Antonio Lepore
Biagio Palumbo
Jean-Michel Poggi *Editors*

Interpretability for Industry 4.0 : Statistical and Machine Learning Approaches

Springer

Interpretability for Industry 4.0 : Statistical and Machine Learning Approaches

Antonio Lepore • Biagio Palumbo •
Jean-Michel Poggi
Editors

Interpretability for Industry 4.0 : Statistical and Machine Learning Approaches

Springer

Editors
Antonio Lepore
University of Naples Federico II
Naples, Italy

Biagio Palumbo
University of Naples Federico II
Naples, Italy

Jean-Michel Poggi
University Paris-Saclay
Orsay, France

ISBN 978-3-031-12401-3 ISBN 978-3-031-12402-0 (eBook)
https://doi.org/10.1007/978-3-031-12402-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Interpretability is a key issue to develop insightful statistical and machine learning (ML) approaches in business and industry. This book aims to provide the readers with a compact, stimulating, and multifaceted introduction to this emerging promising topic.

The contents covered by the volume were stimulated by the ENBIS (European Network for Business and Industrial Statistics, https://enbis.org/) Workshop, Interpretability for Industry 4.0, which was held at the University of Naples Federico II Italy on July 12–13, 2021 https://conferences.enbis.org/event/8/ and offered real-world industrial motivations and deep methodological insights on the topic of interpretability. The workshop was divided into the following three pillars:

- Analyze and propose monitoring tools for additive manufacturing systems.
- Explore the connections between ML tools, sensitivity analysis, and rule-based systems.
- Exploit the contribution of generalized additive models for the development and visualization of interpretable statistical models.

Each half day, devoted to a specific pillar, ended with a roundtable providing a closing discussion challenging the different views of interpretability.

The book collects contributions issued from this workshop and accepted after a review process. It contains four chapters, the first one related to roundtables and three other chapters related to the three pillars mentioned above. Each chapter can be read independently with its own bibliography.

Chapter 1 introduces different views of interpretability in the context of Industry 4.0. It is organized in three different sections, after an introductory discussion about the concepts of explainability and interpretability of ML models. The chapter offers a philosophical discussion about the implications of ML interpretability for scientific and industrial studies and extends the concept in many directions, such as the generalizability of model outputs and implications for Industry 4.0 applications. The last section provides the reader with more specific materials and is dedicated to the connections between ML model interpretability and sensitivity analysis.

Chapter 2 discusses how the advent of artificial intelligence for manufacturing data mining poses new challenges to model interpretability in contrast with the concept of explainability. Starting from a general overview, the chapter focuses on examples of big data mining in additive manufacturing. A real case study focusing on spatter modeling for process optimization is discussed, where a solution based on robust functional analysis of variance is proposed.

Chapter 3 proposes a contribution to interpretability via random forests (RF) and is organized into two sections after an introduction about the need in different applications and domains for interpretable ML models and points out the requirements desired for interpretable methods. The chapter describes an original way to use RF to produce a compact set of rules and offers a thorough overview and analysis of permutation variable importance measures based on RF and describes new variants.

Chapter 4 formally introduces generalized additive models (GAMs) and flexible GAMs for location shape and scale (GAMLSS) as excellent models to achieve interpretability in the model building, as well as for communicating modelling results. Structural assumptions to avoid the curse of dimensionality in the modelling of the effect of a covariate vector on the distribution of a response variable are discussed. In particular, the additive assumption, on which GAMs rely, ensures scalability in the number of covariates and computational convenience in model fitting. The closing section of the chapter shows how to practically apply GAM and GAMLSS models via the `mgcv` and `mgcViz` R packages.

We are grateful to all the authors for their challenging perspectives on yet non-consolidated topics but highly relevant in supporting human decisions. We warmly thank anonymous referees for their conscientious reviews and Eva Hiripi from Springer-Verlag for supporting this project.

Naples, Italy	Antonio Lepore
Naples, Italy	Biagio Palumbo
Orsay, France	Jean-Michel Poggi
April 2022	

Contents

1 Different Views of Interpretability 1
Bertrand Iooss, Ron Kenett, and Piercesare Secchi

2 Model Interpretability, Explainability and Trust for Manufacturing 4.0 21
Bianca Maria Colosimo and Fabio Centofanti

3 Interpretability via Random Forests 37
Clément Bénard, Sébastien Da Veiga, and Erwan Scornet

4 Interpretability in Generalized Additive Models 85
S. N. Wood, Y. Goude, and M. Fasiolo

Chapter 1
Different Views of Interpretability

Bertrand Iooss, Ron Kenett, and Piercesare Secchi

Abstract Interpretability, in the context of machine learning, means understanding the predictions made by the machine learning algorithm, with the aim to support human decisions based on them. In this view, interpretability can involve identifying the input features which drive the predictions. This chapter develops different issues and related methodologies of interpretability of machine learning models. Their implication for scientific and industrial studies are firstly developed. Then, the links between the generalizability of model outputs and interpretability are discussed. Finally, the deep connection between the settings of the machine learning interpretability and the ones of the model output sensitivity analysis is described.

1.1 Introduction

Machine Learning (ML) is one of the substantial branches of artificial intelligence technology and provides a large panel of algorithmic tools to learn from data (e.g., numerical data, images, sounds, texts). However, a severe drawback is that ML algorithms may provide predictions which turn out to be difficult to explain or interpret. From a general point of view, allowing an understandable explanation for any ML model output helps anybody (e.g., an operator, a decision-maker, a statistician or an analyst) to catch the underlying reasoning. Such a property may have positive consequences such as making the debugging process easier, helping for model improvement and acceptability of the tool. Therefore, industrial

B. Iooss (✉)
EDF R&D, Chatou and SINCLAIR AI Lab, Saclay, France
e-mail: bertrand.iooss@edf.fr

R. Kenett
KPA Group and Samuel Neaman Institute, Technion, Israel
e-mail: ron@kpa-group.com

P. Secchi
Department of Mathematics, Politecnico di Milano, Italy
e-mail: piercesare.secchi@polimi.it

A. Lepore et al. (eds.), *Interpretability for Industry 4.0 : Statistical and Machine Learning Approaches*, https://doi.org/10.1007/978-3-031-12402-0_1

deployment of these solutions requires tools together with a panel of best practices to perform explainable and interpretable ML. These two terms “explainability” and “interpretability” will be discussed, hereafter, in the context of ML.

There is a lack of consensus about rigorous definitions of explainability and interpretability of ML models. Indeed, these notions refer to profound cognitive processes related to social sciences and to their different fields of applications (e.g., medical sciences, law and justice, engineering) or scientific communities (e.g., natural language processing and computer vision). Some authors also invoke other fundamental concepts (see, e.g., completeness, fairness, intelligibility, comprehensibility, transparency) to build a proper definition of what “explainable AI” is and what it is intended for [5, 23]. In this chapter, we only focus on the ML interpretability as the property related to the ability of a ML model (or any element related to this model, i.e., inputs, outputs, predictions) to be associated with concepts held by a human being.

Interpretability, in the context of ML, means understanding the predictions made by the ML algorithm, with the aim to support human decisions based on them. In this view, interpretability can involve identifying the input features which drive the predictions. The goal of this chapter is to focus on different important issues and related methodologies of interpretability of ML models.

Firstly, the implication of ML interpretability for scientific and industrial studies are developed (Sect. 1.2). Then, the links between the generalizability of model outputs and interpretability are discussed, providing a high-level view with its implications to Industry 4.0 applications (Sect. 1.3). The approach presented combines an engineering perspective with empirical modeling and soft data in a blended hybrid view which integrates technical and non-technical perspectives. Finally, the deep connection between the settings of the ML interpretability and the ones of the model output sensitivity analysis is described (Sect. 1.4). This connection, which is still underdeveloped, offers rich perspective for cross-fertilizing techniques of both research fields [62].

1.2 Interpretability: In Praise of Transparent Models

The focus of ML is the design of algorithms that learn from a training data set how to associate an input to an output. The training data set must be massive since the learning curve of an ML algorithm increases very slowly with the size of the data set. The performance of a trained algorithm is typically evaluated on the task of prediction and validated with a hold-out data set. If future data are generated by the same population from which the training data set has been drawn, well trained ML algorithms are often excellent predictors.

Interpretability, in the context of ML, means understanding the predictions made by the algorithm, in order to support human decisions based on them. Interpretability involves identifying, for example with the tools of sensitivity analysis, the input features which drive the predictions. In this section we discuss if interpreting predictions, important as they are, is sufficient for science and for industry.

Quoting Carlo Rovelli's recent essay, *Helgoland* [64]: "*The goal of science is not that of making predictions. Science also aims at presenting an image of reality, a conceptual framework where to think about things. This is the ambition which made the scientific thinking successful. If predictions were the only goal of science, Copernicus would not have discovered anything different from Ptolemy: his astronomical predictions were not any better than those of Ptolemy. But Copernicus found the key for rethinking all and for better understanding it*". This passage presents predictions as a partial objective indicating that interpretability should consider a wider scope.

A conceptual framework where to think about things is often required in business and industry. The broader question it therefore when and why, in science as well in business or in industry, do we use data? Briefly, we use data to answer three questions: what happened, what will happen, and what shall be done to make it happen.

1.2.1 What Happened?

This is the question tackled by exploratory and descriptive analyses that, starting from raw data, organize them, fuse different and heterogenous sources impinging upon the same population, sort them out deciding about the relevant and the irrelevant, clean and transform data, graphically represent and summarize the information sufficient for the goal of the analysis, already driving it toward certain hypotheses and conjectures. For being effective, and not mystifying, an exploratory analysis must be open, totally transparent and highly dialectical. Through it the different stakeholders, who promoted the final questions moving the problem tackled with data, should be guided to better formalizations and prioritizations. This is an intensely Human Intelligence (HI) stage, where the data scientists are, explicitly or implicitly, guided by models. For instance, when they discriminate between the features of interest and those that will in fact not be measured and recorded, or when they choose the proper mathematical representation for data. Are the atoms of the analysis time series or functional data? How should time dependence be captured within each datum, explicitly through the autocorrelation function or implicitly by imposing certain smoothness and regularity to their functional representation? Researchers are in fact usually called to decide the specific mathematical space to embed data at hand and thus the geometry that allows for distinct projections and dimensionality reductions, the main mathematical tools for compressing and transferring the sufficient and relevant information.

Picture Galileo entering the Pisa Cathedral and observing the swinging chandelier. Did he observe and record the temperature and humidity of the air in the room in previous days and months, the number of people attending mass, their gender or their age, the phase of the moon, the hour and the day of the year, etc.? In the big data era he might have, but at the end of the sixteenth century [52] Galileo decided to focus only on the periods of the pendulum, the amplitude of the swing, the length of the rod and the mass of the bob. All other data were discarded and considered in advance as non-influential, even before measuring them. Surely this must have been decided based on intuition, a model which was forming in his mind, about the not yet formalized isochronism law of the pendulum, which in the following decades elected it as the disruptive new technological device for timekeeping.

1.2.2 What Will Happen?

This is the stage when we want to make predictions. We use training data, validation data and test data, to build predictor machines and evaluate their performances. These can be transparent models, like generalized additive models mixing endogenous and exogeneous variables, opaque models, like random forests, or black boxes, like deep neural networks. They can be subjected to natural interpretability—at least for the educated data scientist—or they might be inaccessible and require the tools of sensitivity analysis to elicit the contribution of the input features, and their interactions, to generate the final output. Uncertainty quantification is here a must. Different approaches have indeed been cleverly elaborated in the past centuries for the purpose—frequentist inference, Fisherian inference, Bayesian inference, Monte Carlo methods, bootstrap, cross-validation. The very concept of uncertainty has been fragmented many times—aleatoric uncertainty, epistemic uncertainty, forward uncertainty propagation, inverse uncertainty quantification, etc.

1.2.3 What Shall be Done to Make It Happen?

This is the realm of prescriptive analysis and experimental design. Assuming some input data provided by idealized scenarios and given the predictions offered by the models, what actions should be taken in order to generate the desired effects, with a certain degree of certainty. Once more, quantification of uncertainty is a must. But how should sensible scenarios be built? Can they be totally ignorant of the past as captured and represented by the sufficient summaries provided by the exploratory and descriptive analyses? A domain-based HI, transparency and a dialectical perspective are the effective trading tools. Indeed, here again the transparent models—in science, business and industry—are the empowering tools for sharing empirical and «experiential»knowledge, across different teams and units, across generations of scientists, engineers and experts.

In fact, the question is at the basis of experimental design and significance testing. Quoting Fisher [21]: "*We may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us statistically significant results.*"

1.2.4 Patterns and Models

A model is not a magic box. Its value resides in its power to amplify the human thought. This happens if the model is able to represent the interactions and the dependencies among the variables that both the stakeholders and the data scientist believe are describing the system under scrutiny. A model allows for sensible decision making through the action on independent input variables; a model provides scope for simulation and manipulation of the system under scrutiny.

Transparent models, opaque models and black boxes permit interpretability which enable an incremental upgrade of the human knowledge. This requires more than an automated and theory-free data analysis. Contrary to Chris Anderson's dictum on *Wired* [2], "correlation is NOT enough".

Beside predictions, the other selling point of ML algorithms is their ability to find patterns in massive data without the intermediation of a theory, without moving through the slow process of identifying a formal reference system within which questions could be asked and hypothesis could be stated.

As argued by James McAllister [47], any "*empirical data set can be decomposed into any one of all conceivable patterns and an associated noise term*." Hence, only two options seem admissible. Either we assume the ability of the ML algorithm to discern among patterns those that are indicative of real structures of the world and those that are not. Or, in the absence of an automatic criterion implementing this discrimination, we are forced to follow McAllister's argument and "*deny that any such ontologically significant distinction between patterns can be drawn, to admit that all patterns exhibited in empirical data sets correspond to structures in the world, and then to consider the meaning and implications of this claim*."

Without entering any further into this intriguing philosophical debate, let us notice that if one is looking for patterns, the analysis of a big data set formed by the decimal digits in the expansion of Pi, the ratio of any circle's circumference to its diameter, has the potential to fill one's life. Pi is an irrational and transcendent number, whose approximate representation through a decimal expansion is often used as a test for evaluating the power of new supercomputers; the last record has been broken in August 2021, when Pi has been accurately approximated to 62.8 trillion decimal places [73]. By using the MyPiday [75] search engine, one of us found among the digits of Pi his birthday, that of his wife, the day they married and the birthdays of their children… a pattern of a certain relevance, at least to him. An (unproven) conjecture states in fact that Pi is a normal number (see, e.g., Arndt and Haenel [4]), which would also imply that "every finite string of numbers eventually occurs in Pi". Structures representable by finite string of numbers should be able to accommodate the answers to all problems business and industry might want a data scientist to solve, and yet we would not consider as reasonable and practical to search the digits of Pi for finding them. The problem being that, without knowing it in advance, we will not be able to recognize a relevant and correct answer if we met it among the decimal digits of Pi, although we know that is there…

This sounds as an anti-climax for the believers in the automatic heuristic power of the ML algorithms, but it should not. The problem is with the “automatic” qualification. We could indeed use ML to explore data in search for patterns if we admit that our search is driven by our intention, by the objectives of our endeavor. It is the intention driven by the goal of the analysis and framed within a theory, explicitly or implicitly formalized, which generates the relevant conjectures and the hypothesis the data could be challenged with; it is intention which puts the data scientist in the position to decouple the patterns within the data deemed to be relevant from those that are not. Eliciting this intentionality is more easily achieved with a transparent model, where interpretation is—to paraphrase Karl Pearson [70]—“on the table”, but could in principle also be obtained with the tools of Artificial Intelligence, through a stronger and still unusual effort and the development of new mathematical—and transparent—perspectives.

1.3 Generalizability and Interpretability with Industry 4.0 Implications

In this section we focus on the process of moving from numbers to data, to information and insights [31]. In the information quality framework, this is called “generalizability” an expanded form of “interpretability” [36]. The section covers interpretable artificial intelligence (AI), wide angle of statistical generalizability.

1.3.1 Introduction to Interpretable AI

Artificial Intelligence (AI) has focused on predictive analytics with success reflected by sophisticated black box models. In recent years, the need to interpret and explain the factors affecting analytic predictions has risen. To achieve this, various methods have been proposed to help users interpret the predictions of complex models. Lundberg and Lee [44] introduce SHAP (SHapley Additive exPlanations), a unified framework for interpreting predictions. SHAP assigns to each feature in the model an importance value for a particular prediction. It includes the identification of additive feature importance measures and theoretical results showing there is a unique SHAP solution with a set of desirable properties.

Local interpretable model-agnostic explanations (LIME) is a local surrogate interpretable model used to explain individual predictions of black box ML models [49, 63]. Surrogate models approximate the outputs of a black box model [13, 22]. LIME is based on local surrogate models used to explain individual predictions. In a first step, LIME uses the black box model to get model predictions, ignoring the training data. The objective is then to understand why the ML model gives a certain prediction. LIME generates a new dataset using perturbed samples and