Human Factors in Model Interpretability

A Comprehensive Examination of the Application and Perception of ML Interpretability in Industry Settings.

Date: 10/12/2023

Presented by: Mansi Sharma and Mayank Sharma

Authors' background

SUNGSOO RAY HONG, New York University,

JESSICA HULLMAN, Northwestern University,

ENRICO BERTINI, New York University, USA

- This collaborative effort between these esteemed researchers provides a deep dive into the practical implications and challenges of model interpretability in the industry.

Introduction

The ML Landscape

- Machine Learning is rapidly evolving and finding its place across diverse sectors.
- The industry's emphasis is shifting: It's not just about performance anymore; understanding model behaviors is equally crucial.

Background

White-Box vs. Black-Box Models

- White-Box Models: Transparent structures like decision trees and logistic regression, offering inherent interpretability.
- Black-Box Models: Complex structures like neural networks, bringing powerful performance but at the cost of opacity.
- The increased use of black-box models has intensified the call for better model interpretability.

Paper's Objectives

Deciphering Interpretability in Industry

- Understand the perception and application of interpretability among ML experts.
- Highlight areas where current technology falls short in supporting industry practices of interpretability.

Research Methodology

Approaching the Experts

- Utilized open-ended interviews to gather insights from ML professionals.
- Employed qualitative coding to categorize and draw patterns from the shared experiences.

Results Overview

Thematic Structure of Results

Interpretability is multifaceted, encompassing:

- Roles involved.
- Stages of model lifecycle.
- Goals driving the need for interpretability.

Interpretability Roles

Who's Involved?

- Model Builders: Those who design, develop, and test models.
- Model Breakers: Critical evaluators who challenge the models.
- Model Consumers: End-users or stakeholders who rely on model outputs for decision-making.

Interpretability Stages

Lifecycle of a Model

- Planning: Setting objectives and requirements.
- Building: Designing and training the model.
- Deploying: Implementing the model in real-world scenarios.
- Managing: Ongoing monitoring and refinement of the model.

Interpretability Goals

Driving Forces Behind Interpretability

- Ensuring ethical considerations in decision-making.
- Achieving trust and transparency with stakeholders.
- Meeting regulatory and compliance requirements.

Overarching challenges in ML

Roadblocks in Achieving Interpretability

- Varying definitions of interpretability among experts.
- The trade-off between model performance and interpretability.
- Lack of standardized tools and frameworks.

Communication Challenges

Bridging the Gap

- Avoiding misunderstandings between different stakeholders.
- Ensuring the model's outputs and behaviors align with human expectations.
- Overcoming the challenges of making complex models "explainable" to non-experts.

Integration Challenges

Interpretability Tools: A Double-Edged Sword?

- Challenges in aligning academic tools with real-world industry scenarios.
- Integrating interpretability tools into existing complex workflows.
- Overcoming the limitations of "off-the-shelf" solutions.

Limitations of the Study

Acknowledging the Gaps

- Some domains might have been overlooked in this study.
- The study heavily focused on human-consumed model predictions, leaving out automated settings.
- The voices of data scientists were predominant, potentially overshadowing other perspectives.
- Humans themselves don’t fully understand the reasoning behind their decisions

Impact on the Field

Ripples in the ML Community

- This paper shines a light on the discrepancies between academic views and real-world applications of interpretability.
- It emphasizes the value of qualitative insights directly from industry professionals.

Follow-up Work

The Road Ahead

- Deepening the exploration into specific interpretability challenges.
- Studying the nuances of model interpretability in automated settings.
- Bridging the academia-industry gap with collaborative projects.

Personal Thoughts

A Reflective Take

- The paper underscores a crucial but often overlooked aspect of ML: understanding its behavior.
- While comprehensive, there's potential for more domain-specific studies.
- It's a call to action for the community to prioritize model transparency and understanding.
- Intrepretability is an iterative process just like a

Takeaways for the Class

Classroom Insights

- Black-box models, while powerful, bring interpretability challenges.
- The real-world application of interpretability has layers and is more nuanced than textbook definitions.
- The gap between research and practice is significant, urging the need for more applied studies.

Discussion Questions

For Class Discussion

- How can academia contribute more effectively to real-world interpretability challenges?
- Are there alternative methods to enhance interpretability in black-box models?
- How do we ensure that all stakeholders, not just ML experts, understand model behaviors?

Thank you for your attention!

We are open to questions, insights, or any feedback.