for debugging and checking the level of accuracy of the sequences of rules that lead to a conclusion as well as for users, who seek to trust the answers provided by the AI by inspecting the chain of decisions made to support a given conclusion (Preece, 2018). The term "math destruction algorithms" has become synonymous, even among the general population, with the work of the same name by Cathy O'Neil. It presents the potential biases that can be embedded in AI algorithms and the social consequences of these biases (O'Neil, 2017). The regulation authorities for data protection in several countries have expanded this discussion, raising questions about the adequacy of decisions made by these systems, especially regarding issues with a real impact on people's lives (Kenny, Ford, Quinn, & Keane, 2021). In short, the consequences of errors made by a model trained on ML can be severe; for example, if an algorithm classifies an x-ray image as normal, whereas in fact there is a tumor that might threaten the patient's life (Narwaria, 2022).

Brynjolfsson and McAfee (2017) described the problems associated with the opacity of black-box algorithms, bringing this discussion to the nonacademic user. There are, in their view, three risks associated with the opacity of AI systems. The first is the hidden bias, derived from the data used for training the models that cannot be explicitly stated as a rule; for example, a credit-granting system that may take into account factors such as ethnicity, race or gender for its final decision. The second risk is associated with the difficulty of extracting explicit rules from a complex model. For example, a system of neural networks that uses hundreds of thousands of connections, in which each connection has a small contribution to the final decision, means that it is extremely difficult or even impossible to recognize its internal rules. The latter is associated with the difficulty of diagnosing and correcting possible, and sometimes unavoidable, errors in an AI system. Hence, the underlying structures created during the system training can lead to incorrect decisions that are far from the ideal.

AI systems using black-box algorithms are powerful tools in terms of results and predictions, but there is a direct relationship between results (i.e. accuracy) and opacity, which makes it difficult to gain insights into the internal processing (Chen, Yang, Pan, Xu, & Zhou, 2015). The neural network algorithms that are the focus of this research exemplify the opacity in AI systems. To address the opacity issues in ML and, consequently, in AI systems, the term explainable artificial intelligence (XAI) was proposed by van Lent, Fisher, and Mancuso (2004). This can be defined as a system in which the user can not only see, but also understand how the inputs (independent variables) are mathematically mapped in order to generate the outputs (dependent variables) (Adadi & Berrada, 2018). Thus, it seeks to ensure that algorithmic decisions can be explained to final users and other stakeholders in nontechnical terms (Barocas *et al.*, 2018), so that possible biases can be identified and corrected.

Hoffman *et al.* (2018) posed an interesting definition about the function of XAI mechanisms. In their view, explainability artifacts should clarify the mental models behind the analytical model, making it possible to differentiate positive and negative aspects and "shields" that limit the development of new and richer mental models. However, these mechanisms of explainability are not yet widely applied, although AI systems have been increasingly used by organizations. Thus, the following research question was posed: Do

explainability mechanisms for black-box AI systems increase user confidence in system results?

This study aimed to analyze whether the methods of interpretability for the black-box AI algorithms increase user trust in the system results. To reach the objective, a neural network model was developed as well as the explainability mechanisms for their results. In a quasi-experimental design, users' trust was measured when they analyzed the results of the neural network model with and without explainability solutions.

Explainability in AI systems

The implementation of explainability mechanisms, known as XAI, is fundamental to guarantee the practical viability of AI models. However, its adoption has been predominantly post hoc; that is, adopted as an additional analysis after model training, when it should actually be part of the design of the model itself (Narwaria, 2022). It is possible to sort the interpretability models into two groups: global interpretation and local interpretation (Pereira *et al.*, 2017). The goal of global interpretation is to understand how independent variables without transformation have influenced the predictions of the model, generating a general explanation and not a specific solution (explanation by each dataset observation). This type of interpretation does not identify the influence of each independent variable on the prediction but aims to identify general factors influencing the training. Local interpretation, in turn, seeks to understand why certain decisions were made by the model after its estimation, considering the results of the predictions per case of the sample (Adadi & Berrada, 2018). In this study, the mechanisms of local explainability were considered, as they make it possible to understand the results generated by the black-box algorithms on a case-by-case basis.

In addition to global or local level, methods of interpretability can be classified as agnostic or model specific. Agnostic methods allow the generation of explainability solutions independently of the mathematical algorithm used to create the original model. This feature offers the flexibility to explain black-box models. Specific methods have been developed for application to specific algorithms. One of the methods of agnostic local interpretation was proposed by Ribeiro, Singh, and Guestrin (2016). It is called local interpretable model-agnostic explanations (LIME) and is a linear proxy model that proposes the creation of explanations for predictions of any classifier that are interpretable and faithful to the original model prediction. The LIME method was used in the present research but has also been used in other studies regarding explanation of opaque systems (Narwaria, 2022). Other XAI methods are reported in the literature; for example, the contributions oriented local explanations–Hadamard product (COLE-HP) used by Kenny *et al.* (2021) for generating local explanations for a convolutional neural setworks model; the GLocalX agnostic method proposed by Setzu *et al.* (2021), which generates local explanations that, when aggregated, can provide a global explanation for the opaque model; and the Local interpretation-driven abstract Bayesian network (LINDA-BN) method based on Bayesian networks for generating local explanations about conditional dependencies proposed by Moreira *et al.* (2021). Confalonieri, Weyde, Besold, and Moscoso del Prado Martín (2021) proposed the trepan reloaded global explanation model, which considers the ontologies of the dominant knowledge and builds decision trees from black-box models, thus, making it easier to understand the results.

To assess the models, accuracy metrics can be applied to a test dataset. However, this type of assessment may not be an indication that the model is reliable. Inspection of individual forecasts and their explanations is a complementary solution to these metrics, but it is important to suggest which instances should be inspected, especially for large datasets. The LIME method, applied in this research, proposes to provide explanations for single predictions as a solution to "confidence in a forecast" and selects several of these predictions (and explanations) as a solution to the "model confidence" problem.

The explanation is defined as a $g \in G$ model, where G is a class of potentially interpretable models, such as linear models, decision trees or descending rule lists (Wang & Rudin, 2015). A $g \in G$ model can be presented to a user along with visual or textual artifacts. As not every $g \in G$ model can be simple enough to be interpretable, this study used $\Omega(g)$ as a measure of complexity (as opposed to the interpretation) of the explanation $g \in$ G. For example, for decision trees, $\Omega(g)$ can be the tree depth, while for linear models, Ω(g) can be the number of nonzero weights.

The model being explained can be denoted as : $\mathbb{R}^d \rightarrow \mathbb{R}$. In classification models, $f(x)$ is the probability (or a binary indicator) that x belongs to a certain class. Furthermore, this study used $\pi_x(z)$ as a measure of proximity between an instance z to x, in order to define the locality around x. Finally, $\mathcal{L}(f, g, \pi_x)$ was a measure of how unfaithful g is in approximating f in the locality defined by π_x. To ensure interpretation and local fidelity, it was necessary to minimize $\mathcal{L}(f, g, \pi_x)$ by having $\Omega(g)$ sufficiently low to be interpretable by humans. The explanation produced by LIME (Ribeiro *et al.*, 2016) is obtained by equation (1):

$$\xi(x) = argmin_{g \epsilon G}\, \mathcal{L}(f, g, \pi_x) + \Omega\,(g) \tag{1}$$

Source(s): Ribeiro *et al.* (2016, p. 3).

For the explanations to be model agnostic, LIME proposes a loss minimization with locality recognition $\mathcal{L}(f, g, \pi_x)$ without making any assumptions about f. Therefore, to learn the local behavior of f according to the input variables, the approximation of $\mathcal{L}(f, g, \pi_x)$ is made, creating samples, weighted by π_x. The sample instance is created around x' by drawing random nonzero elements of x'.

Given a perturbed sample $z' \in \{0, 1\}^{d'}$, containing a fraction of the nonzero elements of x', the original data sample $z' \in R^d$ is recovered obtaining $f(z)$, which will be used as a label for the explanation model. Given this dataset Z of perturbed samples with the associated labels, equation (1) is applied to obtain an explanation $\xi(x)$.

Trust in AI systems

The adoption of new technologies leads to changes in human behavior, in some cases with loss of control over some functions that were essentially human, sometimes with loss of power of acting over the object. In this sense, the user's confidence in the new technology or system is fundamental so that the feeling of well-being given the loss of control can be restored. If, on the one hand, reliable systems tend to be more frequently used, especially in situations involving risk, on the other hand, the factors that increase trust in a system still need to be explored in depth (Cahour & Forzy, 2009). Trust results from efforts that human beings make to overcome their fears, which are not a rational reaction to uncertainty, but a reaction to the fear generated by such uncertainty (Gléonnec, 2004).

The representation of trust can be summarized as a real relationship between risk and delegation of control to an object. Hence, the emotional comfort of the subjects depends on their level of trust or distrust in the object. When they consider it to be trustworthy (or not risky for them) and predictable, both complexity and uncertainty are reduced by increasing their emotional comfort (Cahour & Forzy, 2009).

The study of trust in AI systems is important because, in order to delegate to the algorithm a task that was originally performed by humans (e.g. credit analysis, performance assessment for a promotion at work, image analysis for disease diagnosis, autonomous driving vehicles, among others), it is necessary to establish a relationship of trust with the algorithm. Since the algorithm studied in this research is that of black box, the difficulty in understanding the reasons that lead to the output can generate emotional discomfort in the user and compromise

confidence in the result presented by the model. Dikmen and Burns (2022) made a counterpoint, arguing that XAI mechanisms help in the interpretation of opaque model results but, at the current stage, they may not yet be human-centered and therefore do not adequately support human decision-making. They carried out a study with AI systems for investment recommendation and found out that even models with explainability were passed over by the users surveyed when there was an investment recommendation made by human experts. Thus, they concluded that the knowledge of human experts, translated into advice, is more decisive for decision-making than the recommendations of an AI system, even those supported by explainability artifacts (Dikmen & Burns, 2022).

In his study, Shin (2021) conducted an experiment with users, presenting them with an AI system for product recommendation; for some of them, an XAI artifact was presented allowing them to identify why a certain recommendation was made by the system. The results showed that the explanation increased the users' confidence, as it allowed them to understand how the recommendation was built by the system; however, their ability to understand the explanation is an important point, as it impacts their emotional trust. In this way, the ability to understand the explanation of the result is as relevant as the explanation itself.

Lewis and Marsh (2022) argued that trust is a kind of human heuristic for decision-making. They then proposed a model in which human judgment about the reliability of the object is fed by subjective and situational factors, and this process is what leads to the behavior of trusting or not trusting the object. Thus, in their view, trust is a complex construct that cannot be achieved by tools such as guides or explainability artifacts, the latter applicable to AI systems.

To measure the trust in a XAI system, Hoffman *et al.* (2018) and Adams, Bruyn, and Houde (2003) proposed two main research questions: Do you trust the machine's outputs? Would you follow the machine's advice? In this current research, the confidence scale proposed by Cahour and Forzy (2009), described below in the methodological procedures, was adapted and applied.

There is still a lack of consensus in the literature on the effectiveness of explainability artifacts in increasing user confidence in AI systems. Additionally, the studies reviewed addressed common user trust in AI systems. However, they do not specifically address skilled users, that is, those professionals involved with ML modeling, for example, data scientists and business intelligence professionals. The present research focused on the study of the trust of these modeling professionals in AI algorithms, as they are responsible for creating the models and for the consequences of possible biases in the results. Thus, the hypothesis tested in the research is:

Ha. Explainability artifacts increase user confidence in a black-box AI system.

Method

Considering the study's aim, the research can be classified as exploratory and descriptive, as it verified whether the explanation artifacts increase a skilled user's trust in black-box models. The research was developed in two phases: 1) creation of the model; 2) assessment of users' trust in the AI system.

In phase one, an open dataset available on the Kaggle website was selected. This dataset related to the retail banking sector and contained information from a telemarketing campaign to offer and sell a financial product (long-term deposits) to bank customers. The artificial neural networks algorithm (multilayer perceptron classifier) was applied to this database in order to predict the customers who would buy the banking product offered in the campaign. The response variable was binary (0 = sale not made, 1 = sale made). After creating the

predictive model, the XAI LIME algorithm was applied to create an explanation for some cases of the test sample, allowing identification of which variables actually influenced the classification of the potential customer into the categories 0 or 1. The characteristics of the dataset and analyzes performed in this step are detailed in the results.

After the estimation of the predictive and explainability models, phase two of the research was performed, which consisted of the assessment of users' trust in the proposed AI system. This phase aimed to test the research hypothesis, evaluating whether offering the explanation about the prediction result increased user trust in the neural networks system. For this, a survey was applied, in a quasi-experimental design, with two groups: an experimental group, which received the explanation artifact along with the neural network results, and a control group, which received only the neural network prediction results. This methodological option met the objective of the research, since it enabled verification of whether there was a difference in the perception of trust between users who have access and those who do not have access to the mechanisms of explanation, making it possible to assess whether XAI has a significant effect on the users' mental model.

The literature reviewed in this research contained results from the assessment of trust in AI systems by nonspecialist users. For this reason, the present research chose to assess the trust of professional users (data scientists, business intelligence professionals) who develop this type of model within real organizations and whose impacts can actually affect their customers' lives and the reputation of their companies. In order to maintain the homogeneity of the experimental and control groups, professional data analysts from Brazilian banking institutions were invited to participate in the survey, given that this sector traditionally makes use of analytical techniques, including black-box models, for financial and risk analysis. Consequently, they were qualified to evaluate the results obtained by the neural network model and by LIME, making it unnecessary for the researcher to offer prior training for the task. In this way, the problem raised by Shin (2021) about the user's ability to understand the result of the model and of the explainability was eliminated. The survey form was composed as follows:

(1) Experimental group: consent form, homepage with information on the management problem being modeled, measurement of user trust in AI systems, presentation of the forecast results' tables without the explanation obtained by XAI and without any report of the meaning of the results, measurement of user trust after exposure to the model.

(2) Control group: consent form, homepage with information on the management problem being modeled, measurement of user trust in AI systems, presentation of forecast results tables with the explanation obtained by XAI and without any report of the meaning of the results, measurement of user trust after exposure to the model.

The trust scale was adapted from the work of Cahour and Forzy (2009), whose items were classified on a five-point Likert scale ranging from strongly agree to strongly disagree. The methodology developed by Cahour and Forzy (2009) emphasizes natural activity, as it is experienced by subjects, as trust is defined as a feeling, and therefore highly subjective and barely observable through the subject's behavior. The scale items can be seen in Table 1.

Data were collected online through the Qualtrics platform. A pilot test was carried out with three professionals from the target audience in order to identify inconsistencies in the questionnaire; however, only minor text corrections were required after the pilot. The potential respondents (banking sector professionals, experienced users with three or more years of experience in the field of data science) totaled 45 people who were individually contacted and invited to participate in the study. Consent was obtained from 17 of them. This sample was then randomly divided into two groups (experimental with nine respondents and

control with eight respondents) and each one was sent the corresponding questionnaire in July 2020. Finally, after data collection and analysis, interviews were undertaken with four participants in the sample, two in the control group and two in the experimental group. This step aimed to understand in greater depth the perceptions of professionals about the reliability of black-box algorithms. The interviews were conducted through video conferences using Skype software and were recorded with the participants' consent and transcribed for analysis.

Results

Estimation of the black-box neural network model and LIME explainability

This research addressed the problem of classifying customers who, through telemarketing calls, could buy long-term deposits. Telemarketing agents could make contact in two ways: (i) through phone calls to a list of clients; or (ii) by approaching the customer while they were contacting the company for another reason. The contact result had only two possible outcomes – unsuccessful or successful contact.

The dataset was available in the Kaggle repository. It had 41,118 observations (only 6,557 of them classified as successful), 20 independent variables were available, and the data were collected from May 2008 to November 2010. Python libraries were used to estimate the neural network, with the following steps: (1) screening and descriptive analysis of data; (2) data preparation; (3) training of neural networks; (4) assessment of model metrics and (5) creation of explainability artifacts.

For the first step (reading and descriptive analysis), pandas and pandas_profiling library were used to read the data source file and to create a report containing the description of the variables and their descriptive statistics (for continuous variables, minimum values, averages and maximums were calculated; for categorical variables, the distribution graphs were developed).

For the second step (data preparation), the pandas library was used to rename the variables to Portuguese. Transformation of categorical variables into dummy variables was also performed. After this step, the training and test data were split using the sklearn library, following the proportion of 80% for training data and 20% for test data. This separation was necessary to respect the proportion of the dependent variable (positive and negative) in both sets (training and test) and was performed without replacement and randomly. It should be noted that the test data were used only to validate the metrics and were not used at any time during the training stage.

For the third stage of neural networks training, the sklearn library was used to estimate the parameters and train the neural network model. To obtain the training parameters of the model, the random search strategy with cross validation was used. In this way, the training dataset was resampled and trained with different parameters in order to obtain the best result. In the case of this study, the model with the best receiver operating characteristic

Variables	I trust in AI systems	Likert scale	5 Strongly agree
	I consider the AI systems work well		4 Slightly agree
	The outputs from the AI systems are predictable		3 Indifferent
	AI is trustworthy. I can trust it to be correct all the time		2 Slightly disagree
	I am sure that when I trust AI systems, I will get the correct answers		1 Strongly disagree
	AI systems can do the task better than a human user		

Source(s): Adapted from Cahour and Forzy (2009)

Table 1. Trust in AI systems scale

(ROC) curve metric in the training data was used. The technique used was the multilayer perceptron classifier.

For the fourth step (assessment of the model metrics), the sklearn library was used to generate the classification performance report for the model trained in the previous step. As this was a black-box model, in this step, a comparison of the metrics of accuracy and area under the ROC curve in the training and test data was performed. These were the only possible measures for evaluating the quality of the model. The precision indicator deals with the percentage of successful cases (code 1) correctly classified among the total of cases classified as positive in the sample and can be obtained by equation (2), in which TP are the true positives (successful cases classified as success) and FP are the false positives (failure cases classified as success).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The recall deals with the percentage of successful cases correctly classified in relation to the total of cases with code 1 in the sample and can be obtained by equation (3), where FN are false negatives (success cases classified as failure).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The F1 considers both recall and precision and is obtained by the harmonic mean of the two measurements multiplied by two; being obtained by equation (4).

$$F1 = 2 \times \frac{recall \times precision}{recall + precision} \tag{4}$$

Finally, the area under the ROC curve varies between zero and one and consists of a performance measure of the binary classification model. In this way, when the threshold approaches zero, there is greater confidence that the case is classified as a failure. Similarly, the closer to one, the greater the confidence, that the case is classified as a success. All the neural network performance measures described are better when they are of a larger type. For the model created in the research, the accuracy results are shown in Table 2. Fair goodness of fit was observed, in which code 0 was the lowest result. However, as its value was close to 50% it can be considered reasonable. Close results on the training and test datasets also suggest the absence of an overfitting problem.

For the fifth and final step (creating the explainability artifacts), the LIME library was used. To create the explanation artifacts, it was necessary to predict the test dataset sorting data from the highest to the lowest value. Then, three observations with the highest probability of success classified as 1 and three observations with the lowest probability of failure classified as zero were selected. Figure 1 illustrates the explainability artifact obtained

Table 2. Neural networks' results for the training and test datasets

	Training				Test			
Class	Precision	Recall	F1	ROC	Precision	Recall	F1	ROC
Success Code 1 Successful sale	0.975	0.896	0.934	0.934	0.967	0.890	0.927	0.931
Failure Code 0 Unsuccessful sale	0.497	0.815	0.617		0.499	0.784	0.610	

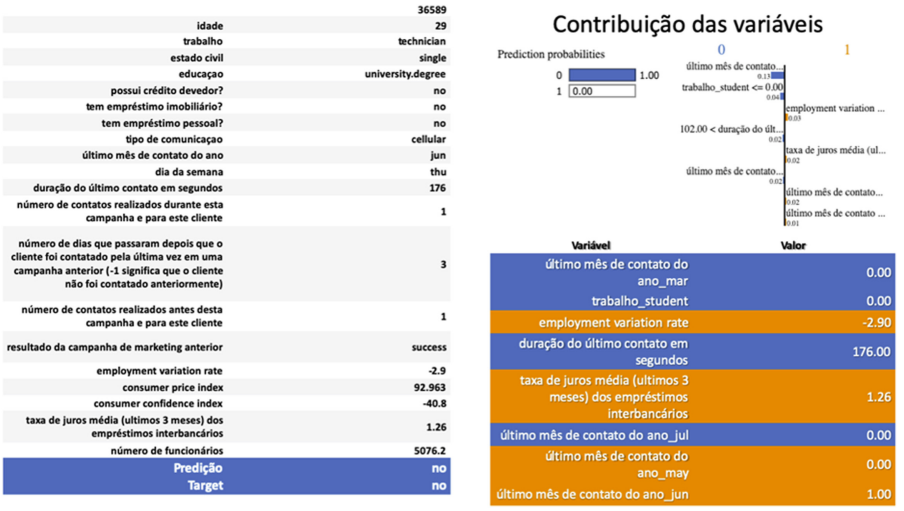

Figure 1.
Example of the LIME explainability artifact for cases of success

by the LIME algorithm for a case that had code zero in the dataset and was correctly predicted with code zero; that is, a case in which the sale was not performed. On the left side of the artifact are the data of the case being explained (case ID 36589 from the database), including its age, occupation, marital status and other corresponding variables. The blue lines on the left side (prediction and target) refer to how the case was predicted by the neural network (prediction) and its real value in the database (target), so this case was correctly predicted with code 0. On the right side of the artifact, the effective explanation of the XAI is observed, the blue lines indicate which variables had an impact and determined the classification of the case as failure; in this case, the last month in which the customer had been contacted by the sales team (dummy variable for months of March and July), the customer's occupation as a student, and the duration in seconds of the last telephone contact made with the customer. Thus, it can be seen that of the vast set of 20 predictor variables inserted in the network, only three were decisive in the classification. The orange lines represent the variables that would indicate the classification of the case as a success, but their weight in the algorithm decision was much lower, so that the classification occurred as a failure. The graphic illustrated in the upper right corner of the artifact shows the strength of the variable's influence and, in this case, it can be seen that the variables in blue were the determining factors.

In managerial terms, one can think of the following implications of the adoption of the neural network created as a tool for classifying potential customers who are likely to buy banking products. Based on the goodness of fit found by the model, it would make a useful tool to find potential customers. However, some possible problems should also be considered; for example, bothering customers who are not likely to buy banking products, since the model has a percentage of misclassification. Nevertheless, being used together with other possible tools in the sales area, the network can add value to the organization, even without the explanation artifact. From a different perspective, this model could be used to evaluate the performance of employees in the sales team; for example, by filtering the data by individual employee and verifying the probability of conversion of each salesperson, eventually rewarding those with the highest probability of success or punishing those less likely to succeed. From this perspective, the use of this model could be harmful, and there may be a bias in the evaluation. When considering only probabilities of success and failure, without

evaluating the determinants of these results, the manager would be ignoring a possible sample bias (some salespeople may have been allocated to a more resistant customer database, for example). The explanation artifact, by opening the classifications of each customer, allows evaluation of the reasons for success and failure and the establishment of different and perhaps more effective approach strategies for the sales team. Poor sales performance does not necessarily imply poor salesperson performance, but perhaps an inadequate customer approach strategy. The explanation artifact allows a very objective assessment of each case individually in a business intelligence action.

Trust in black-box systems

The experimental and control groups, as described in the methodological procedures, completed the trust scale questionnaire before and after observing the neural network results. To assess the difference between the groups, before and after, nonparametric tests of significance were applied, considering the ordinal nature of the Likert scale and the small sample size. Table 3 presents the descriptive statistics for both experimental and control groups, before and after the presentation of the model results. The value of the means demonstrates higher scores in general trust and on the AI systems' functioning.

Wilcoxon's nonparametric test for paired samples was applied to assess the difference in perception before and after seeing the model results. There was no significant difference between both moments (before and after receiving the model results) in both the experimental and control groups (Table 4). These results corroborate the results found in the research carried out by Cahour and Forzy (2009) and Lewis and Marsh (2022).

After the survey, four participants were invited for in-depth interviews, two from each research group. Respondents, when asked directly about trusting AI systems, reported things that exceeded explanation artifacts, such as the dataset quality and the absence of bias in the input data as critical aspects to the reliability of predictions generated by black-box systems. In addition to the concern regarding the methodology applied to the data used in the construction of the model, aspects such as the nature and complexity of the problem, as well as the maturity of the areas that use AI systems for decision-making, were pointed out. When asked about the proper functioning of AI systems, respondents emphasized the importance of applying algorithms to good quality databases. Regarding the predictability of the outputs of AI systems, the creator's knowledge and the ability to adjust the algorithm were raised as

Table 3. Average values for trust

Variable	Control group		Experimental group	
	Before explanation (T_0)	After explanation (T_1)	Before explanation (T_0)	After explanation (T_1)
I trust in AI systems	4.13	4.13	4.00	4.11
I consider the AI systems work well	4.25	4.13	4.00	4.00
The outputs from the AI systems are predictable	3.25	3.38	2.33	2.44
AI is trustworthy. I can trust it to be correct all the time	3.13	3.25	2.44	2.67
I am sure that when I trust AI systems, I will get the correct answers	3.13	3.50	3.22	3.00
AI systems can do the task better than a human user	3.38	3.75	3.56	3.44

Group	Variables compared	*P*-value
Experimental	$P1(T_0) - P1(T_1)$	0.317
	$P2(T_0) - P2(T_1)$	1.000
	$P3(T_0) - P3(T_1)$	0.564
	$P4(T_0) - P4(T_1)$	0.157
	$P5(T_0) - P5(T_1)$	0.577
	$P6(T_0) - P6(T_1)$	0.564
Control	$P1(T_0) - P1(T_1)$	1.000
	$P2(T_0) - P2(T_1)$	0.317
	$P3(T_0) - P3(T_1)$	0.317
	$P4(T_0) - P4(T_1)$	0.655
	$P5(T_0) - P5(T_1)$	0.257
	$P6(T_0) - P6(T_1)$	0.083

Table 4. Non-parametric tests results

influential. On the other hand, the perceptions regarding the predictability and trust in the outputs of AI systems did not appear to be directly related.

The application of AI systems in the financial area was also mentioned as a reliability factor, as financial is an area that has been using analytical methods for a long time. When asked whether the outputs of the systems would always be correct, respondents cited human intervention in the creation of an AI system as a factor that negatively impacts trust, as human bias can be transferred to the model. The error rates of the systems were also cited as a factor that limits trust, but some error is intrinsic to the models; however, if error is high, it undermines trust. On the other hand, a model with great accuracy does not inspire trust, generating suspicion of an estimation problem such as overfitting. Regarding the application of the models to different real contexts, the interviewees showed concern about the adequacy of these models in different situations, emphasizing the need to monitor the system's performance. Another point considered was the comparison of the performance of the AI system against an a priori expectation, and that by confirming the user's expectations the model manages to arouse their trust.

Regarding the ability of an AI system to execute a task performed by a human, the interviewees observed that the computational capacity for information processing and the ability to identify and replicate behavior were advantages of AI in performing tasks. The type of task to be performed by the AI system was cited as an important aspect affecting trust; for example, for punctual and routine tasks, respondents reported trust in the system, but recognized potential problems. Summarizing the data collected in the interviews, it was possible to identify the trust factors highlighted in Table 5.

Trust factors	Information obtained from the interviews
Data	Concern about the quality of input data for estimating AI systems
Method	Estimation steps, choice of the audiences, data and risk mitigation
Maturity	Experience in creating and solving problems using AI
AI system development	Trust in the developers of AI systems
AI system usage	Human intervention during the use of AI systems
Algorithm limitation	Errors associated with the outputs of the algorithms used to create the AI systems
Application context volatility	Changing the application context of AI systems after their estimation

Table 5. Trust factors based on the interviewees' points of view

Conclusions

The research hypothesis, which evaluated the positive influence of trust in black-box models supported by explainability artifacts, was not supported, given the nonsignificance of nonparametric tests. This result, together with the results of the interviews, demonstrates a high level of trust in AI systems, with or without explanation artifacts. The interviews suggested some factors that influence user trust, and explanation artifacts were just one of these factors. It is noteworthy, however, that the sample studied was composed of professionals from financial institutions, working in the area of data analysis and, therefore, they were mature connoisseurs of AI models with well-established mental models about these systems.

The explanation artifacts shown to the research participants did not influence the change in the existing mental model, corroborating Cahour and Forzy's (2009) claim that trust in AI systems is linked to predictability and expectations about the outputs of the systems. The maturity of the financial area participating in this study proved to be a potential factor, so that the difference in trust was not significant, but for other application areas the results may be different. Factors related to the methods used to create an AI system and the representativeness of the data was cited as determinants by the interviewees.

The survey results, in light of the reviewed literature, show that there are different audiences affected by AI. a) The model developers, studied in this research, who have a broad view of modeling from the assessment of data quality, to the objectives of forecasting and its results. For this audience, the artifacts of explanation are just another measure to be observed in the assessment of the model. b) Managers who receive the results of the models to support their decision-making. This audience was not studied in this research, but has a great potential to benefit from the artifacts of explanation, if able to understand the information brought by these tools. c) The final consumer, who is classified by the forecast models, possibly the most vulnerable people of those mentioned, since sometimes they may not even be aware that their demand was classified by an AI algorithm.

From a practical point of view, the research results demonstrate the concern of the professionals who generate the models with the quality of the model and reduction of bias, but when considering managers and final consumers, there is still scope for additional research, which can support regulation policies for the use of this type of system.

It is concluded that the literature on trust factors in AI systems is still in an initial phase when observed from the quantitative aspect. The different applications of these systems affect trust in multiple dimensions. Finally, the study of only one type of audience, the developers of AI systems and only one economic sector, retail banking, in addition to the sample size, are pointed out as limitations of the research. However, due to the high specialization and experience of these professionals, who work in large Brazilian banks, and the homogeneity of the banking sector, a good representation of the results obtained is assumed.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Adams, B. D., Bruyn, L. E., & Houde, S. (2003). Trust in automated systems (report). Ministry of National Defence.

Barocas, S., Friedler, S., Hardt, M., Kroll, J., Venka-Tasubramanian, & Wallach, H. (2018). The FAT-ML workshop series on fairness, accountability, and transparency in machine learning. available from: http://www.fatml.org/

Brynjolfsson, E., & McAfee, A. (2017). Artificial intelligence, for real. *Harvard Business Review*. Available from: https://hbr.org/cover-story/2017/07/the-business-of-artificial-intelligence.

Cahour, B., & Forzy, J.-F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science*, *47*(9), 1260–1270. doi: 10.1016/j.ssci.2009.03.015.

Chen, Y. W., Yang, J. B., Pan, C. C., Xu, D. L., & Zhou, Z. J. (2015). Identification of uncertain nonlinear systems: Constructing belief rule-based models. *Knowledge-Based System*, *73*, 124–133.

Confalonieri, R., Weyde, T., Besold, T. R., & Moscoso del Prado Martín, F. (2021). Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, *296*, 1–20. doi: 10.1016/j.artint.2021.103471.

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, *162*, 1–11. doi: 10.1016/j.ijhcs.2022.102792.

Farquad, M. A. H., Ravi, V., & Raju, S. B. (2012). Analytical CRM in banking and finance using SVM: A modified active learning-based rule extraction approach. *International Journal of Electronic Customer Relationship Management*, *6*(1), 48–60. doi: 10.1504/ijecrm.2012.046470.

Gléonnec, M. (2004). Confiance et usage des technologies d'information et de communication. *Consommations et Sociétés*, *4*, 1–18.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects, *ARXIV*, 1-50. doi: 10.48550/arXiv.1812.04608.

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, *294*, 1–25. doi: 10.1016/j.artint.2021.103459.

Lewis, P. R., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. *Cognitive Systems Research*, *72*, 33–49. doi: 10.1016/j.cogsys.2021.11.001.

Lu, H., Li, Y., Chen, M., Hyoungseop, K., & Serikawa, S. (2018). Brain intelligence: Go beyond artificial intelligence. *Mobile Networks and Applications*, *23*(2), 368–375. doi: 10.1007/s11036-017-0932-8.

Moreira, C., Chou, Y.-L., Velmurugan, M., Ouyang, C., Sindhgatta, R., & Bruza, P. (2021). LINDA-BN: An interpretable probabilistic approach for demystifying black-box predictive models. *Decision Support Systems*, *150*, 1–16. doi: 10.1016/j.dss.2021.113561.

Narwaria, M. (2022). Does explainable machine learning uncover the black box in vision applications?. *Image and Vision Computing*, *118*, 1–4. doi: 10.1016/j.imavis.2021.104353.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown Publishing Group.

Pereira, S., Meier, R., McKinley, R., Wiest, R., Alves, V., Silva, C. A., & Reyes, M. (2017). Enhancing interpretability of automatically extracted machine learning features: Application to a RBM-random forest system on brain lesion segmentation. *Medical Image Analysis*, *44*, 228–244. doi: 10.1016/j.media.2017.12.009.

Preece, A. (2018). Asking 'why' in AI: Explainability of intelligent systems—perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, *25*, 63–72.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should i trust you?' Explaining the predictions of any classifier. *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 1135–1144. doi: 10.1145/2939672.2939778.

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). GLocalX - from local to global explanations of black box AI models. *Artificial Intelligence*, *294*, 1–24. doi: 10.1016/j.artint.2021.103457.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 1–10. doi: 10.1016/j.ijhcs.2020.102551.

van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. *Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence* (pp. 900–907). San Jose, CA: AAAI Press.

Wang, F., & Rudin, C. (2015). Falling rule lists. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, *38*, 1013–1022.

Corresponding author
Daielly Mantovani can be contacted at: daielly@usp.br