Achieving Fairness in Predictive Process Analytics via Adversarial Learning

Massimiliano de Leoni[1] and Alessandro Padella[1]

University of Padua, Italy
deleoni@math.unipd.it, alessandro.padella@phd.unipd.it

Abstract. Predictive business process analytics has become important for organizations, offering real-time operational support for their processes. However, these algorithms often perform unfair predictions because they are based on biased variables (e.g., gender or nationality), namely variables embodying discrimination. This paper addresses the challenge of integrating a debiasing phase into predictive business process analytics to ensure that predictions are not influenced by biased variables. Our framework leverages on adversial debiasing is evaluated on four use cases, showing a significant reduction in the contribution of biased variables to the predicted value. The proposed technique is also compared with the state of the art in fairness in process mining, illustrating that our framework allows for a more enhanced level of fairness, while retaining a better prediction quality.

Keywords: Process Mining · Deep Learning · Predictive Process Analytics · Adversarial Debiasing · Fairness

1 Introduction

Predictive process analytics aims to forecast the outcome of running process instances to identify those requiring specific attention, such as instances risking delays, excessive costs, or unsatisfactory outcomes. By predicting process behavior and outcomes, predictive process analytics enables timely intervention and informed decision-making.

Predictive process analytics naturally needs to rely onto the characteristics of the process being monitored, and performs predictions on their basis. Being that said, this analytics become a problem when predictions are unfair because they are based on characteristics that discriminate in a form that is unacceptable from a legal and/or ethical point of view. For instance, in a loan-application process at a financial institute, one cannot build on the applicant's gender to predict the outcome, namely whether or not the loan is granted. Pohl et al. indicate monitoring, detecting and rectifying biased patterns to be the most significant challenge in Discrimination-Aware Process Mining [5].

Process characteristics are hereafter modelled as process variables. In accordance with the literature terminology [7], we use the term *protected variable* to indicate the variables on which prediction cannot be based. The choice of the set of variables to protect depends on the specific process, and thus needs to be made by the process analysts/stakeholders. Note how simply removing the protected variables from the datasets would not be effective, because the bias would be simply "hidden under the carpet", as it would be possibly just transferred to other variables that are strongly correlated.

While several researchers acknowledge the importance of ensuring fairness in process predictive analytics, very little research has been carried out on this topic (cf. discussion in Section 2 of the extended version in [2]). This paper proposes a framework based on *adversarial debiasing*, which aims to mitigate bias related to protected variables within the predictive models. In a nutshell, the proposed framework is based on the idea of training the model to predict the process' outcome values, constraining accurately predicting the protected variables and reducing bias in its learned representations.

Compared with the current literature in fairness for process' predictive analytics, adversial debiasing aims at more accurate predictions through prediction models that also guarantee higher fairness. However, existing research on adversial debiasing has not focused on process predictive analytics and, more generally, to time series, and cannot be trivially applied in this setting (cf. Section 2 of [2]).

Experiments have been conducted on four use cases to forecast the process-instance total time and whether or not certain activities are going to occur. Protected variables accounted for resources, organization countries, gender, citizenship and spoken languages. The results show that our framework ensures fairness with respect to the chosen protected variables, while the accuracy of the predictive models remains high, also in comparison with the results for comparable research works in literature. Experimental results also highlight that the influence is also reduced for those process variables that are strongly correlated with the protected variables, illustrating that removing the protected variables would just transfer the unfairness to the correlated variables.

2 Preliminaries

The starting point for a prediction system is an *event log*. An event log is a multiset of *traces*. Each trace describes the life-cycle of a particular *process instance* (i.e., a *case*), which is composed by a sequence of *events*, each referring to the execution of a certain activity by a resource at a given timestamps. Additional attributes can be associated to events: the activity cost, outcome, relevant information, etc.

Predictions aims to forecast the outcome value of a running trace, hereafter modelled as an outcome function $\mathcal{K} : \mathcal{E}^* \rightarrow \mathcal{O}$, with $\mathcal{O}$ be the set of potential outcome values. Outcome function $\mathcal{K}(\sigma)$ returns the process-instance outcome observed after observing the sequence σ of its events. Predictive analytics aims to build a **process prediction oracle** $\Psi_{\mathcal{K}} : \mathcal{E}^* \rightarrow \mathcal{O}$ such that, given a running trace σ' eventually completing in σ_T, $\Psi_{\mathcal{K}}(\sigma')$ is a good predictor of $\mathcal{K}(\sigma_T)$.

The literature proposes several Machine- and Deep-Learning techniques, highlighting LSTM's quality (cf. Section 2 of [3]). We instead opted for fully connected neural networks (FCNNs) [1], which are faster to train than LSTM networks but provide similar accuracy results (see our comparison reported in Section 5.5 of [2]). Also known as Feed-Forward Neural Networks, FCNNs are characterized by having every node in one layer connected to every node in the next layer, meaning that every node in one layer receives input from every node in the previous layer.

The training of FCNN models falls into the problem of supervised learning, which aims to estimate a Machine-Learning (ML) function $\Phi : X_1 \times \ldots \times X_n \rightarrow \mathcal{Y}$ where $\mathcal{Y}$ is the domain of variable to predict (a.k.a. dependent variable), and $X_1 \ldots X_n$ are the domains of some independent variables $V_1, \ldots, V_n$, respectively.

To tackle the prediction problem for an outcome function, $\mathcal{Y} = \mathcal{O}$. The values of the independent variables are obtained from the event-log traces: each trace is encoded into a vector element of $X_1 \times \ldots \times X_n$, through a **trace-to-instance encoding function** $\rho_{\mathcal{L}} : \mathcal{E}^* \rightarrow X_1 \times \ldots \times X_n$. Note that the process prediction oracle is thus implemented as $\Psi_{\mathcal{K}}(\sigma) = \Phi(\rho_{\mathcal{L}}(\sigma))$. Section 2 of the extended version of this paper [2] provides further details on how these functions are trained from event logs.

3 An Adversarial Debiasing Framework for Predictive Process Analytics

The overall objective of this paper is to build a process prediction function $\Psi_{\mathcal{K}}$ whose output values are not influenced by the chosen **protected variables**.

The determination of the protected variable depends on the specific use case under consideration (e.g., the gender or nationality of a loan applicant). It is crucial to note that certain variables may be designated as protected in one use case but not in another (e.g., the variable "Gender" might be designated as a protected variable in the context of a loan application process, but it may not hold the same status in the process of hospital discharge). By carefully selecting the protected variables, we aim to ensure that the predictions do not enforce a discrimination that is not ethically and/or morally acceptable.

The framework is visually depicted in Figure 1 where the core component is the prediction model that implements the oracle function $\Psi_{\mathcal{K}}$, capable to of forecasting the outcome of a running trace. Leveraging on neural networks, $\Psi_{\mathcal{K}}$ is obtained through the composition of the trace-to-instance encoding function ρ_L and an ML function $\Phi : X_1 \times \ldots \times X_n \rightarrow \mathcal{O}$, namely for any trace σ, $\Psi_{\mathcal{K}}(\sigma) = \Phi(\rho_L(\sigma))$. The most left gray box in Figure 1 is the encoder ρ_L, which converts the trace into a vector. The second gray box from left depicts the FCNN that implements Φ, along with the decoder represented through the red dot.

Looking from the right in Figure 1, the first gray box depicts the adversarial FCNN, which tackle the debiasing problem to ensure fairness. In particular, let $\overline{V} = \{\overline{V_1}, \ldots, \overline{V_p}\} \subseteq \{V_1, \ldots, V_n\}$ be the set of the protected variables, which are defined over the domains $Z = \overline{X_1}, \ldots, \overline{X_p}$, respectively. Let $N_1, \ldots, N_q$ are the domains of the output of the q nodes that constitute the last layer of the FCNN implementing Φ. The adversarial FCNN implements a function $\Phi_Z : N_1 \times N_n \rightarrow Z$, which aims to predict the values of the protected variables, using the output of the last layer as input.

In accordance with the literature on adversarial debiasing [7], if the neural network that implements Φ - in our case a FCNN - does not build the prediction on the protected variables, then the adversarial network that implements Φ_Z - in our case another FCNN - is unable to predict the protected-variables values from the output of the network implementing Φ.

More formally, let $\hat{y} = \Phi(\vec{x})$ be the predicted value for the running trace σ' that has been encoded $\vec{x} = \rho_L(\sigma')$. Let σ be the real completion of σ' (i.e. σ' is a prefix of σ), with the real outcome $y = \mathcal{K}(\sigma)$. Let $\vec{z} = \Phi_Z(\vec{n})$ be the vector of the values predicted for the protected variables, on the basis of the vector $\vec{n}$ of the output of the last layer of the neural network that implements Φ. The two neural networks are trained so as to minimize the overall loss function: $L_{\overline{V}}(\hat{y}, y, \vec{x}, \vec{z}) = \Delta(\hat{y}, y) - \Delta(\vec{z}, \pi_{\overline{V}}(\vec{x}))$. Symbol Δ indicates the normalized difference between two vectors (or two values), and $\pi_{\overline{V}}(\vec{x})$

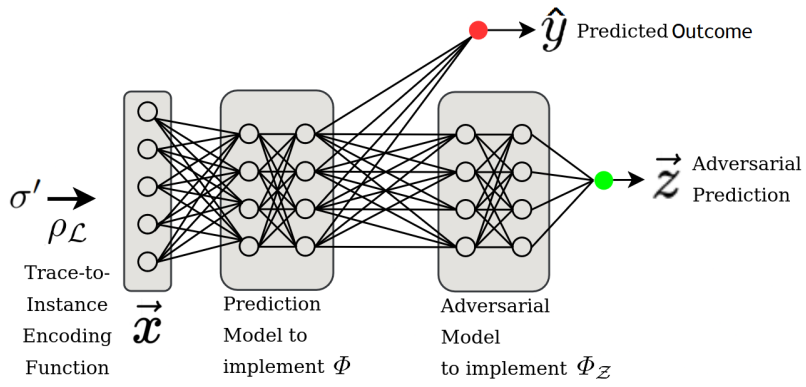

Fig. 1: Overview of our debiasing framework for process' predictive analytics.

is the projection of $\vec{x}$ over $\overline{V}$, namely retaining the dimensions of $\vec{x}$ for the protected variables. The normalization in $\Delta(\hat{y}, y)$ is performed by dividing by the largest outcome value $y = \mathcal{K}(\sigma)$ for all traces σ in the training event log. The normalization in $\Delta(\vec{z}, \pi_{\overline{V}}(\vec{x}))$ is achieved by dividing by the largest vector $\pi_{\overline{V}}(\rho_L(\sigma))$ for all traces σ in the training event log. Minimizing loss function$L_{\overline{V}}$ implies that prediction accuracy is kept reasonably high while the influence of protected variables is minimized.

The whole framework has been implemented through the training of two FCNNs on a stochastic-gradient-descent based algorithm. The inplementation is in Python and available at `https://anonymous.4open.science/r/Fairness-D70B`, leveraging on the *PyTorch* package for FCNN's training and *fairlearn* for other debiasing utilities.

4 Evaluation

The evaluation focuses on evaluating how our framework mitigates the influence of protected variables while still ensuring a good quality. The framework evaluation was carried out by training two FCNNs that implement functions Φ and Φ_Z. In particular, we carried out a grid search to tune the hyper-parameters related to the learning rate, layers shape, epochs, and weight decay, so as to prevent over- and under-fitting problems.

Our debiasing framework was evaluated on four use cases, aiming to assess *(i)* the mitigated influence of the protected variables on the prediction, and *(ii)* the extent of the reduction of the prediction accuracy when our framework was employed. Note that a reduction in accuracy is expected when addressing the fairness problem: if the protected variables have some good predictive power, their exclusion has a natural negative impact on the ML-model accuracy. The baseline of comparison is with the only existing framework by Qafari et al. [6].

4.1 Introduction to Use Cases

Our technique was assessed through three process for which we have identified four use cases. The first and the second use case are from Volvo Belgium and refer to a process

that focuses on an incident and problem management system called VINST.[1] In the first use case, our aim is to predict the **total time** of an execution that is running, while in the second our aim is to predict **whether or not the activity *Awaiting Assignment* will occur** in the future; it is

The third use case refers to the *Hiring* process provided by Pohl et al. in [4]. For this use case we aim to predict the **total time** a running execution.

The last use case is based on the *Hospital* process discussed by Pohl et al. [4]. For this use case our aim is to predict **whether or not the activity *Treatment unsuccessful* will occur** in the future.

For each use case, the available process log has been temporarily split into 70% of the traces that were used for training the prediction and adversarial models and 30% for testing. Protected variables have to be clearly different between the different use cases, since their choice depend on process and is also related to specific fairness-preserving considerations. The different protected variables are summarized in the column 3 of the Table 2 the choices for the four use cases.

4.2 Evaluation Metrics

The evaluation's goal is twofold: it aims to assess the mitigation influence of the protected variables on the prediction and the reduction extent of the prediction accuracy.

For the first and third use cases in which we aim to predict the total time of running traces, i.e. a regression problem, the results are provided in terms of **Absolute Percentage Accuracy** (APA) , which is defined as 100% minus Mean Absolute Percentage Error, between the actual value and the predicted one. For the second and fourth use cases, we aim to test the accuracy prediction on the occurrence for the activities *Awaiting Assignment* and *Treatment unsuccessful*, respectively. This is a classification problem: hence, we choose **F-score** for assessing the accuracy of our predictions.

To assess the reduction in the influence of protected variables, we employ the theory of Shapley values[2], computing them both when our framework is employed and when it is not: our framework is expected to reduce the absolute Shapley value, which corresponds to a lower influence. For classification problems, we also assess an enhanced fairness through the analysis of the false positive rate (FPR) and true positive rate (TPR), and the verification of the **Equalized Odds** criterion [7]: this criterion states that, if we group the samples in the test set by the values of the protected variables, the FPR and TPR should be somewhat similar in all groups. The rationale behind this criterion is that, splitting the test-set samples based on the values of the protected variables, one obtains groups that are statistically equated, including for false and true positive rates, if the model's prediction are not based on the protected variables.

4.3 Evaluation Results

Table 1 illustrates the results in terms of accuracy for the processes, logs and predicted outcomes introduced in Section 4.1. The results are based on a test set that is are constructed as discussed in Section 4.1, and they refer to the work proposed in this paper,

[1] https://data.4tu.nl/articles/dataset/BPI_Challenge_2013_incidents/12693914

[2] More details on our use of Shapley values are given in Section 3.2 of this paper's extended version [2].

Process	Outcome	Methodology	Without	With	Δ
VINST	Total Time	Qafari et al. [6]	69%	60%	9%
		Our Framework	78%	74%	4%
VINST	Occurrence of *Awaiting Assignment*	Qafari et al. [6]	0.71	0.59	0.12
		Our Framework	0.80	0.72	0.08
Hiring	Total Time	Qafari et al. [6]	79.9%	70.02%	9.88%
		Our Framework	83.6%	81.1%	2.5%
Hospital	Occurrence of *Treatment Unsuccessful*	Qafari et al. [6]	0.69	0.58	0.11
		Our Framework	0.78	0.76	0.02

Table 1: Results achieved by our framework and by Qafari et al. [6], in terms of accuracy.

Process	Outcome	Protected Variable	Without	With	Ratio
VINST	Total Time	Resource country	112h	9h	8%
VINST	Occurrence of *Awaiting Assignment*	Organization country	1.8	0.03	1%
Hiring	Total Time	Gender	-463min	-156min	20%
		Religious	-447min	-12min	3%
Hospital	Occurrence of *Treatment Unsuccessful*	Citizen	0.25	0.04	16%
		german_speaking	0.17	0.06	35%

Table 2: Differences in Shapley Values of protected variables with and without the debiasing framework, for the four use cases.

which is then compared with the results that Qafari et al. [6] can achieve, which is considered as baseline. Columns *without* and *within* report on the results when the corresponding techniques doesn't or does aim at achieving fairness, respectively. Column Δ highlights the reduction of accuracy when the techniques aims at fairness. *Our framework consistently obtains higher accuracy for all use cases, if compared with Qafari et al. [6], and also the accuracy reduction is significantly more limited.*

The assessment the effectiveness of our fairness framework to reduce the influence of the protected variables, we computed the Shapley values of the protected variables for the four use cases, both when we employed our framework and when we simply used the FCNN predictor that implements Φ (namely excluding the adversial FCNN for Φ_Z). The results are reported in Table 2. In the use case related the VINST process for predicting the Total-Time outcome, the protected variable *Resource country* is characterized by a Shapley value of 112 hours without using the debiasing framework, and 9 hours using the framework: the use of our framework brought the Shapley value down to 8% of the value without using our framework, which is a remarkable result, given that the Shapley values are directly correlated with the feature importance in the prediction. For the same process, when the outcome was whether or not activity *Awaiting Assignment Occurrence* is predicted to eventually occur, the protected variable *Organization country* was characterized by a Shapley value that dropped from 1.8 to 0.03, when the debiasing framework was employed: the Shapley value has become 1% of the value without debiasing. Similar results can be observed in Table 2 for the other use

		Poland		**Sweden**		**India**		**Brazil**		**Usa**		**Std**	
		Without	With	Without	With	Without	With	Without	With	Without	With	Without	With
Qafari et al. [6]	FPR	0.20	0.18	0.13	0.24	0,11	0,12	0,17	0,17	0,32	0,41	0.143	0.086
	TPR	0.91	0.85	0.78	0.89	0.79	0.89	0.98	0.81	0.89	0.83	0.0641	0.0451
Our framework	FPR	0.04	0.08	0.11	0.09	0,14	0,08	0,02	0,06	0,01	0,06	0.153	0.018
	TPR	0.67	0.61	0.72	0.63	0.62	0.63	0.59	0.65	0.59	0.65	0.052	0.024

(a) VINST use case when aiming to predict the eventual occurrence of *Awaiting Assignment*.

		Citizen						**german_speaking**					
		True		**False**		**Std**		**True**		**False**		**Std**	
		without	with	without	with	without	with	without	with	without	with	without	with
Qafari et al. [6]	FPR	0.30	0.35	0.36	0.41	0.03	0.03	0.31	0.35	0.38	0.40	0.035	0.025
	TPR	0.71	0.67	0.62	0.51	0.05	0.08	0.71	0.67	0.62	0.51	0.09	0.16
Our framework	FPR	0.28	0.26	0.22	0.21	0.03	0.025	0.3	0.24	0.22	0.21	0.04	0.015
	TPR	0.82	0.76	0.77	0.74	0.025	0.01	0.82	0.76	0.77	0.74	0.025	0.01

(b) Hospital-process use case, when aiming to predict the eventual occurrence of *Treatment Unsuccessful*

Table 3: False Positive Rate (FPR) and True Positive Rate (TPR) achieved by the debiasing framework proposed here and by the framework by Qafari et al for two use cases. The standard deviation of the FPR and TPR among groups is also shown.

cases, yielding the conclusion that *observing the significant drop of the Shapley values of the protected value after applying our debiasing framework, the framework is extremely effective to reduce the influence of the protected variables and, thus, enhance the prediction fairness*.

Space limitation does not allow us to show the whole list of Shapley values for the use cases, which are however available in the extended version [2]. If, e.g., we analyze Shapley values for the VINST use case (cf. Figure 3 in [2]), we can see that, indeed, the Shapley value for the protect variable *resource country* has significantly dropped. One could also observe that the Shapley value for variable *organization country* is also significantly reduced, likely because it is correlated with the protected variable. *If we had simply removed the protected variable, the correlated variable organization country would have gained strong influence onto the predictions: the bias would have simply moved from one sensitive variable to another, leaving the prediction model unfair. Conversely, our debiasing framework can also reduce the influence of the unfair variables that are strongly correlated to the one that has explicitly been stated as protected.*

We complete the section by reporting the results with respect to the criterion of Equalized Odds (cf. Section 4.2), which can only be apply to use cases where the set of outcome's values is finite, here namely for the use cases related to VINST process and the Hospital using the activity-occurrence process' outcome.

For the VINST use case related to the occurrence of activity *Awaiting Assignment*, we considered the groups related to top five organization countries, which cover 89% of the instances in the test set (recall that the protected variable is *organization country*): Sweden, Poland, India, Brazil and USA. False positive and negative rates are reported in Table 3a, without and with using the framework, both for our framework and for that of Qafari et al. [6], for all five groups. The last two columns with header *Std* summarizes the standard deviation for FPR and TPR: in case of perfectly meeting the **Equalized Odds** criterion, there would be no difference among the groups, and thus the standard deviation would be zero. For our framework, the introduction of the debiasing phase, the

FPR's standard deviation within the five groups is characterized by a 88% drop, moving from 0.153 to 0.018, whereas the TPR's standard deviation shows a 53% drop (from 0.052 to 0.024). *Using the fairness approach by Qafari et al. [6], the FPR's and TPR's standard deviation* within the five groups show a drop of 53% and 29%, which *is nearly half the drop that our debiasing framework achieves*. We conducted the same analysis for the hospital use case, which is reported in Table 3b. FPRs and TPRs are computed for both protected variables. Also for this use case, our debiasing framework guarantees lower FPR's and TPR's standard deviations for both variables, although the reduction is more limited than what achieved for the VINST use case. However, The framework by Qafari et al. [6] does not reduce the FPR's and TPR's standard deviations for any of the two variables, expect for the FPR for variable *german_speaking*. As a matter of fact, their framework increases the TPR's standard deviation for both of variables, certainly going against the criterion of Equalized Odds.

5 Conclusion

Considerable research efforts have been directed towards predictive process analytics. Literature has shown that the fairness problem has generally been overlooked in predictive process analytics (cf. Section 2 of extended version in [2]). This means that predictions may potentially be discriminatory, unethical, and, e.g., targeting certain ethnics, nationalities and religions. This paper proposes a predictive framework that specializes those based on adversarial debiasing so as to allow sequences (i.e., traces) as input.

Experiments were carried out on three processes and four use cases, and the results show that our debasing framework minimizes the influence of the protected variables onto the prediction. At the same time, we illustrates that the reduction of the prediction quality is limited and lower than what is achieved by an existing framework for fairness-preserving process predictive analytics by Qafari et al. [6].

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
2. de Leoni, M., Padella, A.: Achieving Fairness in Predictive Process Analytics via Adversarial Learning (Extended Version) (2024), `https://arxiv.org/abs/2410.02618`
3. Márquez-Chamorro, A.E., Nepomuceno-Chamorro, I.A., Resinas, M., Ruiz-Cortés, A.: Updating prediction models for predictive process monitoring. In: Advanced Information Systems Engineering. pp. 304–318. Springer International Publishing, Cham (2022)
4. Pohl, T., Berti, A., Qafari, M.S., van der Aalst, W.M.P.: A collection of simulated event logs for fairness assessment in process mining. In: Proceedings of the Best Dissertation Award, Doctoral Consortium, and Demonstration & Resources Forum at BPM 2023. vol. 3469, pp. 87–91. CEUR-WS.org (2023)
5. Pohl, T., Qafari, M.S., van der Aalst, W.M.P.: Discrimination-aware process mining: A discussion. In: Montali, M., Senderovich, A., Weidlich, M. (eds.) Process Mining Workshops. pp. 101–113. Springer Nature Switzerland, Cham (2023)
6. Qafari, M.S., van der Aalst, W.: Fairness-aware process mining. In: On the Move to Meaningful Internet Systems: OTM 2019 Conferences. pp. 182–192. Springer International Publishing
7. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. p. 335–340. AIES '18, Association for Computing Machinery, New York, NY, USA (2018)