Identifying Bias in Machine Learning Algorithms

Machine learning (ML) advanced statistical and mathematical models are used in various partial or fully automated decision-making systems that affect individual lives. Today, these models are not only increasingly used to make important decisions in our financial lives, ranging from retail (closed and open-end products) and wholesale scorecards (application, behavioral, and collection) but also in other aspects such as granting university admission, social benefit assignment, predicting the risk of criminal recidivism (COMPAS model), and part of hiring tools to review job applicants' resumes. In these applications, models are often built using sensitive drivers, also called attributes, such as age, gender, nationality, religion, race, language, culture, marital status, economic condition, zip code, etc. One of the unintended consequences of lax modeling practice is the potential for bias or unfairness in ML models that accentuates our societal stereotypes and contravenes the laws of many jurisdictions as well.

There are a number of recent high-profile incidents reflecting the risk and the difficulty in detecting bias (or unfairness) in ML models. For example, Apple Credit Card offered smaller lines of credit to women than to men due to its ML/AI credit assessment process. Similarly, there was an AI-based talent-hiring tool at Amazon that was biased against women. And possibly the most publicly embarrassing case occurred when Microsoft's racist chatbot plugged into Twitter.

One of the lessons we should all learn from these mistakes is that if this can happen to the best technology firms in the world, there is a risk to every institution in ignoring the risk of bias in ML applications. Furthermore, all institutions should wonder if there are a number of issues in ML models that have gone unnoticed for quite some time. The risk of such bias in ML models is not only a source of major discomfort to people in minority or disadvantaged groups, but also brings reputational risk and potential legal consequences to institutions utilizing them in their decision-making processes.

This bias challenge is inherent in any model development exercise. However, it is accentuated in the ML model development process, which refines, or "trains," a series of algorithms on a large data set by optimizing iteratively as it learns to identify patterns and make predictions for new data. ML essentially imposes much less structure on how data is interpreted compared to conventional approaches in which programmers impose ex-ante rules to make decisions. This leads to difficulty in attributing the results to particular causes or variables.

This article is organized as follows: First, we will start with defining "fairness." We note there are many notions of fairness in practice and various measures of fairness. Next, we review industry best-practice approaches for "fair in-processing" in overcoming unfairness. Then, we discuss some of the key challenges in these approaches. Finally, we conclude the paper with some closing observations.