PhUSE EU Connect 2019

Paper ML05

# Do you know what your model is doing? A look at how human bias influences machine learning.

Jim Box, SAS Institute, Cary, NC, USA
Elena Snavely, SAS Institute, Cary, NC, USA

## ABSTRACT

Machine learning models are becoming more popular and easier to create than ever before. Understanding how the models work and how human bias can creep in and skew results is vital to effectively getting the best models for the business problem you are facing. We explore sources of bias and misunderstanding and discuss ways to combat their influence.

## INTRODUCTION

Machine learning is essentially the practice of creating algorithms that ingest data to detect patterns to predict likely outcomes, identify patterns in the data, categorize like groups in the data or detect unexpected behaviors in the data. Models using these algorithms have been in existence for decades, but as computing power has grown along with the volume, velocity and variety of data available, machine learning is becoming more and more popular. As the execution of this type of modeling gets easier with new software applications and methods, it's important to take a step back and think about the process of using data to make decisions and how human biases can be amplified by applying machine learning.

## BIASED DATA GIVES BIASED RESULTS

The most important part of the machine learning process is not the software, or the algorithm used, it is the data source. All machine learning models are trained on some existing data, and the machine can only learn from experiences the data provides. If the data itself has existing biases, those biases will be amplified by the use of an algorithm. It's vital to understand the biases in the data and to carefully account for it.

In 2014, online retailer Amazon was developing machine learning algorithms to help sort through the hundreds of job applicants they receive for every position (Dastin, 2018). The idea was to use the database of resumes and CVs received over 10 years of hiring to rate every new applicant on a one-to-five star scale, then to focus hiring efforts on the five-star applicants. The algorithm would read through the years' worth of applications that were scored as hired or not hired. This is just the sort of problem for which machine learning is made, and the algorithm was able to find applicants that matched the types that had been hired. The problem was that Amazon, like many other tech companies at the time, had been mostly hiring men. By training itself on this sort of data, the algorithm taught itself that men were preferred candidates, and downgraded applications that were easily identifiable as women (e.g. "member of the women's chess club"). It even went as far as to exclude applicants from two all-women universities. None of this was done explicitly. The model was not designed to favor men. The problem is that the people making decisions in the past had excluded women, and the algorithm picked it up in the training data.

Sometimes, the bias is more subtle. As part of this hiring effort, the Amazon team built models to evaluate posted resumes and profiles on the web to try to identify potential recruitment targets. The algorithm wound up favoring people who used more active and aggressive verbs like "executed," which show up more frequently on profiles about men. Eventually, Amazon abandoned the effort and shut down the project.

The algorithms did their job – they identified candidates that matched the historical data about who the company preferred. The problem was in the training data. Without thoughtful consideration of the data and the goals of the project, it's easy to see how this human bias can impact the results.

The bias in the training data can be less obvious, and can be due simply to the sampling methods. An article in *The Guardian* (Devlin, 2018) points out how this can be a significant issue when it comes to developing genetic tests. The problem lies with concern that the bank of genetic material in the UK Biobank is too ethnically homogenous – the material is collected mostly from white Europeans. Genetic tests developed on this population may not be reliable when generalized to other populations. For example, the article cites a study where "a commonly used genetic test to predict schizophrenia risk gives scores that are 10 times higher in people with African ancestry than those with European ancestry. This is not because people with African ancestry actually have a higher risk of schizophrenia, but because the genetic markers used were derived almost entirely from studies of individuals of European ancestry."

Applying machine learning algorithms on data significantly different from the training data is a sure way to create biased or unreliable results.

## MODELS CAN FIND UNEXPECTED CORRELATIONS

As more machine learning methods get easier to use, it is important to try to understand how they are making predictions. With methods like logistic regressions and decision trees, it's fairly straightforward to see how the model was developed and what types of variables it is taking into account. With more complicated models and deep learning methodologies, this becomes a much bigger challenge. Models can produce results based on confounding variables that can create algorithms explain the training date very well but are not practically useful.

A good example of this concern comes from the field of radiology (Zech, 2018). The article looks at using neural networks to analyze x-rays looking for cardiomegaly (enlarged heart). Images of patients with and without cardiomegaly were run through a neural net to develop scoring code, and the resulting model did a very good job of identifying patients with the condition. Researchers looked at the details of the process to see which parts of the image were most highly correlated to a prediction of a positive result. The expectation was that the areas around the heart in the x-ray would be most predictive, and they were. Surprisingly, there were positive contributions on the edge of the image, far away from the heart. Turns out that there were markings on the edge that indicated what machine was used for the x-ray. The algorithm figured out that a marking that indicated the x-ray was portable was highly correlated to the indication of cardiomegaly. Since portable x-rays were used on patients so sick that they could not travel to the radiology lab, the model essentially figured out that sick patients were more likely to be sick. Factually correct, but not terribly useful in applying the model generally. Other possible confounding variables in this case could be patients from a particular specialist that only saw the harder cases, or clinics that only saw patients who did not have adequate primary care.

It is crucial to understand how a model is using the training data, especially when using an algorithm that is complicated computationally. As more stand alone "black box" software methods are available that claim to run hundreds of models to find the best solution, vigilance on the data used is essential to minimize bias. This is especially important in health care applications. (There are also practical implications to developing complicated models. Netflix [Johnson, 2012] spent $1 Million on an algorithm so complicated that they could not successfully implement it.)

## WHAT TO DO TO MINIMIZE BIAS

As trite as it might sound, awareness of the impact of human bias on machine learning is a huge first step in mitigating its impact. Bias exists in most data collected by humans, and can sometimes be hard to detect. Some steps to consider:

- **Frame the Problem** – Understand the technical question being asked. Make sure that the training data is representative of the population to which you are going to apply the model. Understand the goals – are you trying to find the best possible job candidates, or are you trying to find people like you already have.?

- **Define Fairness** – Think about how you are using the data. Are you using data that won't be applicable to some of your population? Are there regulatory or ethical requirements about the types of inputs you could use? Codifying these decisions can be uncomfortable; be sure people understand why you are doing this.

- **Check for Bias** – Revisit data selection to make sure you are using a representative sample for your problem. Try to identify the sources of human bias that exist in the data. Make sure the model results don't change unexpectedly based on factors like age, race, and gender.

- **Model Interpretability - U**nderstand why the model is producing the results; don't rely on black boxes. This is an iterative process, and you might need several models and cycles to get it right. There are statistical methods that can be employed; we have listed some in the recommended reading.

Education and determination are the keys to reducing and understanding the impact of human biases in the modeling process. It is a big problem, and one that can be easy to ignore and difficult to address. Even companies such as Google struggle with how to manage ethics with machine learning (Piper, 2019).

## CONCLUSION

Selecting the appropriate data for your problem is the most important aspect of applying machine learning. Algorithms created on data with hidden (or overt) biases will make models that are exceptionally adept at applying human biases. It's vital to make sure you have thought carefully about your data sources and how you are applying them to your scientific questions to ensure you minimizing the impact of human biases. Continuing the conversation with colleagues is an essential step in raising awareness of the potential pitfalls of utilizing machine learning without thoughtful oversight.

## REFERENCES

Dastin, J. (2018, October 9). *Amazon scraps secret AI recruiting tool that showed bias against women*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Devlin, H (2018, October 8). *Genetics research 'biased towards studying white Europeans.'* Retrieved from https://www.theguardian.com/science/2018/oct/08/genetics-research-biased-towards-studying-white-europeans

Johnson, C (2012, April 16). *Netflix Never Used Its $1Million Algorithm Due to Engineering Costs.* Retrieved from https://www.wired.com/2012/04/netflix-prize-costs/

Piper, K (2019, April 4). *Google cancels AI ethics board in response to outcry.* Retrieved from https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board/

Zech, J. (2018, July 8). *What are radiological deep models actually learning?* Retrieved from https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

## RECOMMENDED READING

- Machine Learning Primer
- Toward ethical, transparent and fair AI/ML: a critical reading list
- Biased Algorithms Are Everywhere, and No One Seems to Care
- 7 Short-Term AI ethics questions
- Technology Is Biased Too. How Do We Fix It?
- Pay attention to that man behind the curtain
- This is how AI bias really happens—and why it's so hard to fix
- The Mythos of Model Interpretability
- Is your Machine Learning Model Biased?
- Controlling machine-learning algorithms and their biases

## CONTACT INFORMATION

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

Jim Box
SAS Institute
Cary, NC, USA

Email: jim.box@sas.com

Elena Snavely
SAS Institute
Cary, NC, USA

Email: elena.snavely@sas.com