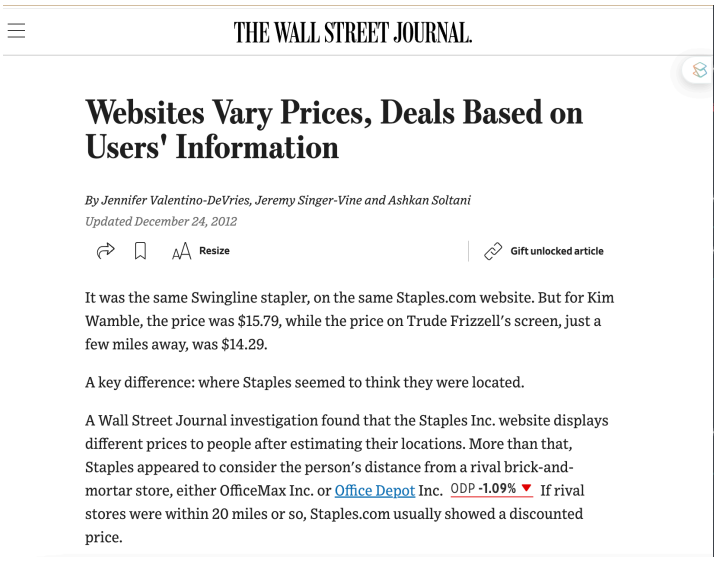
Transparency and Interpretability in Data Science

Introduction **Beyond the Blackbox** **Explanatory Methods**

What is happening?

The extensive application of machine learning models in modern life not only brings convenience but also leads to concerns about how the models work.



THE WALL STREET JOURNAL.

Websites Vary Prices, Deals Based on Users' Information

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani
Updated December 24, 2012

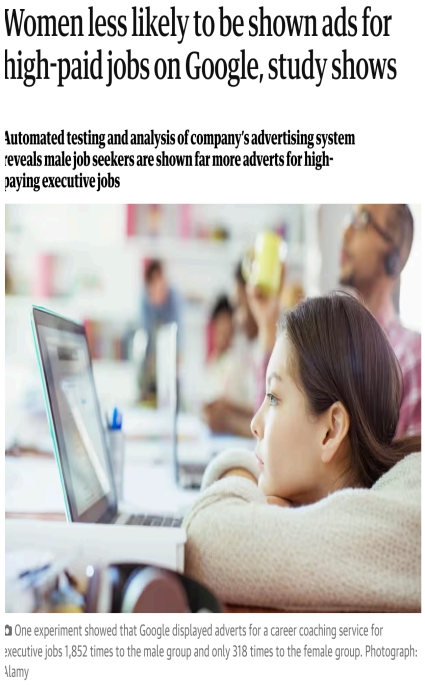
Resize | Gift unlocked article

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

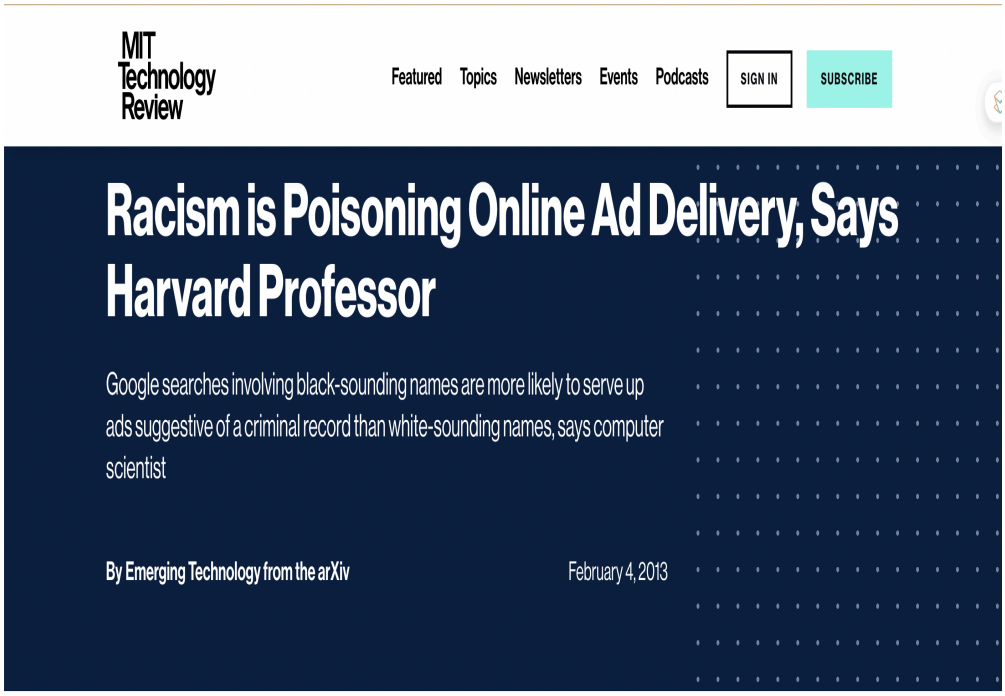
A Wall Street Journal investigation found that the Staples Inc. website displays different prices to people after estimating their locations. More than that, Staples appeared to consider the person's distance from a rival brick-and-mortar store, either OfficeMax Inc. or Office Depot Inc. ODP -1.09% If rival stores were within 20 miles or so, Staples.com usually showed a discounted price.

A Wall Street Journal investigation discovered that Staples.com offered different prices to customers based on their estimated location and proximity to rival stores. The customers, as an indispensable part of this data cycle, are treated differently due to the unknown pricing algorithm.


Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs

One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy


MIT Technology Review

Featured Topics Newsletters Events Podcasts SIGN IN SUBSCRIBE

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

By Emerging Technology from the arXiv February 4, 2013

A study using the AdFisher tool to simulate job seekers that only differ in gender to conduct an experiment about online ad personalization. It found that Google displayed ads for high-paying executive jobs more frequently to males than to females, raising questions about potential discrimination in algorithmic decision-making. Recalling the moment that the pop-up windows from Google asked if you are willing to share your privacy data for further analysis and you clicked “agree”, you might not expect to be treated differently because of gender while you are just laid off and actively seeking a job. Moreover, researchers found that users with black-sounding names are more likely to receive ads about criminal records, compared to users with white-sounding names. These examples strongly emphasize the significance of data transparency.

Achieving Fairness by Opening the Blackbox

Employing models without transparency could impair fairness in gender, race, and other demographic factors. While laws protect everyone’s property from depriving in any means, privacy should be protected from violation as well, by ensuring the model’s fairness, enhancing the interpretability, and providing access to understand the algorithm. Complex models operate as “blackboxes”, with users’ data inputting and decisions made as output. To ensure transparency, we need explanatory methods to peel back the layers to reveal the decision-making process. Interpretability takes transparency further by not just revealing the workings of an AI system but also making it understandable to humans. It is about transforming the complex language of algorithms into a narrative that stakeholders can grasp. To achieve this, we need explanatory methods to further analyze.

This website briefly shows my understanding in the course DS-UA 202 Responsible Data Science