Fair and transparent machine learning at Ahold Delhaize

challenges and research directions

Hinda Haned

April 12th, 2019

About Ahold Delhaize

- 6,769 stores worldwide
- 372,000 associates
- 50 million customers/week

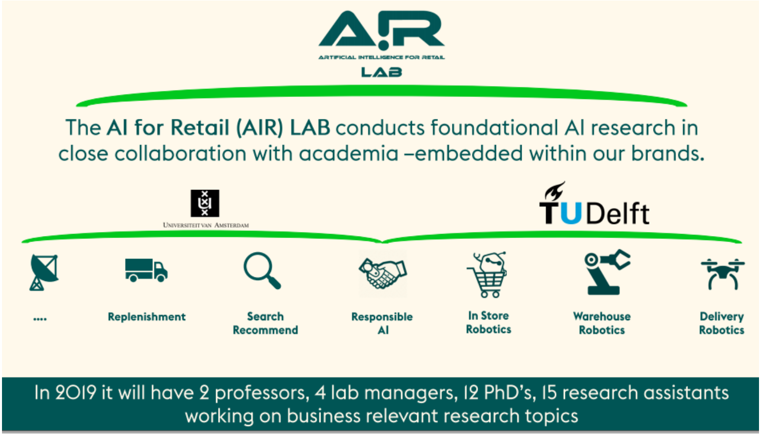
Data science at Ahold Delhaize - to modify/remove

- Main topics:
  - Personalization of promotions and recommendations
  - Alternative healthier products
  - Supply chain and logistics
- Examples of projects:
  - Cybersecurity: vulnerable users
  - Predict congestion in warehouses
  - Effects of associate training on sales
  - Sales forecasting (online and offline)

Research collaborations

- ‘Work to wellbeing’ (2016 – 2020): Methods, software, ethical framework for data gathering and analysis of employee wellbeing
- AI for Retail Lab – 2018: research into socially responsible algorithms that can be used to make recommendations to consumers and into transparent AI technology for managing goods flows

More on AIRLab



Responsible ML in retail: why bother?

- ‘AI hype’ pushed by vendors/consultants
- GDPR 2018: ‘right to an explanation’ & privacy restrictions
- Critical areas in retail, e.g. associate health, recruitment, replenishment, logistics, ...
- Media coverage

Algorithmic (un)fairness


The Telegraph

HOME | NEWS | S

Technology Intelligence

Gadgets | Innovation | Big Tech | Start-ups | Politics of Tech | Gaming | Podcast

Technology Intelligence

Amazon scraps 'sexist AI' recruiting tool that showed bias against women

The Telegraph, October 2018

Motivational example 1

Task: predicting next month's sales

- Current model
  - transaction history
  - auto-regressors
- New model
  - gradient boosting regressor
  - 40+ features

Motivational example 1

Task: predicting next month's sales

- Current model
  - transaction history
  - auto-regressors

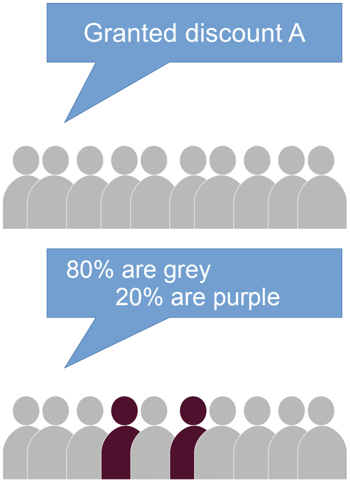
- New model
  - gradient boosting regressor
  - 40+ features

User feedback

- Model perceived as a black-box
- Counter-intuitive results
- Gain in performance vs. loss in interpretability

Motivational example 2

Task: personalized promotions



- Interventions could offer different privileges based on group membership, for example, advantageous promotions or differential pricing
- Classifications are sticky: avoid harmful classifications to stick around

Business questions

Stakeholders requirements

- *"We would like to make sure we are treating our customers* **fairly**"
- *" We need to be able to* **mitigate** *unintended bias in algorithms we develop*"
- *"We should be able to provide meaningful* **explanations** *about recommendations made*"

Goal 1 develop a non-discriminatory decision-making process while preserving as much as possible the quality of the decision.
Goal 2 ensure transparency of algorithmic outcomes, in a way that empower end-users.

State of the art

- Most research is around fairness
  - definitions of fairness vs. protected attributes/groups
  - tradeoff with accuracy
  - applications: credit scores, COMPAS sentencing software, recruitment bias
- Explainability
  - local explanation for particular outcomes
  - applications: model understanding for developers and users

- Fairness in criminal justice risk assessments: The state of the art, Berk et al, 2017
- A Survey of methods for explaining black box models, Guidotti et al, 2018

Challenges existing methods

- Fairness research:
  - assumes access to protected attributes
  - bias can be discovered
  - objective function can be constrained
  -
- Explainable ML research:
  - interpretability vs. explainability
  - assumes homogenous user-base
  - temporal aspect unaccounted for

Challenges: theory vs. real world

- Structural & Organizational
  - undefined or unclear targets
  - available data vs. relevant proxies
  - legal challenges vs. data use
- Data & users
  - mostly sequential information/time series
  - small data sets ($n >> p$)
  - lack of data/models legacy
  - diverse user base

Mitigation algorithms

Methods for fair classification

- Pre-processing: modify the train data
- In-processing: modify the algorithm's objective function to incorporate fairness constraints/penalty
- Post-processing: modifies the predictions produced by the algorithm

Fairness algorithms

Pre-processing	Re-weighing (Kamiran & Calders, 2012)
	Optimized pre-processing (Calmon et al., 2017)
	Learning fair representations (Zemel et al., 2013)
	Disparate impact remover (Feldman et al., 2015)
In-processing	Adversarial debiasing (Zhang et al., 2018)
	Prejudice remover (Kamishima et al., 2012)
Post-processing	Equalized odds post-processing (Hardt et al., 2016)
	Calibrated eq. odds postprocessing (Pleiss et al., 2017)
	Reject option classification (Kamiran et al., 2012)

Fairness algorithms

- For all methods, there is a trade-off to be found between utility and a desired measure of fairness
- Sensitive attributes are known
- Ground truth or observable outcomes are available
- Fairness-preserving algorithms tend to be sensitive to fluctuations in dataset composition, and to different forms of pre-processing (Friedler et al, 2019)

Open questions

- What does it mean for a model to be fair?
- What is an explanation?
- When is a model or an explanation comprehensible?
- What are the problems requiring interpretable/fair models?
- How much are *we* willing to compromise?

Current research projects

- Fairness through awareness
  - fair segmentation to avoid harm of allocation
  - price differentiation vs. unintended discrimination

- Transparency through explainability
  - Contrastive explanations to enhance user trust (Ana Lucic, Maarten de Rijke)
  - Global model insights from local explanations (Ilse van der Linden, Evangelos Kanoulas)
  - Explainability over time: time-series forecasting vs. explanations (Gall.nl)