A justiça na Ciência dos Dados: aplicação de metodologias para a mitigação de vieses sociais em um modelo de Machine Learning para a Concessão de Crédito

Lucas Grossi Vieira Santos
Universidade Federal de Minas Gerais
Avenida Presidente Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, CEP 31270-901, Brasil
lucas.grossi2000@gmail.com

Luiz Henrique Cardoso de Oliveira
Universidade Federal de Minas Gerais
Avenida Presidente Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, CEP 31270-901, Brasil
luizhcoliveira3@gmail.com

Marcelo Azevedo Costa
Universidade Federal de Minas Gerais
Avenida Presidente Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, CEP 31270-901, Brasil
macosta@ufmg.br

RESUMO

A discussão a respeito dos impactos de vieses sociais na justiça de modelos de *Machine Learning* vêm aumentando a sua relevância na atualidade com a maior utilização de *softwares* automatizados onde a interferência humana não é necessária. Este artigo busca fazer uma seleção de metodologias para a mitigação de vieses sociais e aplicá-las ao contexto de concessão de crédito, utilizando para isso uma base de dados brasileira. As metodologias selecionadas utilizam diferentes racionais para a remoção de vieses, incluindo rebalanceamento da base através de reamostragem, alteração da variável resposta para maior equilíbrio da base e remoção das informações relacionadas à variável sensível das variáveis preditoras. Todas as metodologias conseguiram remover, em algum nível, o viés da base de dados. Entretanto, essa remoção é acompanhada de uma perda de desempenho do modelo.

PALAVRAS CHAVE. Vieses Sociais. Justiça. Machine Learning.
Tópico: Mitigação de Vieses Sociais em modelos de Machine Learning

ABSTRACT

The discussion about the impacts of social biases on the fairness of Machine Learning models has been increasing its relevance nowadays with the greater use of automated software where human interference is not necessary. This article focuses on selecting methodologies that attempt to remove social bias and applying them in the context of credit concession utilizing a Brazilian database. The methodologies employed utilize different rationales for bias removal, including resampling to rebalance the dataset, modifying the target variable to achieve greater balance, and removing sensitive variable information from the predictor variables. All methodologies managed to remove bias from the dataset to some extent. However, this removal is accompanied by a loss of model performance.

KEYWORDS. Social Biases. Fairness. Machine Learning.
Paper Topics: Mitigating Social Bias in Machine Learning

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

1. Introdução

Se enganam aqueles que acreditam que o ramo da Ciência de Dados *(Data Science)*, tão relevante no mercado atual, se resume à estatística, computação e matemática. Com a utilização cada vez maior de técnicas sofisticadas de *Machine Learning* para a análise autônoma e auxílio na tomada de decisão em negócios, é necessário um enfoque acerca dos impactos que vieses podem causar sob o ponto de vista social. É inegável o impacto das técnicas contemporâneas na melhoria dos processos decisórios. Conforme destacado por O'Neill [2016] em seu livro *Weapons of Math Destruction*, no passado, a tomada de decisões era inteiramente atribuída aos detentores de poder, baseada em critérios familiares, políticos e demográficos que pouco se relacionavam com os objetivos da análise em questão. No entanto, é crucial reconhecer que, mesmo com os avanços, grande parte dessas decisões humanas ainda encontram espaço nos sistemas computacionais, visto que os modelos são, em essência, um reflexo da sociedade.

Como afirma Beretta et al. [2019] no estudo *The Invisible Power of Fairness. How Machine Learning Shapes Democracy,* a coleta extensiva de informações sensíveis, de cunho pessoal ou sigiloso como gênero, raça e condição social, amplamente utilizadas em áreas como a justiça criminal, publicidade e análise de crédito, por fontes diversas tais quais redes sociais e dados celulares, já levantam uma série de questões éticas e legais. Ainda no texto, os autores relembram que o conceito de algoritmo utilizado por *softwares* automatizados e modelos de *Machine Learning* é transformar as entradas, que neste caso são os dados, em saídas, os resultados esperados, através de um processamento constituído por cálculos e procedimentos computacionais. Caso os dados de entrada possuam vieses, é provável que o resultado final também possua.

Dada a sua pertinência no contexto atual, cercada de ineditismos, o objetivo deste estudo é realizar uma comparação de metodologias de pré-processamento dos dados para a mitigação de vieses sociais em modelos de Machine Learning aplicados ao contexto de concessão de crédito. O estudo busca entender o quão relevante são os vieses relacionados a gênero no banco de dados analisado, além de compreender qual o impacto da retirada desses vieses nos resultados do modelo de Machine Learning ajustado, entendendo o tradeoff entre acurácia e justiça de cada uma das metodologias analisadas.

2. Literatura Relacionada

Khakurel et al. [2022] afirmam que o maior desafio que a área de *Machine Learning* enfrenta está relacionado com ética e justiça na predição de modelos. Como os vieses sociais estão presentes no banco de dados desde o início da coleta destes, o modelo é alimentado com dados já enviesados, gerando previsões também enviesadas. Mesmo em casos em que características sensíveis como gênero, raça e etnia são removidos do banco de dados, o viés imposto por humanos em outros atributos do banco de dados está indiretamente influenciando o algoritmo a realizar previsões enviesadas.

Neste contexto, os autores utilizam um banco de dados de aplicações para cartões de crédito para investigar vieses relacionados ao gênero dos solicitantes, utilizando as métricas de Diferença de Paridade Estatística e Impacto Díspar. A Diferença de Paridade Estatística mede a diferença entre as taxas de resultados favoráveis para grupos privilegiados e não-privilegiados, enquanto o Impacto Díspar calcula a razão entre essas taxas. Um modelo justo possui um Impacto Díspar de 1 e uma Diferença de Paridade Estatística igual a 0.

Para mitigar o viés, os autores propõem uma técnica de pré-processamento que atribui diferentes pesos às instâncias do banco de dados com base no atributo sensível e nos resultados. Essa técnica demonstra ser eficaz na remoção do viés, com uma pequena perda de desempenho do modelo.

Bono et al. [2021] definem três critérios para avaliar a justiça em modelos de crédito: Paridade de Performance, Separação e Suficiência. A Paridade de Performance exige que o modelo seja igualmente preciso em relação a resultados negativos para todos os subgrupos da base de dados. A Separação busca igualdade na taxa de positivos verdadeiros e falsos positivos

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

para cada subgrupo. A Suficiência requer que indivíduos com a mesma probabilidade de receber crédito tenham a mesma chance de inadimplência, independentemente do grupo.

A análise de justiça dos modelos indica que nenhuma das abordagens utilizadas no artigo de Bono et al. [2021] satisfazem de forma completa a justiça estatística em relação aos subgrupos de população definidos, indicando que existe sim um viés nos dados. Essa análise é corroborada quando os autores incluem as características sensíveis como variáveis explicativas do modelo. A inclusão gera um aumento apenas marginal na performance do modelo, indicando que a informação relevante carregada pelas características sensíveis já está presente nas variáveis utilizadas anteriormente. Assim, a conclusão dos autores é que as variáveis sensíveis estão de alguma forma correlacionados com a informação relevante para a previsão de pagamento ou não de um empréstimo, mas que a maior parte da informação relevante que essas variáveis carregam já está presente na base de dados original.

Fu et al. [2021] propõem quatro critérios adicionais para a análise de justiça: Paridade Estatística, Igualdade de Oportunidade, Probabilidades Equalizadas e Equilíbrio da Classe Positiva/Negativa. Esses critérios visam garantir que a probabilidade de receber financiamento seja igual para diferentes subgrupos, que indivíduos com o mesmo perfil de pagamento tenham as mesmas chances, e que a previsão do modelo seja consistente para indivíduos com o mesmo perfil de pagamento em diferentes subgrupos.

Os autores afirmam que os métodos de desenviesamento de dados presentes na literatura apenas reduzem e não removem os vieses, além de estarem limitados a certos algoritmos de Machine Learning e de sacrificarem poder preditivo dos modelos ajustados. Dessa forma, um novo método é proposto a partir da ideia de remover codificação redundante e garantir que as variáveis preditoras sejam independentes das variáveis sensíveis, fazendo com que seja impossível que o modelo faça inferências sobre os atributos sensíveis. Assim, a paridade estatística é garantida quando as variáveis preditoras são ortogonais às variáveis sensíveis.

Após o processo de remoção dos vieses dos dados, os modelos ajustados com dados enviesados e não enviesados são comparados. Enquanto o modelo inicial apresenta vieses relacionados ao gênero e à raça das pessoas, o modelo sem vieses obtém resultados estatisticamente equivalentes para os subgrupos, o que sugere que os vieses presentes na base de dados foram efetivamente removidos. Os autores classificam a perda de desempenho causada pela manipulação dos dados como mínima. Entretanto, é ressaltado que em casos em que exista uma grande correlação entre as variáveis sensíveis e as variáveis preditoras ou entre as variáveis sensíveis e a variável resposta, a perda de desempenho seria maior.

Celis et al. [2020] categorizam os trabalhos anteriores em duas áreas: modificação da base de dados através da reatribuição de atributos sensíveis ou atribuição de pesos, e obtenção de uma nova distribuição de dados que satisfaça as restrições definidas por métricas de igualdade. Os autores também ressaltam a existência de estratégias que buscam alterar a função objetivo do modelo de forma a incluir restrições de justiça, além de métodos que alteram o resultado dos modelos de aprendizado de máquina através de mudanças no limite de tomada de decisão.

O artigo traz as definições de taxa estatística (a razão entre a probabilidade de um indivíduo pertencer a uma determinada classe dado que ele está no grupo sub-representado e a probabilidade de pertencer à mesma classe dado que um indivíduo está no grupo super-representado) e taxa de representação (a razão entre a probabilidade atribuída ao grupo sub-representado e a probabilidade atribuída ao grupo super-representado) são utilizadas para criar as restrições de equidade entre os grupos privilegiados e não-privilegiados. Os autores ressaltam que, idealmente, ambas as métricas deveriam ser iguais a 1 para o modelo ser considerado justo. Entretanto, na prática essas medidas acabam sendo relaxadas.

Feldman et al. [2015] utilizam a métrica de Impacto Díspar para avaliar a equidade dos dados, a qual é definida como a razão entre a probabilidade de um resultado positivo para um indivíduo de um grupo não-privilegiado e a probabilidade de um resultado positivo para um indivíduo de um grupo privilegiado.

Com base nessa métrica, os autores propõem um método de pré-processamento de dados que consiste na alteração da base de dados de forma a preservar o ranking, ou seja, garantir que o

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

modelo continue escolhendo candidatos mais fortes ao invés de candidatos mais fracos. O algoritmo de reparo busca criar um conjunto de atributos não sensíveis modificados de forma a garantir que o atributo sensível se mantenha o mesmo. O modelo proposto por Feldman et al. [2015] é aplicável apenas para valores numéricos ou valores categóricos que possuam ordenação. Outros atributos devem ser removidos da base de dados.

Calders et al. [2009] propõem duas abordagens para remover vieses: massagear o conjunto de dados e ponderar as observações. A primeira abordagem consiste em alterar os resultados de algumas observações para remover a dependência entre o resultado e a classe do indivíduo. A segunda abordagem atribui diferentes pesos às observações para alcançar o mesmo objetivo.

3. Metodologia

O presente artigo busca realizar uma comparação de metodologias de pré-processamento dos dados para a mitigação de vieses sociais em modelos de Machine Learning aplicados ao contexto de concessão de crédito. Para isso, os autores definiram o seguinte fluxo de trabalho: inicialmente, foi realizada uma revisão de literatura extensiva com relação ao tema, buscando compilar as principais fontes de informação sobre técnicas de pré-processamento dos dados para a remoção de vieses sociais. Os resultados da revisão de literatura foram retratados na seção 2.

Após a revisão de literatura, os autores definiram as métricas a serem utilizadas para a identificação de vieses na base de dados e nos resultados do modelo ajustado. As métricas escolhidas e as justificativas para tais escolhas podem ser vistas na seção 3.1. Após a seleção das métricas, os autores passaram para a seleção das metodologias para a remoção de vieses de bases de dados que serão comparadas no estudo. As metodologias escolhidas e suas descrições podem ser encontradas na seção 3.2. Após a aplicação das técnicas de pré-processamento dos dados, foi realizado o ajuste de modelos de Regressão Logística com as bases de dados geradas. Maiores detalhes sobre o modelo de Regressão Logística e suas métricas de avaliação podem ser encontrados na seção 3.3. A descrição do banco de dados utilizado para o ajuste dos modelos está na seção 4. Já na seção 5, é possível encontrar a discussão dos resultados do ajuste dos modelos e suas implicações. Finalmente, as conclusões do presente estudo são consolidadas na seção 6.

3.1. Métricas para a Identificação de Vieses

Em sua revisão de literatura, Pagano et al. [2023] indicam as métricas de Impacto Díspar e de Igualdade de Oportunidade como as métricas mais comumente utilizadas para medir a justiça de modelos de machine learning. O Impacto Díspar é calculado pela razão entre a probabilidade de se receber um resultado positivo do modelo dado que um indivíduo faz parte de um grupo não-privilegiado e a probabilidade de um indivíduo ter um resultado positivo dado que ele faz parte de um grupo privilegiado. Além de ser uma métrica útil para avaliar o modelo de machine learning, também podemos utilizá-la para a avaliação da base de dados. Seja Y a variável resposta do modelo e X o atributo protegido que pode assumir valores de x1 e x2. Além disso, seja Z a variável que informa se um usuário pagaria ou não um empréstimo. O Impacto Díspar pode ser calculado pela fórmula abaixo:

$$ID = P(Y = 1|X = x_1) / P(Y = 1|X = x_2), \text{ em que } Y \in \{0, 1\}$$

Por outro lado, a Igualdade de Oportunidade requer que a probabilidade de um indivíduo que pagaria um empréstimo na vida real receba um resultado positivo do modelo seja igual para os grupos privilegiado e não-privilegiado. Assim, a taxa de verdadeiros positivos deve ser igual para ambos os grupos. A Igualdade de Oportunidade é uma métrica que só pode ser avaliada caso existam informações sobre o resultado real do empréstimo (pagamento ou não pagamento). A razão abaixo denota o nível de Igualdade de Oportunidade. Quanto mais próximo de 1, mais igual é o modelo. Um resultado igual a 1 indica um modelo com Igualdade de Oportunidade.

$IDO = P(Y = 1| X = x_1, Z = 1) / P(Y = 1| X = x_2, Z = 1)$, em que $Y \in \{0, 1\}$ e $Z \in \{0, 1\}$, em que Z representa o resultado real do empréstimo (pagamento ou não pagamento).

3.2. Metodologias para Remoção de Vieses

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

Foram selecionadas quatro metodologias de pré-processamento dos dados para a remoção de vieses. A escolha por metodologias de pré-processamento passa por alguns fatores: inicialmente, é necessário destacar que vieses sociais estão presentes nos dados em si. Balayn et al. [2021] afirmam que uma das principais causas para que sistemas de suporte de decisões baseados em dados apresentem resultados enviesados é justamente o fato de, por vezes, os sistemas utilizarem bases de dados enviesadas como bases de treinamento.

Os autores evidenciam que o fluxo de trabalho típico no desenvolvimento de um sistema de suporte à tomada de decisões baseados em dados não leva em conta a remoção de vieses dos dados. Dessa forma, a utilização de técnicas de pré-processamento evita que os desenvolvedores dos algoritmos tenham que tratar esses vieses internamente ao modelo ou posteriormente à implementação (pós-processamento) A escolha do tipo de técnica a ser utilizada depende do acesso possuído pelo indivíduo. O desenvolvedor do sistema, por exemplo, tende a trabalhar com o processamento dentro do algoritmo. Um tomador de decisão, que só possui acesso aos resultados do modelo, irá focar em métodos de pós-processamento. Além disso, técnicas de pré-processamento possuem as vantagens de não se alterarem de acordo com a base de dados inserida e não serem impactadas pela escolha de classificador utilizado, sendo mais facilmente replicáveis [Celis et al., 2020].

3.2.1. Rebalanceamento dos Dados

A abordagem de rebalanceamento dos dados é citada por Celis et al. [2020], Calders et al. [2009] e Khakurel et al. [2022]. Essa abordagem consiste na atribuição de pesos para cada observação do banco de dados. Observações da classe privilegiada com resultados positivos vão receber pesos menores que observações com resultados negativos, enquanto observações da classe não privilegiada com resultados positivos vão receber pesos maiores que observações dessa mesma classe com resultados negativos. São calculados quatro pesos, um para cada combinação entre atributo protegido e resultado. Os pesos são utilizados para a realização de uma reamostragem do banco de dados a fim de garantir uma representação mais justa. O cálculo dos pesos é feito através da seguinte fórmula:

$$P = P(Y = y) \times P(X = x)/P(Y = y\,|X = x),$$

em que Y assume valores de 0 ou 1 e X assume valores de x_1 ou x_2.

3.2.2. Massageamento dos Dados

A abordagem de massageamento dos dados é citada por Calders et al. [2009]. Essa abordagem consiste na alteração dos resultados de algumas observações da classe privilegiada de positivos para negativos e alterar os resultados do mesmo número de observações da classe não-privilegiada de negativos para positivos, a fim de remover a dependência entre o resultado e a classe do indivíduo. As observações não são alteradas aleatoriamente. É necessário ajustar um modelo inicial e estimar as probabilidades de um resultado positivo para cada indivíduo. No caso deste artigo, a Regressão Logística foi utilizada para o ajuste. As observações inicialmente positivas são selecionadas em ordem crescente de probabilidade, enquanto as observações negativas são selecionadas em ordem decrescente. O número de observações de cada valor que são alteradas é definido pela fórmula abaixo. Assim, são realizadas um total de 2M alterações.

$$M = (N(x1) \times N(x2, Y = 1) - N(x2) \times N(x1, Y = 1))/(N(x1) + N(x2)),$$

em que $N(x_1)$ e $N(x_2)$ indicam o número de ocorrência de cada um dos grupos, e N(x, Y=1) indica o número de ocorrência com resultados positivos de cada grupo.

3.2.3. Alteração de variáveis com preservação de ranking

O método, proposto por Feldman et al. [2015], consiste na alteração da base de dados de forma a preservar ranking, ou seja, garantir que o modelo continue escolhendo candidatos mais fortes ao invés de candidatos mais fracos. O algoritmo de reparo busca criar um conjunto de atributos não sensíveis modificados de forma a garantir que o atributo sensível se mantenha o mesmo. O modelo proposto é aplicável apenas para valores numéricos ou valores categóricos que possuam ordenação. Outros atributos devem ser removidos da base de dados.

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

Todas as variáveis devem ser transformadas em uma escala de 0 a 1, em que 0 é o valor mínimo da variável e 1 é o valor máximo. A escala deve considerar os grupos privilegiados e não privilegiados de forma separada. A transformação utiliza a seguinte fórmula:

(b - max(Bi)|X = x) / (max(Xi)|X = x - min(Xi)|X = x), em que Bi é uma variável preditora, b uma observação desta variável e X a variável sensível que assume valores de x_1 e x_2.

Ou seja, se o valor máximo de uma variável é 1.000 para o grupo não privilegiado e 2.000 para o grupo privilegiado, ambos receberão o valor 1 para aquela variável. Dessa forma, as variáveis passam a ser independentes do atributo protegido. Além disso, é necessário remover a variável sensível do banco de dados para realizar o ajuste do modelo.

3.2.4. Remoção de informação do atributo protegido

De acordo com Fu et al. [2021], os autores propuseram um método prático para mitigar o viés em modelos de *Machine Learning*. O método é aplicável a qualquer algoritmo ou tarefa de previsão, resultando em uma perda mínima de precisão. A abordagem consiste em remover a codificação redundante das características de entrada e torná-las independentes dos atributos sensíveis, a fim de evitar que o modelo infira informações discriminatórias.

As variáveis consideradas no estudo podem ser classificadas em diferentes tipos: contínuas, como renda; categóricas, como status de emprego; categóricas ordinais, como notas de crédito; e censuradas, como quantidades de inadimplentes, que podem ter valores positivos contínuos com uma massa pontual em zero. Para cada tipo de variável, foram desenvolvidas técnicas específicas que adaptam as características modificadas, garantindo a redução do viés discriminatório.

Para atributos contínuos, o método utilizado consiste em estimar inicialmente a função de densidade de probabilidade para cada uma das variáveis do tipo, normalizando-as para uma função de distribuição cumulativa correspondente. Esse processo tem como objetivo tornar os dados mais justos e uniformes. Já para variáveis categóricas, a técnica se baseia em expandir a dimensão do atributo de k valores para 2k valores, usando uma codificação para cada variação do atributo sensível, transformando-os em variáveis numéricas. Por exemplo, no caso de uma variável que representa o estado de residência, em que k = 27, a nova variável assume 54 variações, em que o valor 1 corresponde a "Minas Gerais - Gênero Feminino".

No caso de variáveis categóricas ordinais, o objetivo é evitar tratá-las como variáveis nominais, uma vez que isso resultaria na perda da informação ordinal presente. Para contornar esse problema, os dados são convertidos em valores numéricos seguindo uma distribuição uniforme. Por exemplo, categorias como "baixo", "médio" e "alto" são mapeadas para os valores 1, 2 e 3, respectivamente.

Por fim, o tratamento de variáveis censuradas envolve a transformação do atributo em uma variável binária, em que a massa pontual (representada por um valor que ocorre em grande proporção) assume o valor 0, enquanto os demais valores assumem o valor 1, ou vice-versa. Essa abordagem permite a inclusão da informação sobre a censura nos dados analisados.

Os resultados obtidos demonstram que o método proposto foi eficaz na redução do viés discriminatório em modelos de ML. Além disso, superou abordagens anteriores, preservando a precisão do modelo e promovendo maior igualdade e justiça nos resultados. Essa abordagem tem particular importância em contextos nos quais a discriminação pode ter consequências significativas, como em decisões de contratação, concessão de empréstimos ou no sistema de justiça criminal.

A tabela a seguir apresenta um resumo das metodologias descritas:

Metodologia	Descrição
Rebalanceamento dos Dados	Reamostragem do banco de dados com atribuição de pesos para observações, priorizando a classe não privilegiada e considerando resultados positivos/negativos.

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

Massageamento dos Dados	Alteração não-aleatória dos resultados de algumas observações da classe privilegiada de positivos para negativos e vice-versa.
Alteração de variáveis com preservação de ranking	Alteração da base de dados de forma a preservar ranking, garantindo que o modelo continue escolhendo candidatos mais fortes ao invés de candidatos mais fracos.
Remoção de informação do atributo protegido	Eliminação de redundância e independência dos atributos sensíveis, com tratamento específico para cada tipo de variável de entrada.

Quadro 1: Resumo das Metodologias de Mitigação de Vieses

3.3. Modelo de Regressão Logística

A Regressão Logística é um modelo de Machine Learning Supervisionado, ou seja, um modelo definido pelo uso de conjuntos de dados rotulados para treinar algoritmos que classificam dados. Além disso, a Regressão Logística pode também ser considerada um modelo de Classificação, que tem como objetivo prever o resultado de uma variável binária, que segue a distribuição de Bernoulli.

No contexto deste trabalho, o objetivo é estimar a probabilidade de um indivíduo conseguir ou não um empréstimo com base nas variáveis preditoras. Assim, a regressão logística nos fornece probabilidades de 0 a 1. Para classificação binária, quando trabalhamos com dados balenceados, uma probabilidade menor que 0,5 está associada ao evento 0, enquanto uma probabilidade maior que 0,5 está associada ao evento 1. É importante ressaltar que o tomador de decisão pode alterar esse limite de forma a melhorar o ajuste dos dados e os resultados do modelo. A equação base de um modelo logístico univariado pode ser vista abaixo. [IBM, 2023]

$P(Y = 1|x) = \exp(\beta_0 + \beta_1 \times x) / (1 + \exp(\beta_0 + \beta_1 \times x))$, em que x é o valor de uma observação da variável preditora X e Y é a variável resposta que assume valores de 0 e 1.

3.3.1. Estatísticas de Ajuste para Modelos de Classificação

A Sensibilidade pode ser definida como a taxa de acertos do modelo em relação a resultados positivos (Y=1), enquanto a Especificidade é definida como a taxa de acertos do modelo em relação a resultados negativos (Y=0). Podemos calcular essas métricas a partir das fórmulas abaixo:

$$Sensibilidade = VP/(VP + FN)$$
$$Especificidade = VN/(VN + FP)$$

Os valores para o cálculo da Sensibilidade e da Especificidade podem ser obtidos através da Matriz de Confusão, que descreve o desempenho do ajuste de um modelo comparando o resultado real e o resultado previsto pelo modelo.

Resultado Real/Resultado Previsto	0	1
0	Veradeiro Negativo (VN)	Falso Positivo (FP)
1	Falso Negativo (FN)	Verdadeiro Positivo (VP)

Quadro 2: Matriz de Confusão

Como já informado anteriormente, o tomador de decisão pode alterar o limiar de tomada de decisão, o que afeta os valores de Sensibilidade e Especificidade. A fim de garantir que a avaliação do modelo seja independente do limiar de tomada de decisão, iremos utilizar da métrica de Área Abaixo da Curva (AUC). A AUC mede a área abaixo da curva ROC, que representa o gráfico da Sensibilidade em função de 1 - Especificidade para todos os limiares de corte entre 0 e

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

1. Assim, a curva ROC cruza a taxa de verdadeiros positivos (Sensibilidade) com a taxa de falsos positivos (1 - Especificidade), indicando as taxas de acerto do modelo tanto em prever corretamente observações positivas quanto em prever corretamente observações negativas. Quanto melhor for o modelo ajustado, maior será a AUC. O valor máximo da AUC pode ser 1, quando o modelo consegue prever 100% das respostas para todos os valores de cortes. Isso acontece caso a Sensibilidade e a Especificidade sejam iguais a 1 para todos os valores de corte. Assim, a curva ROC será um quadrado unitário, e a AUC será igual a 1. Por outro lado, o pior resultado da AUC é 0,5, indicando um modelo completamente aleatório.

4. Banco de Dados

O banco de dados utilizado neste estudo foi obtido a partir de uma operadora de cartão de crédito de marca própria de uma empresa de crédito brasileira e suas lojas parceiras, com uma quantidade inicial de 50.000 (cinquenta mil) observações. Esses dados são relacionados aos clientes da empresa e foram coletados por meio de uma aplicação de pontuação de crédito. No entanto, não há informações disponíveis sobre os candidatos rejeitados, o que resulta em uma amostra de dados altamente tendenciosa para a recalibração do sistema. Isso ocorre porque apenas uma parte do mercado é monitorada e tem suas decisões avaliadas como "bem-sucedidas" ou "mal-sucedidas".

A base foi utilizada durante a PAKDD 2010, uma conferência realizada em Hyderabad, Índia, em que foram apresentados os mais recentes avanços em descoberta de conhecimento em mineração de dados. Essa competição, co-organizada pela NeuroTech Ltd. e pelo Centro de Informática da Universidade Federal de Pernambuco, teve como objetivo recalibrar um sistema de avaliação de risco de crédito baseado em dados enviesados. Os participantes da competição foram desafiados a construir um modelo para um sistema de suporte à decisão binária usando a amostra de dados enviesada em questão. A métrica de desempenho utilizada foi a curva ROC, já mencionada e explicada na seção 3.3.1 deste artigo.

A partir desta fonte, foram realizados tratamentos para preparar o banco de dados para análises posteriores. Primeiramente, os dados foram lidos de um arquivo disponível publicamente na plataforma GitHub em formato de texto.

Em seguida, foram aplicados critérios de exclusão às colunas da base de dados. Foram identificadas as colunas que continham apenas um valor de dados ou uma grande proporção de valores nulos, indicando falta de variabilidade e tornando-as irrelevantes para as análises subsequentes. Essas colunas foram removidas da base de dados para reduzir a complexidade e otimizar o desempenho das análises. Por fim, tendo em vista que seria utilizado o modelo de Regressão Logístico, foram removidas as variáveis categóricas com grande variabilidade, como cidade e bairro, para reduzir a complexidade e melhorar a performance computacional do modelo.

Uma análise inicial foi conduzida para examinar a relação entre a variável sensível definida, referente ao gênero, e a variável resposta, que indica se o cliente pagaria o crédito concedido. A variável de gênero foi binarizada, sendo atribuído o valor 1 para o gênero feminino. Ao observar o gráfico de proporção apresentado na Figura 1, é possível identificar uma diferença de dois pontos percentuais em relação à proporção de pessoas em cada gênero propensas a pagar a linha de crédito. Aproximadamente 27% dos homens realizariam o pagamento do crédito solicitado, enquanto o grupo feminino apresentou uma taxa ligeiramente inferior, com 25% do grupo obtendo o crédito.

Para avaliar a relevância dessa diferença proporcional, foi realizado um teste estatístico Qui-quadrado, que compara as proporções estatísticas entre os grupos. Os resultados indicaram suficiência estatística e um tamanho de amostra adequado para afirmar que a diferença entre as proporções é estatisticamente significativa, com um p-valor do teste de aproximadamente 0,0002; tal índice deve ser inferior a 0,05 (5%) para atestar que as proporções não são iguais. Isso sugere que o gênero pode desempenhar um papel importante na concessão de crédito, com possíveis implicações em termos de equidade e justiça.

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

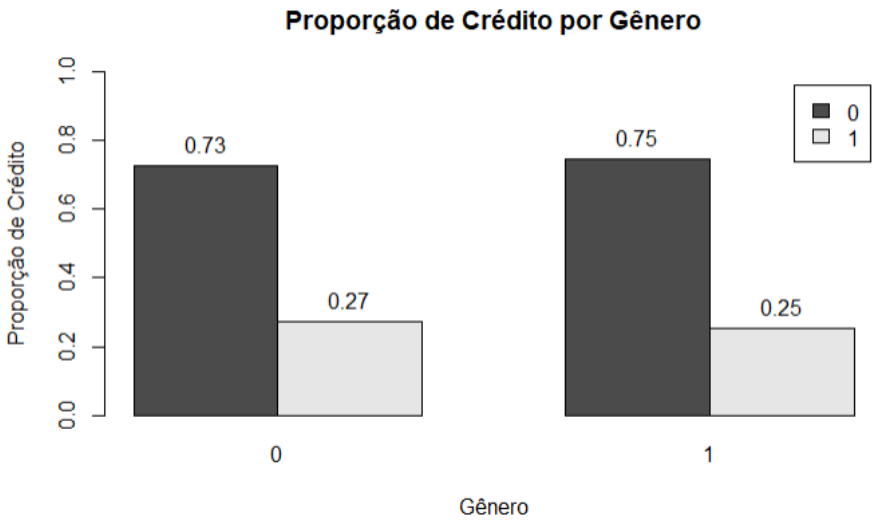


Figura 1: Distribuição da Variável Resposta por Gênero (sendo 1 = F; 0 = M)

Essa análise preliminar fornece uma visão inicial da relação entre gênero e crédito concedido, destacando a necessidade de investigações mais aprofundadas para compreender os fatores subjacentes a essa disparidade. É crucial considerar outras variáveis relevantes, como renda, histórico de crédito e características demográficas, a fim de obter uma compreensão mais completa dos padrões observados e identificar possíveis viéses ou discriminação de gênero no processo de concessão de crédito.

5. Resultados e Discussão

O ajuste do modelo Logístico foi realizado para cada uma das metodologias expostas na seção 3.2. Além das metodologias citadas, foram utilizados dois modelos de referência: o modelo com todas as variáveis, incluindo a variável de gênero, e o modelo com todas as variáveis, apenas removendo a variável de gênero e sem a realização de nenhuma outra manipulação dos dados. Para a realização do ajuste, foi utilizado o *software* R. O cálculo do valor da AUC de cada modelo foi feito com base em uma validação cruzada utilizando o método K-Fold, em que K = 100. Para metodologias em que houve manipulação do conjunto de variáveis preditoras, as bases de dados alteradas foram utilizadas tanto no conjunto de treino quanto no conjunto de teste, uma vez que subentende-se que os dados serão tratados antes de serem inseridos no modelo. Para metodologias que envolvem apenas a alteração da variável resposta, o conjunto de treino utilizado teve como fonte de alimentação os dados manipulados, enquanto o conjunto de teste utilizou a base original.

O cálculo das métricas de igualdade foi feito a partir do ponto de corte do limiar da probabilidade estimada pelo modelo em que a Sensibilidade é igual à Especificidade. Essa escolha foi feita para tornar a comparação mais justa entre os modelos, além de selecionar um ponto de corte em que as proporções de crédito concedido se assemelham às proporções vistas na base original. A avaliação das medidas de igualdade será feita com base na razão entre os valores obtidos para mulheres e homens. Dessa forma, quanto mais próximo de 1, melhor é o desempenho do modelo. Valores maiores do que 1 indicam um favorecimento para as mulheres, enquanto valores menores indicam um favorecimento dos homens. Além da análise do valor bruto das métricas para entender qual gênero é favorecido por qual modelo, a distância entre o valor obtido pelos modelos e o valor unitário também será analisada a fim de entender quais modelos mais se aproximam do ideal de justiça.

A Tabela 3 apresenta os resultados das métricas de igualdade e desempenho para os dois conjuntos de referência, bem como para os quatro modelos nos quais as técnicas de mitigação de viés foram aplicadas. Os valores destacados em preto indicam os modelos que obtiveram melhor

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

desempenho em uma determinada métrica, enquanto o valor destacado em vermelho indica o modelo com pior desempenho na respectiva métrica.

Inicialmente, é observável que todas as abordagens metodológicas obtiveram resultados superiores nos indicadores de igualdade em comparação ao modelo convencional, acarretando, por conseguinte, uma diminuição de até 0,06 na medida da AUC. Além disso, verificou-se que a simples remoção do atributo relacionado ao gênero do modelo não resulta em melhorias significativas em termos de igualdade ou desempenho. Ao contrário, esse modelo em questão revela valores de Igualdade de Oportunidade e Impacto Díspar superiores a 1, indicando a presença de um "discriminação reversa" (conceito distinto daquele utilizado nas Ciências Sociais). Isso ocorre uma vez que ainda existem informações relacionadas ao gênero na base de dados em outras variáveis, fazendo com que apenas parte da informação seja retirada.

Métrica	Base Completa	Base sem Gênero	Base Balanceada	Base Massageada	Base Rankeada	Base Independente
Igualdade de Oportunidade (IDO)	0,72	1,42	0,90	0,97	0,89	0,94
|IDO - 1|	0,28	**0,42**	0,10	**0,03**	0,11	0,06
Impacto Díspar (DI)	0,69	1,17	1,16	1,01	1,02	0,92
|DI - 1|	**0,31**	0,17	0,16	**0,01**	0,02	0,08
AUC	**0,64**	0,63	0,63	0,59	**0,58**	0,63

Tabela 1: Resultados dos Modelos

A análise da tabela revela claramente que o modelo que empregou a técnica de massageamento dos dados alcançou os resultados mais favoráveis em termos de igualdade e mitigação de vieses. No entanto, é importante ressaltar que a sua AUC foi a segunda menor, indicando o impacto negativo que os procedimentos adotados tiveram no desempenho do modelo. Outro aspecto relevante a ser considerado é que essa abordagem é intrusiva, uma vez que promove alterações diretas na variável resposta da base de dados original.

A metodologia que utiliza o rebalanceamento da base (3.2.1) e a que emprega a técnica de retirar toda a informação relacionada a gênero da base de dados (3.2.4) se destacam ao levarmos em conta o *trade-off* entre as métricas de igualdade e o desempenho do modelo. Ambas apresentam uma perda de apenas 0,01 na AUC do modelo, ao passo em que apresentam bons resultados de igualdade, principalmente na métrica de Igualdade de Oportunidade (0,10 e 0,06, respectivamente). Apesar dos bons resultados em métricas de igualdade, o modelo 3.2.3 apresentou resultados bem ruins na AUC (piores resultados de todos). Esse resultado é esperado tendo em vista que a metodologia só considera variáveis numéricas e a base de dados é composta majoritariamente por variáveis categóricas. Dessa forma, grande parte da informação presente no banco de dados não foi usada pelo modelo, resultando em um pior desempenho.

A análise das curvas ROC dos modelos ajustados mostra que para diferentes limites de tomadas de decisão, diferentes modelos apresentam uma melhor performance. Para valores mais altos de Especificidade e mais baixos de Sensibilidade, o modelo da Base Balanceada apresenta os valores mais altos na curva, enquanto o modelo com a base Completa apresenta melhor desempenho para o cenário oposto. Com exceção das bases Ranqueadas e Massageadas, que apresentam valores piores que os outros modelos, é possível observar desempenhos muito similares entre os modelos ajustados, indicando que a perda de desempenho não é tão grande (como evidenciado nos valores da AUC). Dessa forma, dependendo do limite de tomada de decisão, é possível até melhorar o desempenho do modelo ao aplicar as metodologias de remoção de vieses do banco de dados.

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

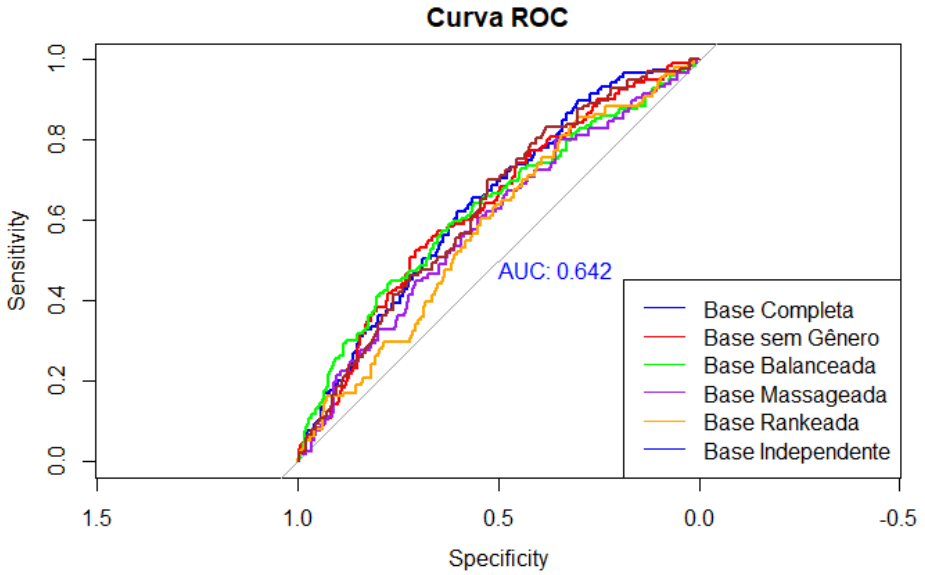


Figura 2: Gráfico de Curva ROC

A decisão de qual a metodologia ideal a ser utilizada depende muito do contexto analisado e do objetivo com o qual a metodologia está sendo usada. Para o contexto específico de Concessão de Crédito, é importante que o desempenho do modelo, medido aqui pela AUC, seja levado em conta na tomada de decisão referente à metodologia escolhida. Afinal, quanto maior a taxa de acerto do modelo, mais viável financeiramente é o modelo.

Com relação às medidas de igualdade, a Igualdade de Oportunidade pode ser vista como a medida mais relevante nesse contexto. Até por questões sociais, é natural que tenhamos grupos privilegiados em situações que favoreçam a concessão de crédito (por exemplo, homens recebendo mais que mulheres e, por consequência, recebendo mais crédito). Dessa forma, o Impacto Díspar vai existir, mas não necessariamente por causa do modelo. Entretanto, ao garantir a Igualdade de Oportunidade, é possível garantir que pessoas que efetivamente pagariam o empréstimo tenham a mesma chance de receber crédito, independentemente de seu gênero. Como ressaltado por Bono et al. [2021], um modelo de concessão de crédito não discriminatório deve refletir a capacidade de pagar crédito que uma pessoa possui, no sentido da chance que ela tem de pagar o crédito concedido. Dessa forma, com base no banco de dados analisado e nos resultados obtidos, os autores acreditam que o modelo de Base Independente é o modelo com melhor custo-benefício entre desempenho e igualdade.

6. Conclusão

Em suma, este estudo ressalta a importância de mitigar vieses sociais ao usar técnicas de Machine Learning. Ao analisar estratégias de pré-processamento de dados para reduzir vieses em modelos de concessão de crédito, foi observado que todas as abordagens resultaram em melhorias significativas na igualdade em relação ao modelo convencional. No entanto, houve um *trade-off* entre igualdade e desempenho do modelo. A metodologia de remoção de informação do atributo protegido foi considerada a mais apropriada, alcançando um equilíbrio satisfatório entre justiça e desempenho.

A métrica mais relevante no contexto da concessão de crédito foi identificada como a Igualdade de Oportunidade. Esses resultados destacam a importância de adotar abordagens de mitigação de vieses em aplicações de Machine Learning, garantindo justiça e equidade em decisões automatizadas. Além disso, enfatiza-se a necessidade de escolher as métricas apropriadas para avaliar os resultados, levando em consideração o domínio e os objetivos específicos do problema em questão.

Esta pesquisa contribui para o avanço do campo de ética em IA, trazendo uma inédita aplicação em uma base de dados brasileira e fornecendo insights valiosos para profissionais e

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br

pesquisadores envolvidos no desenvolvimento de modelos de Machine Learning aplicados à concessão de crédito e outras áreas semelhantes. À medida que a tecnologia evolui, é essencial abordar proativamente os vieses sociais para promover equidade e justiça em nossos sistemas automatizados.

7. Referências

Balayn, A., Lofi, C., & Houben, G. (2021). Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal — The International Journal on Very Large Data Bases*, 30(5), 739–768.

Beretta, E., Santangelo, A., Lepri, B., Vetró, A., & De Martin, J. C. (2019). The Invisible Power of Fairness. How Machine Learning Shapes Democracy. *In Advances in Artificial Intelligence, Proceedings of 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019*, 238–250.

Bono, T., Croxson, K., & Giles, A. (2021). Algorithmic fairness in credit scoring. *Oxford Review of Economic Policy*, 37(3), 585-617.

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints. 2009 IEEE International Conference on Data Mining Workshops, 13-18.

Celis, L. E., Keswani, V., & Vishnoi, N. K. (2020). Data preprocessing to mitigate bias: A maximum entropy based approach. *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, 1349-1359.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*, 259–268.

Fu, R., Huang, Y., & Singh, P. V. (2021). Crowds, Lending, Machine, and Bias. *Information Systems Research*, 32(1), 72-92.

IBM (2016). Webpage. `https://www.ibm.com/topics/supervised-learning`. Acessado 2023-06-21.

Khakurel, U., Abdelmoumin, G., Bajracharya, A., & Rawat, B. D. (2022). Exploring Bias and Fairness in Artificial Intelligence and Machine Learning Algorithms. *Proc. SPIE 12113, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications IV*, 1211324.

KDNuggets (2010). PAKDD 2010: Data Mining Competition. Webpage. `https://www.kdnuggets.com/2010/03/f-pakdd-2010-data-mining-competition.html`. Acessado 2023-06-21.

O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy*. New York, NY: Crown Publishers.

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., et al. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data Cogn. Comput.*, 7, 15.

DOI: 10.59254/sbpo-2024-193549
https://proceedings.science/p/193549?lang=pt-br