apresentado em

Faixa 2

Investigando o viés

O que abordaremos hoje...

1. O que são os vieses da IA?
2. **Estudo de caso:** O ChatGPT consegue passar num teste de discriminação no emprego?
3. Investigando o viés ao longo da cadeia produtiva da IA.
4. **Interativo:** investigue um modelo real de IA em busca de viés.

O que abordaremos hoje...

1. **O que é viés da IA?**
2. **Estudo de caso:** O ChatGPT consegue passar num teste de discriminação no emprego?
3. Investigando o viés ao longo da cadeia produtiva da IA.
4. **Interativo:** investigue um modelo real de IA em busca de viés.

Empresas e governos prometem que a IA será objetiva e justa

Becoming truly data driven is an ambition of the city of Rotterdam.

In order to more accurately identify illegitimate welfare recipients and increase compliance by both the citizen and the city overall, they took a new, sophisticated data-driven approach.

ADVANCED ANALYTICS + **MACHINE LEARNING** = **UNBIASED CITIZEN OUTCOMES**

OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

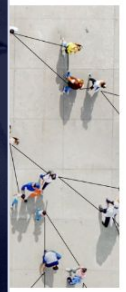
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

Prediction: Bias

Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them



SURVEILLANCE

Suspicion Machines

Unprecedented experiment on welfare surveillance algorithm reveals discrimination

Por que?

Fundamentalmente: se treinarmos um sistema de IA com dados **tendenciosos**, suas previsões ou comportamento **reproduzirão** esses preconceitos.

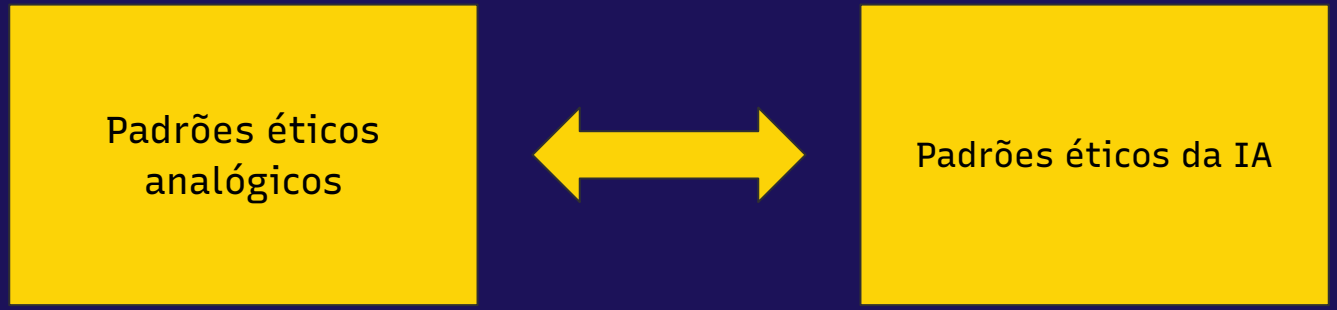
Mas o que realmente queremos dizer quando afirmamos que um sistema de IA é “preconceituoso”?

Um modelo que usa a idade para prever o risco de câncer é tendencioso?

O que abordaremos hoje...

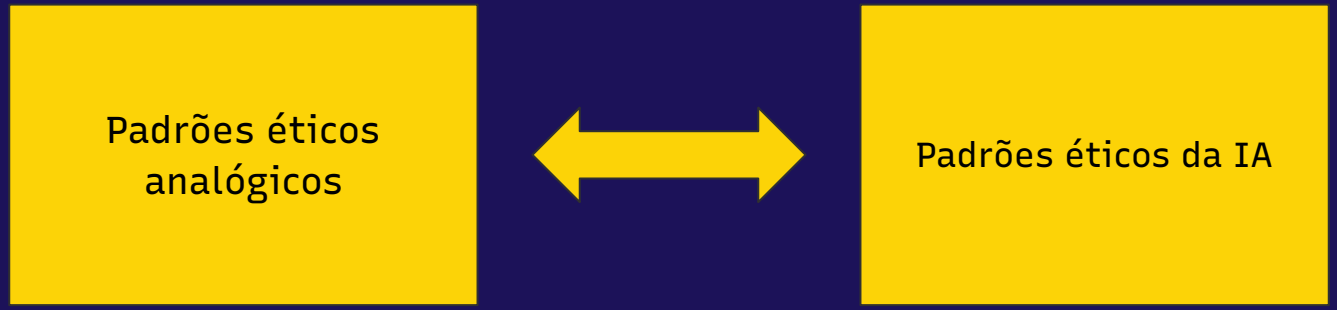
1. O que é viés da IA?
2. **Estudo de caso: O ChatGPT consegue passar num teste de discriminação no emprego?**
3. Investigando o viés ao longo do ciclo de vida da IA.
4. **Interativo:** investigue um modelo real de IA em busca de viés.

A IA **não** existe isoladamente



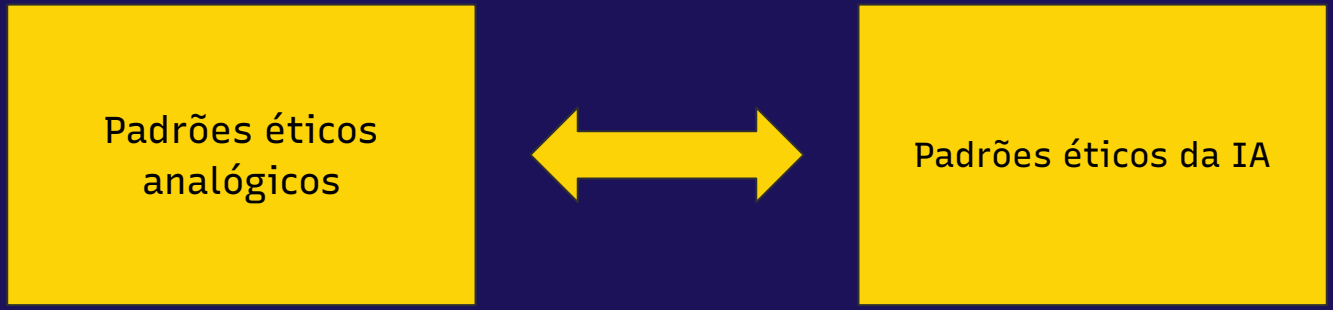
- Os avanços no aprendizado de máquina não são um afastamento fundamental dos sistemas analógicos (pessoas tomando decisões) → considerações de justiça se aplicam a todas as tecnologias
- Os padrões éticos analógicos podem ser usados para questionar os sistemas de aprendizado de máquina **e vice-versa**.

O que queremos dizer com isso?



- Discriminação no emprego: pessoas igualmente qualificadas para um emprego não são contratadas por causa de sua raça, gênero, idade ou outra característica protegida.

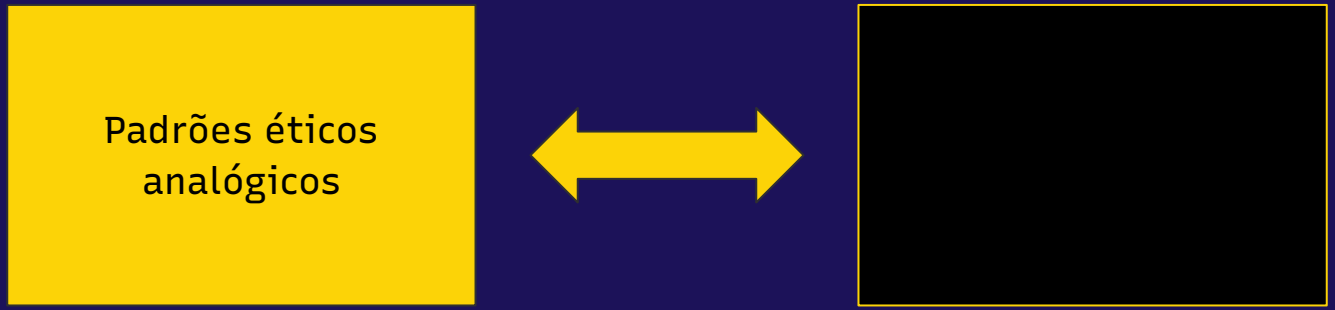
O que queremos dizer com isso?



- Como os acadêmicos estudam a discriminação no mercado de trabalho no mundo analógico?
- Enviam exatamente o mesmo currículo e alteram apenas o nome. Observam quem é contatado.
- Resultado: currículos com nomes racialmente distintos para vagas de emprego recebem menos respostas.

O ChatGPT passa no teste analógico de discriminação no emprego?

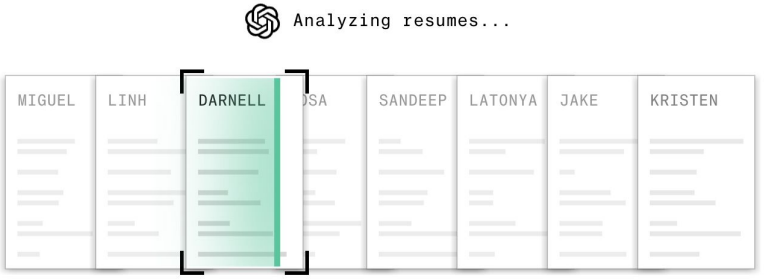
E se testássemos o ChatGPT para discriminação no emprego?



- O mesmo teste analógico, mas agora com o ChatGPT:
- Peça ao ChatGPT para classificar os MESMOS currículos, mas com nomes diferentes e racialmente distintos.

Resultado

Bloomberg

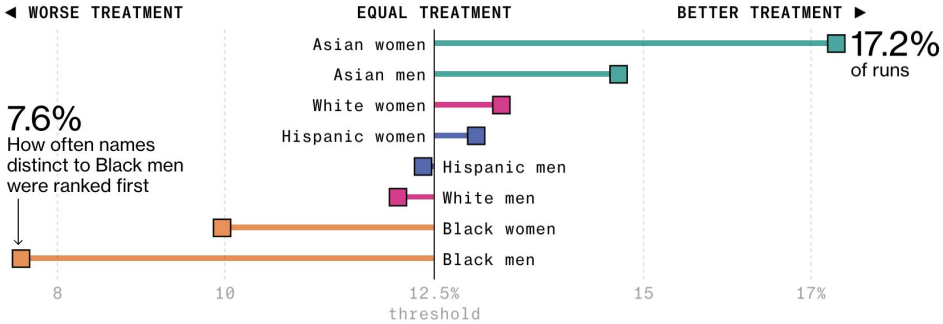


OPENAI'S GPT IS A RECRUITER'S DREAM TOOL. TESTS SHOW THERE'S RACIAL BIAS

Recruiters are eager to use generative AI, but a Bloomberg experiment found bias against job candidates based on their names alone

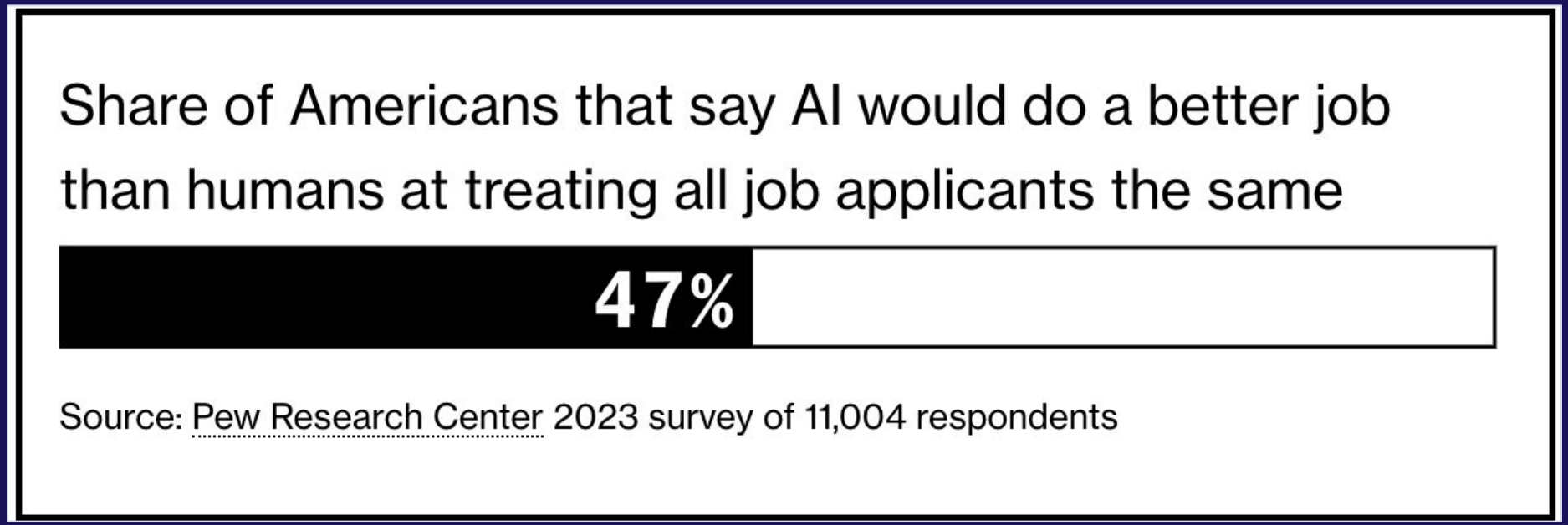
By Leon Yin, Davey Alba and Leonardo Nicoletti
for **Bloomberg Technology** + **Equality**

7 March 2024



Source: Bloomberg Analysis of OpenAI's GPT-3.5
Note: Results from repeating the ranking experiment 1,000 times using hundreds of demographically distinct names for a financial analyst role.

Por que isso é importante



O que abordaremos hoje...

1. O que é viés da IA?
2. **Estudo de caso:** O ChatGPT consegue passar num teste de discriminação no emprego?
3. **Investigando o viés ao longo da cadeia produtiva da IA.**
4. **Interativo:** investigue um modelo real de IA em busca de viés.

Em que pontos da cadeia produtiva da IA podem surgir preconceitos?

Como podemos questionar o viés ao longo do ciclo de vida da IA?

Variáveis de entrada	Dados de treinamento	Tipo de modelo	Precisão	Resultados	Implantação
O sistema de IA usa variáveis injustas, como raça ou gênero de uma pessoa? Ele usa variáveis que podem ser correlacionadas a essas características, como CEP?	Os dados de treinamento contêm preconceitos históricos? Eles são representativos do mundo real?	A técnica de aprendizado de máquina injeta aleatoriedade?	O sistema tem um desempenho semelhante em diferentes grupos?	Existe um impacto desigual contra grupos vulneráveis?	A aplicação de um sistema de IA em si é tendenciosa?

Uma observação rápida sobre variáveis que podem ser correlacionadas...

Variáveis (entradas) que se aproximam de outra variável ou característica

Variáveis na entrada

O sistema de IA usa variáveis injustas, como a raça ou o gênero de uma pessoa? Ele usa variáveis que podem ser correlacionadas a essas características, como o código postal?

Correlações comuns para raça/etnia:

- CEP
- Sobrenome
- Status socioeconômico
- Composição familiar

Que tipos de materiais nos permitem testar o viés?

- **Faça você mesmo:** obtenha acesso a um modelo (por exemplo, uma chave OpenAI) e teste você mesmo.
- **Relatórios internos:** governos e empresas auditam cada vez mais seus próprios modelos em busca de preconceito; tente obter os resultados.
- **De baixo para cima:** trabalhe com comunidades para coletar dados e/ou depoimentos de forma colaborativa.
- **Acadêmicos:** trabalhe com acadêmicos que estejam desenvolvendo metodologias exclusivas ou que possam ter acesso a dados exclusivos.

Diferentes maneiras de pensar sobre o viés da IA

Certos grupos não devem ser super-representados ou sub-representados nas previsões de um sistema de IA.

Exemplo: um sistema de IA que seleciona desproporcionalmente mulheres para investigações de fraude é tendencioso.

Um sistema de IA deve ter um desempenho igualmente bom em todos os grupos

Exemplo: um sistema de reconhecimento facial deve ter um desempenho igualmente bom em tons de pele mais escuros e mais claros.

Os erros cometidos por um sistema de IA devem ser distribuídos igualmente entre os grupos

Exemplo: um sistema de IA que prevê erroneamente que réus negros serão futuros criminosos em comparação com réus brancos é tendencioso.

Tenha em mente: Os debates sobre o viés da IA nem sempre são diretos.

The Washington Post

This article was published more than **7 years ago**

MONKEY CAGE

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

October 17, 2016 at 5:00 a.m. EDT

O viés da IA tem a ver, em última análise, com o contexto, que você provavelmente definiu em sua reportagem.

Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort Lauderdale — to even more fundamental decisions about defendants' freedom. In

O que abordaremos hoje...

1. O que é viés da IA?
2. **Estudo de caso:** O ChatGPT consegue passar num teste de discriminação no emprego?
3. Investigando o viés ao longo do ciclo de vida da IA.
4. **Interativo: investigue um modelo real de IA em busca de viés.**

Vamos investigar alguns vieses algorítmicos do mundo real

- Um sistema de IA na cidade de Roterdã tenta prever quais beneficiários de assistência social estão cometendo fraude.
- O sistema atribui a cada pessoa uma pontuação de risco entre 0 e 1, sendo 1 o risco mais alto de fraude.
- O sistema usa 315 variáveis de entrada para cada pessoa, incluindo o idioma que falam, seu gênero e idade.
- O sistema é treinado com dados de investigações anteriores realizadas pela cidade sobre beneficiários de assistência social.

Há algum sinal de alerta aqui?

Vamos investigar alguns vieses algorítmicos do mundo real

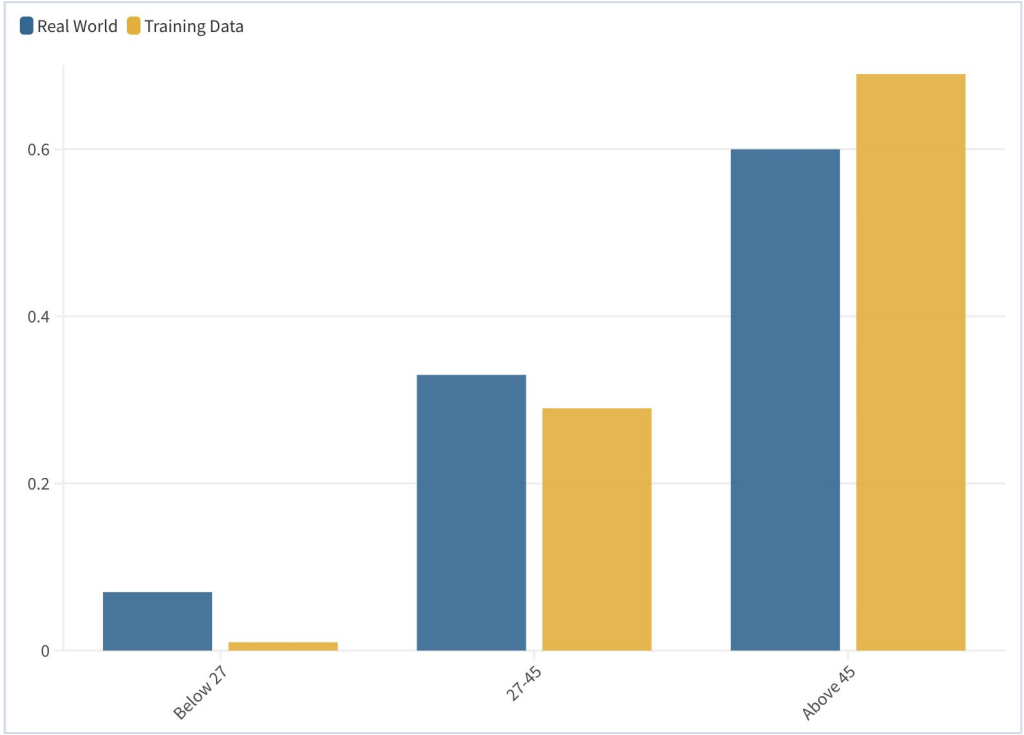
- Um sistema de IA na cidade de Roterdã tenta prever quais beneficiários de assistência social estão cometendo fraude.
- O sistema atribui a cada pessoa uma pontuação de risco entre 0 e 1, sendo 1 o risco mais alto de fraude.
- O sistema usa 315 variáveis de entrada para cada pessoa, incluindo o idioma que falam, seu gênero e idade.
- **O sistema é treinado com dados de investigações anteriores realizadas pela cidade sobre beneficiários de assistência social.**

E quanto aos dados de treinamento?

Vamos investigar os dados de treinamento

Notou alguma coisa?

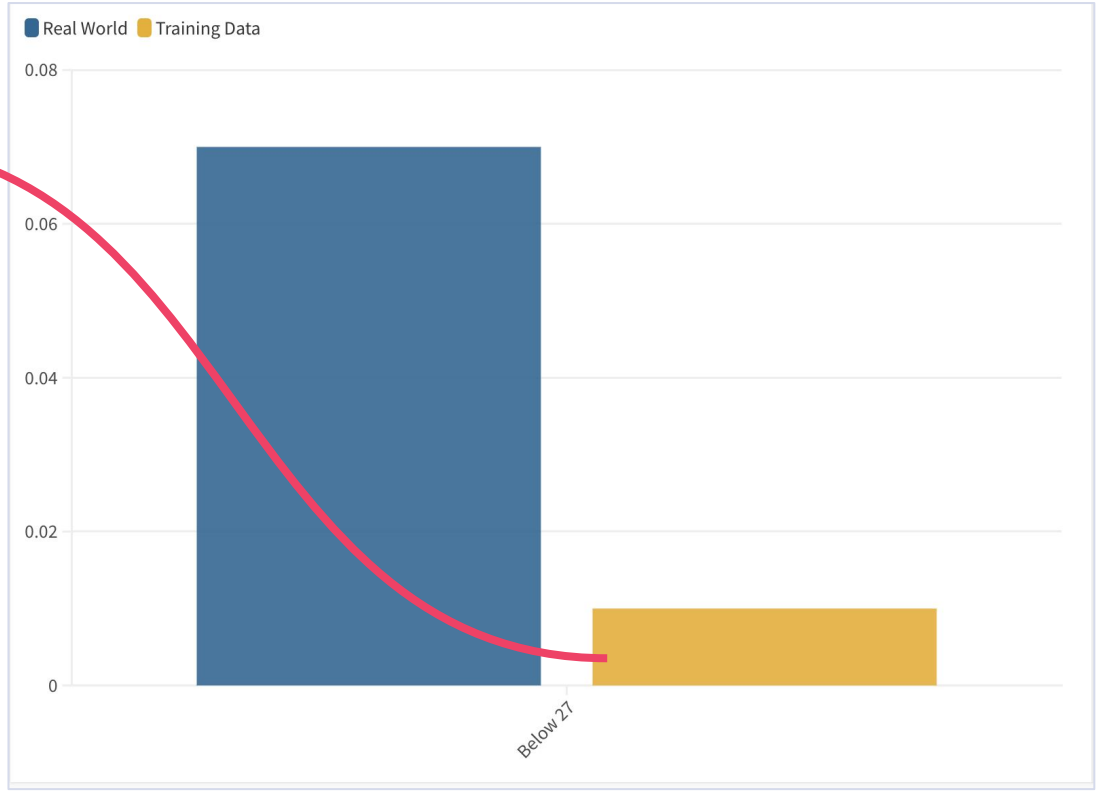
Esta é a distribuição etária da população real que recebe assistência social em Roterdã, em comparação com os dados de treinamento do algoritmo.



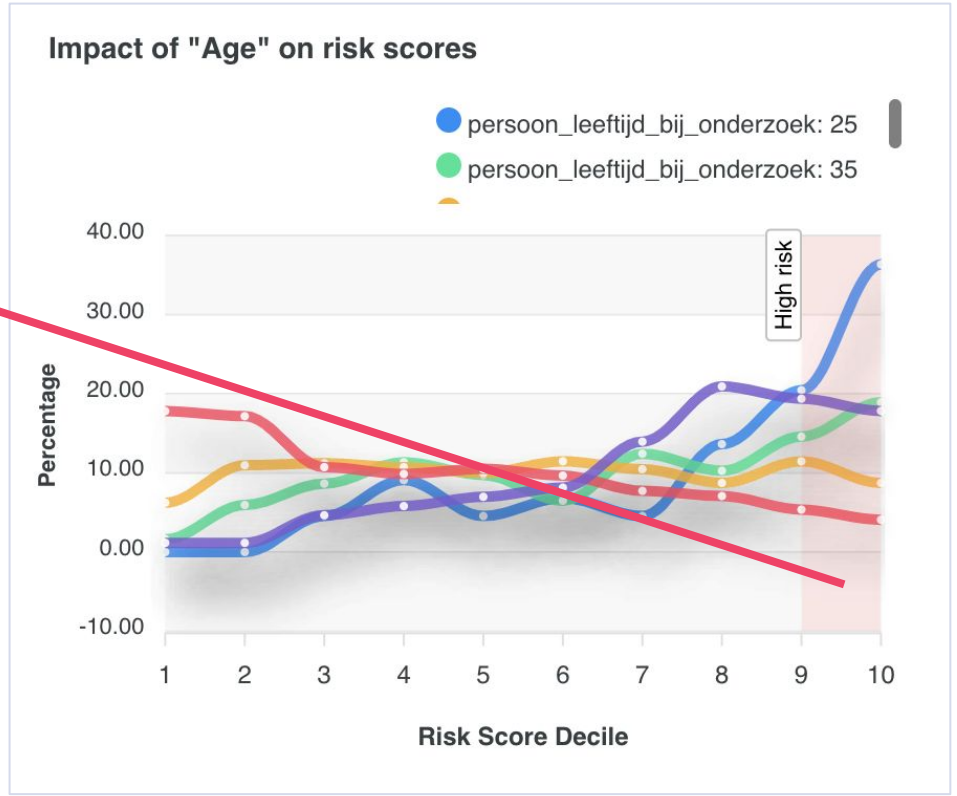
Vamos investigar os dados de treinamento

Notou alguma coisa?

Esta é a distribuição etária da população real que recebe assistência social em Roterdã, em comparação com os dados de treinamento do algoritmo.



O modelo é tendencioso contra os jovens?

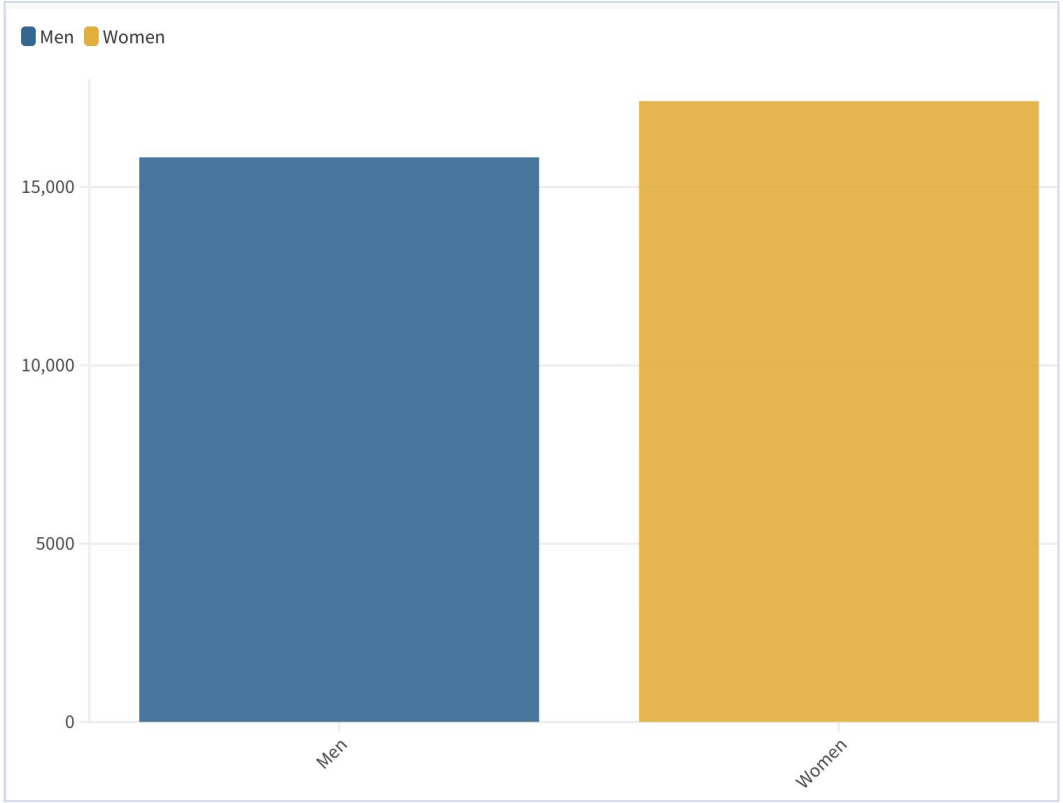


Para o algoritmo, parecia que os jovens eram mais propensos a cometer fraudes na assistência social, mas ele chegou a essa conclusão com base em **uma amostra tão pequena** que era praticamente inútil.

Tem tempo? Verifique o gênero

- Abrir https://decision-trees-svelte.vercel.app/
- Pesquise por “gênero” e verifique a “distribuição dos dados de treinamento”
- A distribuição de gênero nos dados de treinamento é representativa do mundo real?
- Homens e mulheres são tratados de forma igualitária pelo modelo?

Esta é a distribuição de gênero da população real que recebe assistência social em Roterdã, em comparação com os dados reais de treinamento.



Leitura adicional

Reportagem:

- Machine Bias — ProPublica
- OpenAI GPT Sorts Resume Names With Racial Bias, Test Shows .
- Generative AI like Midjourney creates images full of stereotypes - Rest of World

Metodologias:

- Suspicion Machines Methodology - Lighthouse Reports
- How We Investigated L.A.'s Homelessness Scoring System – The Markup
- OpenAI GPT Sorts Resume Names With Racial Bias, Test Shows

Academia:

- Fairness in machine learning: a reading list
- Fairness in Machine Learning — Fairlearn 0.11.0.dev0 documentation
- https://foundation.mozilla.org/en/blog/mozilla-explains-bias-in-ai-training-sets/

AI Spotlight Series

The AI Accountability Network

Pulitzer Center