Fairness in Risk Estimation of Brazilian Public Contracts

Órion Darshan Winter de Lima, Nazareno Andrade

Universidade Federal de Campina Grande, Brazil
orion@copin.ufcg.edu.br nazareno@computacao.ufcg.edu.br

Abstract. Brazilian government agencies are currently using machine learning models to make public contracts audition through risk estimation. Recent works have shown that decision making models, like risk estimation, may be unfair. Despite the fact that risk estimations of public contracts may be unfair, no studies evaluating model fairness have been found. This work contributes by analysing fairness over risk estimation of brazilian public contract. This article found that currently used models are unfair and biased towards a specific class. This means that people within this class may be negatively affected by these decision making models unfairness through risk estimation of their companies.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; I.2.6 [**Artificial Intelligence**]: Learning

Keywords: fairness, machine learning, risk estimation

1. INTRODUÇÃO

Sistemas de apoio a tomada de decisão baseados em aprendizagem de máquina são hoje amplamente utilizados em cenários que afetam a vida de pessoas [Speicher et al. 2018], tal como a análise de risco de crédito [Zan Huang 2004] e estimativa de risco criminal [Julia Angwin and Kirchner 2016]. Com a prevalência desse tipo de sistema em áreas com potencial de alto impacto na vida das pessoas, e de potenciais injustiças nesse impacto, trabalhos recentes têm levantado preocupações sobre possíveis vieses não intencionais de sistemas de aprendizagem de máquina [Saleiro et al. 2018].

No contexto da gestão pública federal brasileira, há interesse e estudos sobre a eficácia do uso de aprendizagem de máquina para estimativa de risco de contratos [Sun and Sales 2018; Silvio L. Domingos and Ramos 2016]. Assim como na análise de risco de crédito e predição de risco criminal em geral, injustiça na estimativa de risco para contratos públicos pode também causar grande prejuízo à vida de pessoas. Por exemplo, se sistemas baseados em aprendizagem de máquina fizerem com que um grupo vulnerável seja mais investigado, além do que seria necessário por seu risco real, as investigações podem gerar prejuízo para o grupo. Este prejuízo pode se dar, por exemplo, através de um ônus para provar a inocência frente a uma investigação. O viés de confirmação do investigador, onde as pessoas tendem a procurar, perceber, interpretar e criar novas evidências de modo a verificar suas crenças preexistentes pode levar investigadores a levantar suspeitas indevidas [Saul M. Kassin and Kukucka 2013], afetando os indivíduos do grupo sensível.

Apesar deste cenário ter potencial impacto em muitas empresas, pessoas e na gestão pública, não encontramos estudos com objetivo de avaliar, além da eficácia, a justiça e vieses no uso de aprendizagem de máquina para a estimativa de risco em contratos públicos brasileiros. Este artigo busca preencher essa lacuna, avaliando se modelos atualmente pesquisados para de estimativas de risco para contratos públicos e empresas são justos. A justiça foi avaliada em relação à características sensíveis das observações, ou seja, das empresa. O conceito de justiça, dentre outros, serão abordado a seguir.

Copyright©2019 Permission to copy without fee all or part of the material printed in KDMiLe is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

2. ESTIMANDO RISCO EM CONTRATOS PÚBLICOS

A aquisição de bens e serviços por parte da Administração Pública direta e indireta são regidos pela lei 8666/93, também conhecida como Lei de Licitações. Os órgãos de controladoria são os representantes da Administração Pública responsáveis por fiscalizar esta contratação, como determina o artigo 67 da lei 8666/93 [Lei 1993]. Apesar de serem entidades governamentais, os órgãos de controladoria, como todas as empresas, trabalham com recurso limitado para suas investigações. Com isso, existe a necessidade de priorizar as investigações, estimando o risco dos contratos, de modo a investigar primeiramente os contratos mais suspeitos [Silvio L. Domingos and Ramos 2016].

Estimativas de risco têm sido utilizadas extensivamente em outras áreas, como análise de crédito, onde os pesquisadores dessa área obtiveram resultados promissores aplicando diferentes métodos estatísticos e de Inteligência Artificial (IA), como aprendizagem de máquina (AM) [Zan Huang 2004]. A estimativa de risco dos contratos pode ser feito de forma manual, mas com o avanço da tecnologia, os órgãos de controladoria precisaram se modernizar e utilizar ferramentas que a tecnologia da informação proporciona, as quais são capazes de detectar padrões sobre as contratações irregulares ou ilícitas [Silvio L. Domingos and Ramos 2016].

A estimativa de risco para o contexto de contratações públicas no Brasil tem como objetivo estimar um risco associado à contratos ou empresas que quantifica a chance destas cometerem irregularidades. Sun e Sales [Sun and Sales 2018], por exemplo, estimaram o risco de empresas cometerem irregularidades severas, a qual é dada quando um contrato é rescindido ou quando uma empresa licitante é proibida de participar de futuras licitações.

3. JUSTIÇA EM MODELOS DE APRENDIZAGEM DE MÁQUINA

Trabalhos recentes têm levantado preocupações sobre a potencial injustiça da utilização de algoritmos de decisões em cenários que podem afetar a vida das pessoas, de acordo com certos grupos sociais ou indivíduos [Speicher et al. 2018]. Angwin et al. [Julia Angwin and Kirchner 2016] identificaram que o sistema de predição de risco criminal COMPAS, utilizado em audiências criminais estadunidenses para estimar o risco de reincidência em crimes, é enviesado do ponto de vista racial. Para o contexto de contratos públicos não foram encontrados trabalhos que avaliassem a justiça de modelos de estimativa de risco.

Nossa análise de justiça é baseada no *toolkit Aequitas*, proposto por Saleiro et al. [Saleiro et al. 2018]. Esse *toolkit* define métricas de justiça e viés para modelos de AM que permitem verificar se existe disparidade no comportamento do modelo para classes sensíveis do conjunto de dados. Classes sensíveis são definidas pelo conjunto das diferentes classes presentes em uma característica sensível do conjunto de dados, as quais não pertencem ao grupo de referência. Uma característica é dita sensível caso o julgamento diferente das classes da característica possa impactar negativamente os indivíduos de certas classes, tocando assim em questões morais. O grupo de referência é determinado para cada característica por uma classe de uma característica sensível do conjunto de dados. Essa classe é uma onde não se espera que haja efeito negativo caso haja um julgamento enviesado. Por exemplo, no caso do sistema de estimativa de risco de reincidência criminal COMPAS [Julia Angwin and Kirchner 2016], uma característica sensível é a raça das pessoas classificadas, e as pessoas da raça negra são uma classe sensível, enquanto brancos são o grupo de referência.

Sejam VP e VN as quantidades de verdadeiros positivos e verdadeiros negativos na classificação de um conjunto de entidades, e FP e FN as quantidades de falsos positivos e falsos negativos na mesma classificação. As principais métricas utilizadas neste trabalho são: a taxa de falsos positivos $TFP = \frac{FP}{FP+VN}$, a taxa de verdadeiros positivos $revocacao = \frac{VP}{VP+FN}$ e a $precisao := \frac{VP}{VP+FP}$. A TFP mede a proporção que o modelo estima erroneamente que é positivo, quando na verdade era negativo. A revocação mede a proporção que o modelo acerta dentre os casos positivos. A precisão mede a proporção que o modelo acerta entre os casos estimados como verdadeiros.

A disparidade diz respeito tanto à métrica avaliada quanto às diferentes classes de uma mesma característica sensível. A disparidade entre uma classe do grupo de referência e uma classe sensível é calculada para uma métrica, pela divisão do valor da métrica para a classe sensível pelo valor da mesma métrica para a classe no grupo de referência (GR). Por exemplo, para a disparidade na TFP entre uma classe s e o grupo de referência GR segundo uma característica c, temos $disp_{TFP,c} = TFP_s/TFP_{GR}$.

4. DADOS USADOS

Os dados utilizados no trabalho são compostos por características de empresas ou contratos e uma variável resposta que determina se houve prestação adequada do serviço prestado ou bem fornecido a um ente público. Nossos experimentos utilizam três bases de dados que descrevem essas entidades em diferentes contextos: empresas em nível federal, empresas em nível municipal e contratos em nível municipal. Os dados usados nos experimentos foram criados a partir de dados utilizados por órgãos de controle externo estaduais e federais fornecidos aos autores e que descrevemos a seguir. Assim como em outros trabalhos da área [Sun and Sales 2018; Silvio L. Domingos and Ramos 2016; Gomes et al. 2017], não descrevemos em detalhes todas as características usadas pelos órgãos de controle para fiscalizar empresas e contratos. Os órgãos entendem que o sigilo da definição das características é estratégico.

A primeira das três bases foi criada pela Controladoria Geral da União (CGU) e é baseada em dados usados para fiscalização de contratos com o Governo Federal. Essa base é resultado de um cruzamento de 7 bancos de dados diferentes que medem características de pouco mais de 10 mil empresas referentes à sua capacidade operacional, perfil de participação em licitações, histórico de punições e descobertas, conflitos de interesses e ligações políticas. Esta base foi criada e utilizada por Sales et. al [Sun and Sales 2018] para avaliar o uso de redes neurais na estimativa de risco de empresas que têm contratos de R$1 milhão ou mais com o Governo Federal. Para cada empresa, existe um rótulo informando se a empresa tem ao menos um contrato que não foi executado de acordo com as suas especificações nos dois anos seguintes à criação dos dados. Por brevidade, nos referimos a essa base como *empresas da esfera federal*.

A segunda base de dados é proveniente do Ministério Público da Paraíba (MPPB) e complementada por nós. Esta base possui características operacionais de cerca de 40 mil empresas do estado da Paraíba e características criadas por especialistas para aferir o risco do governo firmar um contrato com uma empresa. O MPPB usa uma soma ponderada dessas características para auxiliar na análise manual de risco das empresas listadas. Em nosso experimento, o rótulo que define se uma empresa é arriscada no gabarito vem do Sistema de Tramitação de Processos e Documentos do TCE-PB (TRAMITA). Definimos que uma empresa é arriscada se ela teve um ou mais contratos rescindidos em uma data posterior à criação dos dados do MPPB. Para transformar a estimativa de risco manual, a qual possuía valores inteiros, em binária, foi utilizado o limiar de 160 pontos. Este limiar proporcionou que fossem classificadas como arriscadas cerca de 4,3% das empresas com maior estimativa de risco. Chamamos essa base de *empresas da esfera municipal*.

A última base de dados foi criada a partir de dados fornecidos pelo Tribunal de Contas do Estado da Paraíba (TCE-PB), e possui características de cerca de 13 mil contratos celebrados por entes públicos na Paraíba e informações das empresas contratadas no momento da celebração dos contratos. Como forma de rotular contratos arriscados, novamente utilizamos a base de dados TRAMITA, porém agora identificando se o contrato foi rescindido em um momento posterior à celebração do mesmo. Diferente dos casos descritos até aqui, essa base nos permite estudar modelos que estimem risco de contratos, e não de empresas – as quais podem ter múltiplos contratos em um período –, um cenário ao mesmo tempo mais esparso e de mais relevância prática para os órgãos de controle. Essa base de dados foi criada pelo grupo de pesquisa dos autores e será chamada de *contratos municipais na Paraíba*.

Todos os conjuntos de dados utilizados têm um grande desbalanceamento entre observações com

Table I. Características sensíveis dos conjuntos de dados.

Conjunto de dados	**Características sensíveis**	**Classes sensíveis**	**Grupo de referência**
Contratos municipais na Paraíba	Empresa de pequeno porte? e idade da empresa	Empresa de pequeno porte: sim; idade da empresa: jovem e nova	Empresa de pequeno porte: não; idade da empresa: consolidada
Empresas da esfera federal	Empresa de pequeno porte?, empresa do interior? e idade da empresa	Empresa de pequeno porte: sim; empresa do interior: sim; idade da empresa: jovem e nova	Empresa de pequeno porte: não; empresa do interior: não; idade da empresa: consolidada
Empresas da esfera municipal	Empresa do interior?, idade da empresa, porte da empresa e sócio relacionado ao BF?	Empresa do interior: sim; idade da empresa: jovem e nova; porte da empresa: empresa pequeno porte e microempresa; sócio relacionado ao BF: sim	Empresa do interior: não; idade da empresa: consolidada; porte da empresa: demais; sócio relacionado ao BF: não

alto e baixo risco. As bases das empresas da esfera federal, empresas da esfera municipal e contratos municipais na Paraíba possuem 92,7%, 99,6% e 98,9% das observações da classe negativa, respectivamente.

5. METODOLOGIA

Nosso estudo é composto de quatro experimentos que exploram justiça na estimativa de risco com os três conjuntos de dados que usamos. No primeiro experimento, avaliamos a justiça no uso da soma ponderada atualmente disponível para o MPPB para estimar o risco nos dados de empresas na esfera municipal, a qual chamamos de abordagem manual. Nos três experimentos seguintes, utilizamos diferentes modelos de aprendizagem de máquina para estimar risco em empresas da esfera federal, empresas da esfera municipal e contratos da esfera municipal, e avaliamos justiça no resultado das classificações.

As métricas mais relevantes para avaliar justiça no contexto de estimativa de risco em contratos públicos são disparidade na TFP, revocação e precisão. A avaliação da disparidade entre classes foi medida através do *toolkit Aequitas* [Saleiro et al. 2018].

As características definidas como sensíveis foram escolhidas como aquelas que definem empresas mais e menos vulneráveis a efeitos colaterais de investigações e a sanções: porte da empresa, idade da empresa, se a empresa está sediada em cidade de interior e se a empresa tem ou teve um sócio integrante ou ex-integrante do programa bolsa família (BF). Algumas dessas características não estão disponíveis em todas as bases de dados, e a definição das classes sensíveis também é contingente do formato dos dados. O conjunto de características e classes sensíveis e de referências das diferentes bases é detalhado na Tabela I. Em todas as bases a idade da empresa foi limiarizada em nova (menos de 3 anos), jovem (3 a 10 anos) e consolidada (10 anos ou mais).

Para definir os modelos de aprendizagem de máquina a serem utilizados, primeiro testamos seis classes de modelos em cada base de dados, e utilizamos o modelo que teve maior eficácia em cada uma das bases para o estudo da justiça. Para o treinamento dos modelos, nosso processo é similar ao de Sun e Sales [Sun and Sales 2018]: utilizamos validação cruzada com 5 *folds*, balanceando os dados de treino, e particionamos o conjunto em treino e teste em 80% e 20%. Treinamos modelos de diferentes famílias: florestas aleatórias, regressões logísticas, redes neurais, máquinas de vetor de suporte, redes neurais e *k-nearest neighbors*. Os melhores hiperparâmetros para cada modelo foram buscados considerando o *f1-score* [1], tendo em vista o desbalanceamento das classes de risco. Por fim, dado esse mesmo desbalanceamento, treinamos os modelos sem balanceamento, com *undersampling* da classe negativa e com *oversampling* aleatório, e *oversampling* através da técnica SMOTE para a classe positiva. Para todos os conjuntos de dados, os modelos de maior eficácia segundo o *f1-score* são florestas aleatórias com oversampling através da técnica SMOTE. Por conseguinte, os resultados que

[1] $f_1 score = 2 \cdot \frac{precisao \cdot revocacao}{precisao + revocacao}$

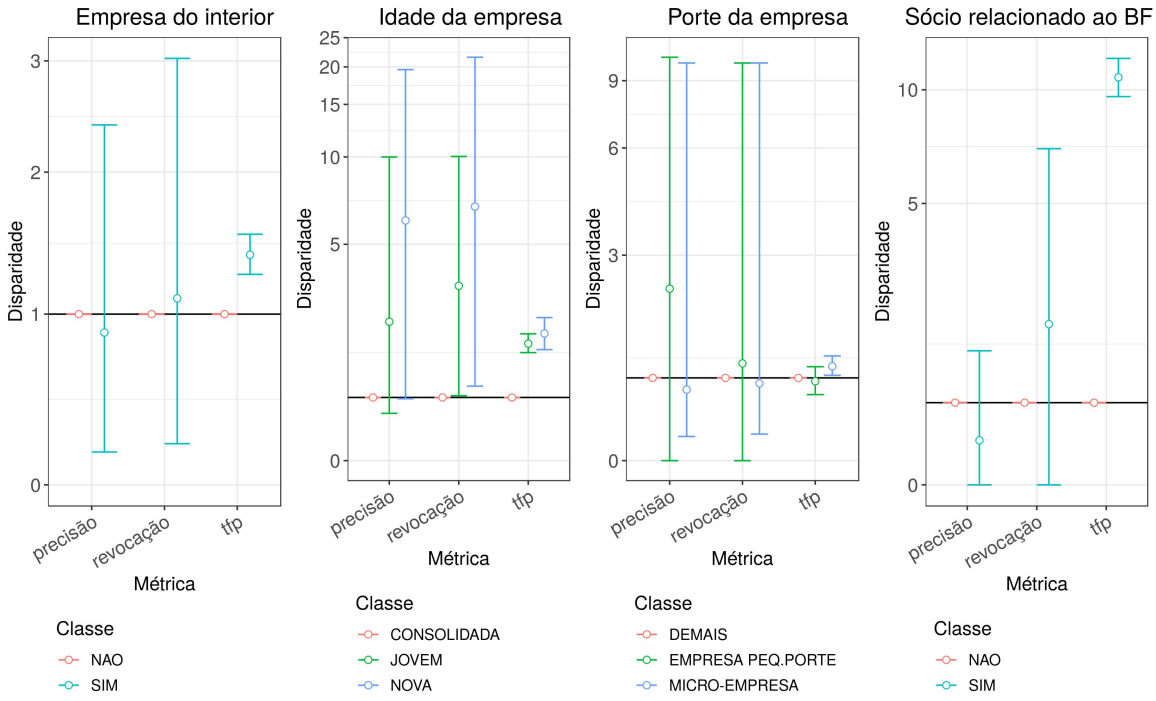

Fig. 1. Disparidade das métricas das empresas da esfera municipal (abordagem manual).

reportamos usam modelos criados segundo esse método.

Como os conjunto de dados analisados são uma amostra das estimativas de interesse, reportamos nossos resultados estimando intervalos de confiança para as métricas de interesse na população das empresas ou contratos. Todos os intervalos são estimados com 95% de confiança através de 10.000 bootstraps. Dado o desbalanceamento das classes sensíveis e da classe positiva na variável de resposta, aplicamos bootstrap estratificado em função da conjunção das características sensíveis com a variável de resposta. Por fim, ao comentar os resultados, optamos por interpretar os intervalos estimados, em lugar de dicotomizar resultado em significativos ou não [Cumming and Calin-Jageman 2016].

6. JUSTIÇA NO ESTADO DA PRÁTICA

A Figura 1 mostra as disparidades na estimativa de risco manual das empresas da esfera municipal. Há uma grande diferença na TFP comparando as empresas que têm ou tiveram sócios vinculados ao bolsa família com as que não tiveram sócios com vínculo; a TFP naquelas foi entre 9,6 e 12 vezes maior que nestas. Além disso, empresas novas e jovens têm TFP entre 1,9 e 2,6, e entre 1,8 e 2,2 vezes maior comparadas às empresas consolidadas, respectivamente. A revocação também foi maior para empresas novas e jovens em relação à empresas consolidadas, mas é difícil quantificar o quão maior, visto que estimamos que é plausível que a disparidade esteja entre 1,2 e 21,5 para as novas e entre 1,03 e 10 para as jovens. A disparidade na precisão das empresas jovens foi entre 0,74 e 10 vezes maior e as novas foram entre 0,98 e 19,6 vezes maior comparada com as consolidadas. Ou seja, a precisão para as empresas jovens e novas pode ser desprezivelmente menor, igual ou muito maior que a da consolidadas. A TFP também é maior para empresas em cidades do interior entre 1,3 e 1,5 vezes, uma disparidade relativamente pequena. Juntos, os resultados apontam que as empresas que tiveram sócios vinculados ao bolsa família, assim como empresas novas e jovens, são mais classificadas como arriscadas erroneamente.

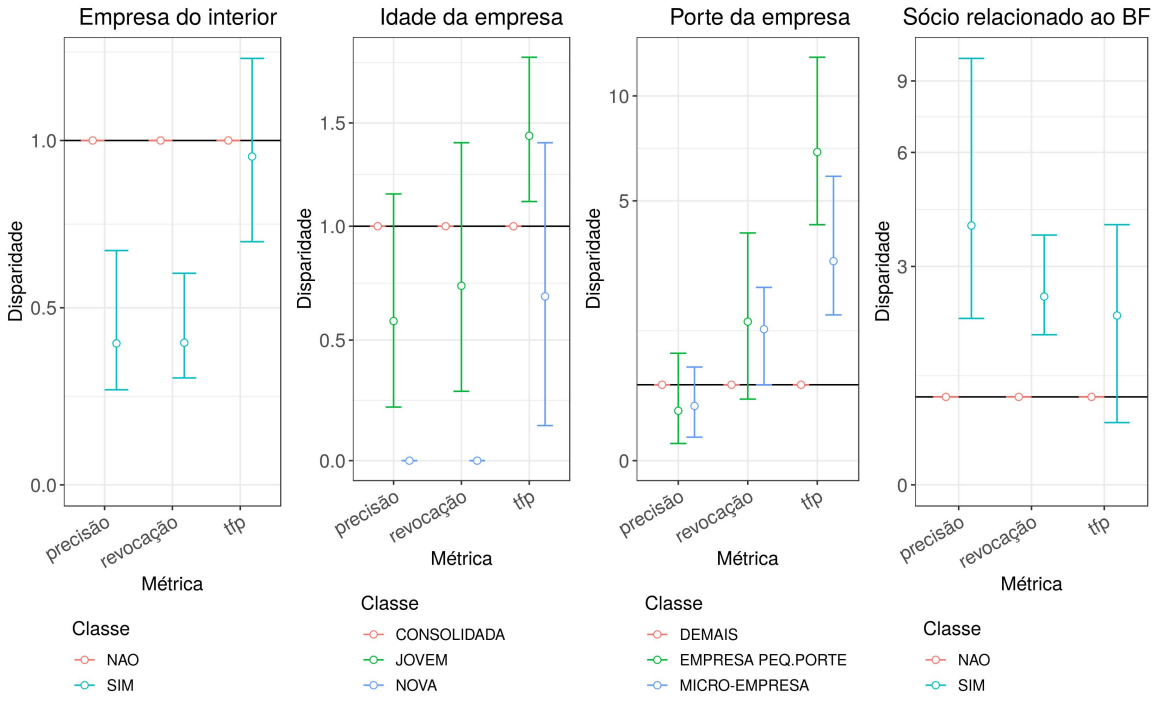

Fig. 2. Disparidade das métricas das empresas da esfera municipal (abordagem AM).

Ao analisar a estimativa de risco com aprendizagem de máquina das empresas da esfera municipal (Figura 2), percebemos que a TFP é maior para empresas de pequeno porte e microempresas em relação às demais. Estimamos que comparadas com grandes empresas, empresas de pequeno porte e microempresas são classificadas erroneamente como arriscadas mais que 4,2 e 2,1 vezes mais, respectivamente. Sobre a idade da empresa, é possível notar que empresas jovens têm maior TFP que empresas consolidadas, entre 1,1 e 1,9 vezes. Em contrapartida, a precisão de empresas jovens se for maior que empresas consolidadas é uma diferença pequena (até 1,2 vezes), mas pode ser muito menor (0,2 vezes).

No caso do modelo de aprendizagem de máquina para empresas da esfera federal, a Figura 3 mostra que ele gera uma disparidade no TFP, precisão e revocação das empresas de pequeno porte. Os valor são semelhantes: [1,5;2,7] [1,4 ;2,8] e [1,2;2,5] vezes, respectivamente. Empresas jovens tiveram tanto TFP quanto precisão maiores em comparação a empresas consolidadas, entre 1,09 e 2,1 vezes e 1,2 e 2 vezes respectivamente, enquanto sua revocação ou foi desprezivelmente menor (0,96 vezes) ou foram consideravelmente maiores (1,7 vezes). Ao contrário do esperado, a TFP foi menor para as empresas do interior, assim como sua revocação, entre 0,46 e 0,88 vezes e 0,45 e 0,94 vezes as taxas de empresas de capitais do Brasil respectivamente.

Por último, como pode ser visto na Figura 4, o modelo de aprendizagem de máquina baseado na base de dados de contratos municipais na Paraíba apontou uma maior TFP para empresas jovens em relação à empresas consolidadas, entre 2,3 e 4,2 vezes.

7. DISCUSSÃO

Nossa análise aponta que existe uma diferença relevante da TFP em pelo menos uma classe sensível para todos modelos analisados. Além disso, em todos os modelos analisados, empresas jovens, com idade entre 3 e 10 anos, tiveram disparidade na TFP comparado a empresas consolidadas. Essas

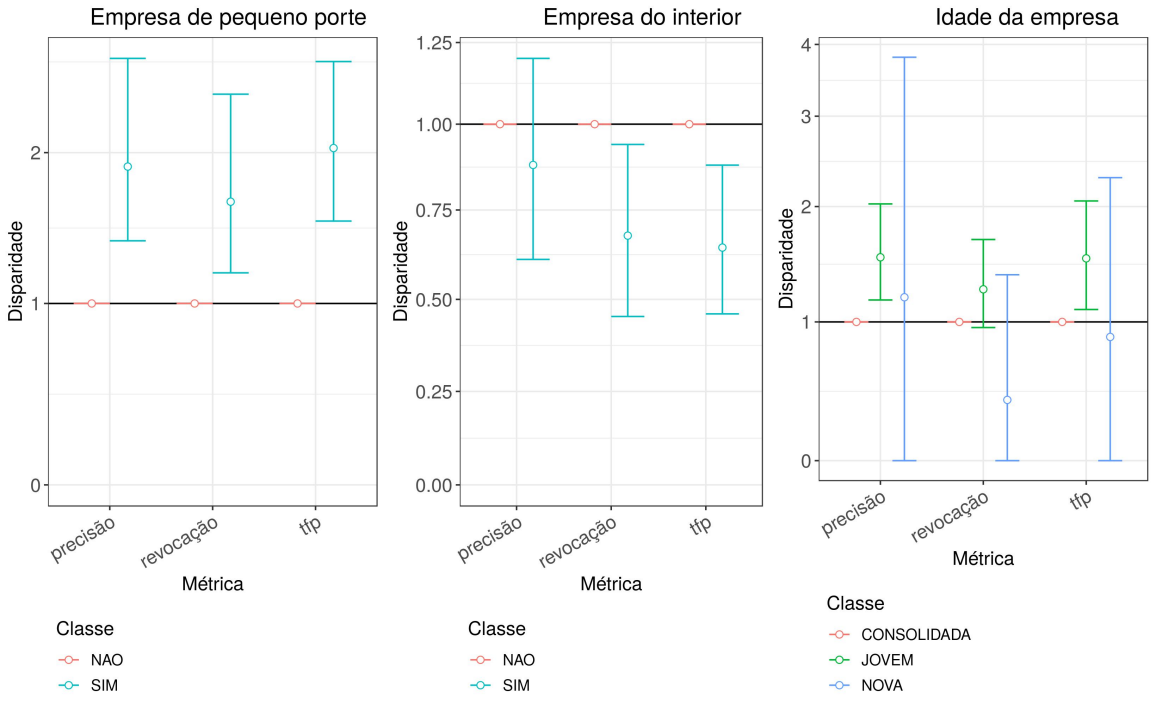

Fig. 3. Disparidade das métricas das empresas da esfera federal.

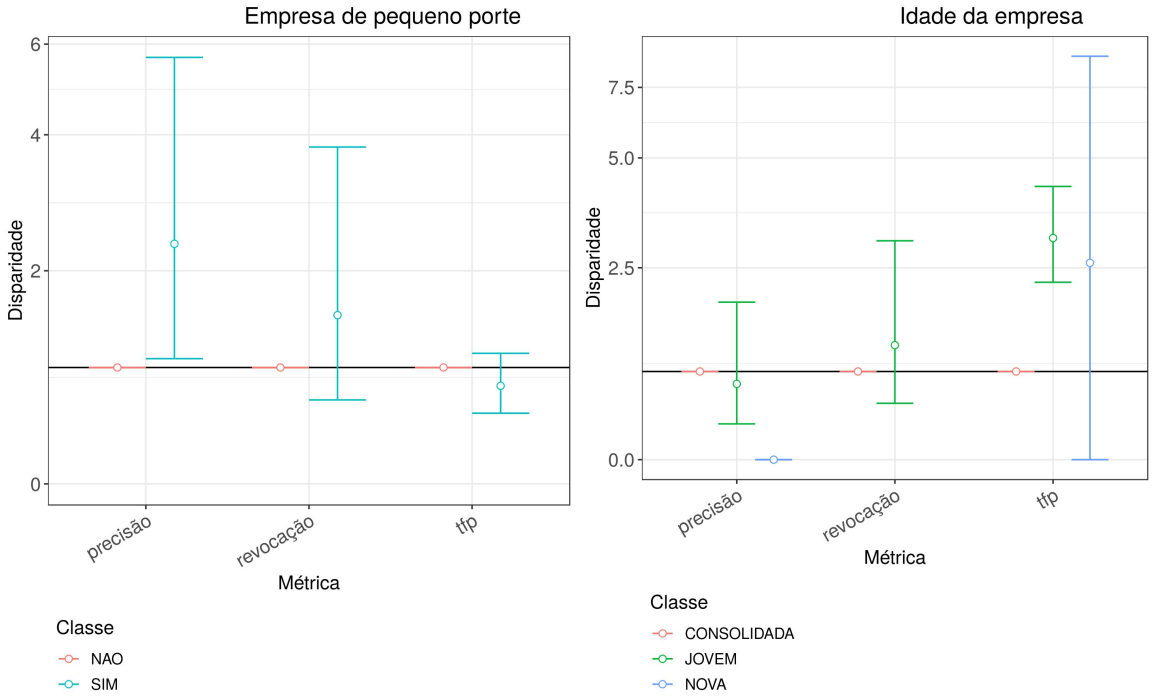

Fig. 4. Disparidade das métricas dos contratos da municipais na Paraíba.

diferenças apontam que empresas de classes sensíveis são estimadas como mais arriscadas de forma injusta em comparação ao grupo de referência.

Ao observar a abordagem manual, fica claro que existe uma grande disparidade com empresas com sócios relacionados ao bolsa família. Muitas dessas são pequenas ou microempresas em que os sócios têm ou tiveram baixa renda. Apesar das dificuldades financeiras que eles passaram, os métodos desconfiam sobremaneira desses sócios, estimando injustamente alto risco.

A disparidade existe tanto na abordagem manual quanto na abordagem de AM, observando a disparidade das empresas da esfera municipal. Existe tanto no âmbito federal quanto municipal, e tanto no nível da empresa quanto do contrato. Em particular, o fato de que modelos treinados tanto com características criadas por especialistas em controle quanto por pesquisadores de AM levaram a vieses semelhantes é marcante. Isso aponta a relevância de considerarmos injustiça na estimativa de risco para controle de contratos públicos. Discriminar empresas jovens, por exemplo, pode desencorajar a participação de startups em licitações, bem como empresas pequenas mais tradicionais.

Ao observar a revocação e precisão juntamente com a TFP, é possível perceber que existem casos em que a TFP é maior para a classe sensível ao mesmo tempo que a precisão ou a revocação é maior. Em outros casos, a diferença na TFP não é seguida de uma diferença da precisão ou revocação. Quando a disparidade não está associada a uma maior precisão ou revocação, temos uma situação de injustiça sem que o modelo tenha mais eficácia classificando a classe sensível. Na situação onde a injustiça está associada com uma maior eficácia, existe um dilema moral: o modelo classifica com maior eficácia a classe na qual ele também produz mais falsos positivos. Esse dilema, por sua vez, aumenta o risco de que um modelo injusto seja posto em prática visando a sua eficácia.

Trabalhos futuros podem investigar métodos para mitigar a injustiça no uso de AM com as variáveis sensíveis que identificamos. Novos experimentos com dados de outros estados também são necessários para avaliar quão generalizáveis são os nossos resultados. Experimentos com bases maiores, incluindo todo o Brasil também melhorariam a precisão dos intervalos de confiança que estimamos.

REFERENCES

Lei 8666. In *Constituição Federal*, 1993.

Cumming, G. and Calin-Jageman, R. *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. Routledge, New York, NY, 10001, 2016.

Gomes, T. A., Carvalho, R. N., and Carvalho, R. S. Identifying anomalies in parliamentary expenditures of brazilian chamber of deputies with deep autoencoders. In *IEEE ICMLA*, 2017.

Julia Angwin, Jeff Larson, S. M. and Kirchner, L. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks., 2016.

Saleiro, P., Benedict Kuester, A. S., Anisfeld, A., Hinkson, L., London, J., and Ghani, R. Aequitas: A bias and fairness audit toolkit. In *eprint arXiv:1811.05577*, 2018.

Saul M. Kassin, I. E. and Kukucka, J. The forensic confirmation bias: Problems, perspectives, and proposed solutions. In *Journal of Applied Research in Memory and Cognition*. Vol. 2. pp. 42–52, 2013.

Silvio L. Domingos, Rommel N. Carvalho, R. S. C. and Ramos, G. N. Identifying it purchases anomalies in the brazilian government procurement system using deep learning. In *15th IEEE ICMLA*. pp. 722–727, 2016.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. A unified approach to quantifying algorithmic unfairness: Measuring individual group unfairness via inequality indices. In *ACM KDD '18*, 2018.

Sun, T. and Sales, L. J. Predicting public procurement irregularity: An application of neural networks. In *Journal of Emerging Technologies in Accounting: Spring 2018*. Vol. 15. pp. 141–154, 2018.

Zan Huang, Hsinchun Chen, C.-J. H. W.-H. C. S. W. Credit rating analysis with support vector machines and neural networks: a market comparative study. In *Decision Support System*. Vol. 37. pp. 543–558, 2004.