Estudo experimental sobre justiça algorítmica aplicada em modelos de análise de crédito

Tiago A. Oliveira[1], João V. L. Oliveira[1], Tarcísio P. Farias[2], Erick W. R. Cruz[2], Leandro J. S. Andrade[3], Robespierre Pita.[1]

[1]Instituto de Computação – Universidade Federal da Bahia (UFBA)

[2]Instituto Federal de Educação, Ciência e Tecnologia da Bahia (IFBA)

[3]Escola de Administração – Universidade Federal da Bahia (UFBA)

{oliveira.t, j.luz}@ufba.br, {tarcisioparaiso, erickwellber}@gmail.com

{leandrojsa, robespierre.pita}@ufba.br

Abstract. *Machine Learning (ML) models for algorithmic decision-making are widely applied to support risk management and credit analysis. However, the significant increase in available data, the complexity of modern models, and public scrutiny surrounding artificial intelligence have intensified the debate on the need to identify and mitigate biases in predictions. This study aims to analyze the relationship between quantitative measures of algorithmic fairness and quality metrics obtained by ML models in credit analysis tasks. Initial results indicate that certain models can achieve promising performance levels without necessarily affecting or deteriorating fairness in their predictions.*

Resumo. *Modelos de Machine Learning (ML) para tomada de decisão algorítmica são amplamente aplicados para suportar a gestão de risco e análise de crédito. Contudo, o sensível aumento de dados disponíveis, a complexidade dos modelos mais modernos e o escrutínio público em torno da inteligência artificial acirraram o debate sobre a necessidade de identificação e mitigação de vieses em predições. Este estudo propõe analisar a relação entre medidas quantitativas de justiça algorítmica e métricas de qualidade obtidas por modelos de ML em tarefas de análise de crédito. Os resultados iniciais indicam que determinados modelos conseguem alcançar níveis promissores de desempenho sem necessariamente afetar ou deteriorar a justiça em suas predições.*

1. Introdução

O serviço de transferência de fundos entre credores e mutuários oferecidos por bancos e instituições financeiras desempenham um relevante papel no desenvolvimento e estabilidade de uma economia moderna [Sathye et al. 2003]. Uma alocação eficiente de crédito depende da capacidade destas instituições de efetivar empréstimos de qualidade, identificando os níveis de inadimplência, baixa diversificação — nível de exposição a um determinado indivíduo ou setor produtivo —, e a aptidão ou interesse de seus clientes em cumprir com suas obrigações contratuais [Bhatore et al. 2020]. Neste sentido, a gestão de risco compreende análises e decisões para reduzir o número de ativos não rentáveis (NPA, do inglês *non-perfomative assets*) ou fraudes, disponibilizando crédito a prospectos comprometidos e legítimos [Akkizidis and Stagars 2015]. Esta avaliação é frequentemente suportada por uma pontuação de crédito (CS, do inglês *credit scoring*), usada para

classificar clientes entre aqueles com e sem capacidade de crédito [Bhatore et al. 2020], utilizando diversas variáveis qualitativas ou históricas dos indivíduos e empresas clientes.

O uso crescente de aprendizado de máquina (ML, do inglês *machine learning*) para análise de crédito, especialmente para SC, se dá pela possibilidade de modelar este como um problema de classificação binária ou multiclasse [Bhatore et al. 2020]. O uso de ML para tomada de decisão algorítmica tem sido potencializada pela crescente disponibilidade de dados e a diversidade de suas fontes de captura [Sadok et al. 2022]. Inúmeros pesquisadores reportam e discutem o resultado da aplicação de árvores de decisão, máquina de vetores de suporte (SVM, do inglês *support vector machine*) [Gyamfi and Abdulai 2018], redes neurais [Derelioğlu et al. 2009, Lai and Zhou 2006], florestas aleatórias (RF, do inglês *random forest*) [Szwabe and Misiorek 2018], árvores com gradiente aumentado (GBT, do inglês *gradient boosted trees*) [Xia et al. 2017], etc. Embora estes modelos alcancem resultados promissores, seu uso acrítico para suportar decisões tão sensíveis pode reproduzir preconceitos e perpetuar estereótipos prejudiciais [He et al. 2020], demandando métodos auxiliares capazes de identificar ou mitigar injustiças que afetem grupos desfavorecidos.

Este estudo propõe analisar a relação entre medidas quantitativas de justiça algorítmica (JA) e métricas de qualidade obtidas por modelos de ML em tarefas de SC. Os resultados de acurácia, precisão e pontuação F1 obtidos pelos modelos RF e GBT são em bases de dados de benchmark usados SC amplamente utilizadas na literatura.

2. Trabalhos relacionados

Os estudos sobre JA em modelos de ML, inclusive aplicados para SC, podem ser categorizados entre os de abordagem quantitativa ou mitigatória. Em seguida, apresentaremos alguns estudos encontrados na literatura que se dedicam a avançar o conhecimento nestes métodos ou aplicá-los no contexto de análise de crédito. Para permitir uma análise comparativa destes trabalhos, consideraremos apenas o desempenho alcançado na *German credit score dataset* [Hofmann 1994], expresso pela medida de porcentagem das observações corretamente classificadas (POCC).

Diversos autores na relatam a comparação do desempenho de modelos de ML em bases de dados de *benchmark* para SC [Guidolin and Pedio 2021]. Um dos trabalhos mais tradicionais [Baesens et al. 2003] demonstra que métodos mais simples e explicáveis, tais como regressão logística e (LR, do inglês *logistic regression*) e análise discriminante linear (LDA, do inglês *linear discriminant analysis*), alcançam uma POCC de 74,6%, um resultado similar aos 73.7% obtidos por modelos modernos como os baseados em redes neurais (NN, do inglês *neural networks*). Diversas bases de dados publicamente disponíveis são usadas na avaliação, tornando o trabalho um importante *baseline* para avaliação futura de modelos. Trabalhos mais recentes se dedicam a avaliar arquiteturas mais complexas de NN e sua aplicação em SC [West et al. 2005], superando substantivamente o *baseline* estabelecido em [Baesens et al. 2003].

Modelos *ensemble* alcançam frequentemente resultados promissores quando comparados com métodos anteriores. Combinações de redes neurais [West et al. 2005] ou modelos SVM [Zhou et al. 2010] alcançam POCC entre 70 e 77%. Em seu trabalho, [Lessmann et al. 2015] atualiza os resultados de [Baesens et al. 2003], trazendo novos modelos e mais detalhes de parametrização. Desta vez, modelos *ensemble* baseados em árvores como RF e GBT são destacados, obtendo POCC de 86% e 81%, respectivamente.

Apesar da extensa literatura avaliando o desempenho dos modelos de ML, poucos artigos se dedicam a avaliar a relação entre as métricas de qualidade da predição em SC e a JA reproduzida. Trabalhos como [Kozodoi et al. 2022] e [Kasmi 2021] se dedicam a avaliar um número limitado de bases ou métricas de justiça. Em seu trabalho, [Hardt et al. 2016] propõe um método de mitigação de vieses capaz de aproximar os resultados de precisão $\tilde{Y} = Y$, sendo $\tilde{Y}$ a classe predita após o uso do método e Y a classe correta. Em seu estudo de caso, a proposta foi aplicada num classificador proprietário usado para estimar a inadimplência numa base de dados publicamente disponível. O método alcançou 99.3% do resultado ótimo (máxima rentabilidade), superando outras estratégias.

3. Justiça algorítmica em aprendizagem de máquina

Justiça algorítmica em ML consiste na capacidade destes modelos de evitar predições enviesadas, favorecendo um grupo, subgrupo ou indivíduo a partir de suas características [Mehrabi et al. 2021]. A identificação e quantificação desta propriedade depende frequentemente das métricas de qualidade alcançadas em diferentes estratos populacionais existentes numa base de dados, caracterizada por variáveis demográficas protegidas [Caton and Haas 2024].

As diversas definições para JA podem ser categorizadas em três categorias: individuais, de grupo ou subgrupo [Mehrabi et al. 2021, Caton and Haas 2024]. Uma noção de justiça individual define como adequadas as predições similares para indivíduos com características semelhantes [Dwork et al. 2012]. Esta definição garante que, caso o valor de um *atributo protegido* (como "sexo", "raça", "estado civil"ou "faixa etária") seja alterado, mantendo-se os valores das demais variáveis que não são causalmente independentes, a predição não mude [Caton and Haas 2024]. Alternativamente, pesquisadores podem considerar justas as predições que tratam igualmente diferentes os grupos representados nos dados. A noção de subgrupos, por sua vez, inclui estratégias para identificação de estratos populacionais ou amostras desfavorecidas pelo classificador, possibilitando usar esta informação para o ajuste iterativo dos modelos de ML. Na seção a seguir serão descritas brevemente quatro métricas que podem ser utilizadas para quantificar estes vieses e serão utilizadas no experimento: *Statistical parity difference* (SPD), *Disparate Impact* (DI), *Average Odds Difference* (AOD), *Equal Opportunity Difference* (EOD).

3.1. Statistical Parity Difference (SPD)

Usada para avaliar a equidade de um modelo de aprendizado de máquina. Essa métrica verifica se um resultado (por exemplo, conseguir um empréstimo) ocorre na mesma proporção para diferentes grupos demográficos [Ruback et al. 2022].

A SPD mede a diferença na proporção de decisões positivas entre dois grupos demograficamente diferentes (grupo protegido e grupo não protegido). Se o resultado for 0, indica paridade estatística, ou seja, ambos os grupos recebem decisões positivas na mesma faixa. Um valor diferente de 0 indica disparidade na distribuição de decisões entre os grupos.

3.2. Disparate Impact (DI)

É a razão entre a taxa de resultados favoráveis para um grupo protegido e a taxa de resultados favoráveis para um grupo de referência [Ruback et al. 2022]. Portanto,

$DI = \frac{Pr(\tilde{Y}=1|D=protegido)}{Pr(\tilde{Y}=1|D=referencia)}$, onde $Pr(\tilde{Y} = pos_label|D = desfavorecido)$ e $Pr(\tilde{Y} = pos_label|D = favorecido)$ se referem às taxas de seleção dos grupos protegido e de referência, respectivamente.

3.3. Average Odds Diference (AOD)

É calculada como a média das diferenças nas taxas de verdadeiros positivos e falsos positivos entre um grupo protegido e um grupo de referência [Ruback et al. 2022].

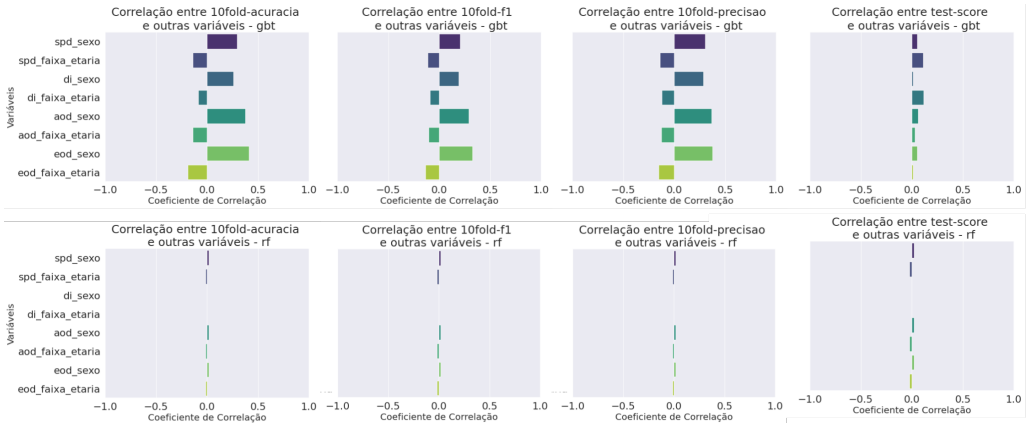

Figura 1. Correlação entre métricas de desempenho e de justiça algorítmica

3.4. Equal Opportunity Difference (EOD)

É a diferença entre as taxas de verdadeiros positivos para grupos protegidos e grupos de referência, assegurando que indivíduos qualificados de diferentes grupos demográficos tenham a mesma chance de conseguir um resultado positivo [Ruback et al. 2022].

4. Materiais e métodos

Na segunda etapa foram conduzidas as tarefas de engenharia de dados. Em geral, os nomes de variáveis foram alterados para suprimir espaços em branco, o nome das variáveis dependentes foram substituídas por ”*class*”e variáveis categóricas foram codificadas para números, mantendo a sua característica nominal para envio aos modelos de ML. Definimos ”sexo”e ”idade”como variáveis protegidas, presentes em todas as bases. Criamos uma variável binária ”faixa etária”a partir da ”idade”, categorizando indivíduos como ”jovens”(abaixo de 24 anos) e ”adultos”(24 anos ou mais). Consideramos a população favorecida como sendo do sexo masculino ou adultos. Em bases com classes desbalanceadas, aplicamos a técnica de *oversampling* minoritária sintética (SMOTE) para aumentar os registros da classe minoritária. Realizamos testes para garantir que o SMOTE não alterou a distribuição da base, utilizando a diferença média padronizada.

Em seguida, na terceira etapa, foram definidos os modelos e os conjuntos de hiperparâmetros usados para executá-los sobre cada uma das bases disponíveis. Ainda nesta etapa são coletadas as métricas de qualidade do modelo. Seguindo a tendencia dos trabalhos mais recentes na literatura, usamos a técnica do *10-fold cross validation* e calculamos a pontuação F1, a precisão para os dados de treinamento. Calculamos a acurácia nos 30% previamente separados para teste de cada experimento. Os modelos ensemble utilizados

foram o RF e o GBT com 240 e 600 conjuntos de parâmetros, respectivamente. No total $(240 * 5) + (600 * 5) = 2400$ experimentos foram executados.

Na última etapa do nosso desenho de experimento, os resultados de cada um das 2400 predições foram submetidas para o cálculo das medidas de justiça apresentadas na Seção 3. As variáveis de sexo e faixa etária (categorizada a foram consideradas

Tabela 1. Resultados de desempenho e justiça algorítmica obtidos

		medidas de qualidade dos modelos															
		10fold-acuracia				10fold-f1				10fold-precisao				test-score			
base	modelo	min	mean	max	stddev	min	mean	max	stddev	min	mean	max	stddev	min	mean	max	stddev
CreditCardApproval	gbt	0,934	0,972	0,986	0,015	0,934	0,972	0,986	0,015	0,917	0,964	0,983	0,019	0,933	0,973	0,987	0,015
CreditCardApproval	rf	0,745	0,806	0,855	0,039	0,744	0,805	0,855	0,040	0,713	0,782	0,844	0,046	0,742	0,806	0,866	0,041
GitHubLoanContest	gbt	0,665	0,734	0,819	0,043	0,399	0,556	0,739	0,149	0,665	0,748	0,826	0,056	0,723	0,777	0,849	0,029
GitHubLoanContest	rf	0,665	0,694	0,722	0,013	0,399	0,410	0,419	0,005	0,665	0,694	0,722	0,013	0,623	0,686	0,755	0,031
HomeCreditDefaultRisk	gbt	0,931	0,944	0,953	0,006	0,931	0,944	0,953	0,006	0,946	0,959	0,969	0,006	0,927	0,946	0,956	0,006
HomeCreditDefaultRisk	rf	0,786	0,813	0,834	0,015	0,785	0,812	0,833	0,015	0,748	0,790	0,819	0,022	0,785	0,815	0,844	0,015
german	gbt	0,711	0,736	0,786	0,019	0,635	0,671	0,728	0,024	0,773	0,795	0,825	0,013	0,677	0,740	0,777	0,021
german	rf	0,680	0,699	0,719	0,009	0,405	0,411	0,418	0,003	0,680	0,699	0,719	0,009	0,657	0,703	0,747	0,020
taiwan	gbt	0,804	0,815	0,822	0,005	0,669	0,679	0,687	0,005	0,592	0,643	0,673	0,023	0,803	0,815	0,825	0,006
taiwan	rf	0,799	0,812	0,824	0,008	0,579	0,637	0,682	0,037	0,674	0,694	0,732	0,014	0,797	0,812	0,829	0,008

		medidas de justiça algorítmica para variável sexo															
		spd_sexo				di_sexo				aod_sexo				eod_sexo			
base	modelo	min	mean	max	stddev	min	mean	max	stddev	min	mean	max	stddev	min	mean	max	stddev
CreditCardApproval	gbt	-0,209	-0,242	-0,276	0,014	0,526	0,573	0,624	0,021	-0,004	-0,024	-0,065	0,015	-0,008	-0,027	-0,073	0,014
CreditCardApproval	rf	0,270	0,312	0,384	0,029	1,808	2,012	2,393	0,120	0,111	0,173	0,278	0,043	0,068	0,131	0,217	0,029
GitHubLoanContest	gbt	-0,002	0,051	0,242	0,082	0,905	1,082	1,464	0,129	-0,001	0,036	0,207	0,083	-0,001	0,026	0,189	0,077
GitHubLoanContest	rf	0,000	0,000	0,000	0,000	1,000	1,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
HomeCreditDefaultRisk	gbt	0,161	0,186	0,209	0,012	1,363	1,423	1,486	0,032	0,026	0,040	0,056	0,007	-0,002	0,011	0,029	0,008
HomeCreditDefaultRisk	rf	0,137	0,159	0,182	0,010	1,336	1,405	1,479	0,036	0,027	0,060	0,083	0,011	0,035	0,071	0,107	0,015
german	gbt	-0,012	-0,080	0,015	0,050	0,492	0,755	1,070	0,140	-0,013	-0,068	0,042	0,062	-0,005	-0,082	0,090	0,093
german	rf	0,000	0,000	0,002	0,000	1,419	1,419	1,419	0,000	0,000	0,000	0,002	0,000	0,000	0,000	0,004	0,000
taiwan	gbt	-0,015	-0,030	-0,043	0,007	0,952	0,966	0,983	0,007	-0,004	-0,022	0,003	0,011	-0,004	-0,017	-0,026	0,006
taiwan	rf	0,006	0,018	0,035	0,006	1,007	1,021	1,040	0,007	0,000	0,013	0,032	0,009	0,001	0,009	0,021	0,004

		medidas de justiça algorítmica para variável faixa_etaria															
		spd_faixa_etaria				di_faixa_etaria				aod_faixa_etaria				eod_faixa_etaria			
base	modelo	min	mean	max	stddev	min	mean	max	stddev	min	mean	max	stddev	min	mean	max	stddev
CreditCardApproval	gbt	-0,199	-0,224	-0,252	0,014	0,626	0,655	0,690	0,017	-0,005	-0,023	-0,066	0,016	0,000	-0,009	0,005	0,012
CreditCardApproval	rf	-0,208	-0,272	-0,321	0,028	0,548	0,621	0,707	0,039	-0,092	-0,141	-0,199	0,022	-0,083	-0,142	-0,215	0,024
GitHubLoanContest	gbt	-0,016	-0,086	0,061	0,074	0,695	0,891	1,086	0,093	-0,001	-0,052	0,070	0,068	-0,008	-0,020	0,123	0,073
GitHubLoanContest	rf	0,000	0,000	0,000	0,000	1,000	1,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
HomeCreditDefaultRisk	gbt	0,281	0,366	0,456	0,041	1,546	1,714	1,905	0,080	-0,002	0,175	0,337	0,079	-0,003	0,009	0,045	0,024
HomeCreditDefaultRisk	rf	0,232	0,329	0,409	0,040	1,543	1,722	1,899	0,083	0,048	0,261	0,457	0,077	-0,004	0,030	0,124	0,043
german	gbt	-0,033	0,170	0,366	0,094	0,888	1,762	3,283	0,477	-0,029	0,126	0,363	0,108	-0,029	0,099	0,437	0,129
german	rf	-0,008	0,000	0,000	0,001	0,000	0,000	0,000	0,000	-0,014	0,000	0,000	0,002	-0,029	0,000	0,000	0,003
taiwan	gbt	-0,029	-0,060	-0,093	0,015	0,895	0,931	0,967	0,017	-0,013	-0,052	-0,101	0,022	-0,004	-0,032	-0,058	0,013
taiwan	rf	-0,009	-0,042	-0,089	0,020	0,901	0,954	0,990	0,022	0,000	-0,033	0,011	0,024	-0,002	-0,022	0,005	0,011

Foram utilizadas as implementações de referência disponíveis nas bibliotecas *Scikit-Learn* [Pedregosa et al. 2011] e *AI Fairness 360* [Bellamy et al. 2019]. Buscamos com este estudo responder a seguinte pergunta de pesquisa: *Existe uma (cor)relação entre o desempenho obtido nos modelos tradicionais de ML e a magnitude da justiça algorítmica alcançada?*. A seção seguinte se dedica a comunicar nossos resultados.

5. Resultados

Os resultados gerais coletados de nossos experimentos estão sumarizados na Figura 1 e Tabela 1. Seguindo as informações fornecidas em [Ruback et al. 2022], os grupos privilegiados são favorecidos quando as métricas SPD, AOD e EOD são menores que zero. No caso da métrica DI, o intervalo que denota este favorecimento é abaixo de um. Desta forma, busca-se verificar se os resultados de justiça algorítmica circulam o zero (ou um, no caso da DI) enquanto se mantém uma boa qualidade na predição.

Ao inspecionar a Tabela 1, nota-se que, em geral, os resultados obtidos pelos modelos estão compatíveis com os intervalos observados na literatura. Adicionalmente, também não há diferença importante no desempenho dos diferentes modelos. Contudo, a Figura 1 expõe indícios de que diferentes modelos conseguem alcançar níveis similares de desempenho sem necessariamente afetar ou deteriorar a justiça em suas predições. Este achado permite subsidiar uma discussão ainda em aberto sobre a possibilidade de incluir métricas de justiça algorítmica e métodos de mitigação de viéses como uma propriedade de qualidade no *pipeline* de análise de crédito.

6. Conclusão e trabalhos futuros

Neste estudo, conduzimos uma abordagem empírica para verificar se há relação entre o desempenho de modelos de ML e a produção de víeses na predição. Verificou-se que determinados modelos tem o potencial de alcançar níveis promissores de qualidade sem propagar injustiças.

Este trabalho é parte de uma pesquisa ainda mais ampla que visa avaliar se modelos inovadores de análise de crédito, baseados em psicometria, são capazes de mitigar vieses por concepção. Desta forma, como trabalho futuro imediato, os pesquisadores pretendem aprofundar a discussão e aplicar estes resultados em bancos de dados reais que podem ser disponibilizados por cooperativas de crédito ou grupos de pesquisa em colaboração.

Agradecimentos. Este trabalho foi possível devido ao apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) através de uma bolsa de Iniciação Científica voltada para alunos ingressantes através de ações afirmativas, programa PIBIC-AF (25124/2023), do Plano de Trabalho da UFBA "*Investigação de justiça algoritmica aplicada em modelos de análise de crédito*"e do projeto entitulado "*Inovar para pessoas negras*", executado sob o convênio 364/2022, celebrado entre a UFBA, IFBA e o Nubank.

Referências

Akkizidis, I. and Stagars, M. (2015). *Marketplace lending, Financial Analysis, and the Future of credit: Integration, Profitability, and risk management*. John Wiley & Sons.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *JORS)*, 54(6):627–635.

Bellamy, R. K., Hind, M., Mojsilović, A., et al. (2019). Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM JRD*, 63(4/5):4–1.

Bhatore, S., Mohan, L., and Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, 4(1):111–138.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38.

Derelioğlu, G., Gürgen, F., and Okay, N. (2009). A neural approach for sme's credit risk analysis in turkey. pages 749–759. Springer.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd ITCS*, pages 214–226.

Guidolin, M. and Pedio, M. (2021). Sharpening the accuracy of credit scoring models with machine learning algorithms. In *Data Science for Economics and Finance: Methodologies and Applications*, pages 89–115. Springer.

Gyamfi, N. K. and Abdulai, J.-D. (2018). Bank fraud detection using support vector machine. In *2018 IEEE 9th IEMCON*, pages 37–41. IEEE.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., editor, *NIPS*, volume 29. Curran Associates, Inc.

He, Y., Burghardt, K., and Lerman, K. (2020). A geometric solution to fair representations. In *Proceedings of the AAAI/ACM*, AIES '20, page 279–285, USA. ACM.

Hofmann, H. (1994). Statlog (german credit data).

Kasmi, M. L. (2021). Machine learning fairness in finance: An application to credit scoring. *Diss., Tilburg University*.

Kozodoi, N., Jacob, J., and Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *EJOR*, 297(3):1083–1094.

Lai, K. K. and Zhou, L. (2006). Neural network metalearning for credit scoring. In *International Conference on Intelligent Computing*, pages 403–408. Springer.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *EJOR*, 247(1):124–136.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830.

Ruback, L., Carvalho, D., and Avila, S. (2022). Mitigando vieses no aprendizado de máquina: Uma análise sociotécnica. *iSys*, 15(1):23–1.

Sadok, H., Sakka, F., and El Maknouzi, M. E. H. (2022). Artificial intelligence and bank credit analysis: A review. *Cogent Economics*, 10(1):2023262.

Sathye, M. et al. (2003). *Credit analysis & lending management*. Wiley.

Szwabe, A. and Misiorek, P. (2018). Decision trees as interpretable bank credit scoring models. In *BDAS 2018, Poland, September 18-20, 2018, Proceedings 14*, pages 207–219. Springer.

West, D., Dellana, S., and Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *COR*, 32(10):2543–2559.

Xia, Y., Liu, C., Li, Y., and Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert systems with applications*, 78:225–241.

Zhou, L., Lai, K. K., and Yu, L. (2010). Least squares support vector machines ensemble models for credit scoring. *Expert systems with applications*, 37(1):127–133.