Paper SAS102-2022

What's your Model Really Doing? Understanding Human Biases in Machine Learning

Jim Box, SAS Institute Inc.

ABSTRACT

Machine learning and artificial intelligence are having a tremendous impact on our day-to-day lives, covering areas like financial opportunities, health care, job selection and promotion, and even how we interact with vehicles on the street. It's tempting to think that because a model came up with a suggestion, it must be based on science that has been fairly conducted, but that is far from the case - human biases have tremendous impacts on how machine learning algorithms come up with predictions. As programmers, we have a responsibility to understand the sources and impacts of these biases, and to look at ways we can mitigate the potential harm.

INTRODUCTION

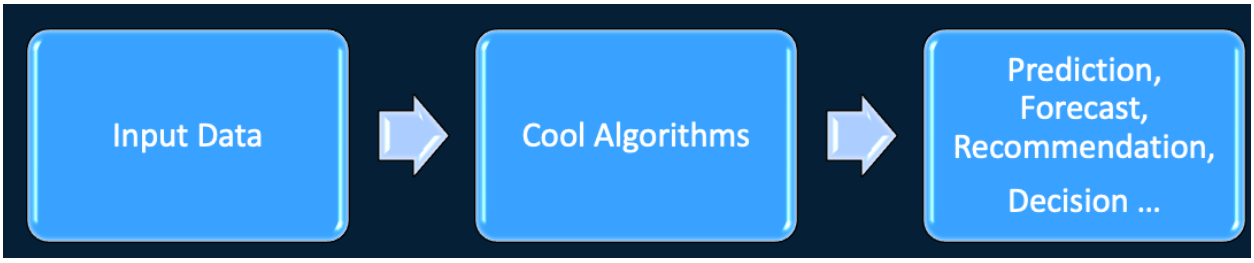
Artificial Intelligence (AI) systems are popping up everywhere, across all industries. It's important to understand what that encompasses. For our purposes, AI is the science of training systems to emulate human tasks through learning and/or automation. AI has existed for decades; a web-based randomization system can be considered a rudimentary AI system, as it asks for prompts and automatically guides you to a resolution. When we think of AI systems now, we think of things a little more complicated, usually powered by technologies such as machine learning (ML), natural language processing (NLP) or the like. This paper is focusing on the ML models sometimes used to power AI systems, and how human biases can impact their performance. AI systems have a dramatic impact on several aspects on daily life, usually without you even knowing it.

Machine Learning is a branch of AI based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. There are many types of ML models, from tree-based systems that predict binary outcomes, to advanced neural networks that attempt to diagnoses diseases from radiology scans. AI/ML applications have become much more popular due to several reasons, including:

- Vast amounts of data are now available, and easier to store, join and utilize
- Powerful computing systems, with much faster computational capabilities and utilizing graphics processers
- Powerful algorithms and easy to use software

There is a strong drive for organizations and governments to implement AI solutions. A primary factor is the cost savings or income generation – these systems may allow companies to better attract and retain customers, or to identify employees who are at risk of leaving. Another factor is that AI systems can five off the impression of impartiality – it's the algorithm that is making the decision about who gets approved for a loan, or who gets a harsher sentence, which can give the impression of impartiality. There's a term for letting the computer take responsibility for making these sorts of decisions – it's called mathwashing. Cathy O'Neil covers this in detail in her book *Weapons of Math Destruction* (2016).

To understand the impacts of bias in machine learning, we'll look at the three steps typically involved with creating an ML application:



Each step has ways for bias to seep into the process (or for unexpected outcomes to pop up).

BIAS IN TRAINING DATA

The foundation to building a ML model is the training data – this is all the information used by a model to develop later predictions. The most important thing to realize is that **biased data will give you biased results**. It's vital to understand that if you have data collected by humans and/or about humans, then there is bias present in that data, and it's important to identify the ramifications.

One example (of many, many possible examples) details efforts made by Amazon (Dastin, 2018) to build a system to screen through job applicant resumes. The idea was to look at the resumes of all the people Amazon had hired and use that information to rank the job applicants, and to only interview the highest-ranking ones. The problem here was that historically, Amazon had not been hiring women for the engineering roles, and there was an overwhelming percentage of men in the training data. The ML algorithm picked up on that and gave lower scores to any candidate it identified as a woman. The developers were able to institute workarounds to prevent the algorithm from considering specific words (e.g. women, she, her) and correlations like attending an all-women's school, but ultimately the algorithm was so successful at finding ways to exclude women, the team was forces to scrap the system. The algorithm itself did exactly what was asked of it, the problem was it was being trained on data with inherent biases.

The key takeaway here is that models trained on biased data will excel at applying that bias – perhaps even more efficiently than humans. Your responsibility when building or implementing an AI/ML solution is to question the data:

- Where did it come from?
- What historical biases does it include?
- How representative is the data?
- How did it get labeled (why are some records successes and others failures)?
- Is there a feedback loop for addressing problems?

ALGORITHM PERFORMANCE

Machine Learning algorithms do exactly what you ask them to do, which might not be what you meant them to do. There is no nuance or conscience involved, unless specifically programmed in. This may sound obvious, but the implications may be subtle. You may give an algorithm a bunch of labeled pictures of dogs and wolves and train it to tell the difference. Although a person may spend time looking at features such as snout length, coloring, and ear shapes, the algorithm might have an easier time sorting the pictures into mostly correct groups by looking at background image features, like the presence of snow (Ribeiro et al, 2016). It is important to try to understand why an algorithm is making a specific prediction, which can be very trick to ascertain with complex, black-box models.

This can cause real problems, because unless you investigate why a model made the decision it did, you may be implementing decisions based on sub-optimal information. A good example of this was done on a study of chest x-rays used to diagnose cardiomegaly (enlarged heart) by a neural network (Zech, 2018). On first glance, the algorithm seemed to be working fine – it was trained on images that were labeled as yes/no for cardiomegaly, and a validation set was examined. The model performed well, giving high predictive scores to images of patients with the condition and lower scores to those without. Neural networks are complicated, and it is not particularly obvious why the model was scoring the images as it was, but it seemed accurate enough, so it is not hard to imagine it being put into use diagnosing patients. Only when the author of the paper took the effort to use heatmaps to see what parts of the image were most important that he noticed something – the algorithm was looking at parts of the image that had nothing to do with the condition of the patient – specifically, it was looking at some words that were printed on the upper right of the image that indicated the machine used to take the image was a portable x-ray. So, the model was taking into account that the patient was too sick to travel to the imaging room, which is not really information that should be used to make an image-based diagnoses. Essentially, it was cheating a little by knowing the answer before looking at the data. For someone looking at just the model results, though, it would have seemed like an excellent model doing its job.

The key thing to remember is that models are lazy, but effective. They will perform the task without considering context, unless specifically programmed to do so. Your responsibility when building or implementing an AI/ML solution is to question the results:

- Why did the model make this specific prediction for this specific case?
- What are the key inputs being used to make predictions?
- What, exactly, was the model set up to do?

APPLYING THE MODEL TO NEW DATA

The third area of concern is how a model will be applied to new data. Models are only accurate on data similar to what they were trained on, even though they will provide results either way. There can be serious problems with applying a model to novel data. An article int eh Guardian (Devlin, 2018) outlines a relevant concern. A large genetics database was set up in the UK, and machine learning models were used to develop risk scores to predict the onset of schizophrenia. The test produced scores 10 times higher for people of African ancestry than those without. This happened not because there was actually a higher risk of disease, but because the genetics database was built almost entirely upon markers collected from people of Northern European descent. Because the model did not have enough training data on patients of diverse background, the results were wildly inaccurate. Problems of this nature can be difficult to notice, because the model will give results that seem to be based on science, but have little actual value. Your responsibility when building or implementing an AI/ML solution is to question the application:

- Is this data similar to what the model was trained on?
- Do different population subgroups get different predictive results?
- Is there a human feedback loop that allows for training the model?

CONCLUSION

As practitioners of Machine Learning, it is our responsibility to understand how biases can impact the types of models we build and implement. It is altogether too easy to fall into the trap of thinking that because an algorithm gave a result, we eliminated the impact of humans on the decision. The main points to remember are:

- Models can pick up patterns in the data we are not explicitly trying to teach them
- There is a lack of awareness by most data scientists and statisticians about how historical and societal biases may be present in different aspects of data modeling, including
    - How we collect and classify data
    - The problems we are trying to solve with ML
    - The Data we choose to train models on and apply results to
    - How we asses accuracy
    - How we present and implement results

REFERENCES

Dastin, Jeffrey. “Amazon scraps secret AI recruiting tool that showed bias against women.” Reuters, October 9, 2018. Available https://www.reuters.com/article/us-amazon-com-jobsautomation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-againstwomen-iduskcn1mk08g

Devlin, Hannah. “Genetics research ‘Biased towards studying white Europeans.’” The Guardian, October 8, 2018. Available https://www.theguardian.com/science/2018/oct/08/genetics-research-biased-towardsstudying-white-europeans

O’Neil, Cathy. September 6, 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing.

Zech, John. “What are radiological deep models actually learning?” Medium, July 8, 2018. Available https://medium.com/@jrzech/what-are-radiological-deep-learning

ACKNOWLEDGMENTS

This paper and presentation are based on work I did with two excellent colleagues: Elena Snavely, Senior Manager of Corporate Analytics & Insights at SAS and Hiwot Tesfaye, Technical Advisor for the Office of Responsible AI at Microsoft.

RECOMMENDED READING

Box, Jim, Snavely, Elena and Tesfaye, Hiwot. SAS Global Forum 2020. Paper SAS4506-2020.

"Human Bias in Machine Learning: How Well Do You Really Know Your Model?" Available https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2020/4506-2020.pdf

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jim Box
SAS Institute
Jim.box@sas.com
https://www.linkedin.com/in/jwbox/