Research Project Report on

Explainability, Fairness and Bias in Algorithmic Decision Making

Sini Suresh

M1 - DATAAI

Supervised by : **Sophie CHABRIDON, Amel Bouzeghoub**

IP Paris | Télécom SudParis | Télécom Paris

February 2021

Contents

Abstract

With the recent high emergence and demand of machine learning models, it has become a necessity that we focus on the ethical concepts that affects the public/users of these models. As we know that the accuracy of these systems is 95-98 percentage, if used in a larger space, there is a concern of fairness and bias. This project targets to understand the level of fairness achieved in machine learning models by studying Automated hiring systems.

To learn more about fairness and bias involved in these systems, we take a closer look on the need for explainability. There are various factors to be considered under each topic - explainability, fairness and bias; and generalizing the solution is not possible. Every scenario has to be studied on a case by case basis and evaluated in depth to understand the root cause of discrimination. This project focuses on the major features and challenges in explainability, fairness and bias and aims to find the relation between explainability, fairness and bias.

Chapter 1

Introduction

Machine Learning and various related algorithms are undergoing advances day by day. As early as 1952, machine learning stepped into making human life more easier in different aspects by introducing models and agents. However, with advent of such models and agents, there is an important fact that needs to develop alongside. The legalities related to using these models and agents. To have a better legal structure set up, we need to understand the working of the AI models and systems. This is a major concern as it raises a question regarding the ethical factors. The users should be aware of the internal functions of the system they use to trust the system.

As machine learning models are invading our day-to-day lives, it has become more important for the developers of these models to give the explanation on its technical features and reasoning understandable by anyone who uses it.

In this research, the focus is mainly on identifying the active research areas and challenges in Explainability and deriving the relation between Explainability, Fairness and Bias. A lot of researches are being carried out in the area of Explainability, popularly know as XAI (Explainable Artificial Intelligence). This research project report progresses with automated hiring systems as a usecase to understand Explainability, Fairness, Bias and the relation between them.

Automated Hiring System - Automated algorithms are being used to assess and classify applicants' potential, while only the highest-ranked applications make it into human hands [1]. In such systems the decision processes are written not by a person, but are trained using data.

In the following sections we discuss the need of XAI - why is it important, features related to explainability - interpretability, completeness and justification, explanations, social explanations and challenges of XAI.

We then focus on Fairness and Bias - the importance of fairness, the difference between group fairness and individual fairness, tradeoffs and discrimination.

1.1 Automated Hiring System (AHS)

Automated Hiring System [1] is the case study that will be referred to in the coming sections. As mentioned in the introduction, the decision processes of AHS are trained using data. In the training process, a number of features of importance are selected, such as degree name, institute name, grade-point average, work experience, etc. Then, taking a huge stack of prior applications and the final decisions of these applications (hire or not hire), a mathematical model is automatically derived, using a machine learning algorithm that predicts the likelihood of a person being hired. It does this by discovering patterns in the underlying data.

Key concerns about Automated Hiring Systems include the lack of transparency and potential limitation of access to jobs for specific profiles [2]. AHSs claim to detect and mitigate discriminatory practices against protected groups and promote diversity and inclusion at work. These are rarely scrutinized and evaluated, and when done so, have almost exclusively been from a US socio-legal perspective because most of these models are developed in the U.S. In this report, the study is based on a perspective outside the U.S. by critically examining how three prominent AHS in regular use in the UK, HireVue, Pymetrics and Applied function to understand and attempt to mitigate bias and discrimination and deduce the need for explainability and fairness.

Pymetrics[1] - a vendor of hiring technology that performs a pre-employment assessment of candidates with games tests that are based on neuroscience research. The software uses unsupervised learning clustering algorithm and generates metrics of cognitive, social and emotional traits. It follows 4/5th rule of Equal Employment Opportunity Commission (U.S.)

HireVue[2] - is a product to automate the pre-interview assessment of candidates from a pool. It performs automated video interview and games to profile candidates. The games and questions are designed based on Industrial Organization psychology research. The tool extracts three types of indicator data from applicants: categorical, audio and video. Applies 4/5th rule and trains the model using clustering methods until there is no bias detected. The strategy also consists of modifying the learning algorithm to account for fairness. In machine learning, the objective function is a mathematical expression of how well the model is fitted to the data. It guides the learning algorithms in the process of learning from data and creating data transformations that contribute to improving accuracy.

Applied[3]- is a hiring platform specialized in promoting diversity and inclusion in recruitment. The system includes a numerical, analytical and problem-solving testing platform called Mapped[4] that designs the tests by excluding patterns that are found to negatively impact on different demographic groups and improve pass rates of candidates of different group.

Limitations claimed by the authors of [1] include: (1) Access to relevant information

[1] https://www.pymetrics.com/

[2] https://www.hirevue.com/

[3] https://www.beapplied.com

[4] http://www.get-mapped.com/

explaining the bias mitigation in hiring is not available for thorough analysis. (2) More work is needed looking specifically at systems developed in the UK and the EU that also connects these to the actual practices and experiences of employers and candidates to get a sense of how AHSs shape those interactions. (3) The study is done based on publicly available data - information relating to code, data sets, features design, trained models, or even the application user interface was not possible. (4) The study states that algorithm to find the best fit based on the existing employees brings in bias - but without access to the source code or training model, it is not clear whether it can bring in biased groups.

But what approach to transparency is required by EU law, remains itself vague. Even when considering one statistical definition for bias such as the error rate balance amongst groups, the understanding and implication of that approach radically varies with the context and the consequential decisions that are driven by the algorithmic output. All socio-technical systems, even when designed to mitigate biases, are designed with use cases in mind that may not hold in all scenarios.

Issues in AHSs also include:

1. attempts at mitigation within AHSs run into on-going concerns with computational fairness.
2. attempts at bias mitigation in AHSs within a UK context also show problems with accountability.
3. the lack of information about how AHSs work,
4. the approach they take to tackling discriminatory hiring practices,
5. where and how they are used around the world is therefore a significant problem.

Chapter 5

Conclusion

This research project is primarily based on explainability, fairness and bias. These topics are covered with the use case, Automated Hiring System (AHS). Pymetrics, HireView and Applied are the AHSs studied. These models were considered because other major AHSs had least information available publicly. Pymetrics, HireView and Applied are also leading AHSs in the UK, however, access to their source code or information on the input parameters were unavailable. Hence the research is carried out on the available information and this is one of the major challenges faced in the explainable AI (XAI) industry.

We focused on the need of XAI, interpretability, justification and completeness which are the major features or basic requirement in an XAI. The goal of interpretability is to describe the internals of a system in a way that is understandable to humans and The goal of completeness is to describe the operation of a system in an accurate way [2]. To make the AI system more reliable, and to improve transparency and accountability, we need to interpret the internal functions of these systems in Laymen's language. Talking about the challenges, Miller [3] explains three main challenges of XAI being: opaqueness, causality and human centeredness.

Then we extended our research to fairness and bias where we discuss about the difference between group fairness and individual fairness. We also see the discussions on tradeoffs and discrimination. The problem of algorithmic discrimination has no one-size-fits-all solution as different fairness definitions have different meanings in different contexts and not all fairness criteria can be simultaneously fulfilled in one decision process.

Then we conclude our research by relating explainability, fairness and bias with the use case on automated hiring system where we learn that to over come the unfairness and bias in the existing models, it is indeed necessary to focus on the explainability of AI algorithms.

This is an active research field and every AI engineer should focus on these ethical aspects so that we achieve accountability and transparency at every level.

Bibliography

[1] Javier Sánchez-Monedero et al., (2020). What does it mean to ‘solve’ the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems 1, 2, 9

[2] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal (2019). Explaining Explanations: An Overview of Interpretability of Machine Learning 2, 5, 13

[3] Tim Miller (2019). "But Why?” Understanding Explainable Artificial Intelligence, 1528-4972/19/03 4, 6, 13

[4] Tim Miller (2018). Explanation in Artificial Intelligence: Insights from the Social Sciences 5, 6

[5] B. Herman, “The promise and peril of human evaluation for model interpretability,” arXiv preprint arXiv:1711.07414, 2017. 5

[6] D. J. Hilton, Conversational processes and causal explanation, Psychological Bulletin 107 (1) (1990) 65–81. 5

[7] P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplement 27 (1990) 247–266. 5, 7

[8] F. Heider, The psychology of interpersonal relations, New York: Wiley, 1958. 6

[9] B. F. Malle, How the mind explains behavior: Folk explanations, meaning, and social interaction, 6

[10] Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. Proceedings of the National Academy of Sciences, 108(17):6889–6892 8

[11] Barocas, S. and Selbst, A. (2016). Big Data ’ s Disparate Impact. California law review, 104(1):671–729 8

[12] Cynthia Dwork et al., (November 29, 2011) Fairness Through Awareness - Dwork, Microsoft Research Silicon Valley, Mountain View, CA, USA. 8, 9

[13] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. pages 1–42. 9

[14] Songül Tolan (2018), European Commission, Joint Research Centre (JRC), Seville, Spain. Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges 10

[15] Barocas, S. and Selbst, A. (2016). Big Data ' s Disparate Impact. California law review, 104(1):671–729. 10

[16] Green, B. and Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning. 10

[17] Crawford, K. (2013). Think again: Big data. Foreign Policy, 9. 10

[18] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., and Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010. 11

[19] Drosou, M., Jagadish, H., Pitoura, E., and Stoyanovich, J. (2017). Diversity in big data: A review. Big data, 5(2):73–84. 11