Understanding and Addressing Bias in Analytics

Susan Currie Sivek, Ph.D.
Data Science Journalist, Alteryx


the Game of LIFE
SPIN TO WIN!
With Instant Setup & Easy Play!
Hasbro Gaming
AGES 8+
2-4 PLAYERS


MB
Milton Bradley
The Game of LIFE
Over 30 years of family fun!
Ages 9 to Adult
Get the most out of Life by collecting money and Life Tiles!
2 to 6 Players

The algorithmic game of life

“The game simulates a person's travels through … life, from **college** to **retirement**, with **jobs**, **marriage**, and possible **children** along the way.”

Which aspects of the “game of life” do **predictive algorithms** shape today? Which might they affect in the future?

Type your answers in the chat!

Life changes...

Fair play

Predictive models play a growing role in the “game of life”

They “feel” objective, but can perpetuate unfair aspects of our everyday social world

Need to generate decisions (and use them) in ways that support equity

Avoid creation of “feedback loops” that disadvantage some repeatedly

The required list of algorithmic failures

Algorithms can make the “game of life” a few levels harder for some:

- Criminal justice algorithms may predict greater odds of recidivism for people of color
- Skin-cancer detection models may perform worse on darker skin
- High-paying jobs have been advertised more to men than women
- Search results may offer stereotypical images of race/gender
- Recruiting tools may analyze facial expressions and voice patterns in ways that reinforce discrimination
- … and other examples happening now that we don’t yet know about

Reasons for concern

Even if not sentencing people to prison or determining their career path...

You still want your models to make fair and equitable decisions and recommendations

Reach your audience and price products ethically

Avoid potentially discriminatory and offensive behavior

“Numbers can’t speak for themselves, and **data sets — no matter their scale — are still objects of human design**. The tools of big-data science … do not immunize us from skews, gaps, and faulty assumptions. Those factors are particularly significant when big data tries to reflect the social world we live in, yet we can often be fooled into thinking that the results are somehow more objective than human opinions. **Biases and blind spots exist in big data** as much as they do in individual perceptions and experiences.”

— Kate Crawford, Microsoft Research/NYU

Designing data collection

“Redundant encoding”: potential effects of proxy variables, even if avoiding use of demographic variables

Sampling bias: not representing the real world due to your sampling strategy

Caution in using secondary data; need insight into its sources

Deleting data assumed to be “irrelevant”

Choosing and defining your target variable

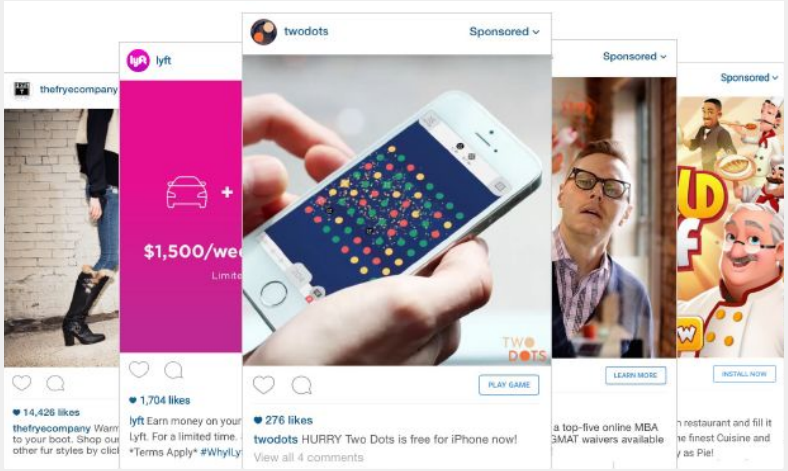
Interpretability and explainability of models

An initial step: What's going on in that model?

Interpretability:

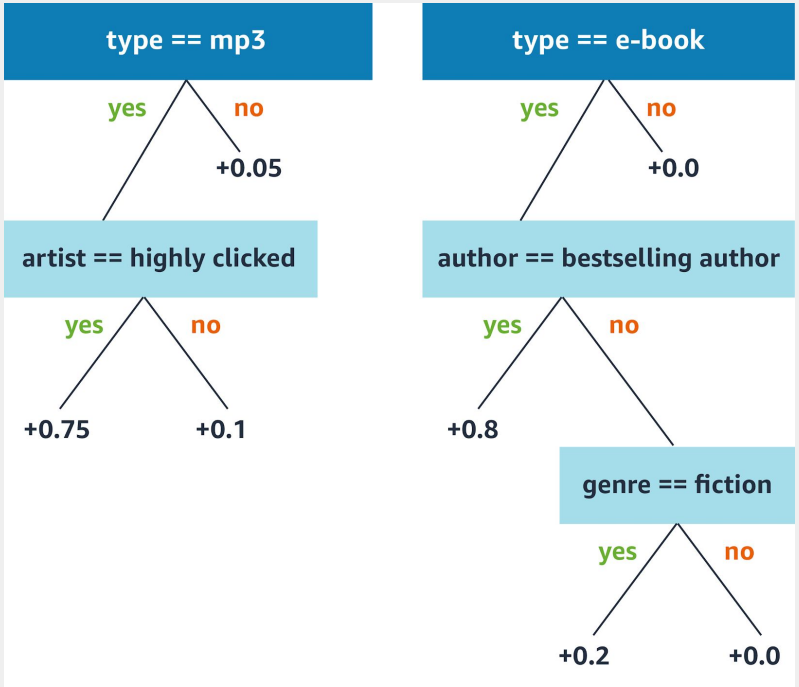
- How the model makes decisions in general (*global*)
- How the model made a specific decision for an individual (*local*)

Explainability: can you communicate all this to another human, at varying levels of technical detail?



White vs. black box models

Some kinds of models are easier to interpret and explain



The appeal (and risk) of black box models

Complex models may do some tasks very well …

… but at the cost of interpretability and explainability

Growth of “Explainable AI” (aka XAI) tools to aid in understanding models

Simpler models may perform just as well in many cases, or close enough

Benefits of interpretability and explainability

Interpretable and explainable models:

- Offer a window into their operation
- Build trust with stakeholders
- Anticipate likely growth of regulatory oversight
- Enhance effectiveness and creativity because what works is identifiable and can be built upon
- Provide a chance to detect and intervene in potential bias

Tools for explaining models

Feature importance

SHAP (SHapley Additive exPlanations): learn how a feature contributes to the model's output

LIME (Local Interpretable Model-agnostic Explanations): focuses on local interpretability

Interpretable CNNs (Convolutional Neural Networks): layers of the network each recognize specific components of images

Methods for implementing with various tools

Designing and assessing models

Strength of interpretable and explainable models: identifying potential issues of bias

Significance of diverse teams who may recognize concerns, ensure adherence to values

Ethics checklists

Formal “AI audit” procedures by trained experts

Monitoring fairness after models are in production

Involving humans

Upskilling everyone involved in process so they can understand these issues

Retaining humans in the loop: keep human wisdom in the decision-making process, especially in ethically fraught or uncertain situations

Participatory machine learning: a strategy to gather input from people affected by models’ usage

Resources

Interpretable Machine Learning, Christoph Molnar (free ebook)

Ethics.fast.ai: ethics of AI ‘syllabus’ with videos, readings, etc., plus labs for coders

The Assessment List for Trustworthy Artificial Intelligence (ALTAI), European Commission document

Race After Technology, book by Ruha Benjamin, Princeton

Thank you!

Susan Currie Sivek, Ph.D.

susan.sivek@alteryx.com

@susansivek