Explainable AI in Industry

Krishna Gade
Fiddler Labs
krishna@fiddler.ai

Sahin Cem Geyik
LinkedIn
sgeyik@linkedin.com

Krishnaram Kenthapadi
LinkedIn
kkenthapadi@linkedin.com

Varun Mithal
LinkedIn
vamithal@linkedin.com

Ankur Taly
Fiddler Labs
ankur@fiddler.ai

ABSTRACT

Artificial Intelligence is increasingly playing an integral role in determining our day-to-day experiences. Moreover, with proliferation of AI based solutions in areas such as hiring, lending, criminal justice, healthcare, and education, the resulting personal and professional implications of AI are far-reaching. The dominant role played by AI models in these domains has led to a growing concern regarding potential bias in these models, and a demand for model transparency and interpretability [6]. In addition, model explainability is a prerequisite for building trust and adoption of AI systems in high stakes domains requiring reliability and safety such as healthcare [1] and automated transportation, and critical industrial applications with significant economic implications such as predictive maintenance, exploration of natural resources, and climate change modeling.

As a consequence, AI researchers and practitioners have focused their attention on explainable AI to help them better trust and understand models at scale [8, 9, 19]. The challenges for the research community include (i) defining model explainability, (ii) formulating explainability tasks for understanding model behavior and developing solutions for these tasks, and finally (iii) designing measures for evaluating the performance of models in explainability tasks.

In this tutorial, we will present an overview of model interpretability and explainability in AI [4], key regulations/laws, and techniques/tools for providing explainability as part of AI/ML systems [7]. Then, we will focus on the application of explainability techniques in industry, wherein we present practical challenges/ guidelines for using explainability techniques effectively and lessons learned from deploying explainable models for several web-scale machine learning and data mining applications. We will present case studies across different companies, spanning application domains such as search and recommendation systems, sales, lending, and fraud detection. Finally, based on our experiences in industry, we will identify open problems and research directions for the data mining/machine learning community.

ACM Reference Format:
Krishna Gade, Sahin Cem Geyik, Krishnaram Kenthapadi, Varun Mithal, and Ankur Taly. 2019. Explainable AI in Industry. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332281

1 OUTLINE OF THE TUTORIAL

This tutorial is aimed at attendees with a wide range of interests and backgrounds, including researchers interested in knowing about model interpretability and explainability in AI, key regulations/laws, and explainability notions/techniques as well as practitioners interested in implementing explainable models for web-scale machine learning and data mining applications. We will not assume any prerequisite knowledge, and present the intuition underlying various explainability notions and techniques to ensure that the material is accessible to all KDD attendees.

The tutorial will consist of two parts: foundations including motivation, definitions, models, algorithms, tools, and evaluation for explainability in AI/ML systems (1.5 hours) and case studies across different companies, spanning application domains such as search and recommendation systems, sales, lending, and fraud detection (1.5 to 2 hours).

Tutorial webpage:
https://sites.google.com/view/kdd19-explainable-ai-tutorial

1.1 Foundations

In the first part of the tutorial, we will cover the following key topics:

- Motivation
  - Need for Transparency and Explainability in AI
  - Model Validation: Validation metrics, such as classification accuracy, are an incomplete description of most real-world tasks.
  - Scientific Consistency (beyond statistical consistency)
- Definitions
  - Feature Importance
  - Model Internals
  - Explaining by Examples
  - Intrinsic Interpretable Models
- Tasks: Post-hoc Explainability
  - Explaining Model Behavior Globally [12]. A global surrogate model is an interpretable model that is trained to approximate the predictions of a black box model.

KDD ’19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332281

  - Explaining Model Behavior Locally. Local surrogate models [15] are interpretable models that are used to explain individual predictions of black box machine learning models.
  - Example-based explanation methods select particular instances of the dataset to explain the behavior of machine learning models or to explain the underlying data distribution.
  - Explaining Model Differences
- Tasks: Explainability By Design
  - Designing explainable models for prediction [3, 10, 11, 13, 20].
- Tasks: Explainability by Examples: Prototypes, Criticisms, Adversarial examples, Counterfactual
- Algorithms: LIME [16] and its variants such as xLIME, Anchors [17], SHAP [2], GAMs, Feature Importance (Random Forest)
- Evaluation of Explainability: Accuracy, Coverage / Representativeness, Complexity, Human friendliness (concepts easier for humans to understand)

1.2 Case Studies

In the second part of the tutorial, we will present case studies across different companies. These studies span several domains such as the following:

- Search and Recommendation systems: Understanding of search and recommendations systems, as well as how retrieval and ranking decisions happen in real-time [14]. Example applications include explanation of decisions made by an AI system towards job recommendations, ranking of potential candidates for job posters, and content recommendations.
- Sales: Understanding of sales predictions in terms of customer up-sell/churn
- Lending: How to understand/interpret lending decisions made by an AI system [5]
- Fraud Detection: Examining and explaining AI systems that determine whether a content or event is fraudulent.

We will focus on the practical challenges and lessons learned during development and deployment of these systems, which would be beneficial for researchers as well as practitioners working on explainable AI. Finally, we will discuss open challenges and research directions for the community.

2 BRIEF BIOGRAPHIES OF THE PRESENTERS

Krishna Gade is the founder and CEO of Fiddler Labs, an enterprise startup building an explainable AI engine to address problems regarding bias, fairness, and transparency in AI. An entrepreneur and engineering leader with a strong technical experience of creating scalable platforms and delightful consumer products, Krishna previously held senior engineering leadership roles at Facebook, Pinterest, Twitter, and Microsoft.

Sahin Cem Geyik has been part of the Careers/Talent AI teams at LinkedIn over the past three years, focusing on personalized and fairness-aware recommendations across several LinkedIn Talent Solutions products. Prior to LinkedIn, he was a research scientist at Turn Inc., an online advertising startup which was later acquired by Amobee, a subsidiary of Singtel.

Krishnaram Kenthapadi is part of the AI team at LinkedIn, where he leads the fairness, transparency, explainability, and privacy modeling efforts across different LinkedIn applications. He also serves as LinkedIn’s representative in Microsoft’s AI and Ethics in Engineering and Research (AETHER) Committee.

Varun Mithal is an AI researcher at LinkedIn, where he works on jobs and hiring recommendations. He has developed several algorithms to identify rare classes and anomalies using unsupervised change detection as well as supervised learning from weak labels.

Ankur Taly is the head of data science at Fiddler Labs, where he is responsible for developing and evangelizing core explainable AI technology. Previously, he was a Staff Research Scientist at Google Brain where he carried out research in explainable AI, and was most well-known for his contribution to developing *Integrated Gradients* [18], a new interpretability algorithm for deep networks.

REFERENCES

[1] M. A. Ahmad, C. Eckert, and A. Teredesai. Interpretable machine learning in healthcare. In *ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2018.

[2] L. Antwarg, B. Shapira, and L. Rokach. Explaining anomalies detected by autoencoders using SHAP. *arXiv preprint arXiv:1903.02407*, 2019.

[3] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*, 2015.

[4] F. K. Došilović, M. Brčić, and N. Hlupić. Explainable artificial intelligence: A survey. In *IEEE International convention on information and communication technology, electronics and microelectronics (MIPRO)*, 2018.

[5] R. M. Grath, L. Costabello, C. L. Van, P. Sweeney, F. Kamiab, Z. Shen, and F. Lecue. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245*, 2018.

[6] D. Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[7] H. Lakkaraju, E. Kamar, R. Caruana, and E. Horvitz. Discovering unknown unknowns of predictive models. In *NeurIPS*, 2016.

[8] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Interpretable & explorable approximations of black box models. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT-ML)*, 2017.

[9] Z. C. Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10), 2018.

[10] Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In *KDD*, 2012.

[11] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *KDD*, 2013.

[12] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. Explainable AI for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*, 2019.

[13] D. A. Melis and T. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*, 2018.

[14] D. Qiu and Y. Qian. Relevance debugging and explaining at LinkedIn. In *OpML*, 2019.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin. Model-agnostic interpretability of machine learning. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2016.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin. “Why should I trust you?” Explaining the predictions of any classifier. In *KDD*, 2016.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, 2018.

[18] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[19] S. Tan, R. Caruana, G. Hooker, P. Koch, and A. Gordo. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640*, 2018.

[20] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. Designing theory-driven user-centric explainable AI. In *CHI*, 2019.