



Função de custo para aprendizado robusto à injustiça

Ygor Canalli

Orientador: Geraldo Zimbrão da Silva

01 de setembro de 2021

PESC/COPPE/UFRJ

Table of contents

1. Função correção de injustiça
2. Metodologia
3. Resultados no *Income dataset*
4. Resultados no *German dataset*

Função correção de injustiça

Correção de função de custo

É possível criar robustez a ruído alterando a função de custo^{1,2}.

- Estima-se uma matriz de transição T capaz de descrever o ruído (NAR) de acordo com a classe
- A correção *forward* de uma função de custo $\ell()$ é definida como

$$\ell^{\rightarrow}(P(\tilde{Y}|X)) = \ell(T^T P(\tilde{Y}|X))$$

- A correção *backward* por sua vez é definida como

$$\ell^{\leftarrow}(P(\tilde{Y}|X)) = T^{-1} \ell(P(\tilde{Y}|X))$$

¹Giorgio Patrini et al. *Loss Factorization, Weakly Supervised Learning and Label Noise Robustness*. Tech. rep. 2016. URL:

<http://proceedings.mlr.press/v48/patrini16.pdf>.

²Giorgio Patrini et al. “Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach”. In: July 2017, pp. 2233–2241. doi: [10.1109/CVPR.2017.240](https://doi.org/10.1109/CVPR.2017.240).

Função correção de injustiça

Sejam

- X a matriz de atributos
- \tilde{Y} a matriz de classes previstas em formato *one hot encoding*
- S um atributo sensível em X
- $s_i \in S$ o conjunto dos grupos delimitados em S de tamanho $|S|$
- X_{S_i} um vetor binário indicando se cada instância em X pertence ao grupo s_i
- T_i a matriz de transição correspondente ao grupo $s_i \in S$
- ℓ uma função de custo

A função $\ell_S()$ de correção de injustiça sobre o atributo S é dada por

$$\ell_S(P(\tilde{Y}|X)) = \sum_{i=1}^{|S|} \ell(T_i^T P(\tilde{Y}|X)) X_{S_i}$$

Metodología

- Experimentos nos datasets *german credit* e *census income*
 - Classificação binária
 - Gênero como atributo sensível
 - Protegido: Feminino
 - Não-protégido: Masculino
- Uma matriz de transição para cada grupo
- Testar diferentes combinações de matrizes de transição
- Medir acurácia
- Medir métricas de injustiça

Parâmetros de na matriz de transição

- Para cada matriz de transição basta definir os valores da diagonal secundária, de forma que cada linha some 1.
- A taxa α é o fator de incerteza da previsão positiva, com efeito prático de rebaixar a classificação
- A taxa β é o fator de incerteza da previsão negativa, com efeito prático de promover a classificação

$$T = \begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix}$$

Busca por parâmetros

- Tomamos uma matriz de transição T_1 para a grupo não-protégido e T_2 para o protégido
- rebaixamento do grupo não-protégido $\alpha_1 \in \{0, 0.1, 0.2, \dots, 1.0\}$
- promoção do grupo não-protégido β_1 fixado em $\alpha_1/10$
- promoção do grupo protégido $\beta_2 \in \{0, 0.1, 0.2, \dots, 1.0\}$
- rebaixamento do grupo protégido α_2 fixado em $\beta_2/10$

$$T_1 = \begin{bmatrix} 1 - \alpha_1 & \alpha_1 \\ \beta_1 & 1 - \beta_1 \end{bmatrix}, \quad T_2 = \begin{bmatrix} 1 - \alpha_2 & \alpha_2 \\ \beta_2 & 1 - \beta_2 \end{bmatrix}$$

Resultados no *Income dataset*

Acurácia no Income Dataset

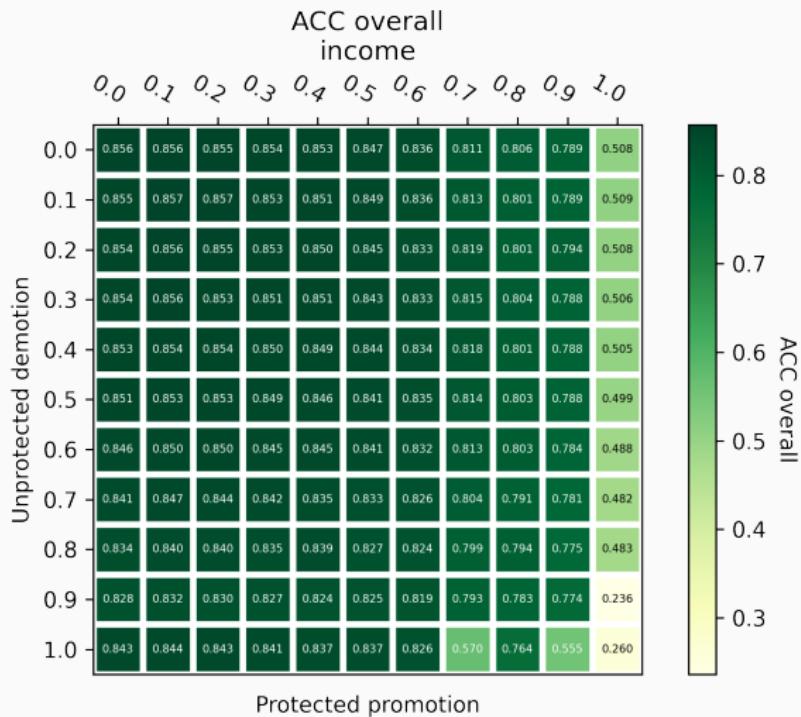


Figure 1: Acurácia geral

Accuracy Equality no Income Dataset

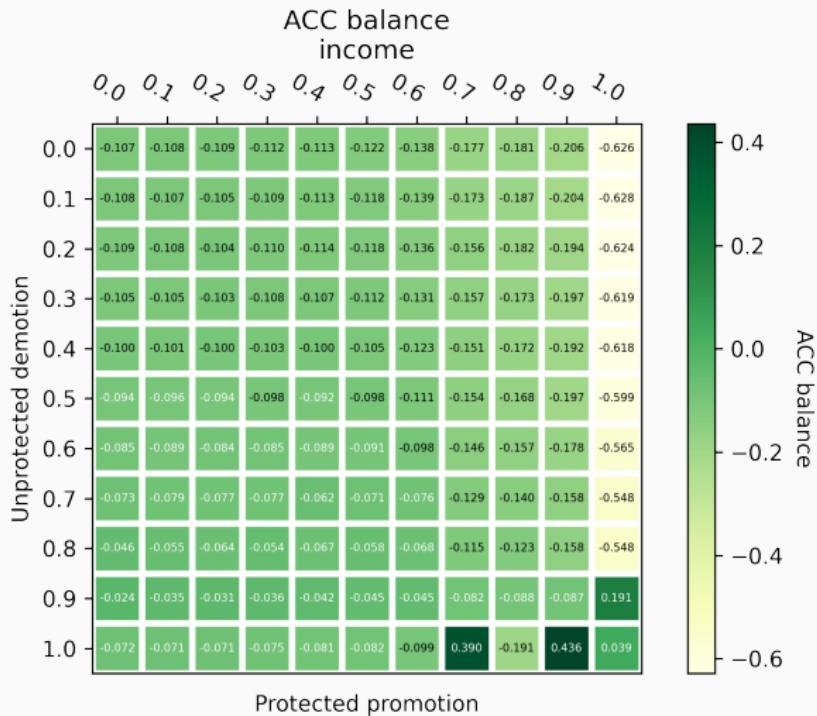


Figure 2: Balanço da acurácia, diferença entre não-protégido e protegido

Conditional Accuracy Parity no Income Dataset

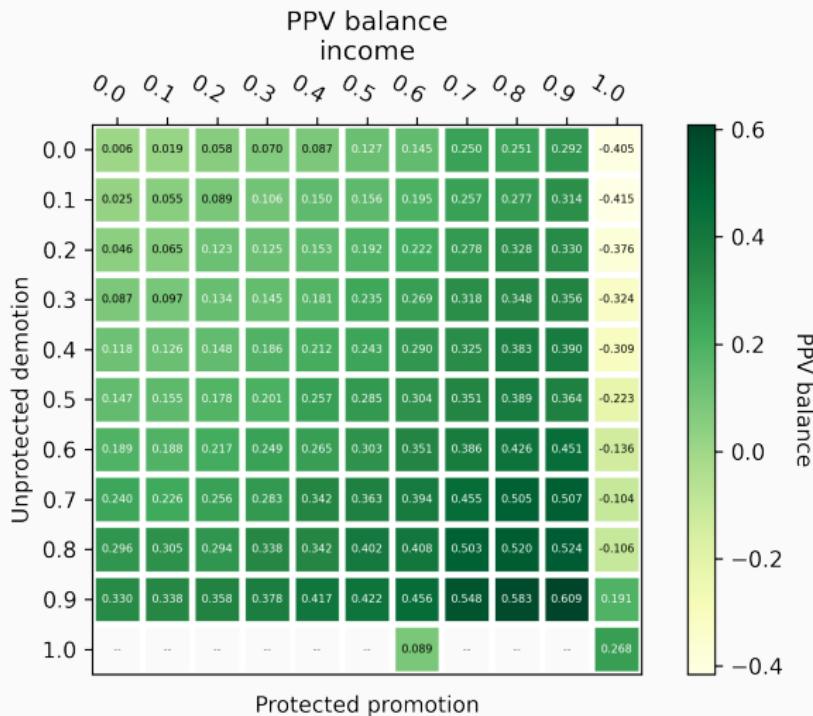


Figure 3: $PPV = \frac{TP}{TP + FP}$, diferença entre não-protégido e protegido.

Conditional Accuracy Parity no Income Dataset

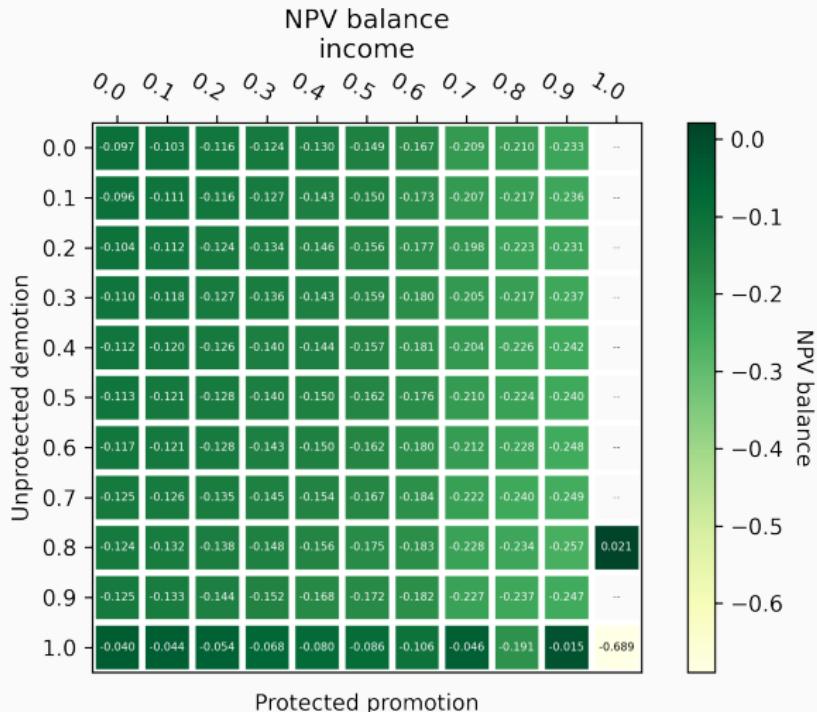


Figure 4: $NPV = \frac{TN}{TN + FN}$, diferença entre não-protégido e protegido.

Statistical Parity no Income Dataset

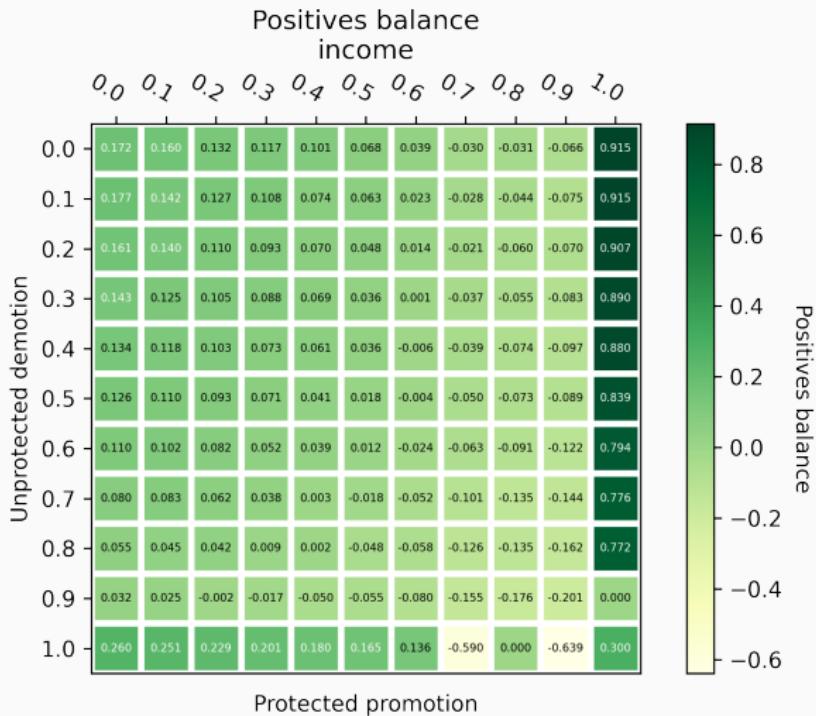


Figure 5: Balanço da taxa de previsões positivas, diferença entre não-protégido e protegido

Predictive Quality no Income Dataset

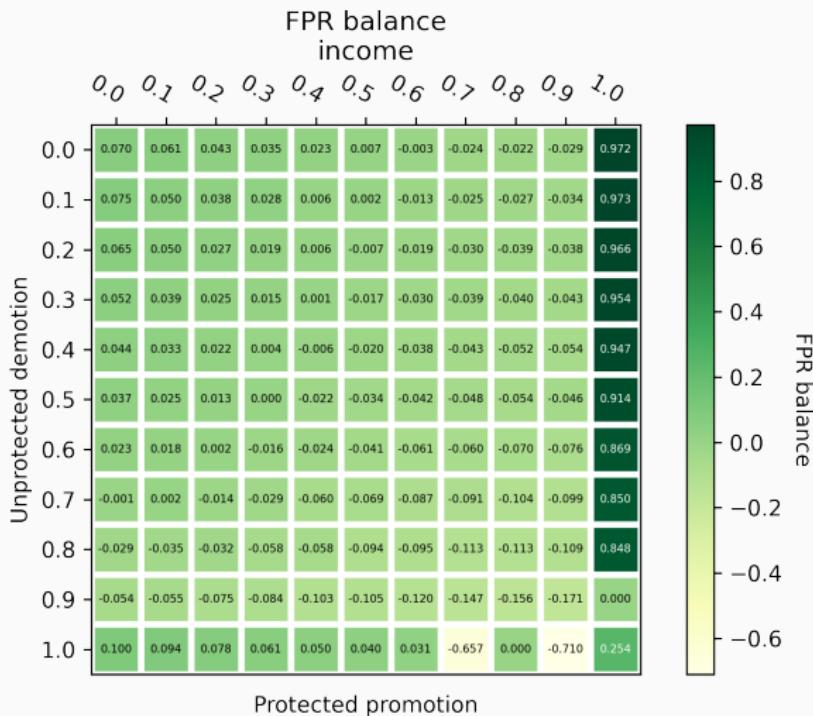


Figure 6: $FPR = \frac{FP}{TN + FP}$, diferença entre não-protégido e protegido.

Equal Opportunity no Income Dataset

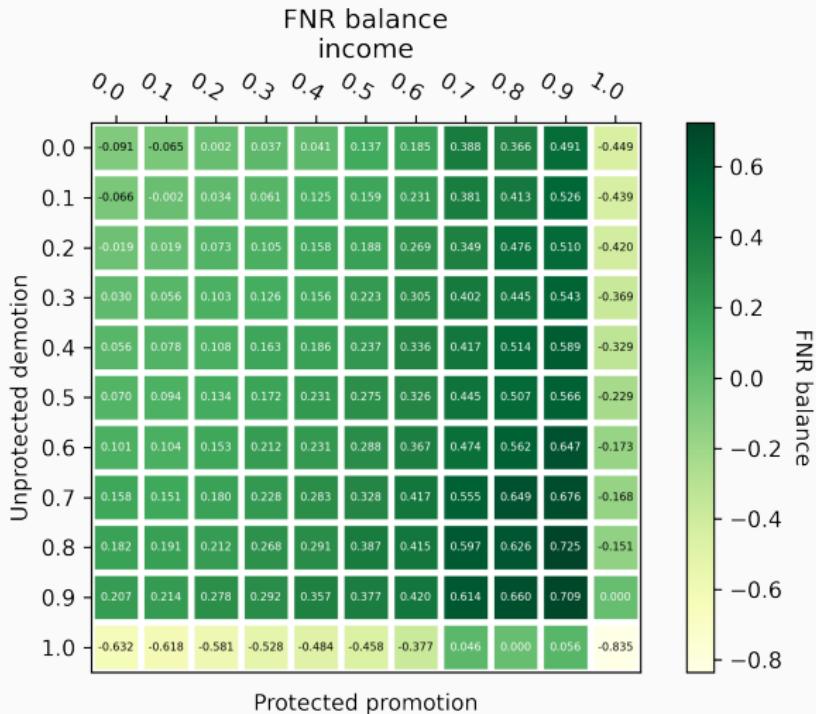


Figure 7: $FNR = \frac{FN}{TP + FN}$, diferença entre não-protégido e protegido.

Coeficiente de Correção de Matthews no Income Dataset

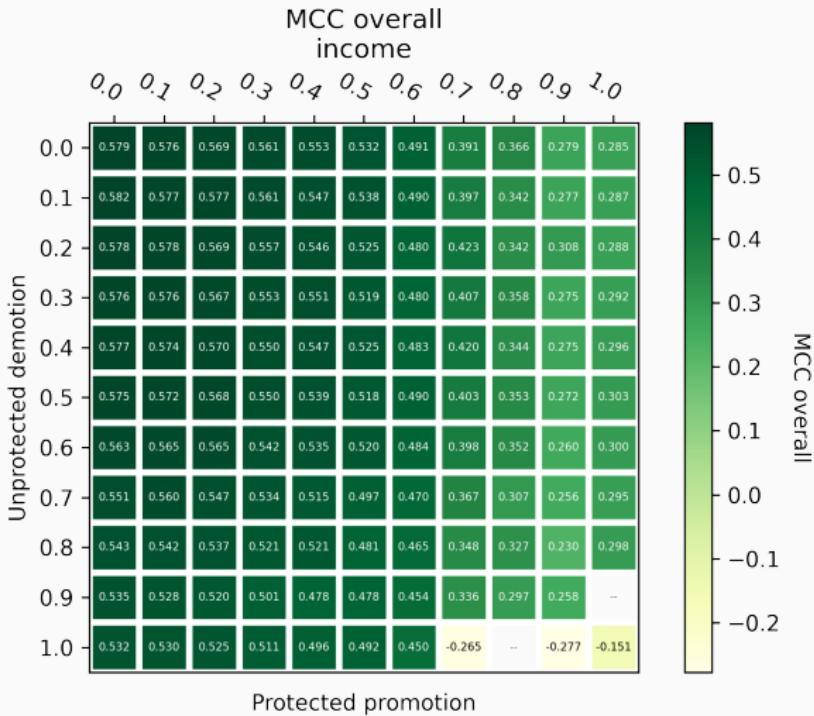


Figure 8: MCC geral, $MCC = \frac{((TP * TN) - (FP * FN))}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$

Coeficiente de Correção de Matthews no Income Dataset

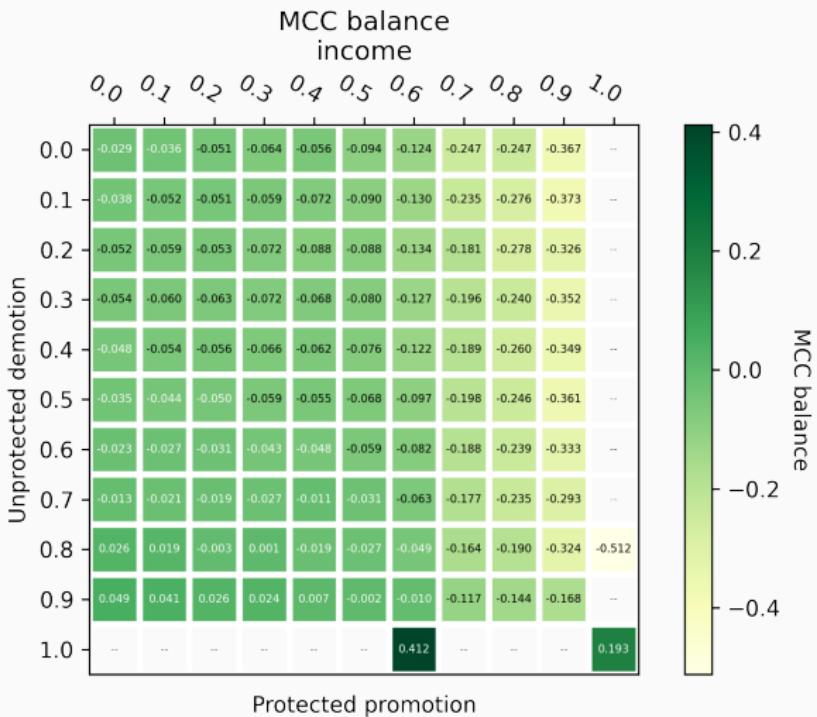


Figure 9: Balanço do MCC, diferença entre não-protégido e protegido

Resultados no *German dataset*

Acurácia no German Dataset

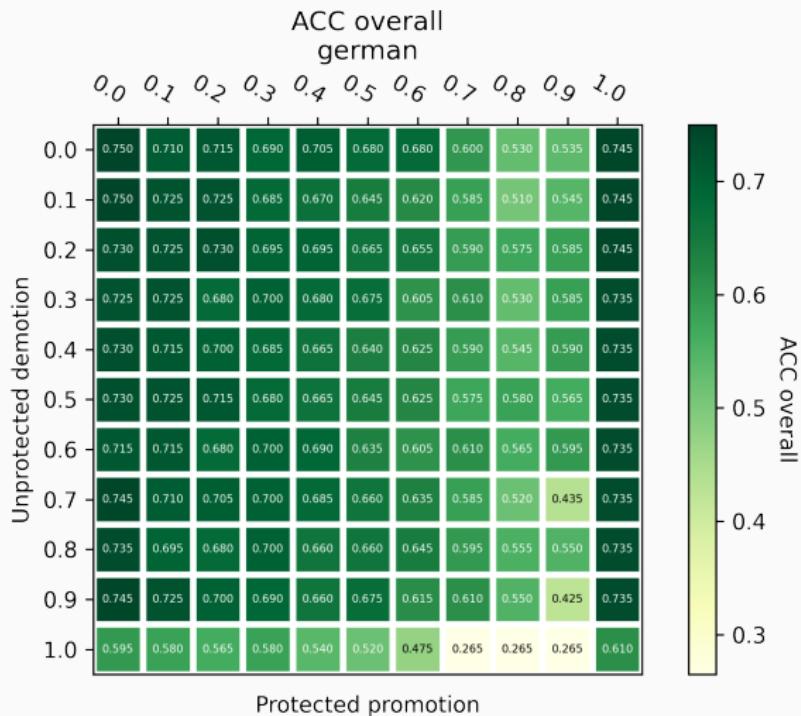


Figure 10: Acurácia geral

Accuracy Equality no German Dataset

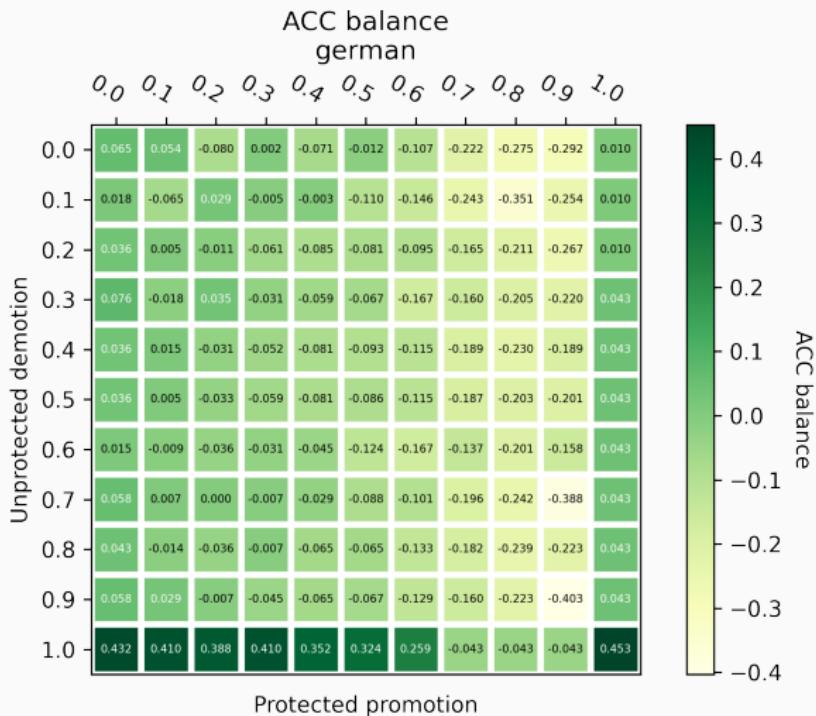


Figure 11: Balanço da acurácia, diferença entre não-protégido e protegido

Conditional Accuracy Parity no German Dataset

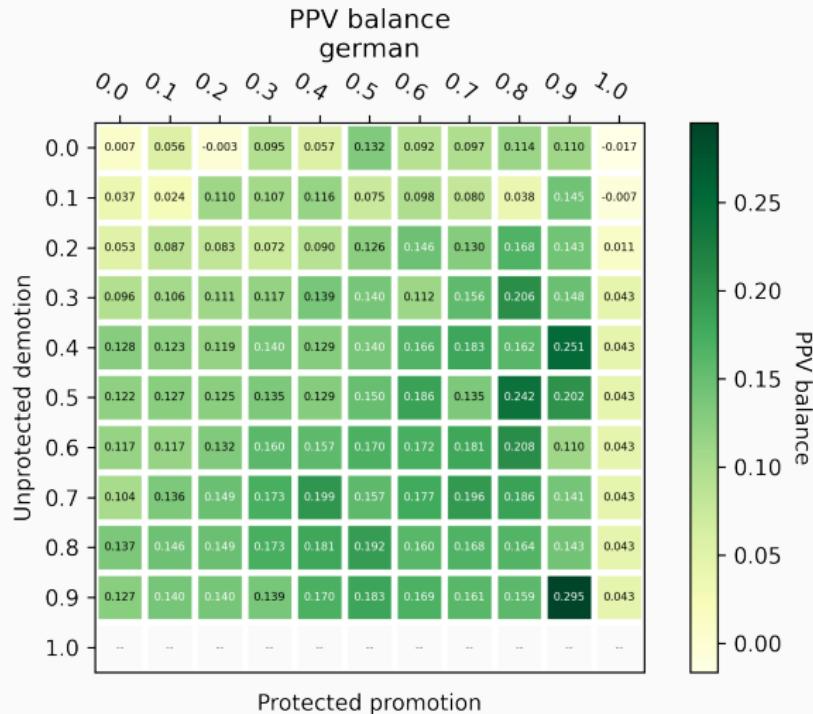


Figure 12: $PPV = \frac{TP}{TP + FP}$, diferença entre não-protégido e protegido.

Conditional Accuracy Parity no German Dataset

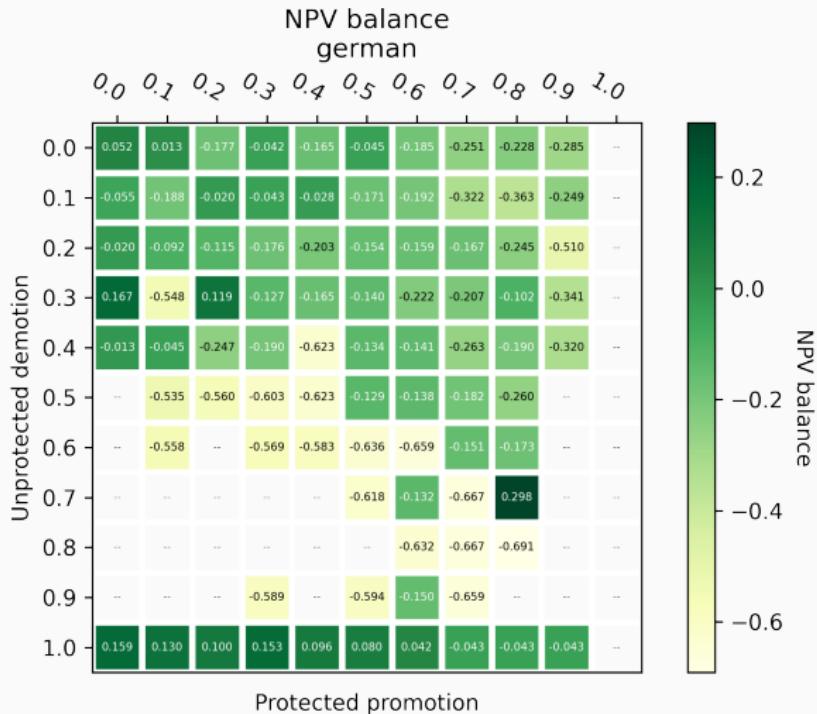


Figure 13: $NPV = \frac{TN}{TN + FN}$, diferença entre não-protégido e protegido.

Statistical Parity no German Dataset

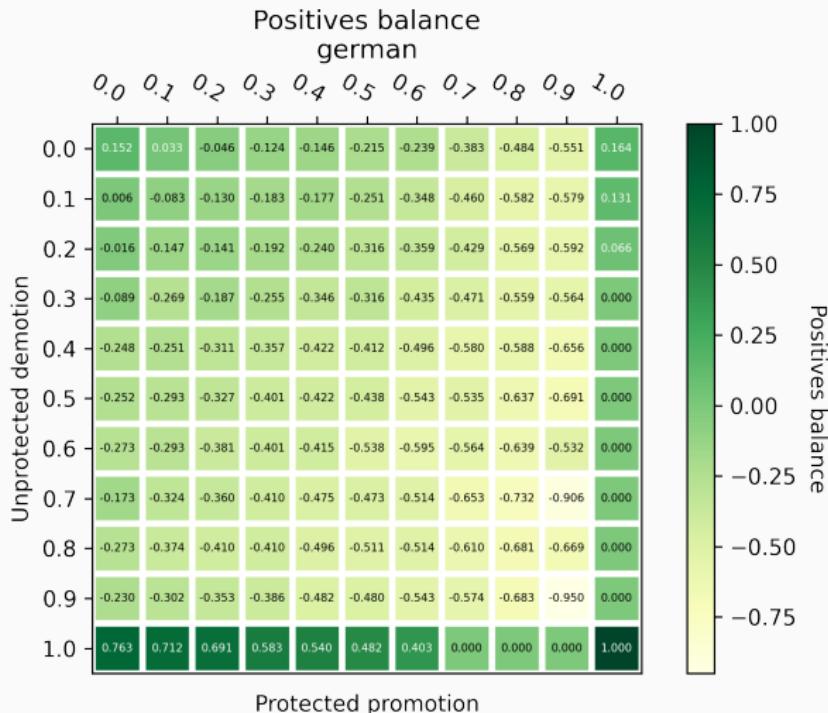


Figure 14: Balanço da taxa de previsões positivas, diferença entre não-protégido e protegido

Predictive Quality no German Dataset

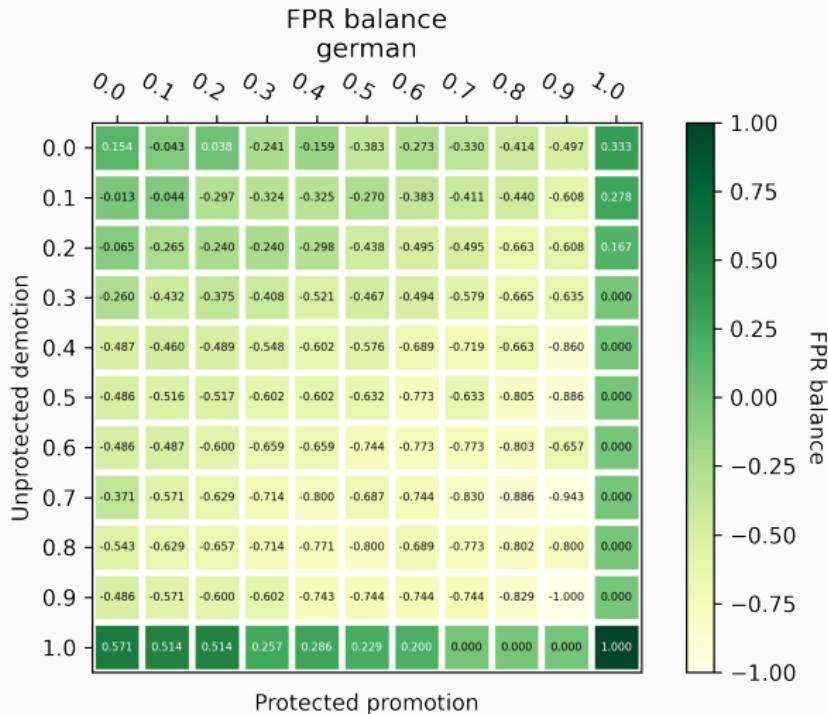


Figure 15: $FPR = \frac{FP}{TN + FP}$, diferença entre não-protégido e protegido.

Equal Opportunity no German Dataset

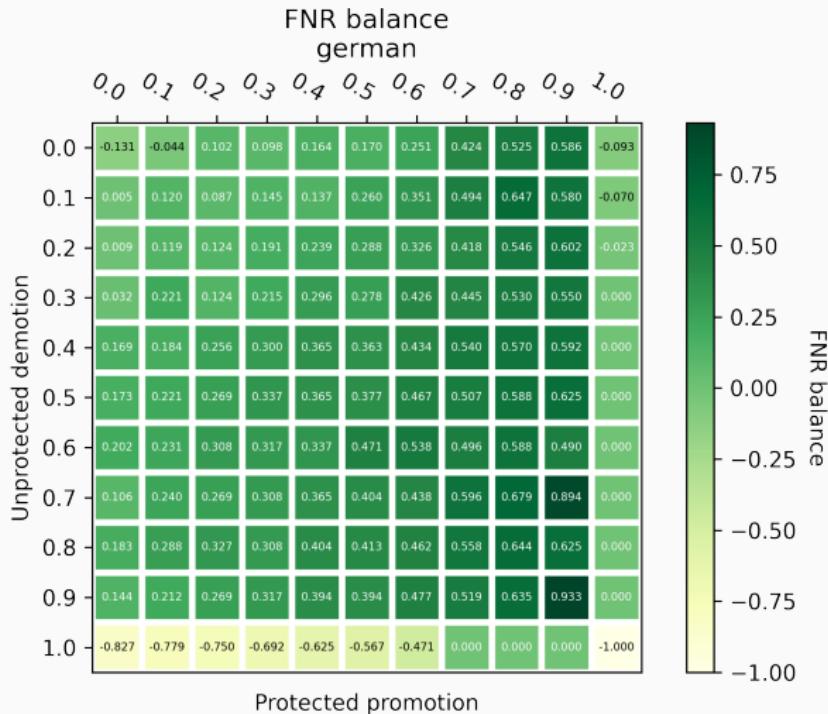


Figure 16: $FNR = \frac{FN}{TP + FN}$, diferença entre não-protégido e protegido.

Coeficiente de Correção de Matthews no German Dataset

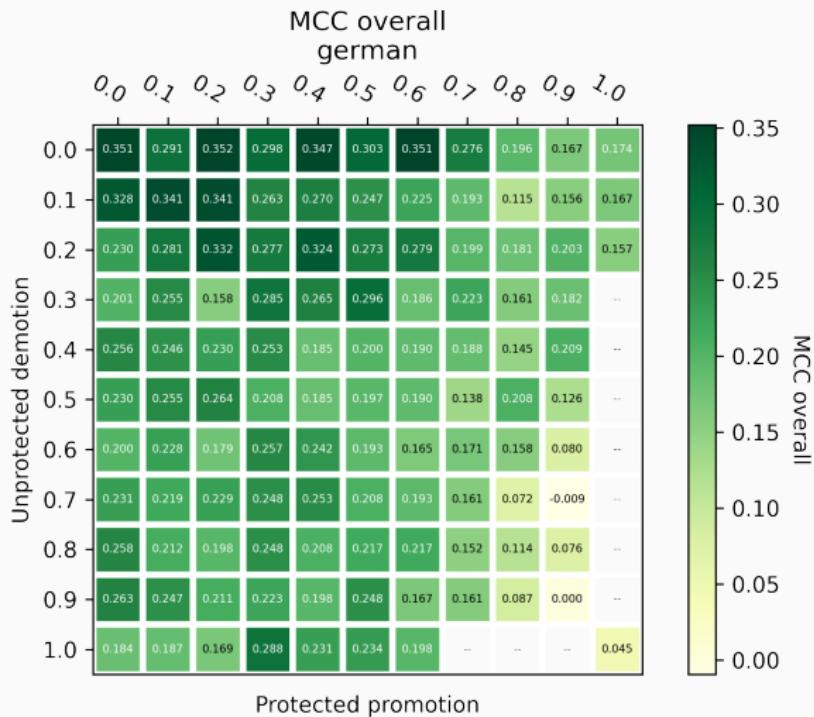


Figure 17: MCC geral, $MCC = \frac{((TP * TN) - (FP * FN))}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$

Coeficiente de Correção de Matthews no German Dataset

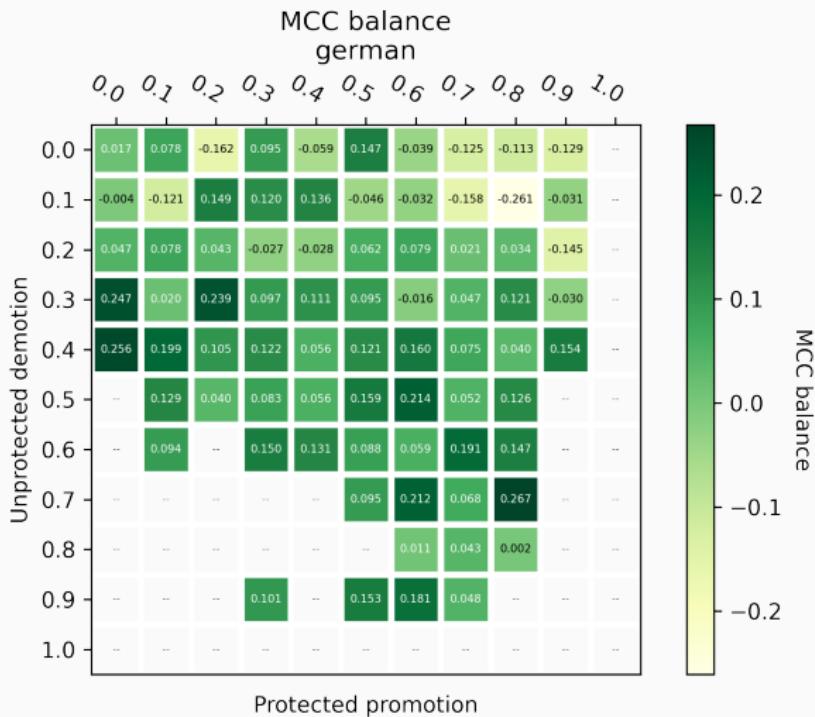


Figure 18: Balanço do MCC, diferença entre não-protégido e protegido