Week 3: Literature Review - The Case for Interpretability

(Karan) RK Rajkumar

Paper Information

Primary Paper 1: "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead" (Rudin, 2019)
Primary Paper 2: "Artificial intelligence in human resources management: Challenges and a path forward" (Tambe et al., 2019)

1. Background and Motivation

With AI adoption becoming more ubiquitous by the year in enterprise and organizational decision making the need for more transparency and employee trust in the implementation of **black box** and/or AI solutions.

IBM's Global Adoption index recently estimated in 2023 that **42%** of IT professionals at large organizations have already implemented AI solutions [1]. The survey maintains that **23%** of the organizations that have eschewed these solutions have done so on the grounds of *ethical concerns* [1]. IT professionals on this survey also rated **Data Privacy** and **Trust/Transparency** as more substantial inhibitors towards the adoption of these black box solutions than the lack of trained employees to operationalize these technologies.

These inhibitors are not merely *theoretical anxieties*; they are grounded in a documented history of algorithmic failure in **high stakes domains**. One of the most prominent occurrences of this was underscored by **Tambe et al.** [2] in having found that **Amazon Inc.** had retired their black box prediction model which was formerly used for screening new hires when it was discovered that women were being hired at conspicuously lower rates using this technology since the Machine learning algorithm used in this endeavor had been trained on decades of data that correlated capable job performance with *Caucasian and male identity*.

Obermeyer et al. (2019) [3] inform us that the harm rendered by black box modelling is not confined to multinational corporations alone. In domains like **healthcare** the *proxies* used in

algorithmic risk models often obscure demographic realities. Specifically, they found that when projected **healthcare costs** are used to assign risk scores, the model frequently overlooks the *lack of access to healthcare* as a key facet of public health risk, effectively penalizing patients for being underserved.

The aforementioned examples illustrate the central thesis presented by **Rudin (2019)** [4]; her own work furthermore substantiates the dangers of black box modeling with striking examples of **typographical errors** in data entry erroneously influencing the decisions meted out by algorithmic models like **COMPAS** used in the **US Justice System** and proprietary algorithmic models like Google's **BreezoMeter** (which was used during the **California Wildfires of 2018**) incorrectly predicting *good air quality* for outdoor activities when commensurate interpretable models like those used by the **EPA** would have rightly suggested otherwise. These cases reinforce the necessity for **inherently interpretable models**, where the logic is visible to domain experts before high-stakes deployment.

2. Methods Used

Rudin (2019) [4] grounds her findings in comparative case studies to provide the reader with a *vis a vis* juxtaposition of **black box models** against their **interpretable counterparts**. Most salient in her study was the line of contrast drawn between the **COMPAS** model (a recidivism risk prevention tool used widely in the **US Justice System** that relies on **expert designed surveys** to aggregate over **130 factors**, such as socioeconomic information and criminal history) and the **CORELS** model (an algorithmic model that uses more elementary **conditional branching logic** derived from **optimal pattern mining in data** to forecast outcomes) in which Rudin determined via algorithmic demonstration that there was **no statistically significant difference** in the accuracy of the classifications produced by these models. Rudin also made use of **pairwise comparison** in her exploration of the true efficacy of the **BreezoMeter** model discussed in Section 1.

Rudin's employ of **mathematical formalism** in arguing for the existence of interpretable models in novel scenarios is also noticeable in her invocation of the "**Rashomon Set Argument**" which holds that for any dataset which supports a finite number of accurate interpretable forecasting models, there must exist *at least* one such data model which is interpretable.

Tambe et al. [2] compile their study by facilitating a **single day workshop** in the fall of 2018 with **Data Science faculty** from various universities and **Workforce Analytics professions** from over a dozen **major US corporations**. The workshop included dedicated sessions on the management and (possibly interdepartmental) acquisition of HR data in enterprises, the viability of using employee **social media feeds** as training data for predictive modelling, the **fairness and ethics** of HR decisions made using algorithmic solutions and the role and purpose of **employee recommendations** and feedback in workforce analytics. Prior to the workshop the authors had seeded these 4 workshop foci by mailing out a **short survey** to

workshop participants that featured **open ended questions** about initiatives, barriers and expected breakthroughs.

Following the workshop, the authors crystallized their findings into a paradigm termed "**The life cycle of an AI-supported HR practice**" which segmented Workforce analytics operations into **4 distinct but sequential stages**: **Operations**, **Data Generation**, **Machine Learning** and **Decision Making**. The authors raise and explore several recurring problems faced by enterprises at each of these stages as examples and explicitly attribute the construction of these case examples and challenges as being sourced from the feedback and engagement derived from the single day workshop they conducted.

3. Significance of the Work

The significance of **Rudin's** work is best observed in her systematic deconstruction of the "**interpretability vs accuracy**" consensus prevalent among data professionals [4]. Rudin identifies the paucity of clarifying information returned by black box models along with their outputs and further states that the clarifying information returned by these models to date has often been rendered unintelligible with omissions or selective displays of information. Rudin illustrates this with the example of **saliency maps** which are regarded as explanatory counterparts to black box image classification models, highlighting that these maps allow the onlookers to define which parts of the image were excluded from analysis but fall short of elucidating exactly *how* the regions of the image that were retained for analysis were ultimately employed for the classification task.

Rudin urges decision makers, enterprises and authorities to strengthen **governance standards** as it pertains to the usage of black box models in **high stakes decision spaces** like hiring and medicine. To this effect, Rudin offers an operable maxim that states that **no black box model is to be deployed** in a high stakes scenario where there exists an interpretable model of commensurate accuracy and performance [4]. Rudin's regulatory call and emphasis that opacity in modeling protects **intellectual property and trade secrets** at the expense of fairness has found wide reaching impact in the broader academic community having been cited over **10,000 times** and is widely credited with reorienting the academic dialogue on black box modeling from *explicability* of modeling to the *feasibility* of black box analyses in the first place [5].

Similar to Rudin, the contributions of **Tambe et al.** [2] identify a chasm between the optimistic expectations and the implementational reality of AI (and black box modelling) into enterprise level HR operations. The authors aver that the nature of the most formidable barrier to AI implementation in HR is **legal, not technical** and point to the inadequacy of the existing framework of predictive modeling (and machine learning) which is largely based on **correlational rather than causal analysis** (e.g. defending against **disparate impact claims** in employment law would be more feasible if the litigated employment decisions obtained algorithmically were based on causal relationships between the variable of interest and

independent variables. **Tambe et al. argue that unlike opaque correlation-based models, causal models offer the inherent explainability required to satisfy these legal and ethical standards.**).

Also noteworthy is the authors' mention of the **collider effect** as this underscores the necessity of thoughtful consideration on behalf of HR professionals and executives in the administration of algorithmic models; Tambe et al. illustrate this with an example of an algorithmic model used for hiring decisions. They warn that if an employer implements a hiring algorithm that favors conscientiousness and college grades, it would soon find its workforce depleted of employees who were neither conscientious nor had good grades, potentially creating **spurious negative correlations** between traits among the remaining staff. As a remedial measure, Tambe recommends "**turning off**" HR algorithms periodically to introduce **randomness** in training data.

Crucially, the authors also state that unlike in the physical sciences, even the most sophisticated algorithms used to guide HR decisions are prone to **gamification** by employees and candidates who are aware of their workings (e.g. employees sharing positive traits as "weaknesses" on job application forms can skew the predictive results of hiring algorithms that parse job applications). Tambe et al. conclude their study with the recommendation that AI shouldn't replace human consideration in workforce decision making but rather serve as a form of **augmented intelligence** in supplemental capacity to the work of HR professionals and policy experts.

The work of Tambe et al. has been well received, garnering over *1000 citations** [6], which is a rarity for an article in a niche area of study like Workforce Analytics & Management.

4. Connection to Other Work

Rudin, in her work, has explicitly positioned herself at odds with the dominant trend of Explainable AI (or "xAI") in her advocacy for the prioritization of interpretable models arguing that the field has meandered down the path of fashioning post-hoc explanations for opaque models [4]. The theoretical underpinnings of Rudin's approach can be gleaned from seminal work authored by **Leo Breiman** in 2001 [7]. Rudin instrumentalizes Breiman's logic on **Rashomon sets** (which were discussed in Section 2) to assert the existence of interpretable models for a dataset where black box models exist yet departs from Breiman's conclusion that interpretability must be sacrificed for accuracy. While Breiman championed opaque **algorithmic models** (like Random Forests) as a necessary evolution from the sclerotic **data models** of 1990s statisticians, Rudin argues that modern algorithms can now optimize for both accuracy and interpretability. Rudin also cites **Angwin et al. (2016)** which was a ProPublica study on COMPAS in her comparative analysis in an effort to demonstrate the propensity black box models have in concealing significant biases and failing to provide statistically significant advantages to their interpretable counterparts [8].

Tambe et al. also distinguish themselves from their peers in HR analytics (who champion predictive accuracy) by shifting the focus to the **causal validity** of the model outputs in themselves by remarking on the potential legal pitfalls in the application of AI and black box modelling to HR decision making and practice [2]. They substantiate this position using **Judea Pearl's (2018)** work in their assertion that cause and effect modelling is requisite for legal defensibility [9]. They also utilize the work of **Dietvorst et al. (2015)** on **algorithm aversion** to demonstrate that employees or the subjects of algorithms are less forgiving of erroneous judgment from algorithms than they are of human arbiters [10].

5. Relevance to Capstone Project

The literature reviewed herein provides the intellectual framework for the methodological choices of my Capstone where I will de-prioritize the implantation and use of complex ensemble methods like random forests and XGBoost with the exception of utilizing these algorithms to benchmark the interpretable models I will use (Logistic Regression, Decision Trees etc.). Rudin's 'Rashomon Set' argument offers me the theoretical justification to construct interpretable models of competitive accuracy with regards to their black box counterparts without compromising the transparency that is requisite of human capital decisions [4]. Furthermore, Tambe et al.'s emphasis on 'causal validity' validates my project's objective to generate intelligible model coefficients (like odds ratios for logistic regression) and actionable interventions, rather than surmising model output as a scalar 'flight risk' score [2]. Consequently, my Capstone will diverge from the industry obsession with state of the art (SOTA) modelling complexity, instead positioning interpretability as the primary metric for success to ensure my findings are legally defensible , organizationally actionable and procedurally just in enterprise settings.

References

[1] IBM Corporation, "Data suggests growth in enterprise adoption of AI is due to widespread deployment by early adopters." IBM Newsroom, Jan. 2024. Available: https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters

[2] P. Tambe, P. Cappelli, and V. Yakubovich, "Artificial intelligence in human resources management: Challenges and a path forward," *California Management Review*, vol. 61, no. 4, pp. 15–42, 2019.

[3] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[5] Association for the Advancement of Artificial Intelligence, "Duke computer scientist wins $1 million artificial intelligence prize, a 'new nobel'." AAAI News, Oct. 2021. Available: https://aaai.org/duke-computer-scientist-wins-1-million-artificial-intelligence-prize-a-new-nobel/

[6] Google Scholar, "Google scholar citation profile: Artificial intelligence in human resources management." Google Scholar Database, 2026. Available: https://scholar.google.com/citations?user=UXmpersAAAAJ

[7] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[8] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." ProPublica, May 23, 2016. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[9] J. Pearl and D. Mackenzie, *The book of why: The new science of cause and effect*. New York, NY: Basic Books, 2018.

[10] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, 2015.