Intelligible and Explainable Machine Learning: Best Practices and Practical Challenges

Rich Caruana
rcaruana@microsoft.com
Microsoft Research
Redmond, WA

Scott Lundberg
scott.lundberg@microsoft.com
Microsoft Research
Redmond, WA

Marco Tulio Ribeiro
marco.correia@microsoft.com
Microsoft Research
Redmond, WA

Harsha Nori
hanori@microsoft.com
Microsoft Corporation
Redmond, WA

Samuel Jenkins
sajenkin@microsoft.com
Microsoft Corporation
Redmond, WA

ABSTRACT

Learning methods such as boosting and deep learning have made ML models harder to understand and interpret. This puts data scientists and ML developers in the position of often having to make a tradeoff between accuracy and intelligibility. Research in IML (Interpretable Machine Learning) and XAI (Explainable AI) focus on minimizing this trade-off by developing more accurate interpretable models and by developing new techniques to explain black-box models. Such models and techniques make it easier for data scientists, engineers and model users to debug models and achieve important objectives such as ensuring the fairness of ML decisions and the reliability and safety of AI systems. In this tutorial, we present an overview of various interpretability methods and provide a framework for thinking about how to choose the right explanation method for different real-world scenarios. We will focus on the application of XAI in practice through a variety of case studies from domains such as healthcare, finance, and bias and fairness. Finally, we will present open problems and research directions for the data mining and machine learning community.

What audience will learn:

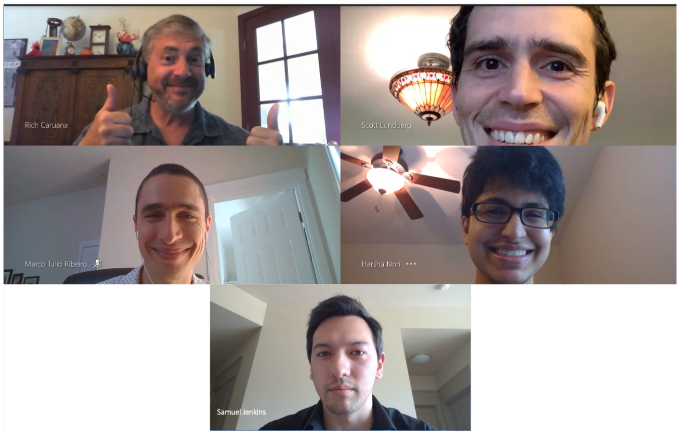
- When and how to use a variety of machine learning interpretability methods through case studies of real-world situations.
- The difference between glass-box and black-box explanation methods and when to use them.
- How to use open source interpretability toolkits that are now available

KEYWORDS

interpretability, intelligibility, responsible AI

ACM Reference Format:

Rich Caruana, Scott Lundberg, Marco Tulio Ribeiro, Harsha Nori, and Samuel Jenkins. 2020. Intelligible and Explainable Machine Learning: Best Practices and Practical Challenges. In *26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3394486.3406707



BIO

Rich Caruana is a senior principal researcher at Microsoft Research. Previously, he was on the faculty in the Computer Science Department at Cornell University, at UCLA's Medical School, and at Carnegie Mellon University's Center for Learning and Discovery. Rich received an NSF CAREER Award in 2004 (for meta clustering); best paper awards in 2005 (with Alex Niculescu-Mizil), 2007 (with Daria Sorokina), and 2014 (with Todd Kulesza, Saleema Amershi, Danyel Fisher, and Denis Charles); co-chaired KDD in 2007 (with Xindong Wu); and serves as area chair for Neural Information Processing Systems (NIPS), International Conference on Machine Learning (ICML), and KDD. His research focus is on learning for medical decision making, transparent modeling, deep learning, and computational ecology. He holds a PhD from Carnegie Mellon University, where he worked with Tom Mitchell and Herb Simon. His thesis on multi-task learning helped create interest in a new subfield of machine learning called transfer learning.

Marco Tulio Ribeiro is a researcher at Microsoft Research, in the Adaptive Systems and Interaction group. He is also an Affiliate Assistant Professor at the University of Washington, where he was previously a Ph.D student advised by Carlos Guestrin and Sameer

KDD '20, August 23–27, 2020, Virtual Event, USA
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7998-4/20/08.
https://doi.org/10.1145/3394486.3406707

Singh. Marco's research focus is helping humans interact with machine learning models meaningfully. That involves interpretability, trust, debugging, feedback, etc.

Scott Lundberg is a senior researcher at Microsoft Research. Before joining Microsoft, he did his Ph.D. studies at the Paul G. Allen School of Computer Science & Engineering of the University of Washington working with Su-In Lee. His work focuses on explainable artificial intelligence and its application to problems in medicine and healthcare. This has led to the development of broadly applicable methods and tools for interpreting complex machine learning models that are now used in banking, logistics, sports, manufacturing, cloud services, economics, and many other areas.

Harsha Nori is a data scientist at Microsoft who works on responsible AI software. He is a co-founder and core developer of InterpretML, an open source toolkit for training interpretable machine learning models and explaining blackbox systems.

Samuel Jenkins is a data scientist at Microsoft who works on responsible AI software. He is a co-founder and core developer of InterpretML, an open source toolkit for training interpretable machine learning models and explaining blackbox systems.