

Ferramentas para mitigar vieses em IA

Gabriela de Queiroz, @gdequeiroz
Paolla Magalhães

Quem Somos?



Paolla Magalhães

Cientista de Dados, iFood



Gabriela de Queiroz

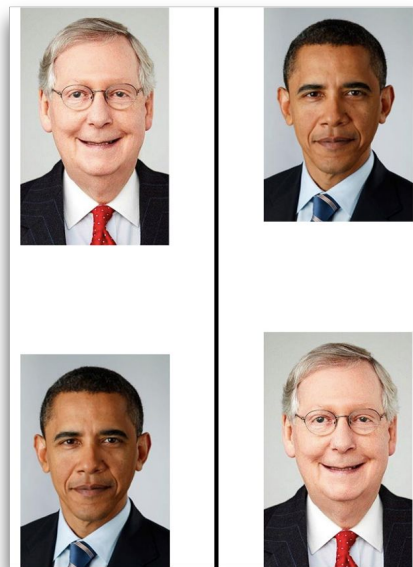
Chief Data Scientist, IBM

Agenda

- Motivação
- Justiça em Machine Learning
- Terminologia
- Definição do Problema
- Viés/Justiça
- Avaliação de Justiça
- Métricas
- Mitigação do Viés
- AIF 360
- Atividade prática

Motivação

- Problema: Modelos discriminatórios
- Facilitar a busca por conhecimento nessa área
- Abrir a caixa preta!



Twitter investigates racial bias in image previews

Justiça em Machine Learning

Discriminação em ML

Em sua grande maioria vem do viés existente nos dados.

- Discriminação histórica
- Pouca informação sobre minorias
- Procedimentos de anotação dos dados

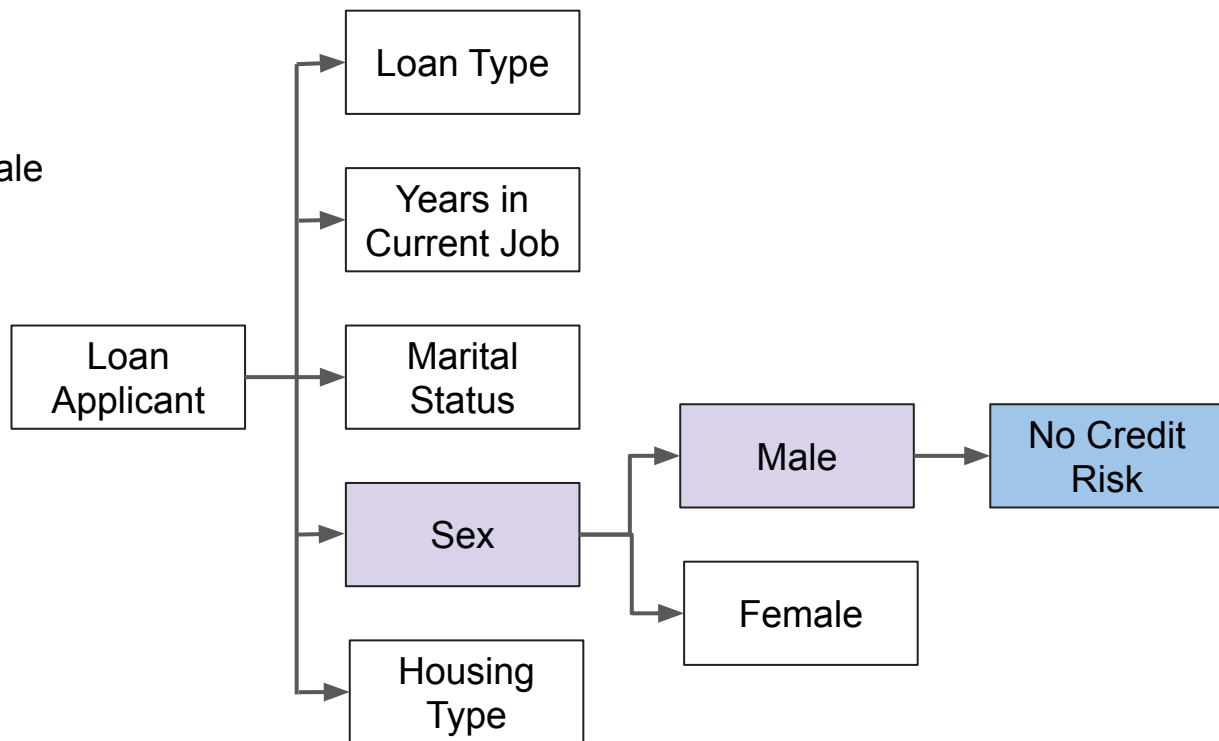


AI Fairness 360 (AIF 360)

Terminologia

- **Favorable label:** rótulo cujo valor corresponde a um resultado que oferece uma vantagem (como receber um empréstimo, ser contratado para um emprego, não ser preso)
- **Protected attribute:** um atributo que divide uma população em grupos cujos resultados devem ter paridade (como raça, sexo, casta e religião)
- **Valor privilegiado** (de um atributo protegido) (privileged value) : Um valor de atributo protegido que indica um grupo que historicamente tem uma vantagem sistêmica
- **Discriminação / viés indesejado:** Quando específicos grupos privilegiados são colocados em uma vantagem sistemática e específicos grupos não privilegiados são colocados em desvantagem sistemática. Isso está relacionado a atributos como raça, sexo, idade e orientação sexual.

- Favorable label: No Credit Risk
- Unfavorable label: Credit Risk
- Protected Attribute: Sex
- Privileged Protected Attribute: Male



Definição do Problema

Definição do Problema

Como decidir o **atributo sensível** (protected attribute)?

- Vai depender do problema
- O atributo é sensível ou é necessário?

Como decidir as classes **privilegiadas e não privilegiadas**?

- Pesquisa científica do histórico de discriminação geral na sociedade
- Análise estatística pelo balanceamento do atributo sensível (Ex: quantidade de pessoas do grupo menos privilegiado que tiveram resultado favorável vs do grupo mais privilegiado)

Como definir o **resultado favorável**?

- Se o resultado for favorecer um determinado indivíduo, ou seja, se for ajudar ou privilegiar um determinado indivíduo.

Exemplo

CARGO EM ENGENHARIA

HOMENS

MULHERES

HOMENS



Exemplo

CARGO EM PEDAGOGIA

MULHERES

HOMENS

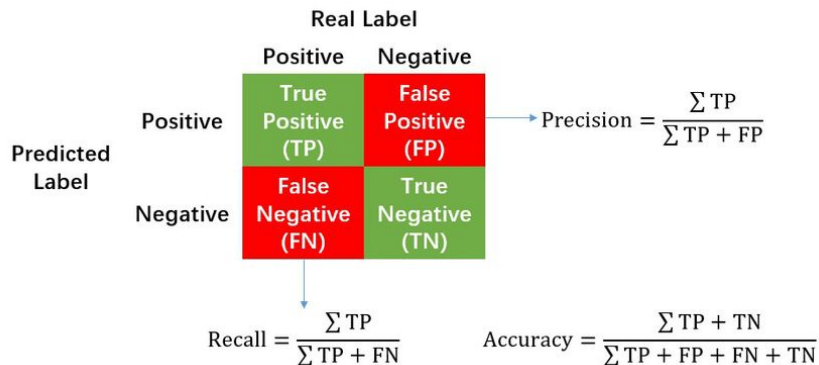
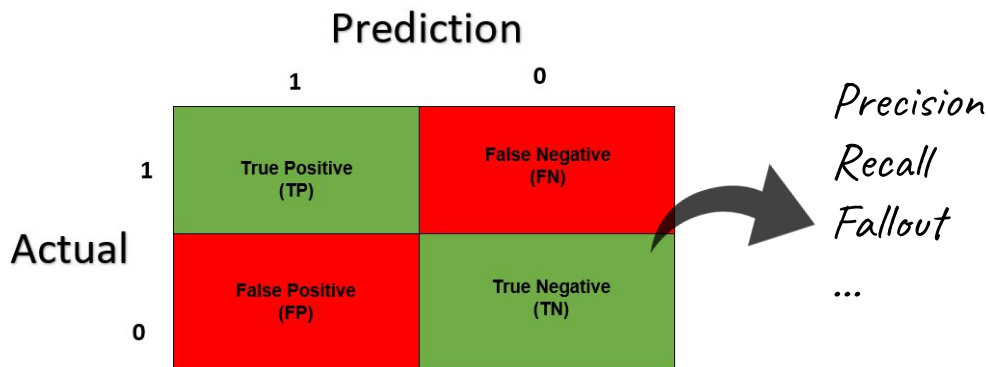
MULHERES



**Como fazer uma máquina entender o
que é viés/justiça?**

Quantificação de Justiça

Utilizam-se **probabilidades** ou métricas retiradas da **matriz de confusão**.



Quantificação de Justiça

IGUALDADE DE OPORTUNIDADE

$$Fallout_{unpriv} = Fallout_{priv}$$

$$Recall_{unpriv} = Recall_{priv}$$

PARIDADE ESTATÍSTICA

$$P_{unpriv}(1) = P_{priv}(1)$$

PARIDADE PREDITIVA

$$Precision_{unpriv} = Precision_{priv}$$

Avaliação de Justiça

Avaliação de Justiça

Métricas de Performance

ACCURACY

PRECISION

RECALL

F1



Métricas de Justiça

RELAÇÃO DE IMPACTO DESIGUAL
(Disparate Impact)

DIFERENÇA DE PARIDADE ESTATÍSTICA
(Statistical Parity Difference)

DIFERENÇA DE MÉDIA DE PROBABILIDADE
(Average Odds Difference)

DIFERENÇA DE IGUALDADE DE OPORTUNIDADE
(Equal Opportunity Difference)



Métricas

RELAÇÃO DE IMPACTO DESIGUAL
(Disparate Impact)

$$0.80 < \frac{P_{unpriv}(1)}{P_{priv}(1)} < 1.25$$

DIFERENÇA DE MÉDIA DE PROBABILIDADE
(Average Odds Difference)

$$-0.1 < \frac{(Fallout_{unpriv} - Fallout_{priv}) + (Recall_{unpriv} - Recall_{priv})}{2} < 0.1$$

DIFERENÇA DE PARIDADE ESTATÍSTICA
(Statistical Parity Difference)

$$-0.1 < P_{unpriv}(1) - P_{priv}(1) < 0.1$$

DIFERENÇA DE IGUALDADE DE OPORTUNIDADE
(Equal Opportunity Difference)

$$-0.1 < Precision_{unpriv} - Precision_{priv} < 0.1$$

Métricas

RELAÇÃO DE IMPACTO DESIGUAL (Disparate Impact)

- Valor ideal: 1.0
- Valor < 1 : maior benefício para o grupo privilegiado
- Valor > 1 : maior benefício para o grupo não privilegiado

DIFERENÇA DE MÉDIA DE PROBABILIDADE (Average Odds Difference)

- Valor ideal: 0
- Valor < 0 : maior benefício para o grupo privilegiado
- Valor > 0 : maior benefício para o grupo não privilegiado

DIFERENÇA DE PARIDADE ESTATÍSTICA (Statistical Parity Difference)

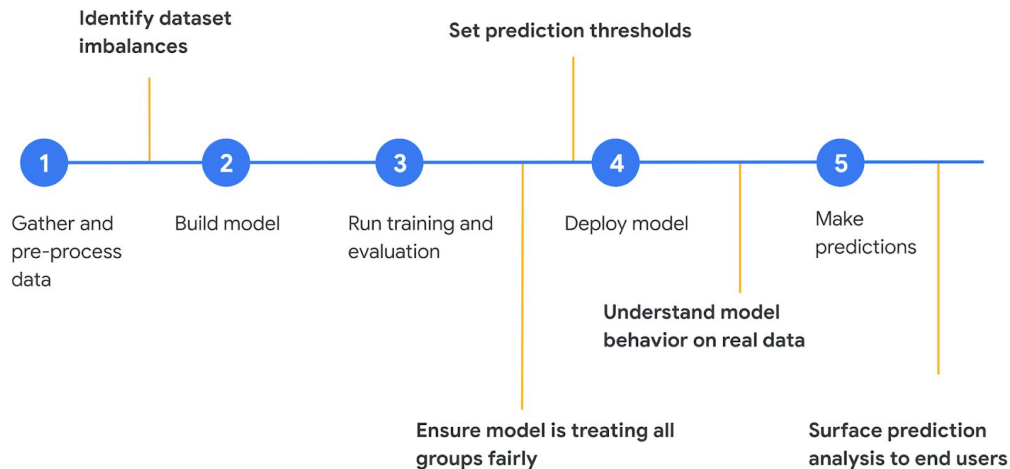
- Valor ideal: 0

DIFERENÇA DE IGUALDADE DE OPORTUNIDADE (Equal Opportunity Difference)

- Valor ideal: 0
- Valor < 0 : maior benefício para o grupo privilegiado
- Valor > 0 : maior benefício para o grupo não privilegiado

Mitigação de viés

Mitigação de viés



CLASSES DE ALGORÍTMOS

1. PRÉ-PROCESSAMENTO
2. EM PROCESSAMENTO
3. PÓS-PROCESSAMENTO

FONTE:
<https://strategicfocus.com/2019/09/25/building-ml-models-for-everyone-understanding-fairness-in-machine-learning/>

Técnicas de mitigação de viés



PRÉ PROCESSAMENTO

- Suppression
- Massaging the dataset
- Disparate impact remover
- **Reweighting**
- Optimized preprocessing
- Learning fair representations



EM PROCESSAMENTO

- **Adversarial Debiasing**
- Prejudice remover
- Meta Fair Classifier
- ART Classifier



PÓS PROCESSAMENTO

- Equalized odds
- **Calibrated equalized odds**
- Reject option classification

Ferramenta

AI Fairness 360 (AIF360)

Website: <http://aif360.mybluemix.net/>

Github: <https://github.com/Trusted-AI/AIF360>

Documentation:

<https://aif360.readthedocs.io/en/latest/>

- Toolkit extensível e de código aberto de avaliação e mitigação de viés
- Facilidade na assimilação (similar ao scikit-learn)
- Utilização em dados que você já tem
- Disponível em Python e R
- Compatibilidade com Scikit-Learn

Caso de Uso

Dataset

Dados retirados do Census bureau database de 1994 por Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics).

A tarefa de predição é determinar se a pessoa recebe mais que 50 mil dólares no ano.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype
---  -
0   age                   32561 non-null  int64
1   workclass             32561 non-null  object
2   fnlwgt               32561 non-null  int64
3   education             32561 non-null  object
4   education.num         32561 non-null  int64
5   marital.status        32561 non-null  object
6   occupation            32561 non-null  object
7   relationship          32561 non-null  object
8   race                  32561 non-null  object
9   sex                   32561 non-null  int64
10  capital.gain          32561 non-null  int64
11  capital.loss          32561 non-null  int64
12  hours.per.week        32561 non-null  int64
13  native.country        32561 non-null  object
14  income                32561 non-null  int64
dtypes: int64(8), object(7)
memory usage: 3.7+ MB
```

OBRIGADA!



Paolla Magalhães

<https://www.linkedin.com/in/paollamota/>



Gabriela de Queiroz

<https://www.linkedin.com/in/gabrieladequeiroz/>