MO810 - Imparcialidade

Aline C. C. S. Azevedo
Instituto de Computação
UNICAMP
Campinas, Brasil
a189593@dac.unicamp.br

Beatriz C. Nascimento
Instituto de Computação
UNICAMP
Campinas, Brasil
b247403@dac.unicamp.br

Daniel G. Gratti
Instituto de Computação
UNICAMP
Campinas, Brasil
d214729@dac.unicamp.br

I. Introdução

No cenário atual, a avaliação e classificação de áreas com índices de criminalidade elevados têm emergido como um desafio crítico e complexo. A aplicação de técnicas de aprendizado de máquina permite auxiliar na análise desses índices através de uma vasta quantidade de dados. A capacidade de identificar e compreender essas regiões desempenha um papel fundamental para a formulação de estratégias de segurança pública. No entanto, é necessário também considerar fatores sociais para que a utilização de modelos de aprendizado de máquina não contribua ainda mais com viéses discriminatórios.

Apresentamos uma proposta que visa aprofundar a análise dos índices de criminalidade nas diferentes regiões de São Paulo, adotando uma abordagem ética. Nosso objetivo é treinar um modelo de classificação de criminalidade em regiões da cidade de São Paulo, classificadas como Perigosas ou Não-Perigosas de acordo com a distribuição de Boletins de Ocorrência feitos naquela região. Para garantir que o modelo treinado seja livre de possíveis vieses inclusos em nosso *dataset*, utilizamos técnicas de mitigação de vieses e balanceamento do tratamento estipulado pelo classificador com o objetivo de minimizar a disparidade de classificação entre regiões com maior ou menor poder aquisitivos.

II. Trabalhos relacionados

Realizamos uma revisão bibliográfica sobre estudos que empregam técnicas de imparcialidade em sistemas de aprendizado de máquina, predominantemente no contexto da criminalidade. O nosso foco recaiu sobre a análise detalhada de cinco artigos específicos, os quais discutimos a seguir.

O artigo *"Enforcing fairness using ensemble of diverse Pareto-optimal models"* [1] discute a aplicação de técnica de otimização multi-objetivo durante a fase de treinamento de modelos de aprendizado de máquina, utilizando conjunto de modelos Pareto-optimo, a fim de escolher um sistema mais justo sem reduzir muito a acurácia. Portanto, considerando o *trade-off* entre *fairness* e acurácia, é gerado uma fronteira de Pareto com um conjunto de modelos, minimizando o erro de aprendizagem e de discriminação. Depois esses modelos são agregados usando conjunto de filtros e procedimentos de votação. Essa agregação de modelos com diferentes níves de benefícios para cada grupo melhora a robustez em relação a performance e *fairness*.

O artigo *"Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Optimal Transport and Conformal Prediction Sets"* [2] explora a aplicação de algoritmos de avaliação de risco na justiça criminal. Utiliza-se a abordagem contrafactual para melhorar a equidade ao tratar membros de classes desfavorecidas como se fossem de classes privilegiadas no algoritmo. Um classificador de aprendizado de máquina é combinado com um ajuste ótimo de transporte para lidar com viés de entrada. O estudo usa tabelas de confusão e previsões conformes para avaliar a equidade na estimativa de risco. Os resultados mostram melhorias significativas na equidade e para classes legalmente protegidas.

Na área de *individual fairness*, encontramos um método para viabilizar esse paradigma, superando o desafio da especificação humana de uma métrica de similaridade utilizada para avaliar "quem se assemelha a quem". Para isso, foi empregado um modelo de otimização matemática que combina a entrada X com o grafo de *fairness* G, resultando em uma representação unificada. Esse resultado é alcançado ao formular um problema de otimização que relaciona o *embedding* do grafo G com o aprendizado da representação [3].

Já o artigo *"Counterfactual Fairness"* [4] aborda o impacto do aprendizado de máquina em setores como seguros, empréstimos, contratações e policiamento preditivo, ressaltando as implicações legais e éticas da automação de decisões nessas áreas. Essa pesquisa propõe um framework para a modelagem de *fairness*, utilizando ferramentas de inferência causal, onde é aplicado em um contexto prático, focado na previsão equitativa do sucesso acadêmico em faculdade de direito. Além disse apresenta dois cenários como exemplo de aplicação de equidade contrafactual (*counterfactual fairness*, onde um desses cenários é sobre regiões com alto índice de criminalidade.

Encontramos também, uma pesquisa que utiliza redes neurais profundas no artigo *"Auditing the fairness of place-based crime prediction models implemented with deep learning approaches"* [5], o qual discute modelos de previsão de crime baseados em locais implementados com aprendizado profundo. Esses modelos utilizam padrões espaço-temporais de crimes históricos e fatores do ambiente construído para prever volumes agregados de incidentes criminais em locais específicos. No entanto, há preocupações com vieses nos dados de incidentes criminais e na mobilidade humana coletada por telefones celulares, podendo afetar a equidade desses modelos.

O estudo analisa a justiça desses modelos em múltiplas cidades nos EUA, explorando possíveis causas de vieses nos dados de crime, mobilidade humana e nos algoritmos preditivos. Os resultados revelam que os modelos de previsão de crime apresentam viéses devido aos dados enviesados de crimes, e a inclusão de dados de mobilidade de celulares pode diminuir a equidade sob certas circunstâncias, com parte dessa perda de equidade sendo explicada pelo viés nos dados de crimes e algoritmos preditivos.

III. Metodologia

A. Modelo e métricas de desempenho

Para este estudo, empregamos os modelos de Regressão Logística e XGBoost, conforme descritos no trabalho anterior, para fins de comparação com os resultados obtidos após a implementação de métodos de redução de viés. Além disso, mantivemos as mesmas métricas de avaliação de desempenho do modelo, a acurácia balanceada, área sob a curva ROC (ROC AUC) e *brier score*.

B. Métricas de imparcialidade

As métricas de imparcialidade desempenham um papel fundamental no aprendizado de máquina, sendo essenciais para identificar e mitigar o viés e a discriminação presentes nos algoritmos.

Para resolver o nosso problema específico, optamos por utilizar duas métricas de imparcialidade que se alinham com a visão de mundo mais adequada ao nosso caso, conhecida como *"what you see is what you get"*, que assume que existem disparidades e desigualdades no espaço construído, e isso é verificado no espaço observado sem a necessidade de um viés no processo de medição.

Escolhemos a métrica de Diferença Média de Probabilidades (*Average Odds Difference*) e a Consistência (*Consistency*). A primeira está relacionada à justiça de grupo (*group fairness*), ou seja, refere-se à imparcialidade aplicada a um grupo de pessoas, em nosso problema, encontramos que a variável relacionada ao poder aquisitivo da região possuia uma distribuição entre os rótulos, conforme pode ser observada na Figura 1, que não possuía a mesma distribuição quando condicionada à saída do modelo, mostrando um possível viés do modelo. Por outro lado, a segunda métrica está relacionada à justiça individual (*individual fairness*), concentrando-se na equidade aplicada a cada indivíduo separadamente, sob a ótica de que indivíduos semelhantes merecem tratamentos semelhantes.

Na Figura 2 é possível observar os diferentes formatos entre a distribuição esperada, conforme os dados e a distribuição obtida pelo classificador. Para analisar os efeitos das técnicas de mitigação de vieses, utilizaremos os gráficos de distribuições das variáveis sensíveis condicionadas à saída dos modelos e a métrica de AOD para diferentes cortes. Cada corte é representado por um ponto no gráfico em que se mede a disparidade de performance entre as amostras cuja variável sensível é menor que o representado pelo corte e as amostras com valores maiores na mesma variável.

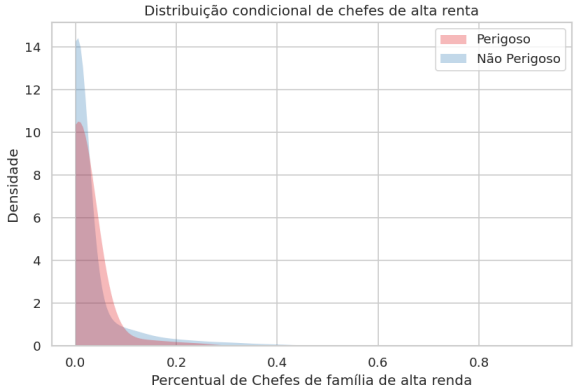

Figura 1. Distribuições referentes à variável sensível do problema condicionada aos rótulos.

Esperamos que, com as técnicas de mitigação de viés, as distribuições se assemelham mais a distribuição contida nos dados de treinamento, indicando que o classificador não possua preferência de tratamento.

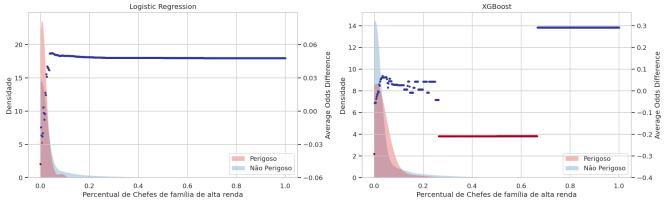

Figura 2. Distribuições da variável sensível condicionadas à saída do modelo. A métrica AOD está apresentada no gráfico na forma de *scatter plot*, indicando o valor da métrica para diferentes cortes.

C. Método de pré-processamento

Os métodos de mitigação de parcialidade no pré-processamento em aprendizado de máquina referem-se a técnicas aplicadas aos dados antes de treinar um modelo para mitigar o viés e garantir mais equidade nos resultados. Foram escolhidos dois métodos de pré-processamento de acordo com a nossa aplicação, o método de *augmenting dataset* e o método *fair score transformer*.

- ***Augmenting Dataset:*** Inclusão de dados fictícios gerados a partir de inversão de atributos protegidos para equilibrar conjuntos de dados (como justiça contrafactual), promovendo uma representação mais equitativa durante o aprendizado do modelo. Em nossa implementação, as amostras adicionadas são amostras do conjunto de treino cujos valores da variável sensível são reamostrados.
- ***Fair Representations:*** A ideia principal nesta técnica de pré-processamento é mapear cada indivíduo, representado como um ponto em um determinado espaço de entrada, para uma distribuição de probabilidade em um novo espaço de representação. O objetivo desta nova representação é retirar qualquer informação que possa identificar se a pessoa pertence ao subgrupo protegido,

mantendo o máximo possível de outras informações. Essas representações também são otimizadas para que quaisquer tarefas de classificação que as utilizem sejam extremamente precisas. [6].

D. Método de processamento

Já os métodos de mitigação de parcialidade no processamento referem-se a técnicas aplicadas durante o treinamento do modelo ou na própria arquitetura do algoritmo para mitigar o viés e a discriminação, promovendo resultados mais justos e equitativos em suas predições. Escolhemos dois (três) métodos de processamento que consideramos adequado à nossa aplicação, o método *optimization based on equal opportunity*, *optimization based on average odds* e um método de otimização multi-objetivo utilizando *Non-Inferior Set Estimation* (NISE).

- ***Optimization based on Equal Opportunity:*** foca em assegurar igualdade nas taxas de verdadeiros positivos entre diferentes grupos, garantindo oportunidades iguais de previsões corretas independentemente de sua afiliação a um grupo específico.
- **Otimização Multi-Objetivo utilizando NISE:** A otimização multi-objetivo visa maximizar ou minimizar várias funções objetivas conflitantes simultaneamente, buscando soluções que não sejam dominadas por outras, constituindo um equilíbrio entre os diferentes objetivos. Esse processo resulta em um conjunto de soluções Pareto-ótimas, conhecido como fronteira de Pareto, cabendo ao tomador de decisão escolher a solução final. Por meio do método NISE, baseado em otimização multi-objetivo com técnica de soma ponderada, é possível gerar uma Fronteira de Pareto com os melhores modelos e selecionar a melhor solução que reduza a desigualdade entre grupos sem prejudicar significativamente o desempenho.

E. Método de pós-processamento

No pós-processamento, os métodos de mitigação de parcialidade são aplicados após o treinamento do modelo para corrigir possíveis vieses encontrados nas previsões ou garantir a equidade nas decisões. Para nossa aplicação optamos por utilizar os métodos *fair score transformer* e

- ***Confusion Balance:*** Esse método tenta equilibrar a saída do modelo utilizando as probabilidades condicionais da saída do modelo com os rótulos, e então, as saídas são ponderadas com essas distribuições, modificando os *scores* das amostras de forma a balancear as classes.
- ***Threshold Optimization:*** A *threshold optimization* é usada para atribuir rótulos de classe aos *scores* de probabilidade de saída de um modelo. O *threshold* ótimo ou melhor é aquele que maximiza o *score* de uma métrica de desempenho especificada, no caso, acurácia balanceado com restrição em Equalized Odds [7].

F. Ajuste dos Hiperparâmetros

Para garantir uma boa performance em nossos modelos, que são estimadores paramétricos de classificação, é preciso escolher hiperparâmetros de forma a garantirem capacidade de generalização e evitar *overfitting*, tanto para o método de aprendizado de máquina, quanto para os métodos de mitigação de viés. Neste trabalho, cada estimador foi treinado e avaliado por validação cruzada (K-fold), com 10 pastas, para cada tentativa de escolha de hiperparâmetro e utilizamos a acurácia balanceada dos estimadores nos conjuntos de validação para guiar nosso otimizador de hiperparâmetros. A otimização foi feita através do framework Optuna, e cada estimador foi submetido a um espaço de busca próprio, como mostrado na Tabela I.

Tabela I
ESPAÇO DE BUSCA DE HIPERPARÂMETROS PARA OS ESTIMADORES TREINADOS

Estimator	Hyperparameter	Details	Search space
Logistic Regression	C	Regularization	$[10^{-4}, 10^{4}]$
	Penalty	Penalty	{L1, L2}
XGBoost	Max depth	Maximum tree depth	$\{2, 4, 5, 6, 8\}$
	Learning Rate	Learning rate	$[0.01, 0.1]$
	N estimators	Number of trees	$[50, 200]$
	Subsample	% used for training	$[0.5, 1]$

IV. RESULTADOS

Após a otimização de hiperparâmetros em 100 tentativas, obtemos os melhores modelos para cada um dos métodos, avaliamos sua performance através de um conjunto separado, obtendo as métricas na Tabela II. A métrica de *Average Odds* foi calculada para o ponto de corte 0.07 na variável sensível de porcentagem de chefes de família com alta renda. Esta escolha foi feita com base na distribuição dos dados de treinamento, pois trata-se de um ponto em que as densidades de probabilidade de ambas as distribuições condicionais são as mesmas. Marcamos em negrito os melhores resultados para aquela métrica no estimador em questão (Regressão Logística ou XGBoost), exceto pela *Average odds*, cuja medida de performance é subjetiva. Portanto, escolhemos marcar o modelo com performance que menos se afastou da média entre os métodos para um mesmo estimador e, também, baseado nas distribuições condicionadas à saída do modelo, que podem ser encontradas no Apêndice.

A. Mitigação de viés no pré-processamento

As técnicas utilizadas de pré-processamento se baseavam em tratar dos vieses contidos nos dados antes da otimização do modelo. A primeira técnica consistia em criar amostras adicionais ao conjunto de dados já existente, numa forma de aumentação de dados. Os resultados desta técnica se mostraram satisfatórios ao melhorarem a performance de ambos os classificadores.

No entanto, a técnica de *Fair representations* não performou tão bem. Apesar da consistência e *Average odds* performarem melhores, é possível observar pelas distribuições que a técnica não conseguiu capturar informações suficiente dos dados de treinamento, obtendo um classificador que somente classifica Não-perigoso, para o Regressor Logístico, e um classificador pior que o baseline para o XGBoost.

B. Mitigação de viés no processamento

Estas técnicas se baseiam em remover vieses através da formulação de problemas de otimização que forçam restrições de *fairness*. Estes métodos formulam problemas de classificação por Regressão Logística cuja função objetivo é alterada e o problema de otimização passa a ter restrições. Apesar da possibilidade de adaptar estes métodos para árvores de decisão, esta possibilidade não foi explorada por conta de sua complexidade ao tratarmos problemas não-convexos.

A primeira técnica, baseada em *Equal opportunity*, garante que o modelo trate indíviduos de diferentes classes, de acordo com a variável sensível, com uma mesma performance. Esta técnica se mostrou superior ao baseline apenas em acurácia balanceada e melhorou as distribuições de probabilidade, conforme a Figura 5 indica, ao distribuir melhor a distribuição de regiões perigosas ao longo da variável sensível, indicando um menor viés do modelo em relação a *baseline*.

O segundo modelo contrói um problema de otimização diferente do anterior, se baseando num problema multi-objetivo, em que se busca minimizar a *loss* de ambos os grupos sensíveis ao mesmo tempo. Para isso, modelos de Regressão Logística são treinados com diferentes ponderações para cada uma das classes e modelos são gerados proceduralmente de forma a popular a fronteira de Pareto do problema. Como a penalização L1 não é convexa, e portanto a fronteira também pode não ser, removemos este tipo de penalização da busca de hiperparâmetros e modificamos o intervalo de valores para o hiperparâmetro C, de forma a variar entre 10^4 e 10^{15} para garantir uma função objetivo convexa.

Após a formação de uma fronteira de Pareto, o modelo final é escolhido como aquele que atingiu a maior acurácia balanceada dentre todos os modelos não-dominados. A performance geral do melhor modelo teve métricas parecidas com o modelo treinado por *Equal oportunity*, no entanto, o víes nas distribuições não parece ter sido corrigido, provavelmente pelo fato do modelo ainda ser um regressor logístico padrão e possuir as mesmas limitações que o modelo *baseline*.

C. Mitigação de viés no pós-processamento

Nesta categoria a otimização é feita sobre modelos já treinados, no entanto, realizamos a escolha de hiperparâmetros considerando a saída pós-processada. Esta categoria também foi a mais desafiadora, com o menor número de algoritmos disponíveis.

A primeira técnica trata de uma abordagem simples de escolha do threshold adequado para a binarização das probabilidades que o modelo gera. Esta otimização leva em conta as métricas de fairness e tenta maximizar a Acurácia Balanceada enquanto impõe uma restrição em *Average odds*. Como resultado, o modelo final tem a maior acurácia balanceada dentre todos os demais, no entanto, as demais métricas sofrem um pequeno decréscimo devido a compensação significativa em acurácia. Um ponto a se notar, na Figura 7, a distribuição de Não-perigosos não é vazia, mas está concentrada em 0, o que indica que o modelo aprendeu uma classificação que melhora acurácia sendo menos justa e, portanto, influenciando no alto valor de *Average odds*.

O último modelo testado, utiliza de uma outra lógica para melhorar a acurácia do preditor. Seu objetivo é balancear a saída do modelo utilizando das distribuições da saída do modelo condicionadas aos rótulos e então ponderar as probabilidade conforme estas distribuições. Sua performance é parecida com o Data Augmentation, garantindo métricas melhores que o modelo *baseline*.

V. Discussão

Este trabalho teve como objetivo explorar técnicas de redução de vieses através de diferentes métodos de tornar a classificação justa para grupo de indíviduos de acordo com uma variável sensível, Renda para nosso problema, e também garantindo justiça individual. Em geral, encontramos que técnicas de Pré-processamento, como a aumentação de dados, tiveram impactos positivos para o modelo, atingindo resultados esperados para a mitigação de vieses sem piorar a performance do modelo. Acreditamos que estes métodos são importantes para mitigar problemas logo no início do ciclo de vida da solução, evitando que vieses se propaguem durante o desenvolvimento ou outros sejam criados no meio do processo.

Métodos de pós-processamento foram considerados os piores neste quesito, provavelmente devido a pouca disponibilidade destes métodos e de sua efetividade em consertar problemas éticos em modelos com vieses cuja inserção não é certa.

De forma equilibrada, métodos *in-processing* se mostraram com grande potencial de mitigar vieses, no entanto, acreditamos que foram pouco explorados e resultados melhores podem ser obtidos com estratégias mais espertas das que selecionamos.

Por fim, métodos de *fairness* mostraram o grande potencial de se treinar preditores mais éticos, no entanto, sentimos uma falta na disponibilidade destes algoritmos em larga escala, dificultando seu uso.

Referências

[1] Vitória Guardieiro, Marcos M Raimundo, and Jorge Poco. Enforcing fairness using ensemble of diverse pareto-optimal models. *Data Mining and Knowledge Discovery*, pages 1–29, 2023.

[2] Richard A Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *Sociological Methods & Research*, page 00491241231155883, 2021.

[3] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439*, 2019.

[4] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[5] Jiahui Wu, Saad Mohammad Abrar, Naman Awasthi, and Vanessa Frías-Martínez. Auditing the fairness of place-based crime prediction models implemented with deep learning approaches. *Computers, Environment and Urban Systems*, 102:101967, 2023.

[6] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[7] Joffrey L Leevy, Justin M Johnson, John Hancock, and Taghi M Khoshgoftaar. Threshold optimization and random undersampling for imbalanced credit card data. *Journal of Big Data*, 10(1):58, 2023.

APÊNDICE

Tabela II
PERFORMANCE DOS ALGORITMOS DE FAIRNESS

Model	Fairness method	Bal Acc	ROC AUC	Brier Loss	Consistency	Average odds
Logistic Regression	Baseline	52.5%	**0.920**	**0.020**	0.973	0.028
	Data Augmentation	55.1%	**0.920**	**0.020**	0.971	**0.058**
	Fair representations	50.0%	0.608	0.024	**0.975**	0.000
	Equal Opportunity	57.7%	0.890	0.022	0.970	0.087
	NISE	56.7%	0.894	0.021	0.970	0.076
	Threshold Optimization	**63.9%**	0.913	0.021	0.869	0.270
	Confusion Balance	52.5%	**0.920**	**0.020**	0.973	0.028
XGBoost	Baseline	54.0%	0.915	**0.021**	0.970	**0.049**
	Data Augmentation	54.2%	**0.916**	0.022	0.973	**0.048**
	Fair representations	50.5%	0.723	0.024	**0.975**	0.011
	Threshold Optimization	**62.9%**	0.887	**0.021**	0.872	0.257
	Confusion Balance	54.1%	0.914	0.022	0.971	**0.048**

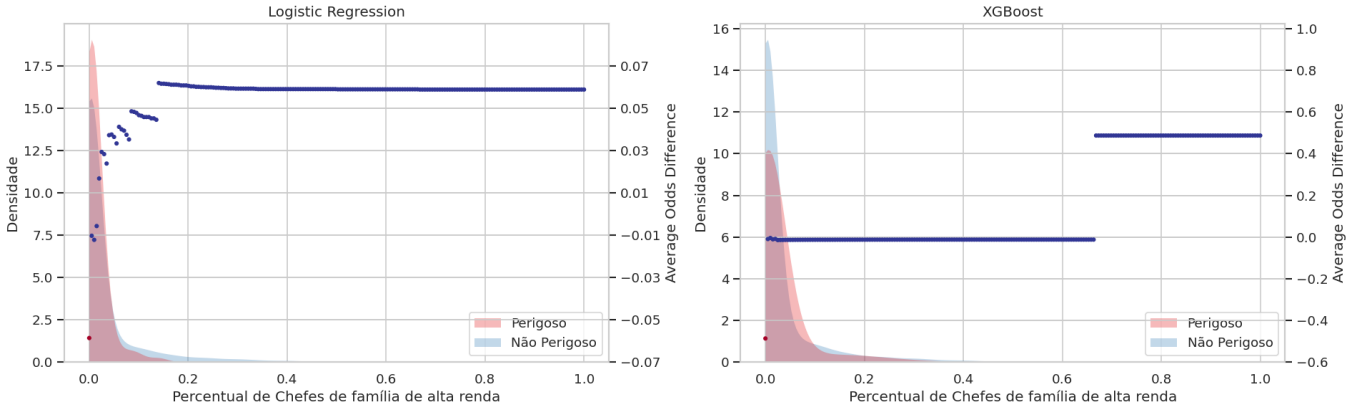

Figura 3. Distribuições obtidas após treinamento em dados aumentados.

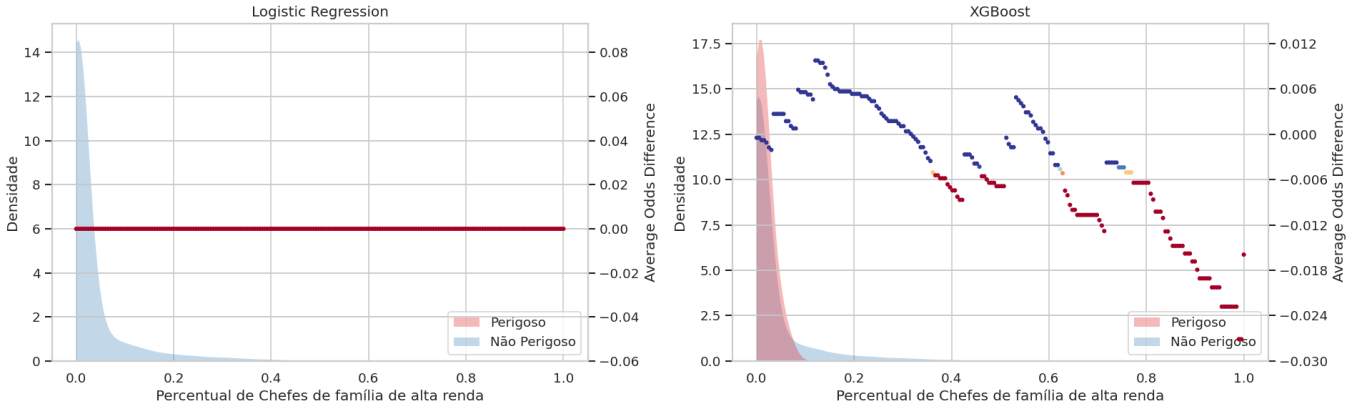

Figura 4. Distribuições obtidas após técnica de *Fair Representations*.

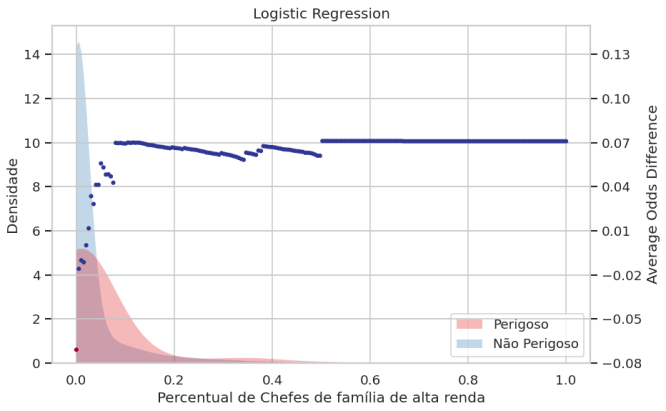

Figura 5. Distribuições obtidas pelo modelo de *Equal opportunity*.

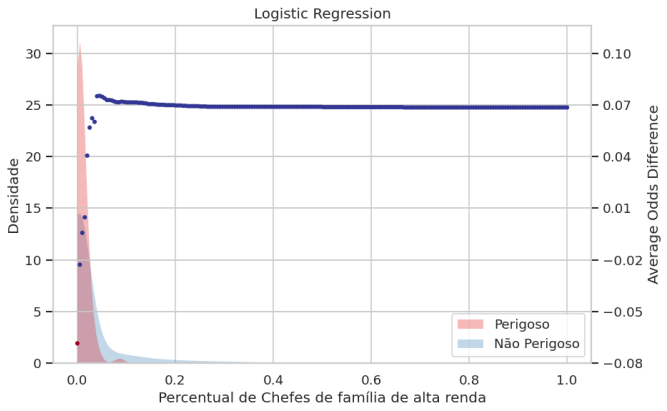

Figura 6. Distribuições obtidas pela regressão logística multi-objetivo.

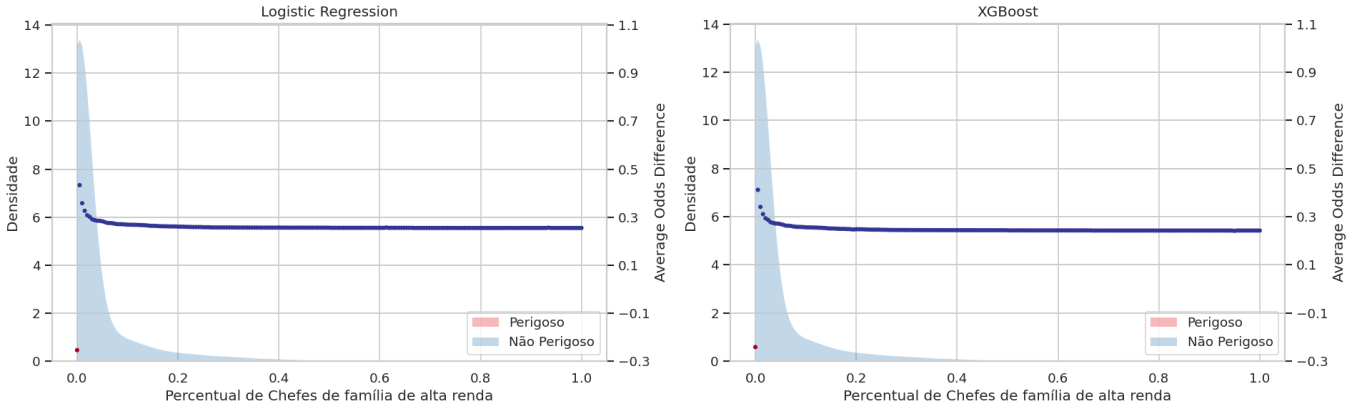

Figura 7. Distribuições obtidas com otimização de *thresholds*.

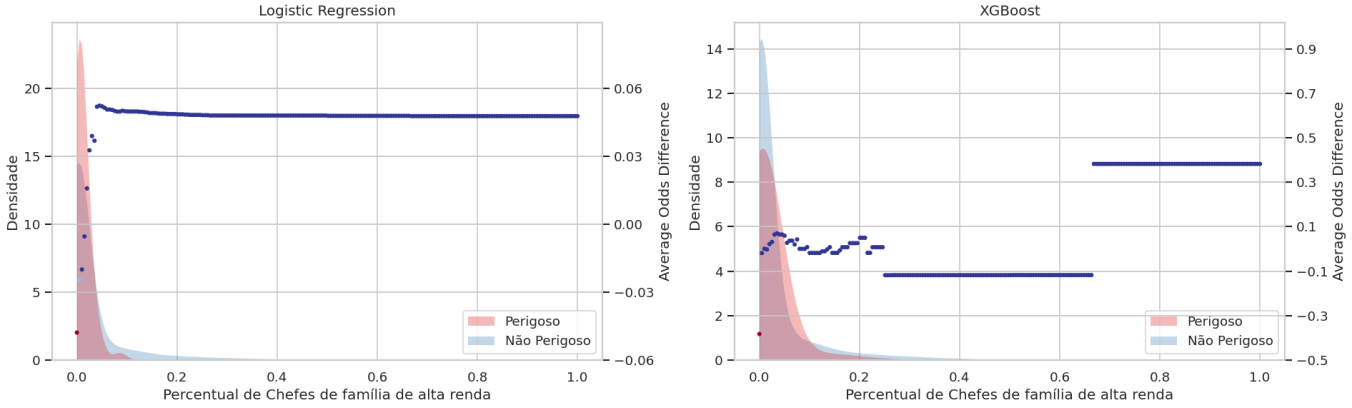

Figura 8. Distribuições obtidas após balanceamento.