Artificial Intelligence and Law (2021) 29:111–147
https://doi.org/10.1007/s10506-020-09268-y

ORIGINAL RESEARCH

Evaluating causes of algorithmic bias in juvenile criminal recidivism

Marius Miron[1] · **Songül Tolan**[1] · **Emilia Gómez**[1,2] · **Carlos Castillo**[2]

Published online: 7 June 2020

Abstract

In this paper we investigate risk prediction of criminal re-offense among juvenile defendants using general-purpose machine learning (ML) algorithms. We show that in our dataset, containing hundreds of cases, ML models achieve better predictive power than a structured professional risk assessment tool, the Structured Assessment of Violence Risk in Youth (SAVRY), at the expense of not satisfying relevant group fairness metrics that SAVRY does satisfy. We explore in more detail two possible causes of this algorithmic bias that are related to biases in the data with respect to two protected groups, foreigners and women. In particular, we look at (1) the differences in the prevalence of re-offense between protected groups and (2) the influence of protected group or correlated features in the prediction. Our experiments show that both can lead to disparity between groups on the considered group fairness metrics. We observe that methods to mitigate the influence of either cause do not guarantee fair outcomes. An analysis of feature importance using LIME, a machine learning interpretability method, shows that some mitigation methods can shift the set of features that ML techniques rely on away from demographics and criminal history which are highly correlated with sensitive features.

Keywords Criminal recidivism · Machine learning · Algorithmic fairness · Risk assessment · Criminal justice · Automated decision making

Marius Miron and Songül Tolan have contributed equally to this work.

✉ Marius Miron
miron.marius@gmail.com

[1] European Commission's Joint Research Centre, Seville, Spain

[2] Universitat Pompeu Fabra, Barcelona, Spain

1 Introduction

In recent years there is an increasing use of Machine Learning (ML) to assist decision making in areas of high societal relevance such as criminal justice (Berk et al. 2017; Goel et al. 2018). ML models are able to learn rules from large datasets and may improve decision processes by being more accurate and avoiding human cognitive biases (Langley and Simon 1995; Kleinberg et al. 2017). However, ML models also tend to inherit rules from previously discriminating decisions that are reflected as biases in data (Barocas and Selbst 2016) and may automate decisions that discriminate against certain minority groups or populations (Angwin et al. 2016; Chouldechova 2017; Barocas and Selbst 2016). In algorithmic fairness the features related to these populations, such as gender, race, nationality or religion are known as *protected* or *sensitive* features, and ideally they should not affect the outcome.

Statistical methods to inform criminal decision making are not new but have become subject to intense evaluation especially since the implementation of ML systems (Berk et al. 2017). These are increasingly used to inform decision-making situations in the criminal justice system, such as probation or bail decisions, sentencing, or corresponding arrangements (Christin et al. 2015; Monahan and Skeem 2016; Goel et al. 2018). The use of these methods causes changes in judge and case worker decisions and subsequently defendants' lives, incarceration and public safety levels (Cowgill 2018; Corbett-Davies et al. 2017). Moreover, the usage of machine involvement itself in these processes can cause substantial changes in criminal decision making that may bypass jurisdictive protocols (Green 2018). It is therefore crucial for policy makers involved in criminal justice reform to understand the extent of the changes in criminal justice outcomes. A comparison in terms of predictive power and relevant fairness metrics between a "human-in-the-loop" empirically informed assessment and a pure statistical risk-assessment sheds some light on this question.

This research extends the work done in Tolan et al. (2019), which compares off-the-shelf ML for juvenile recidivism prediction in terms of predictive performance and group fairness metrics (introduced in Sects. 2.1 and 5.1.3) with a risk assessment tool, that supports structured professional judgement (SPJ), the Structured Assessment of Violence Risk in Youth (SAVRY) (Hilterman et al. 2014). These results (Tolan et al. 2019) suggest that ML methods can achieve better predictive performance than SAVRY but at the expense of unfairer outcomes. The main goal of this paper is to investigate why that happens. More precisely, we explore the impact of two data-related biases that may translate to algorithmic discrimination: (1) the difference in the prevalence of recidivism between protected groups, and (2) the use of protected features or those that are correlated with protected features in the algorithm's training process.

In addition, we design evaluation experiments for different feature sets (Sect. 4.2) and we discuss the importance of static versus dynamic features (Sect. 4.4). Our assumption is that static features, such as demographic characteristics and past criminal history, have a higher correlation with the protected features and induce more disparity between groups on the studied group fairness metrics than the dynamic features, which include current substance abuse, peer rejection, or hostile behavior.

This is particularly relevant for technology driven reforms in jurisdiction, since sentence-course decisions that only depend on static features preclude defendants from being able to actively affect their own sentencing. Such a deterministic decision support system would contradict a criminal justice system that aims at stimulating societal reintegration efforts from defendants.

To explore the effects of the two data-related biases, we propose in Sect. 4.3 a stratified oversampling algorithm to equalize base rates. We compare this method with a state of the art algorithm which reduces bias in data by removing the impact of protected features (Zemel et al. 2013). In the field of Fair, Accountable, and Transparent computing (FAT) these two methods are considered forms of pre-processing algorithmic bias mitigation.

The results presented in Sect. 5.2 show that the two mitigation strategies, (1) equalizing base rates and (2) removing the impact of protected features (Zemel et al. 2013), improve commonly used metrics of group fairness. However, neither method guarantees group fairness. To that extent, fixing disparity between groups for a given group fairness metric does not ensure that the ML model satisfies other group fairness metrics (Chouldechova and Roth 2018). Moreover, we find that both methods have a negative impact on predictive accuracy. We embed these results into findings from the existing literature in Sect. 6.

Our quantitative analysis gives insights on the impact of algorithmic bias on juvenile criminal recidivism prediction. Namely, we show that biases that are often observed at the data level have a direct impact on the disparity observed in ML systems. Considering that ML systems are often opaque for end users and policy makers (Pasquale 2015), data analysis may expose potential disparities between groups within ML systems. Although our direct contributions are statistical, we discuss the social and policy implication of this in Sect. 7.

Most research in the literature that evaluates the impact of ML in criminal risk assessment is US based. This study contributes to this type of analysis in a European context using data from Catalonia in Spain. Moreover, this paper complements the fair machine learning literature that proposes unfairness mitigation techniques (e.g., Zemel et al. 2013; Žliobaitė and Custers 2016; Zafar et al. 2017; Agarwal et al. 2018) by testing some of the sources of unfairness mentioned by, among others, Barocas and Selbst (2016). Finally, we make use of a state-of-the-art ML explainability method (Ribeiro et al. 2016), to contribute to the literature that shows the negative consequences of unfairness mitigation techniques (Corbett-Davies et al. 2017; Kallus and Zhou 2018; Liu et al. 2018).

2 Background

2.1 Algorithmic fairness

Algorithmic fairness, a part of the research in Fairness, Accountability and Transparency (FAT) in ML, is concerned with discrimination happening within algorithmic systems. Here, we study group fairness under which a process or decision is

considered fair if it does not discriminate against people on the basis of their membership to a protected group.

Fairness is a value driven concept with roots in ethics and law. For instance, the United States forbids discrimination based on sex, race, color, religion, national origin (Civil Rights Act of 1964), citizenship (Immigration Reform and Control Act), age (Age Discrimination in Employment Act of 1967). Similarly, the European Convention on Human Rights (Article 14) forbids discrimination by "sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status." Because fairness is a value driven concept and not a technical feature of ML models, it is difficult to implement and evaluate. While discrimination occurs within different institutions and at different parts of the process, here we address fairness at the decision stage and we do not deal with other types of fairness such as process fairness. The discrimination in our specific case is evidenced by a measurable disparity between the outcomes of people from different groups using well-established metrics. However, fulfilling a single metric solely ensures parity between groups with respect to that metric. It does not ensure that the algorithm is "fair." In fact, the literature mentions at least 21 definitions of fairness (see, e.g., Berk et al. 2017; Narayanan 2018; Tolan 2018 for an overview on different definitions of algorithmic fairness), and proves some group fairness criteria are incompatible with each other (Chouldechova 2017; Kleinberg et al. 2016). Thus, fairness largely depends on context and value choices.

There are inherent trade-offs between accuracy and group fairness metrics, but we note that in some cases, for instance if a classifier is manually crafted instead of optimized for accuracy, one in principle could achieve higher accuracy and improve group fairness in a particular metric simultaneously. Hence, the trade-off between accuracy and a particular group fairness metric needs to be empirically observed on a case by case basis (Chouldechova and Roth 2018; Kleinberg et al. 2016).

2.2 Mitigating unfairness

Part of the research in the FAT field deals with developing fair algorithms that maintain the higher accuracy of ML decision support systems. These methods are applied at different parts of a ML pipeline. *Pre-processing* methods modify training data before training the model, *in-processing* optimizes a ML model considering a group fairness objective during training, and *post-processing* modifies the outcomes of a ML model (Corbett-Davies and Goel 2018; Žliobaitė and Custers 2016; Lipton et al. 2018; Hardt et al. 2016; Zafar et al. 2017; Agarwal et al. 2018).

In many cases the data collection pipeline and the data itself have problems which are better solved by other structural measures, rather than at the algorithmic level. One example is the difference in base recidivism rates as the result of years of structural discrimination long before the present data is recorded (Green and Hu 2018). Related to this is the problem of biased labels, e.g., if overpolicing of minority groups causes a large part of the minority group to be wrongfully labelled as re-offenders (Chouldechova 2017). Other issues refer to feedback loops, skewed data distributions, and uninformative (or unequally informative) features. To that extent,

risk prediction systems may be part of feedback loops which reinforce previous biases. These loops occur in environments characterized by complex interactions such as policing systems (Lum and Isaac 2016). Furthermore, a group of people may be underrepresented in the data or the features used might not be representative of that group (Corbett-Davies and Goel 2018).

We look at two particular issues that occur within the training data, and that are potentially problematic for specific group fairness metrics: unequal base rates and the use of input features strongly correlated with the protected features. While we do not advocate for trying to mitigate these conditions at the algorithmic level for the reasons previously described, we deploy two methods, (1) equalizing base rates (EBR, details in Sect. 4.3) and (2) "learning fair representations" (LFR) (Zemel et al. 2013), for diagnostic purposes only and not advocating them as "unfairness mitigation" techniques. Instead, we use of these two methods to trace specific sources of disparity in the data.

2.3 The SAVRY risk assessment tool

This section gives a general overview of SAVRY. SAVRY is a violence risk assessment tool for juvenile offenders which is designed as a "Structured Professional Judgment" (SPJ) (Borum et al. 2011). SAVRY was developed in 2003 (Bartel et al. 2003) based on research and literature on adolescent development as well as on violence and aggression in youth.[1] SAVRY has been found to have moderate to high predictive accuracy for recidivism (AUC$\approx$ 0.7), performing similarly to other instruments for juvenile risk assessment such as YLS/CMI (Ortega-Campos et al. 2020). As opposed to COMPAS [a risk assessment tool for adult offenders (Northpoint, Inc. 2012)], SAVRY is an open and interpretable assessment that actively guides the evaluating expert through the individual features that make up the overall risk assessment. As such it leaves a high degree of involvement by individual expert assessments.

Compared to the numerous studies published on COMPAS, the literature in juvenile criminal justice is still scarce (Hilterman et al. 2014) and it is unclear whether SPJs like SAVRY could have discriminating outcomes. An analysis of SAVRY for racial bias against black defendants in Pennsylvania found that SAVRY did not predict significantly different risk scores as a function of race (Perrault et al. 2017). However, this result does not extrapolate to completely different institutional settings, such as our case of juvenile delinquents in Catalonia.

Within SAVRY, juvenile justice professionals give scores on three levels of severity (low, moderate high) using 24 risk factors and six protective factors. These risk factors are divided into three categories: Historical, Individual, and Social/Contextual. The scoring mechanism is described in the SAVRY manual: "[I]n coding the History of Violence item, a youth would be coded as 'Low' if he had committed no prior acts of violence, 'Moderate' if he was known to have committed one or two

[1] Note that the authors of SAVRY and the authors of this study do not overlap.

violent acts, and 'High' if there were three or more. Protective factors are simply coded as present or absent" (Borum et al. 2003). The total risk score is a simple sum of the 24 risk factors (SAVRY sum), while other risks are sums of different sets of factors. The six protective factors are recorded as present/absent. After the scores are computed, an expert assigns a final overall score on the same three levels of risk that indicates the defendants risk of violent recidivism (Expert). This final evaluation is a professional judgment of the case worker, that is informed by the realizations of risk and protective factors. Thus, it is not algorithmically determined by the total score of risk factors.

The 24 risk factors comprise static (e.g., historical factors such as "history of violence" or "past supervision/intervention failures") and dynamic (e.g. "peer rejection" or "substance-use difficulties") factors. The focus of the practitioners is to choose the proper treatment to change the dynamic factors. Although in our experiments we compare two feature sets corresponding to the 24 SAVRY risk factors and to the demographics and criminal history Non-SAVRY features, we use ML interpretability to determine the important features in resulting ML models. By these means, we determine the proportion of static and dynamic features within the high-rank important features.

Note that experts are informed about substantial sex-differences in the response to specific risk factors (Björkqvist et al. 1992; Rowe et al. 1995; Wright et al. 2007) and are aware that SAVRY is designed mostly for male defendants. Thus, the risk factors that may apply differently to males and females (Borum et al. 2003). Meta studies show a good predictive validity for the SAVRY expert evaluation with a median AUCROC of 0.71 (Olver et al. 2009; Singh 2014) and the SAVRY sum with mean weighted AUCROC values of 0.71 (Guy 2008).

SAVRY, as a simple sum of factors, has not been developed through machine learning (e.g., using logistic regression), and it has not been algorithmically optimized for accuracy. Additionally, the performance of SAVRY and SAVRY-informed expert evaluation have not been evaluated from the perspective of algorithmic fairness. Hence, we would like to understand how these methods compare against machine learning in terms of accuracy and group fairness metrics.

3 Dataset and data pre-processing

This analysis is based on a dataset of juvenile offenders who were incarcerated in the juvenile justice system of Catalonia ($N = 4753$) and who were released in 2010.[2] The offenders committed the corresponding crimes between 2002 and 2010 when they were aged 12–17 years. Their recidivism status (after their release in 2010) was followed up on December 31, 2013 and December 31, 2015 (independent of their association to the juvenile or adult justice system). In other words, we observe recidivism behaviour between 2010 and 2015. The focus of our analysis is a sub-sample

[2] Provided by the Centre for Legal Studies and Specialised Training (Blanch et al. 2017), available at http://cejfe.gencat.cat/en/recerca/opendata/jjuvenil/reincidencia-justicia-menors/index.html.

of 855 defendants whose risk of recidivism was assessed with SAVRY towards the end of their sentences in 2010. The SAVRY assessment did not impact the sentence that the defendant received for the main crime. We use the recidivism status by December 31, 2015 as outcome label. In this research we use a pre-processed version of the data. The data-preparation code and the resulting dataset are available on the repository described in Sect. 5.1.5. The SAVRY assessment is represented by both the SAVRY sum of 24 risk factors as well as the final expert assessment (see Sect. 2.3). We elaborate on their dependency in Sect. 4.2.

Table 2 in Appendix shows descriptive statistics by recidivism status in 2015 of the relevant features of the analysis. We distinguish between two sets of input features: features that are not encoded in SAVRY, including protected features, and features that are encoded in SAVRY. Note that SAVRY features are not restricted to the 24 risk factors but also include indicators on the presence of six protective factors (see Sect. 2.3). The top panel depicts additional statistics for the protected features. For the present analysis we look at the following protected features: male/female, as well as Spaniard/foreign. Among foreigners, we look at two relatively large subgroups: Latin Americans and Maghrebis – other groups are too small. The fact that many non-SAVRY features as well as almost all SAVRY features significantly differ between the group of recidivists and non-recidivists emphasizes the empirical relevance of the input features used in this analysis. The table further shows that this also accounts for almost all protected features. Finally, we observe substantial differences in the base recidivism rates (the prevalence of recidivism within each group) across protected group features.

4 Methodology

Here we propose a methodology to study the causes of algorithmic discrimination when using common ML classification algorithms to predict juvenile criminal recidivism. We evaluate different algorithms, feature sets, and biases in training data on metrics related to predictive performance and group fairness. Note that our methodology includes the data analysis and the data pre-processing procedures described in Sect. 3.

4.1 Learning algorithms

Recidivism risk assessment is usually modeled as a ML classification problem with discrete risk classes (low risk, medium risk, high risk), although it could be modeled as a regression problem (risk score). A ML model outputs a probability of recidivism or probabilities for each risk category. To simplify the evaluation in terms of group fairness, we consider a binary classification scenario, similar to the majority of the algorithmic fairness literature. In this case, we predict "High risk" and "Low risk" based on "Recidivist" and "Non-Recidivist" labels. Data for the time between 2002 and 2010 is used as input and recidivism is predicted for the period between release in 2010 and December 31, 2015. There is a certain imbalance between

classes, as the positive class (recidivists) comprises about one-third of the cases, while the negative class (non recidivists) comprises the other two-thirds.

Our experimental evaluation assumes a k-fold cross validation (Robert 2014) in which data is partitioned into k folds and k different learning rounds are executed. A different fold of the dataset is used for testing in every round, and the remaining $k-1$ folds are used for training. We subsequently split this training data by keeping 10% random elements for validation. The validation set is used to tune the ML model's hyper-parameters and to pick the binarization threshold for the prediction of the ML models.

We evaluate the following supervised ML algorithms: logistic regression (*logit* in the following tables and figures), multi-layer perceptron (*mlp*), support vector machine with a linear kernel (*lsvm*), K-nearest neighbors (*knn*), random forest (*rf*), decision tree (*dt*), and naive Bayes (*nb*) (Robert 2014). We report predictive performance metrics in terms of area under the curve (AUC, defined in Section 5.2) for all ML models. However, to better visualize the group fairness plots across multiple experiments we solely analyze the top two performing models in terms of group fairness metrics, which correspond to logistic regression (*logit*) and multi-layer perceptron (*mlp*).

4.2 The influence of feature sets

In order to determine the influence of features on the disparity between groups and on the predictive performance we train the ML models on different subsets of features. In a first setting, denoted "SAVRY," we take all the SAVRY risk items as input variables. We select the final expert evaluation, the 24 risk items, the corresponding summary scores, the six protective features, the five average scores on individual characteristics as well as the program that the defendant was in (internment or probation) during the SAVRY assessment. "SAVRY" features include both static (home violence, school performance) and dynamic factors (achievement, personality). That is, SAVRY features are not limited to criminal history but also contain individual and social/contextual features. The full list of "SAVRY" features is detailed in Table 2.

For the second setting, "Non-SAVRY," we choose demographic and criminal history features which are part of the dataset and not directly included into the "SAVRY" features.[3] The third setting, "All," includes "SAVRY" and "Non-SAVRY" features sets.

The baselines include the sum of all SAVRY risk items, using no machine learning, denoted in the following by "SAVRY Sum," in addition to the final expert evaluation, denoted by "Expert." While "SAVRY Sum" does not represent the final professional judgment, it is a good proxy as a meta-study shows summed scores and professional judgments in risk assessments are not significantly different in terms of predictive power (Chevalier 2017). Figure 1 supports this finding,

[3] Note that SAVRY implements criminal history as risk factors into its framework.

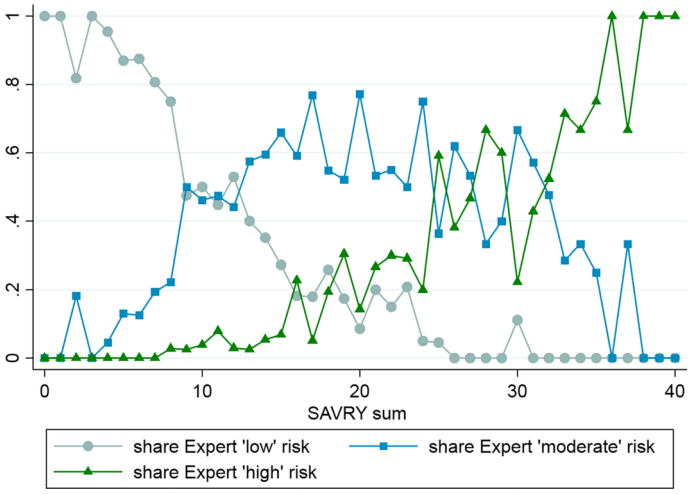

Fig. 1 Plot of expert assessment, represented as shares of "low", "moderate", and "high" risk categorization against "SAVRY Sum", the summed score of all 24 SAVRY risk factors. We observe that in general people with a low "SAVRY Sum" get a "low risk" expert evaluation and people with a high "SAVRY Sum" get a "high risk" expert evaluation

comparing the distribution of "SAVRY sum" in cases where the expert indicated low risk, medium risk, and high risk. We observe a clear trend to higher "SAVRY sum" for cases having higher expert assessments of risks.

We limit the comparison between experiments and baselines to the 855 people for which the SAVRY items are available.

4.3 The influence of data bias with respect to the protected features

Next, we look at whether the bias in the data with respect to the protected features yields disparity between groups. To perform this analysis, we propose a comparison of the baseline, unrestricted data settings with two mitigation methods that address these issues: (1) equalized base rates (EBR) and (2) learning fair representations (LFR). Differences in predictive performance, group fairness and the set of features, which are important for prediction, provide further insights in the role that these conditions play with respect to group fairness.

EBR assumes the following stratified oversampling procedure. Considering the input features x and the outcomes y, a protected feature f' has I possible value corresponding to different groups: $\{g_1, \ldots, g_I\}$, comprising a reference group g_{ref}. For each group g_i, different to the reference group, we compute the number of positive or negative condition samples $|S_i|$ we need to add to the training data, where $|S_i|$ is the absolute value of $S_i = P(x(f') = g_i)(P(y|x(f') = g_i) - P(y|x(f') = g_{ref}))$. If S_i is negative then we add to the data $|S_i|$ negative condition samples $P(y = 0|x(f') = g_i)$, while for S_i positive we add $|S_i|$ positive condition samples

$P(y = 1|x(f') = g_i)$ of group g_i. This procedure yields equal base rates between the group g_i and the reference group g_{ref}, namely $P(y|x(f') = g_i) = P(y|x(f') = g_{ref})$.

LFR (Zemel et al. 2013) is a pre-processing method. It transforms the data before training a ML model in such a way that the outcomes are more balanced between groups with respect to particular features and also between individuals. Namely, LFR is based on an optimization procedure which transforms the data with respect to a protected feature (e.g. race) by removing any information about membership with respect to the protected group.

By including LFR along EBR we can determine whether the cause of disparity between groups lies in the difference in base rates, if there are other characteristics of the data that LFR is able to fix and EBR is not, or if the measured disparity is due to other causes like the type of input features and the ML algorithm.

Pre-processing methods such as LFR achieve fair outcomes with respect to a protected feature. However, just like EBR, it may be the case that this induces further disparity on another related feature. For example, mitigating group fairness issues with respect to race may induce disparities with respect to sex, particularly when women are under-represented in the training data.

4.4 The importance of static versus dynamic features

While the "Non-SAVRY" feature set contains static features, "SAVRY" has predominantly dynamic features. Our goal is to determine whether static features are more important than the dynamics ones, particularly when we analyze the combination of the two feature sets, "All."

We determine the influence on group fairness outcomes of specific features in the feature sets by using ML interpretability (Lipton 2016). ML algorithms such as logistic regression or decision trees have the advantage of being easily interpretable. For instance, the coefficients learned by the *logit* model correspond to feature weights. Other models such as neural networks are not as easy to interpret.

To overcome the lack of interpretability of such models we obtain local explanations for each data point prediction using a state of the art interpretability framework, LIME (Ribeiro et al. 2016). Note that the interpretations are in this case linear approximations because LIME fits a linear model each time it perturbs a feature in order to determine its importance.

Given a test dataset comprising N pairwise observations $(\mathbf{X}^{N\times F}, \hat{\mathbf{Y}}^N)$ and their associated ground-truth binary labels $\mathbf{Y}^N$, we derive N feature importance vectors $\mathbf{E}^{N\times F}$ with LIME. The global importance vector is computed by aggregating the local explanations with the formula:

$$\mathbf{e}(f) = \sqrt{\sum_{n=1}^{N} E(n,f)} \tag{1}$$

where $n = 1 \ldots N$ denotes the explanations and $f = 1 \ldots F$ the features.

5 Evaluation

5.1 Experimental setup

5.1.1 Data encoding

There are different reasons for encoding the data numerically: to make sure numerical relations between data are encoded correctly when data is not numeric, to ensure the compatibility of data when using different ML algorithms, to ensure standardization and robustness.

The encoding depends on the nature of the input features. Numeric features are standardized to have a mean of 0 and standard deviation of 1. Categorical features are encoded using one-hot encoding if they do not have a numerical relation or numerically (e.g. High, Medium, Low are encoded as 1, 0.5, 0). One-hot encoding has the goal to remove any undesired numerical relation between the categories by creating a set of dummy features for a single feature.

5.1.2 Predictive performance evaluation metrics

Thresholding is the way to obtain class labels from the class probabilities yielded by a ML model. The simplest way is to set this threshold always at 0.5. However, particularly in the cases where there is an imbalance between classes, the threshold may be determined by considering the cost-benefit trade-off, maximizing accuracy or any other metrics. Since we are interested in the predictive performance for both of the classes, we use the threshold for which balanced accuracy, defined as $BA(\theta) = 0.5(TPR(\theta) + TNR(\theta))$, is maximum; where $\theta \in (0, 1)$ is the varying threshold, TPR is the true positive rate, and TNR is the true negative rate.

Because we want to get a measure of the predictive performance for all possible thresholds, we compute the area under the ROC curve (AUCROC) which trades-off false positive rate and true positive rate for all the thresholds $\theta \in (0, 1)$.

The "SAVRY Sum" method generates an integer in [0, 40], which we normalize to lie in the same [0, 1] interval as the output of the ML methods. Similarly to the ML models, the threshold for "SAVRY Sum" is set to maximize balanced accuracy. The "Expert" evaluation has three possible outcomes: high risk, moderate risk, and low risk. Given we want to maximize balanced accuracy, we assign a non-recidivist prediction to moderate or low risk and a recidivist prediction to high risk.

5.1.3 Fairness evaluation metrics

Metrics related to the disparity between groups were previously used in analyzing algorithmic fairness in criminal recidivism. We computed and looked at eleven

further group fairness metrics but found these to be highly correlated, in line with findings by (Friedler et al. 2018; Miron et al. 2020.

Similarly to Chouldechova (2017) and guided by Aequitas' (Saleiro et al. 2018) fairness metrics decision tree,[4] we consider two group fairness metrics that are useful to look at and have been widely used in the context of criminal recidivism. *False positive rate disparity* and *false negative rate disparity* measure the disparity between two groups as the ratio of people wrongfully labeled as recidivists and non-recidivists. Note that both metrics measure different objectives and their relevance may differ for different stakeholders. A false positive occurs if a system classifies as high risk someone who will not recidivate. That is, a false positive has a detrimental (direct) effect on the defendant. In contrast, a false negative occurs if the scoring system classifies as low risk someone who will recidivate. If a low risk classification normally prevents detainment, a false negative could cause a more detrimental (direct) effect on public safety. To acknowledge this crucial difference, and simplify policy-relevant interpretations, we show results for both metrics in separate tables. The trade off between these two metrics is known as error-rate balance and it is impossible to bring to zero when the base rates are unequal (unless the automatic classifier can perfectly predict the outcomes, which is not the case in real-world conditions).

Given a protected feature f which has the values $\{g_1, \ldots, g_I\}$, group fairness metrics are reported for all the protected groups g_i with respect to the reference group g_{ref}. We denote the outcome as $\mathbf{Y}$, where $\mathbf{Y} = 1$ if the defendant recidivated, 0 otherwise. We denote the number of defendants of group i labeled with $\mathbf{Y} = 1$ as LP_i. We denote the number of defendants of group i labeled with $\mathbf{Y} = 0$ analogously as LN_i. The predicted outcome is represented by $\hat{\mathbf{Y}}$. The ML algorithm classifies someone as high risk for recidivism, i.e. $\hat{\mathbf{Y}} = 1$ if the risk score R surpasses a predefined threshold (θ), i.e. $R > \theta$. We denote the number of defendants of group i predicted positive for recidivism as PP_i. We denote the number of defendants of group i predicted negative for recidivism as PN_i. Equivalently, we denote the number of group-specific false positives (FP_i), false negatives (FN_i), true positives (TP_i), and true negatives (TN_i).

The criterion of error rate, i.e., equal false negative rates and equal false positive rates, Chouldechova (2017) is achieved when people from the protected group g_i have the same probability of falsely being classified as recidivist (or non-recidivist) than people from the reference group with attribute g_{ref}.

The error rate balance is computed using false positive rate and false negative rate of group g_i (FPR_i,FNR_i), from which we derive false positive rate disparity and false negative rate disparity in relation to the reference group g_{ref} ($FPRD_i$,$FNRD_i$):

$$\begin{aligned} FPR_i &= FP_i/LN_i \quad FPRD_i = FPR_i/FPR_r \\ FNR_i &= FN_i/LP_i \quad FNRD_i = FNR_i/FNR_r \end{aligned} \tag{2}$$

[4] http://www.datasciencepublicpolicy.org/projects/aequitas/.

Intuitively, if a group g_i has $FPRD_i = 2$ means that someone with attribute g_i is twice as likely to be wrongly classified as recidivist as someone from the reference group with attribute g_{ref}.

5.1.4 Hyper-parameters' tuning and model selection

Because we want to obtain predictions for all the people in the dataset, we use a k-fold cross validation experimental design with $k = 10$. This assumes training successively on non-overlapping data folds and testing on the remaining data. Moreover, to account for variability in the data, we repeat each cross-fold experiment 20 times for a different random seed which controls the initialization of the parameters and the random split between training, validation, and testing.

At each seed and each fold we compute the best values for the hyper-parameters of the ML algorithms. We train 30 models with different hyper-parameters we choose the model which achieves the best performance in terms of AUCROC on the validation set.

The hyper-parameters depend on the ML algorithms. For *logit* we pick the inverse of regularization strength from a uniform distribution $\mathcal{U}(0.1, 10)$. For *mlp*, we use a two layer network with the sizes $(F, L * F), (L * F, (L + 1) * F), (L * F, 1)$, where F is the number of input features and L is chosen randomly from an uniform distribution $\mathcal{U}(1, 10)$. In addition we experimentally determined the batch size to be 64, we update parameters using the stochastic gradient descent for 100 epochs. The cost function for *mlp* classification is binary cross entropy, with an $\mathcal{L}_2$ penalty on weights of 0.01 to avoid over-fitting. For *knn* the number of neighbors and the distance metrics are picked randomly between (3, 20) and between Minkowski, Euclidean and Manhattan. For the SVM we trained a linear and radial kernel separately. The kernel radius and gamma are drawn from uniform distributions $\mathcal{U}(0.1, 10)$. For the *rf* we randomly pick the number of estimators to be between (10, 50), the maximum depth between (5, 50) and the minimum number of samples per leaf between (1, 10).

5.1.5 Software implementation details

We bootstrap experiments for 20 random seeds to ensure robustness and reproducibility. This research complies with research reproducibility principles. Code in Python, including pointers to the machine learning libraries used, as well as the processed datasets, are made available as a part of a framework.[5]

[5] https://gitlab.com/HUMAINT/humaint-fatml.

Table 1 AUCROC for each experiment and for the ML models, including mean and standard deviations aggregated across 20 random seeds

	logit		*mlp*		*knn*		*lsvm*		*dt*		*nb*		*rf*	
	Mean	std.dev.	Mean	std.dev.	Mean	std.dev.	Mean	std.dev.	Mean	std.dev.	Mean	std.dev.	Mean	std.dev.
SAVRY	.68	.0036	**.69**	.0071	.63	.0186	.68	.0049	.60	.0187	.66	.0014	.67	.0074
SAVRY EQB S	**.68**	.0042	**.68**	.0059	.62	.0148	**.68**	.0043	.58	.0196	.66	.0015	.67	.0082
SAVRY EQB F	**.68**	.0052	**.68**	.0076	.62	.0150	.67	.0071	.59	.0189	.66	.0017	.66	.0092
Non-SAVRY	**.72**	.0037	.71	.0072	.66	.0121	.69	.0099	.61	.0158	.71	.0031	.68	.0117
Non-SAVRY EQB S	**.70**	.0041	**.70**	.0062	.65	.0074	.61	.0119	.61	.0158	.69	.0046	.69	.0072
Non-SAVRY EQB F	**.71**	.0036	.70	.0123	.65	.0086	.60	.0142	.60	.0176	.71	.0038	.67	.0085
Non-SAVRY LFR S	**.66**	.0103	.64	.0143	.64	.0163	.66	.0133	.62	.0154	.64	.0149	.64	.0143
Non-SAVRY LFR F	**.66**	.0132	.64	.0187	.63	.0122	.66	.0145	.62	.0106	.64	.0141	.64	.0129
All	.72	.0050	.72	.0110	.66	.0150	**.73**	.0043	.61	.0151	.70	.0018	.71	.0083
All EQB S	**.71**	.0044	**.71**	.0062	.65	.0109	**.71**	.0052	.62	.0155	.69	.0016	.71	.0070
All EQB F	.71	.0064	.71	.0068	.65	.0173	**.72**	.0070	.62	.0155	.70	.0022	.71	.0087
All LFR S	.65	.0245	**.65**	.0282	.62	.3880	.6	.0426	.58	.0416	.61	.0360	.61	.0376
All LFR F	.64	.0269	**.64**	.0278	.60	.0473	.59	.0437	.57	.0464	.60	.0273	.60	.0338

For comparison, the baselines “SAVRY Sum” and “Expert” achieve AUCROC of .64 and .66. The scores for the top two ML methods are marked in boldface

5.2 Results

5.2.1 Predictive performance

The metrics for predictive performance in terms of AUCROC are presented in Table 1. Note that the performance of off-the-shelf ML methods on this dataset is similar to the recidivism prediction on other datasets: 0.67 for a 5-variables random forest classifier (Green and Chen 2019), 0.68–0.71 for COMPAS (Northpoint, Inc. 2012), 0.65–0.66 for the Public Safety Assessment (DeMichele et al. 2018), 0.57–0.74 in a meta-study of various risk assessment used in the US (Desmarais et al. 2016).

We report results for the machine learning methods *logit*, *mlp*, *knn*, *lsvm*, *dt*, *nb*, *rf*. To assess the influence of different feature sets we compare "SAVRY," "Non-SAVRY," and "All." For each feature set we further explore whether equalizing the base rates ("EQB") and encoding sensitive features in data with LFR for sex and foreigner status affects the predictive performance. Note that LFR is not applied to the "SAVRY" feature set for which the features related to sex and nationality are missing. These features are essential for the data transformation within LFR.

We observe that training on "SAVRY" features leads to 0.02 lower performance than the "Non-SAVRY" feature set. In addition, combining "SAVRY" and "Non-SAVRY" feature sets leads to a 0.01 increase in performance for most of the machine learning methods except logistic regression. We note that the majority of the ML methods have better predictive performance than "SAVRY Sum" and "Expert."

Despite recent hype of black-box deep learning methods, we note that interpretable methods such as logistic regression and deterministic methods such as naive bayes achieve on-par performance with the multi-layer perceptron, particularly when using "Non-SAVRY" features. Decision tree has on average $0.08 - .1$ less in AUCROC than other methods, with random forest performing better. The support vector machine with a linear kernel has similar performance to the top two performing methods, *logit* and *mlp*. Moreover, a simple classification method such as k-nearest neighbors has similar performance to "SAVRY Sum" and "Expert" if the input features are predominantly from the "Non-SAVRY" set including demographic features and criminal history ("Non-SAVRY," "All").

Equalizing base rates (EBR) between men and women or between foreigners and nationals introduces on average a 0.01 decrease in performance. Conversely, applying LFR mitigation decreases the performance with 0.06 on average across all ML methods and feature sets and has higher standard deviation. While here we do not use EBR and LFR explicitly for mitigation, we confirm a common finding in the FAT literature (Corbett-Davies and Goel 2018), the fact that ML methods trade off accuracy for group fairness.

5.2.2 Group fairness

We measure the impact of ML algorithms, of "SAVRY Sum," and "Expert," different feature sets, and equalized base rates on the disparity between protected groups for "sex" and "nationality."

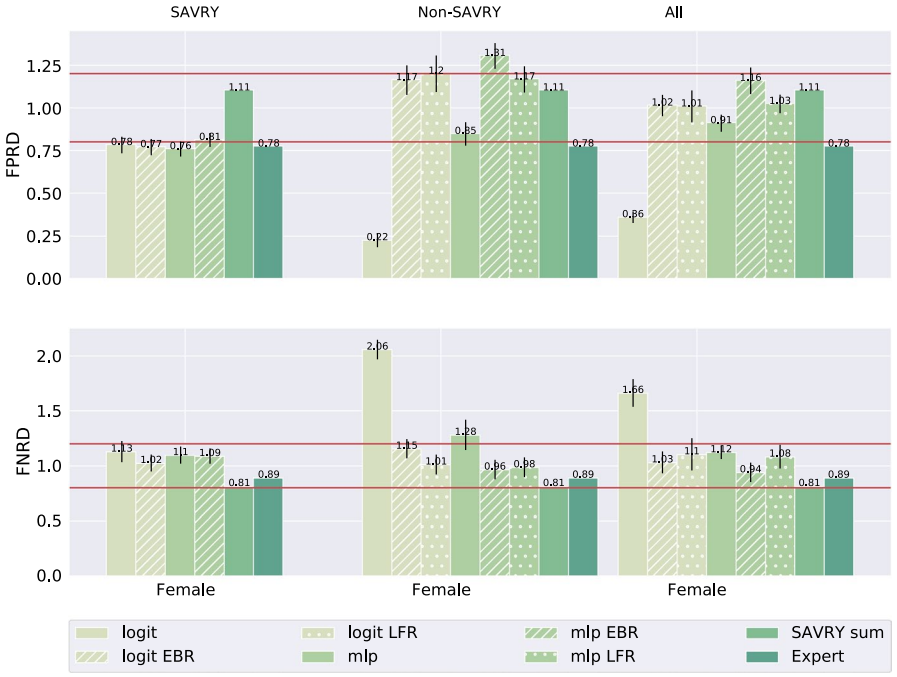

Fig. 2 Comparison of group fairness metrics using sex as the protected attribute. The reference group are men

The group metrics defined in Sect. 5.1.3 are reported in Figs. 2, 3 and 5. Error bars represent 95% confidence intervals across the 20 seeds. The feature sets "SAVRY," "Non-SAVRY," "All" are separated by vertical blue lines. The EBR and LFR results are presented in dashed texture bars. We delimit thresholds for disparity with horizontal red lines between (0.8, 1.2) similar to ProPublica's COMPAS analysis (Angwin et al. 2016). Note that despite the fact that these thresholds are derived from US laws and they do not hold any legal status in Catalonia, they are present in most of the FAT literature. From a visualization point of view, thresholds are useful in claiming that a method is unfair when perfect parity between groups is difficult to observe. Note that by definition, the disparity is a ratio between the metrics of a group and the reference group. Hence, we do not plot FPRD and FNRD for the reference group which is by definition equal to 1.

We present the group fairness metrics for the protected feature "sex" in Fig. 2. In this case, the reference group are men. We note that "SAVRY Sum" achieves low disparity according to FPRD and FNRD while the "Expert" evaluation is slightly below the threshold in terms of FPRD. We remark that the experts are informed that SAVRY is mainly designed for juvenile men and it is possible that when dealing with juvenile women they interpret SAVRY scores differently or rely more on their own expert opinion.

In general, ML does not satisfy group fairness criteria when used for classification in this setting. The ML methods achieve better predictive performance than

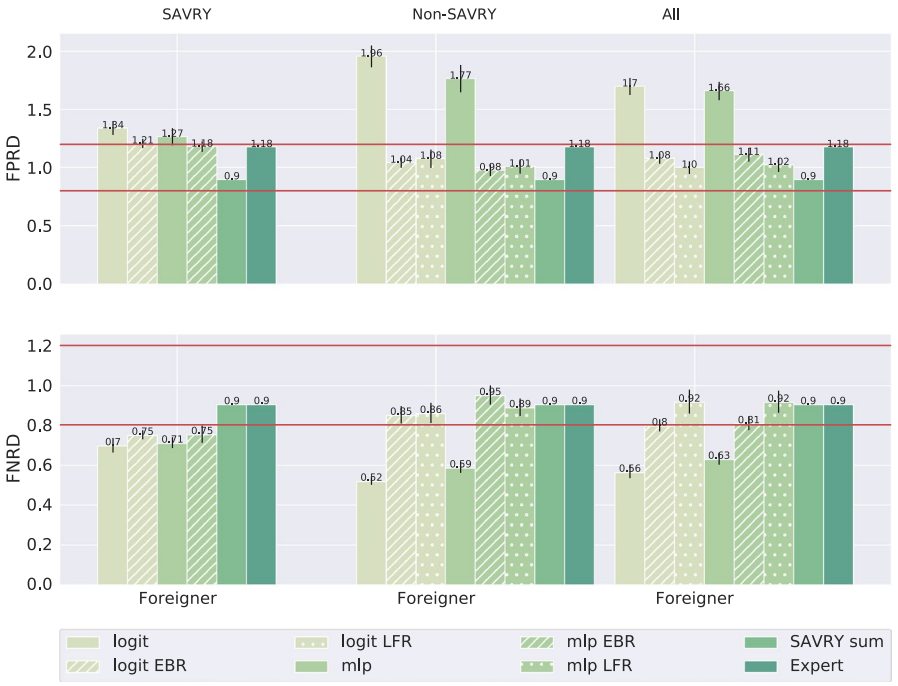

Fig. 3 Comparison of group fairness metrics in terms of nationality. The reference group are Spanish nationals

"SAVRY Sum" and "Expert" as seen in Sect. 5.2.1, but bring disparities according to error rates. For instance, when using logistic regression, men are three or four times more likely to be wrongfully classified as recidivists across "Non-SAVRY" or "All" feature sets.

There is a clear difference in terms of disparity when looking at the ML models trained with different feature sets. Models trained with "SAVRY" features yield less disparity than models with "Non-SAVRY" features. Furthermore, when combining "SAVRY" and "Non-SAVRY" features, the disparity between men and women is reduced.

Equalizing the base recidivism rates (using EBR) of men and women reduces the disparity between these two groups. Encoding sensitive features (using LFR) also reduces disparity on the group fairness metrics. EBR is particularly useful when having predominantly static features as input, such as "Non-SAVRY" and "All" feature sets. Neither EBR nor LFR, data pre-processing procedures, ensure group fairness criteria are satisfied for all feature sets. In some cases, such as FPRD for "Non-SAVRY" feature set, performing EBR or LFR shifts the disparity between men and women in the opposite direction, discriminating against women. A possible cause is the fact that there are considerably less women in the training data set. Thus, oversampling data or encoding data transforms the training data in unexpected ways. This is in line to the discussion on the problematic use of mitigation methods in Sect. 2.2.

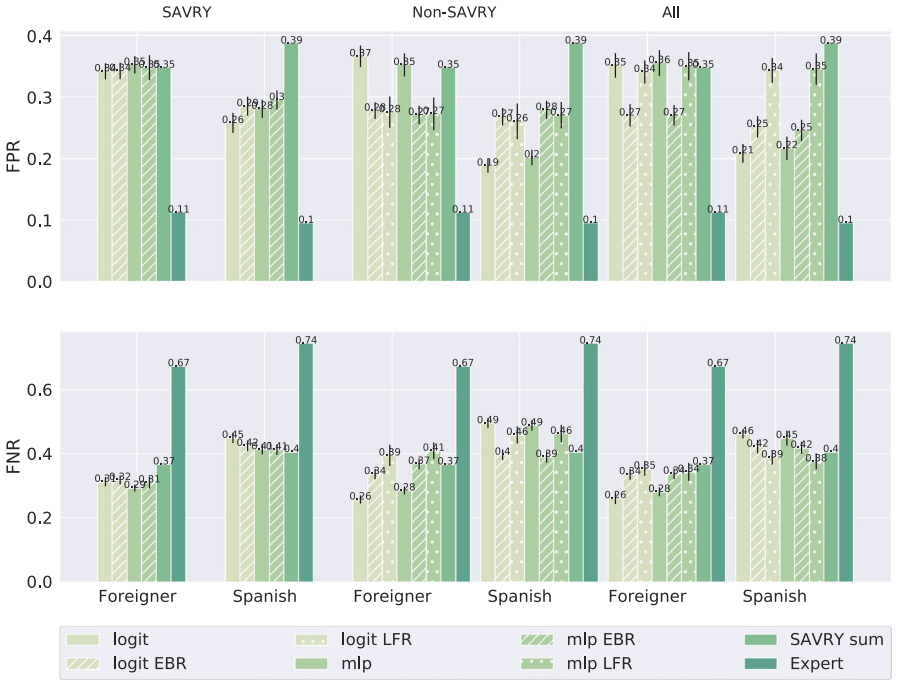

Fig. 4 Comparison of FPR and FNR metrics in terms of nationality

In Fig. 3 we present results in terms of group fairness for "nationality." Here the reference group are Spanish nationals and the protected group are foreigners. Similar to the "sex" protected feature, "SAVRY Sum" and "Expert" in general satisfy group fairness criteria while unmitigated ML methods fail to do so, particularly when including predominantly static features like the "Non-SAVRY" and "All" feature sets. To that extent, *logit* and *mlp* produce worse outcomes for foreigners across all the feature sets. Foreigners are more likely to be wrongfully labeled as recidivists while Spanish nationals are more likely to be wrongfully labeled as non-recidivists. The disparity between groups decreases when including "SAVRY" features.

EBR and LFR reduce disparities with a similar amount when the dataset has features correlated to nationality. Although the data distribution is the same, EBR is not as effective when using "SAVRY" input features. While the difference in base rates can be considered the main cause of disparity, it is not the only cause. In this case, disparity is caused by the ML algorithm considering interaction between the input features and characteristics of the data.

To better understand how EBR and LFR work, in Fig. 4 we plot the FPR and FNR for Spanish nationals and foreigner instead of the disparity between the two groups. We observe that EBR and LFR increase error rates for the group which has been initially positively discriminated. This happens when EBR and LFR are effective, on "Non-SAVRY" and "All" feature sets. In this case, Spanish nationals are more likely to be falsely labeled as recidivists and less likely to be wrongfully labeled as non-recidivists. In addition, when combining "SAVRY" and "Non-SAVRY" features

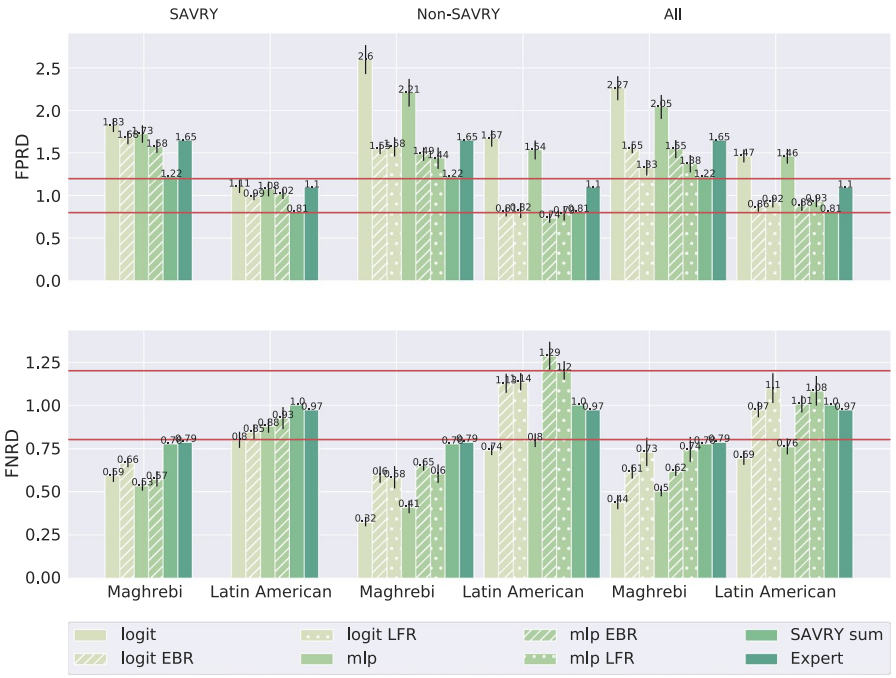

Fig. 5 Comparison of group fairness metrics in terms of national groups. The reference group are Spanish nationals

("All" feature set), LFR increases FPR for Spanish nationals with a larger amount than EBR, while not decreasing FPR for the foreigners. This finding suggests that although LFR and EBR have a similar impact on disparity, these methods may work in a very different way. We further analyze this using interpretability in Sect. 5.2.3.

On another note, Fig. 4 shows that "SAVRY Sum" and "Expert" have different false positives rates and false negatives rates, despite their disparity being similar as shown in Fig. 3. Experts are more "lenient," less likely to wrongfully predict that someone will recidivate, and more likely to wrongfully predict that someone will not recidivate.

Towards finding the influence of feature sets, equal base rates, and ML method on disparity between national groups, we subsequently split foreigners into Maghrebi, Latin American, non-Spaniard European and Others. The intuition behind this decision is that a ML model can balance a fair outcome towards a group by discriminating a particular subgroup and positively discriminate other subgroups. In Fig. 5 we present group fairness results for the selected nationality subgroups. Note that for "EBR" the base rates are equalized between Spanish nationals and foreigners, and not necessarily for the presented subgroups. We exclude small subgroups with less than 50 people like non-Spaniard European (37 people) and Others (13 people) for which a few different decisions may drastically impact the overall group fairness outcome. We note that European and Others tend to be positively discriminated in our experiments.

While in Fig. 3 "SAVRY Sum" and "Expert" in general satisfy group fairness criteria, here we observe disparity close to the threshold for Maghrebis. Experts are more likely to wrongfully label them as recidivists. On the other hand, "SAVRY Sum" and "Expert" in general do not have group fairness issues with respect to Latin Americans. ML methods, *logit* and *mlp*, have discrepancies between groups for Maghrebi and Latin Americans when using "Non-SAVRY" features. Particularly, the Maghrebi subgroup experiences worse outcomes than the main group, foreigners.

In a similar trend to the previous analysis on sex and national groups, the models trained on "SAVRY" features achieve better group fairness. When looking at the feature sets, ML methods yield more disparity towards Maghrebi and Latin Americans for all metrics when including non-SAVRY features in training. Training with SAVRY items has slightly higher disparity than the "SAVRY Sum." This disparity is within the acceptable bounds for Latin Americans and surpassing the bounds for Maghrebi.

While EBR does not target Maghrebi and Latin Americans in particular, these groups are more numerous than the Europeans and Others. Even if the prevalence of the recidivism for Maghrebi is higher, there are more Maghrebi non-recidivists than European non-recidivists. Thus, as a side-effect of targeting EBR to foreigners, we note a drop in the disparity between Maghrebi and Spanish nationals. However, in all the cases the drop of disparity is not within the accepted limits and Maghrebi are still discriminated against. Furthermore, the disparity shifts in the opposite direction, positively discriminating Latin Americans when using *mlp* on the "Non-SAVRY" feature set. Reasons why EBR and LFR do not always work are related to the group size. Similarly to the case of women, sampling from a small population enhances the noise. Furthermore, ensuring fair outcomes for a big subgroup does not guarantee group fairness for corresponding subgroups.

EBR and LFR achieve similar metrics across all experiments. The fact that the metrics change more substantially when non-SAVRY features are included, is consistent with the hypothesis that both methods are more effective when including features which are more correlated with the protected feature they target. This requires a further analysis on the feature importance when combining static and dynamic features, which we do in the next Sect. 5.2.3.

5.2.3 Feature importance

As discussed in Sect. 4.4, we use ML interpretability to see which features are important for different ML models. Particularly, we analyze the ratio of static and dynamic features in the top ten most important features for each ML model. Because the "SAVRY" feature set contains mostly dynamic features and the "Non-SAVRY" mostly static features, we focus on their combination, "All."

The top ten most important features, the mean and standard deviation over 20 seeds of the global feature importance computed with LIME (Ribeiro et al. 2016) (defined in Eq. 1) are given in Table 3 for *logit* and in Table 4 for *mlp* in "Appendix".

The top ten features ranked in terms of importance for *logit* and *mlp* are static and personal history features. Sex and nationality, age, province of residence, criminal history are important when doing off-the-shelf ML. This finding is particularly concerning because more informative dynamic features are not deemed to be important. In other words, the offender cannot change key traits used by such a ML system.

In Sect. 5.2.2 we show that EBR and LFR have a similar impact on the group fairness outcomes. However, when looking at the features which are important for *logit* and *mlp* after applying EBR and LFR, we observe that the static features are the most important for the EBR, and the dynamic features for LFR. This finding points out to the difference between the two algorithms. LFR encodes all the features correlated to sex and foreigner in such a way that these features are not helpful in the prediction. The fact that LFR relies on a different set of features explains why LFR has considerably lower AUCROC than the baseline, while this is not the case for EBR. Basically, LFR relies on a different set of features which are not as predictive and do not yield as much disparity.

6 Relation with previous work

This paper expands on the work in Tolan et al. (2019) and aims at studying the impact of using ML in criminal justice on group fairness and predictive performance. The main contribution of this paper is determining the causes residing in data of disparity between groups for ML juvenile recidivism prediction: unequal base rates between protected groups and different feature sets. With respect to the former, we introduce an oversampling method, EBR, to equalize base rates between protected groups. We show that equal base rates partially explain the observed disparity between groups. Towards studying the influence of the correlations between the input features and the protected features we compare EBR with a more complex and computationally intensive method, LFR (Zemel et al. 2013). Our experiments using different feature sets show that including features correlated with the protected features may explain disparity between the protected groups.

In contrast to research done in the field of algorithmic fairness for machine learning, we do not propose a mitigation framework such as Zemel et al. (2013), Žliobaitė and Custers (2016), Zafar et al. (2017), Agarwal et al. (2018). We use EBR to explain the influence of prevalence among groups on two group fairness metrics. Similarly, including LFR (Zemel et al. 2013) in the analysis gives insight on the influence of training features. Although a simple oversampling method, EBR, reduces unfairness on the selected metrics, at a comparable rate to LFR without a significant reduction of the predictive performance, we show that the usage of mitigation techniques may be problematic.

Our analysis on the impact of equal base rates on group fairness metrics and the impossibility to achieve parity between different subgroups in terms of the reported metrics is similar to Chouldechova (2017). In addition, we use a state-of-the-art ML explainability method (Ribeiro et al. 2016), to contribute to the literature that shows

the negative consequences of unfairness mitigation techniques (Corbett-Davies et al. 2017; Kallus and Zhou 2018; Liu et al. 2018).

In comparison to the previous research on juvenile recidivism (Olver et al. 2009; Singh 2014; Guy 2008; Ortega-Campos et al. 2020), we evaluate the algorithmic fairness problems within a ML based system. To the best of our knowledge, such a system is not used in practice at this moment. However, risk assessment tools tend to become more actuarial (Hilterman et al. 2014) and less clinical. With the recent adoption of deep learning in decision making application, using ML for juvenile recidivism prediction is plausible.

7 Discussion

The evaluation section points out the trade-off between predictive performance and group fairness. More precisely, the application of ML over the risk factors yields a more accurate prediction than the simple SAVRY sum. The ML methods achieve on average 0.04 points of AUCROC more than non-ML methods, depending on the input features used. Yet, ML introduces issues of group fairness that a simple SAVRY sum does not have. This issue gets more pronounced as we include Non-SAVRY factors such as the actual group features, further demographic information and data on criminal history. Beside gains in accuracy the additional features cause an even higher disparity in the error rates. Furthermore, we find that processing the data before training the algorithm in a way that removes the predictive power of protected features can in some cases help to remove the observed group-unfairness regarding the protected features but neither do these methods guarantee the removal of group unfairness nor do they necessarily achieve overall group fairness along unspecified protected features or intersections of these.

These findings have to be interpreted in light of some limitations. First, despite the relatively random selection of the sample by release year in 2010, we find that the selection into the SAVRY assessment of 855 defendants is on average targeted to defendants with a higher violence risk. Still, the sample remains fairly heterogeneous. The second problem is sample bias. The outcomes of the ML analysis could mostly be driven by the largest protected group, in this case Spaniard males. However, we repeated the ML analysis allowing for group-specific features and found no substantial differences to the baseline analysis.

We have to consider the potential for measurement error because we measure recidivism with rearrests. That is, the base rate of a particular minority group could be upwards biased if policing tends to be more strict with this protected group, causing biased labels.[6] Despite these data limitations the overall setting remains relevant

[6] Note that the studied decision-making setting does not suffer from the selective labels problem, i.e. the case where the observed outcome depends on the decision of the case workers. The selective labels problem appears in the context of pre-trial bail decision making (Kleinberg et al. 2017), in which we cannot observe potential breach of bail on those who are denied bail. However, in this case the release in 2010 does not depend on the risk assessment of the decision maker, and recidivism behaviour can be observed for every defendant in the sample.

as long as it occurs similarly in real world settings. This has important policy implications which we discuss further below. To that extent, it is important to understand the underlying mechanisms that eventually lead to group unfairness as structured human decision-making processes are replaced (or just even extended) by ML applications.

In the following, we discuss the two potential sources of group unfairness that are investigated in the current analysis: unequal base rates and the use of features sets highly correlated with the protected features. Subsequently, we discuss more general limitations of unfairness mitigation techniques. We conclude this section with a discussion of the policy implications.

7.1 Sources of group unfairness

Predicting recidivism is not a trivial problem and no ML method achieves perfect accuracy. The selected top two ML models trade off predictive performance for group fairness. One explanation can be that the base rates (i.e. the prevalence of recidivism) differ between various groups, as seen in the top panel of Table 2. Another explanation could be that we use input features, such as demographics and criminal history, which are correlated to protected features. This was the case of "Non-SAVRY" and "All." We discuss each explanation subsequently.

7.1.1 Unequal base rates

The literature has shown extensively that base rates substantially affect the outcomes of group-fairness measures (Berk et al. 2017). Most prominently, Kleinberg et al. (2016) and Chouldechova (2017) show that, when base rates differ, it is mathematically impossible to fulfill multiple measures of group fairness simultaneously. In this dataset, the recidivism rate for men is 40%, while the recidivism rate for women is 20%. Also, the recidivism rate for foreigners is 46%, (specifically for Latin Americans it is 45% and for Maghrebi 55%), while for Spaniards it is 32%. This is emphasized by the differences in the group composition between recidivists and non-recidivists. In detail, Table 2 shows that compared to non-recidivists, recidivists are significantly more likely to be male, foreign and specifically Maghrebi or Latin American but less likely to be female or Spaniards. ML methods pick up on these empirical correlations when producing predictions on recidivism. Under these conditions, it is clearly difficult to achieve similar classification rates for both groups.

To test the role of unequal base rates in the unfair classification, we oversample from the under-represented group until its base rate matches the base rate of the reference group. Then, we repeat the accuracy, fairness and interpretability analysis under equalized base rates with respect to sex and nationality. The results are presented in Sect. 5.2. In fact, the results show that in the case of disparity between foreigners and Spanish nationals and between males and females, the method leads to a level change of FPRD as well as FNRD to presumably acceptable levels (between 0.8 and 1.2) showing that unequal base rates do in fact contribute to issues in group fairness metrics. However, the accompanying reduction in accuracy reveals that part

of this equalization in error rates is achieved through increases in error rates for one group. In the case of increased FPR, this increases harms for the affected group as it could lead to increased wrongful convictions. Similarly, a higher FNR negatively affects public safety as it may lead to increased wrongful releases. This shows that improving on fairness metrics can lead to increased harms at other places.

Furthermore, we see in Fig. 2 the case of "Non SAVRY" and *mlp* that EBR is not a guaranteed solution to unfair classification as the measure could overcorrect from a low FPRD to a FPRD that is too high. A reason for this could be the very small sample size of female defendants ($N_{female} = 108$) as in small samples little changes in input can lead to large changes in marginal outcomes. Moreover, the very small changes due to EBR when using the "SAVRY" feature set reveals further limits to how much of the unfair classification base rates can explain. Similarly, Fig. 5 reveals that equalised base rates on a larger group (in this case foreigner status) does not address the issue for smaller subgroups, such as different nationalities among foreigners. This confirms the finding in Chouldechova (2017).

A closer look at the interpretability results in Tables 3 and 4 reveals that compared to the unconstrained dataset, equalised base rates do not significantly affect the order of feature importance when predicting. In fact, the protected feature along which the base rates are equalised becomes the most important. That is, instead of reducing the role of the protected feature, a supposed mitigation measure could aggravate the situation otherwise. In addition, this suggests that not only base rates but also confounders, i.e., features that are correlated with the protected feature as well as the outcome play a role in explaining unfair algorithmic decision making.

7.1.2 Demographic and criminal history input features

Besides unequal base rates another, more obvious explanation for unfair classification is the influence of input features correlated with the protected features. This issue has been discussed extensively in the literature (Barocas and Selbst 2016; Hardt et al. 2016) and we discuss a more ad-hoc mitigation measure, i.e. removing these features from the input, in Sect. 7.2. The interpretability results in Tables 3 and 4 reveal that we face the same issue since the protected features, "sex" and "foreigner," are the most important in the unconstrained dataset.

In order to investigate the part that the contribution of protected input features can have in explaining unfair classification, we repeat the same analysis for a third time with the LFR mitigation measure.

This measure removes the explanatory power of protected features as well as the correlation of other features with the protected features and the outcome in prediction. As per the construction of this issue, it cannot explain unfair classification based on the "SAVRY" feature set as this set does not contain the actual protected features. However, the Figs. 2, 3 and 5 show that this measure works very similar to the "ERB" measure in terms of FPRD and FNRD. It explains the fairness metrics disparities for the groups that it is addressing without affecting the disparities in smaller subgroups.

The interpretability results in Tables 3 and 4 show that this measure further addresses the problem of confounding as with "LFR" the protected features as well

as the correlated static and demographic features are being outranked by dynamic SAVRY features in explaining recidivism.

Finally, it should be highlighted that both unequal base rates as well as the influence of protected features in input explain parts of unfair classification outcomes. However, neither method guarantees a decision making system that does not yield any disparity.

7.2 Mitigation

After having discussed the impact of two specific sources of unfairness by applying mitigation measures, we move to a general discussion on limitations of mitigation methods for specific unfairness metrics.

7.2.1 "Color blind" methods

Removing protected attributes from the input, thus creating a "color-blind" model, is often suggested as a potential solution to fairness-related issues in ML algorithms. In this case this would not affect the results of "SAVRY" and "SAVRY ML," as none of these settings use protected attributes, but it could have an effect in the case of "Non-SAVRY" and "All ML." The idea behind this solution is that, in order to not discriminate against men/women or foreigners, the machine (or human) decision maker should not take these sensitive attributes into consideration (i.e. should remain blind to these attributes). However, in general, avoiding disparate treatment of different groups (by not including respective features in training) does not guarantee the absence of discriminating outcomes for affected groups. As a matter of fact, these two fairness goals (fair treatment and fair outcomes) often trade each other off (Corbett-Davies and Goel 2018; Žliobaitė and Custers 2016; Lipton et al. 2018). The core problem is that removing the protected feature from the input does not remove the correlation of the protected feature with the outcome. The remaining features that are correlated with the sensitive attributes will pick up on this correlation (Hardt et al. 2016; Barocas and Selbst 2016). Even a drastic reduction of the algorithm to only two input features in training (age, and number of previous crimes) can pick up on these correlations that make an algorithm produce unfair results (Dressel and Farid 2018; Jung et al. 2017). In other words, if the protected attributes and the outcome of recidivism correlate (see Table 2), this correlation will not disappear if you remove the protected features. Besides, the results for "Savry ML" in Figs. 3 and 5 for Maghrebi show that just using ML methods, without including protected attributes, produces unfair outputs as well.

7.2.2 Different models or different thresholds

As a consequence of the issues with color blind algorithms, other unfairness-mitigating methods arise that deal with the problem of disparate outcomes. The underlying idea of these methods is that instead of avoiding the sensitive attributes in training,

we consider them in particular by using different prediction models or different risk thresholds for members of different protected groups. For instance, treating women like men, and ignoring findings that show that in juvenile justice females react differently from males to specific risk factors (Wright et al. 2007), can cause considerable harm to women (Skeem et al. 2016). In contrast, sacrificing equal treatment for the benefit of equal outcomes can be interpreted as holding different groups to different standards, which is equally hard to justify in a very sensitive context such as criminal justice, as has been shown in the US Supreme Court lawsuit Ricci v. DeStefano (United States Supreme Court 2009). In addition, equalizing outcomes for one protected group does not guarantee equally fair outcomes for respective subgroups (Chouldechova 2017). In any case, adjusting the outcomes of a decision process, potentially at the expense of accuracy, produces public costs that will either have to be paid in the form of sacrificing public safety or making innocent people subject to costly criminal justice interventions (Corbett-Davies et al. 2017).

7.2.3 Algorithm adjustments

In-processing methods to achieve fairness in machine learning are the ones that produce "fairness-aware" models by adjusting the objective that is optimized during the training phase, typically by adding additional fairness constraints. In this case for instance, one could introduce an extra term that penalizes classifiers that yield different error rates (Hardt et al. 2016; Zafar et al. 2017; Agarwal et al. 2018). The results would be similar to applying different thresholds to different groups. Therefore, adjusting a classifier without understanding the underlying issues that cause unfairness can have similar problems as the ones discussed in the previous section. For instance, in order to address unbalanced false positive rates, the adjusted classifier would reduce the risk-assessment of the high-risk group at random, causing releases of individuals of uncertain risks. The consequence of this could be a bad track records for minority groups, creating downwards dynamics in terms of risk assessment for the respective group (Kallus and Zhou 2018; Liu et al. 2018). Therefore, understanding the causes of recidivism and taking into account the feedback of the implementation of the algorithm in the real world is more beneficial than bruteforcing the decision-making system to fit certain fairness metrics.

7.3 Policy implications

The findings of our study have considerable policy implications. First, the application domain for ML methods, criminal recidivism prediction, can be considered a high-risk application. More specifically, the White Paper on AI of the European Commission (EC) (COM 2020) defines the risk level of AI applications based on two criteria: the first one depends on the sector of exploitation (e.g. health, transport and public sectors are considered to be of high-risk) of AI and the second one on the particular AI application. Criminal risk recidivism fits both criteria. It belongs to the public sector, which includes judicial decisions, a sector where significant risks can

be expected to occur. In addition, recidivism assessment is linked to legal or similarly significant effects for the rights of an individual.

Apart from the relevance of the selected use case, the reported accuracy gains indicate that the potential for ML to improve decision making in criminal justice cannot be ignored. However, at the same time the findings on group fairness highlight that the involvement of ML in decision making processes still requires human oversight to avoid discrimination and ensure that relevant (not necessarily highly correlated) risk factors are considered in decision processes related to recidivism. Thus, policy makers have to ensure that we achieve decision making mechanisms in criminal justice that reap the benefits that come from accuracy gains but also avoid the dangers of algorithmic discrimination.

The EC White Paper discusses options to navigate this trade-off between the promotion and uptake of AI and the associated risks with certain uses of AI. This also involves the creation of an "ecosystem of trust," i.e., a policy mechanism that ensures that AI systems comply with the rules that protect fundamental rights and consumers' rights, in particular for AI systems that pose a high risk. In line with this, our study highlights that bias and discrimination can be present in human decision-making processes as well as decision making processes supported by ML models. Our focus on the group fairness issues of the ML system relates to the fact that the same bias that is present in human decision making could have much larger effects in an ML-supported process, due to the scaling effects of automated systems. Moreover, these processes, if not tested properly, remain hidden due to the fact that algorithms may be opaque, complex, unpredictable and partially autonomous (COM 2020).

The methods presented in this study contribute to the understanding of the mechanisms and consequences of ML driven decision making and consequently provide a way to deal with issues of opaqueness and check compliance with rules of existing EU law meant to protect fundamental rights. In the context of the future regulatory framework for AI in Europe, the White Paper discusses types of legal requirements that can be imposed on relevant actors in the context of high-risk AI applications. The requirements discussed in the White Paper relate to those addressed in our study, including the evaluation and consideration of bias in datasets used for training AI systems, the evaluation of the performance of AI systems (here measured as accuracy) and limitations. Further requirements involve the naming of necessary conditions for the correct functioning of AI systems, transparency about the information provided to final users, and enabling human oversight before, after and during the AI operation.

Our study shows that ML can help by better informing experts in the correct weighting of different case-related information. At the same time a possible automation of unfair decision making processes due to certain biases pose a high risk to persons affected by the system. Ensuring human oversight through a machine-human interactive decision framework, i.e. a human-in-the-loop framework could be an appropriate response. Yet, further research on the dynamics of human-in-the-loop frameworks is necessary to inform the design and policy making of such decision processes.

8 Conclusions and future work

In this paper we study the causes of disparity between groups of different sex and nationality, equal base rates, input features, and ML algorithms. We discover that unequal base rates lead to disparity between several groups on the considered group fairness metrics. However, base rates alone can not fully explain the observed disparity. Additional disparity between groups is due to the nature of input features and the type of ML algorithm used. With respect to the input features, ML models exhibit more disparity when using demographic and criminal history features and less disparity with risk items taken from a risk assessment tool, SAVRY. These risk items contain a high proportion of dynamic features, which are not highly correlated to static features such as sex or nationality.

We observe that the usage of mitigation methods may be problematic. To that extent, the importance of the SAVRY risk items increases when we use a mitigation method, LFR (Zemel et al. 2013), that encodes the input features to remove the correlations with the protected features. However, in this case we notice a drop in predictive performance. Besides the increase of error rates, mitigation along a group fairness metric with respect to a given protected feature may not resolve group fairness issues for another protected feature. In addition, some subgroups may be discriminated against while others positively discriminated causing an average "fair" outcome for the whole group. We provide a comprehensive discussion on the impact of unfairness mitigation methods.

In terms of feature importance, we compare dynamic and static features and note that the more a model relies on static features, the less group fairness it has. The dynamic features are related to items that a defendant can change and reflect the validity of the treatment for the defendant at the moment of the SAVRY evaluation. A judicial system using a ML model relying solely on items a defendant can not change, can be considered coercive, rather than transforming.

Our study contributes to understanding the extent of the changes in criminal justice outcomes that come with machine learning involvement in criminal risk-assessment. Our comparison in terms of predictive power and relevant group fairness metrics between a "human-in-the-loop" empirically informed assessment and a pure statistical risk-assessment, sheds light on this question and informs policy makers about some of the risks associated with the implementation of ML in criminal justice processes.

In our study we use SAVRY data from Catalonia between 2002 and 2010 with recidivism observed in 2013 and 2015. The conclusions of this study should be further assessed on recent data from Catalonia or other regions/countries which use SAVRY and further compared with other risk assessment tools.

Acknowledgements Funding was provided by Joint Research Centre (Grant No. HUMAINT).

Appendix

See Tables 2, 3 and 4.

Table 2 Descriptive statistics

	Base rate	Not recidivated		Recidivated		Difference	
		Mean	Std.Dev.	Mean	Std.Dev.	Diff	Std.Dev.
Protected features							
Male	40.03%	0.839	0.368	0.931	0.253	0.093***	0.021
Female	20.37%	0.161	0.368	0.069	0.253	− 0.093***	0.021
Spanish	32.06%	0.667	0.471	0.523	0.499	− 0.143***	0.035
Foreign	46.22%	0.333	0.471	0.477	0.499	0.143***	0.035
Latin American	44.52%	0.161	0.368	0.215	0.411	0.054*	0.028
Maghrebi	55.12%	0.107	0.309	0.218	0.413	0.111***	0.027
European	32.35%	0.043	0.203	0.034	0.182	− 0.009	0.013
Other	20.00%	0.022	0.148	0.009	0.096	− 0.013	0.008
Static features (not SAVRY)							
Age maincrime		16.011	1.009	15.720	1.060	− 0.292***	0.074
Prior crimes		0.700	0.458	0.863	0.344	0.163***	0.028
Prior crimes frequency							
1 incident		0.586	0.493	0.604	0.489	0.018	0.035
2 incidents		0.243	0.429	0.215	0.411	− 0.028	0.030
3 or more incidents		0.170	0.376	0.181	0.385	0.010	0.027
Maincrime violent		0.609	0.488	0.611	0.488	0.002	0.034
Maincrime category							
Nonviolent against property		0.251	0.434	0.265	0.441	0.014	0.031
Violent against property		0.264	0.441	0.293	0.455	0.029	0.032
Against persons		0.345	0.475	0.318	0.466	− 0.027	0.033
Other		0.140	0.347	0.125	0.330	− 0.016	0.024
Maincrime program sentence							
Technical sentence		0.060	0.237	0.262	0.440	0.202***	0.027
Mediation and reparation		0.021	0.142	0.025	0.156	0.004	0.011
Enforcement measure		0.919	0.272	0.713	0.452	− 0.206***	0.028
Internment (no probation)		0.142	0.349	0.265	0.441	0.122***	0.029
Days to sentence start		481.803	269.611	364.579	276.429	− 117.224***	19.343
Sentence duration (days)		285.058	190.887	235.536	233.089	− 49.522***	15.411

Table 2 (continued)

	Base rate	Not recidivated		Recidivated		Difference	
		Mean	Std.Dev.	Mean	Std.Dev.	Diff	Std.Dev.
Year of main crime							
2006 or earlier		0.064	0.244	0.069	0.253	0.005	0.018
2007/2008		0.672	0.469	0.449	0.497	− 0.224***	0.034
2009/2010		0.264	0.441	0.483	0.5	0.219***	0.034
SAVRY							
Final expert evaluation		0.315	0.330	0.530	0.359	0.215***	0.025
SAVRY summary scores							
Total (automatic)		14.150	8.424	18.262	8.712	4.112***	0.608
Historical factors		5.762	3.941	7.084	3.890	1.322***	0.276
Social factors		3.803	2.562	5.050	2.701	1.246***	0.187
Individual factors		4.584	3.438	6.128	3.703	1.543***	0.255
Protective factors		2.152	1.861	2.956	1.877	0.805***	0.132
SAVRY 24 risk items							
Previous violent offenses		0.428	0.404	0.536	0.402	0.108***	0.028
History nonviolent offending		0.344	0.368	0.472	0.386	0.128***	0.027
Early violence (below 14)		0.182	0.326	0.254	0.364	0.072***	0.025
Past intervention failures		0.184	0.317	0.280	0.365	0.097***	0.025
Self-harm/suicide history		0.099	0.248	0.132	0.268	0.033*	0.018
Home violence		0.254	0.383	0.263	0.379	0.009	0.027
Childhood mistreatment		0.239	0.352	0.290	0.379	0.051**	0.026
Criminal parent/caregiver		0.163	0.310	0.196	0.347	0.033	0.024
Childhood care giving disruption		0.285	0.389	0.335	0.402	0.050*	0.028
Poor school achievement		0.705	0.351	0.783	0.319	0.078***	0.023
Delinquency in peer group		0.364	0.362	0.525	0.365	0.161***	0.026
Rejection by peer group		0.110	0.230	0.154	0.282	0.045**	0.019
Stress and poor coping		0.390	0.350	0.438	0.373	0.048*	0.026
Poor parental management		0.456	0.351	0.578	0.368	0.122***	0.026
Lack of personal/social support		0.286	0.340	0.419	0.380	0.133***	0.026
Community disorganization		0.297	0.381	0.411	0.394	0.114***	0.027

Table 2 (continued)

	Base rate	Not recidivated		Recidivated		Difference	
		Mean	Std.Dev.	Mean	Std.Dev.	Diff	Std.Dev.
Negative attitudes		0.279	0.304	0.397	0.326	0.118***	0.022
Risk taking/impulsive		0.369	0.343	0.469	0.349	0.100***	0.025
Substance abuse		0.317	0.347	0.416	0.371	0.098***	0.026
Anger management issues		0.334	0.340	0.410	0.341	0.075***	0.024
Low empathy		0.282	0.326	0.393	0.342	0.111***	0.024
Attention deficit hyperactivity		0.202	0.297	0.262	0.323	0.059***	0.022
Poor compliance		0.202	0.294	0.313	0.348	0.111***	0.023
Low commitment to school		0.306	0.366	0.405	0.402	0.099***	0.027
SAVRY 6 protective factors							
Pro-social activities		0.485	0.500	0.333	0.471	− 0.152***	0.034
Pro-social support		0.700	0.458	0.536	0.499	− 0.165***	0.034
Pro-social support (by adult)		0.676	0.468	0.592	0.491	− 0.084**	0.034
Positive attitude		0.837	0.369	0.741	0.438	− 0.096***	0.029
High interest in school/work		0.669	0.471	0.508	0.500	− 0.161***	0.035
Positive/resilience characteristics		0.481	0.500	0.333	0.471	− 0.148***	0.034
SAVRY 5 factors model							
Antisocial behavior		0.548	0.449	0.747	0.445	0.199***	0.032
Family dynamics		0.470	0.527	0.542	0.558	0.072*	0.039
Personality		0.561	0.422	0.720	0.446	0.159***	0.031
Social support		0.540	0.425	0.738	0.468	0.198***	0.032
Treatment susceptibility		0.565	0.351	0.727	0.377	0.162***	0.026
N		534		321			

Descriptive statistics of input features by recidivism status. Not displayed: province of residence, province of sentencing

Standard errors in parentheses. *, ** and ***Denotes significance level of 10%, 5% and 1%, respectively

Table 3 Feature importance for predicting recidivism, *logit* on "All" feature set

	Mean	Std
Foreigner	242.83	13.82
Sex*	233.00	22.87
National group*	221.75	13.98
Province of residence*	192.16	13.21
No. of maincrimes*	167.58	9.87
Age maincrime*	167.92	11.23
Province of execution*	165.53	9.38
Maincrime program sentence*	163.08	7.74
No. of prior crimes*	167.63	12.34
Maincrime program duration*	160.10	6.53
EBR on attribute "sex"		
Sex*	210.16	10.90
Foreigner*	189.62	5.43
Maincrime violent*	174.24	11.05
No. of maincrimes*	170.70	8.58
No. of prior crimes*	172.40	10.31
Province of execution*	167.44	8.68
Maincrime category*	171.52	13.91
Maincrime program*	164.10	7.60
Maincrime program duration*	160.09	6.51
Says between crime and program*	158.18	6.79
EBR on attribute "nationality"		
Foreigner*	228.52	18.53
National group*	213.45	12.75
Sex*	218.75	21.10
Province of residence*	187.01	13.35
No. of prior crimes*	155.59	8.86
No. of maincrimes*	155.28	9.76
Province of execution*	152.80	8.04
Maincrime program*	152.35	8.99
Age maincrime*	157.60	15.19
Maincrime program duration*	149.42	7.44
LFR on attribute "sex"		
Positive attitude	123.33	23.26
Antisocial behaviour	122.28	22.96
Family dynamics	122.29	23.03
Attention deficit hyperactivity	121.87	22.83
Expert score	122.21	23.26
Positive/resilience characteristics	122.08	23.18
Community disorganization	122.09	23.36
Pro-social support (by adult)	121.15	22.56
Stress and poor coping	122.10	23.58
Poor parental management	122.15	23.68

Table 3 (continued)

	Mean	Std
LFR on attribute "nationality"		
Antisocial behaviour	123.50	18.96
Attention deficit hyperactivity	123.29	18.85
Anger management issues	122.70	18.39
Treatment susceptibility	122.58	18.30
High interest in school/work	123.38	19.20
Pro-social support (by adult)	122.10	18.13
Substance abuse	122.50	18.64
Personality	122.54	18.73
Family dynamics	122.38	18.66
Positive/resilience characteristics	121.59	17.95

*Denotes Non-SAVRY features

Table 4 Feature importance for predicting recidivism, *mlp* on "All" feature set

	Mean	Std
Sex*	155.03	13.99
Foreigner*	164.13	29.42
National group*	137.40	18.22
Maincrime program duration*	123.67	15.05
Days of program*	123.58	15.04
SAVRY program*	121.38	13.06
Days between crime and program*	124.14	16.05
Maincrime program*	124.25	16.22
Expert score	121.61	14.04
Province of residence*	120.44	13.38
EBR on attribute "sex"		
Sex*	158.61	16.44
Foreigner*	160.53	24.77
National group*	124.58	13.93
SAVRY program*	120.64	10.42
No days of program*	120.64	10.75
SAVRY total score	121.07	11.22
Historical factors	120.18	10.47
Individual factors	120.43	11.01
Expert evaluation	121.20	11.87
Social factors	120.68	11.38
EBR on attribute "nationality"		
Foreigner*	165.33	13.89
Sex*	155.99	12.03
National group*	135.81	13.97
No. of maincrimes*	117.60	8.41
Province of residence*	118.45	10.26
Maincrime program*	116.46	9.82
No. of prior crimes*	117.39	10.80
Maincrime violent*	116.22	10.59
Maincrime category*	115.52	10.04
Province of execution*	115.94	10.63
LFR on attribute "sex"		
High interest in school/work	114.39	24.57
Family dynamics	114.09	25.02
Social support	114.65	25.70
Stress and poor coping	113.05	24.17
Personality	113.67	24.88
Anger management issues	113.08	24.44
Risk-taking impulsivity	113.84	25.30
Positive/resilience characteristics	113.61	25.08
Pro-social activities	113.22	24.77
Lack of personal/social support	114.37	25.94

Table 4 (continued)

	Mean	Std
LFR on attribute "nationality"		
Antisocial behaviour	115.08	17.83
Low commitment to school	114.99	17.79
Positive/resilience characteristics	115.89	18.81
Family dynamics	115.72	18.82
Pro-social support (by adult)	115.49	18.58
Low empathy	113.45	16.60
Poor compliance	115.05	18.26
Social support	114.48	18.11
High interest in school/work	114.83	18.54
Attention deficit hyperactivity	114.25	17.97

*Denotes Non-SAVRY features

References

Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H (2018) A reductions approach to fair classification. arXiv preprint arXiv:1803.02453

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. ProPublica, May, 23 2016

Barocas S, Selbst A (2016) Big data's disparate impact. Calif Law Rev 104(1):671–729. https://doi.org/10.15779/Z38BG31

Bartel PA, Forth AE, Borum R (2003) Development and concurrent validation of the Structured Assessment for Violence Risk in Youth (SAVRY). Unpublished manuscript

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2017) Fairness in criminal justice risk assessments: the state of the art, pp 1–42. arXiv:1703.09207

Björkqvist K, Lagerspetz K, Kaukiainen A (1992) Do girls manipulate and boys fight? Developmental trends in regard to direct and indirect aggression. Aggress Behav 18(2):117–127

Blanch M, Capdevila M, Ferrer M, Framis B, Ruiz U, Mora J, Batlle A, López B (2017) La reincidència en la justícia de menors. CEJFE

Borum R, Bartel P, Forth A (2003) Manual for the structured assessment of violence risk in youth. University of South Florida, Tampa

Borum R, Lodewijks H, Bartel PA, Forth AE (2011) Structured assessment of violence risk in youth (SAVRY). In: Handbook of violence risk assessment. Routledge, pp 73–90

Chevalier CS (2017) The association between structured professional judgment measure total scores and summary risk ratings: implications for predictive validity, Ph.D. thesis

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163

Chouldechova A, Roth A (2018) The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810

Christin A, Rosenblat A, Boyd D (2015) Courts and predictive algorithms. Data & Society. Retrieved from http://www.datacivilrights.org

COM (2020) White paper on artificial intelligence—a European approach to excellence and trust. Tech. rep., European Commission (2020)

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv preprint arXiv:1808.00023

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 797–806

Cowgill B (2018) The impact of algorithms on judicial discretion: Evidence from regression discontinuities. Technical Report. Working paper

DeMichele M, Baumgartner P, Wenger M, Barrick K, Comfort M, Misra S (2018) The public safety assessment: a re-validation and assessment of predictive utility and differential prediction by race and gender in Kentucky. Available at SSRN 3168452
Desmarais SL, Johnson KL, Singh JP (2016) Performance of recidivism risk assessment instruments in us correctional settings. Psychol Serv 13(3):206
Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580
Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton E, Roth D (2018) A comparative study of fairness-enhancing interventions in machine learning. arXiv preprint arXiv :1802.04422
Goel S, Shroff R, Skeem JL, Slobogin C (2018) The accuracy, equity, and jurisprudence of criminal risk assessment. Equity Jurisprud Crim Risk Assess
Green B (2018) 'fair'risk assessments: a precarious approach for criminal justice reform. In: 5th Workshop on fairness, accountability, and transparency in machine learning
Green B, Chen Y (2019) Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments. In: ACM conference on fairness, accountability, and transparency
Green B, Hu L (2018) The myth in the methodology: towards a recontextualization of fairness in machine learning. In: Proceedings of the machine learning: the debates workshop
Guy LS (2008) Performance indicators of the structured professional judgment approach for assessing risk for violence to others: a meta-analytic survey. Ph.D. thesis, Dept. of Psychology-Simon Fraser University
Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp 3315–3323
Hilterman EL, Nicholls TL, van Nieuwenhuizen C (2014) Predictive validity of risk assessments in juvenile offenders: comparing the SAVRY, PCL:YV, and YLS/CMI with unstructured clinical assessments. Assessment 21(3):324–339. https://doi.org/10.1177/1073191113498113
Jung J, Concannon C, Shroff R, Goel S, Goldstein DG (2017) Simple rules for complex decisions. Available at SSRN 2919024
Kallus N, Zhou A (2018) Residual unfairness in fair machine learning from prejudiced data. arXiv preprint arXiv:1806.02887
Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. In: Proceedings of innovations in theoretical computer science (ITCS)
Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human decisions and machine predictions. Q J Econ 133(1):237–293
Langley P, Simon HA (1995) Applications of machine learning and rule induction. Commun ACM 38(11):54–64
Lipton Z (2016) The mythos of model interpretability. arXiv preprint arXiv:1606.03490
Lipton Z, McAuley J, Chouldechova A (2018) Does mitigating ML's impact disparity require treatment disparity? In: Advances in neural information processing systems, pp 8125–8135
Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M (2018) Delayed impact of fair machine learning. arXiv preprint arXiv:1803.04383
Lum K, Isaac W (2016) To predict and serve? Significance 13(5):14–19
Miron M, Tolan S, Gómez E, Castillo C (2020) Addressing multiple metrics of group fairness in data-driven decision making. arXiv preprint arXiv:2003.04794
Monahan J, Skeem JL (2016) Risk assessment in criminal sentencing. Annu Rev Clin Psychol 12:489–513
Narayanan A (2018) 21 fairness definitions and their politics. Tech. rep., Conference on fairness, accountability and transparency 2018, tutorial
Northpoint, Inc. (2012) Compas risk and need assessment system. Northpoint, Inc, Tech. rep
Olver ME, Stockdale KC, Wormith J (2009) Risk assessment with young offenders: a meta-analysis of three assessment measures. Crim Just Behav 36(4):329–353
Ortega-Campos E, García-García J, De la Fuente-Sánchez L, Zaldívar-Basurto F (2020) Predicting risk of recidivism in spanish young offenders: comparative analysis of the SAVRY and YLS/CMI. Psicothema 32(2):221–228
Pasquale F (2015) The black box society. Harvard University Press, Harvard
Perrault RT, Vincent GM, Guy LS (2017) Are risk assessments racially biased? Field study of the SAVRY and YLS/CMI in probation. Psychol Assess 29(6):664

Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1135–1144

Robert C (2014) Machine learning, a probabilistic perspective, pp 62–63

Rowe DC, Vazsonyi AT, Flannery DJ (1995) Sex differences in crime: Do means and within-sex variation have similar causes? J Res Crime Delinq 32(1):84–100

Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, Rodolfa KT, Ghani R (2018) Aequitas: a bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577

Singh JP et al (2014) International perspectives on the practical application of violence risk assessment: a global survey of 44 countries. Int J Forensic Ment Health 13(3):193–206

Skeem J, Monahan J, Lowenkamp C (2016) Gender, risk assessment, and sanctioning: the cost of treating women like men. Law Hum Behav 40(5):580

Tolan S (2018) Fair and unbiased algorithmic decision making: current state and future challenges. JRC Digit Econ Work Pap 10

Tolan S, Miron M, Gómez E, Castillo C (2019) Why machine learning may lead to unfairness: evidence from risk assessment for juvenile justice in Catalonia

United States Supreme Court (2009) Ricci v. DeStefano, 557 U.S. 557. https://supreme.justia.com/cases/federal/us/557/557/

Wright EM, Salisbury EJ, Van Vanik P (2007) Predicting the prison misconducts of women offenders: the importance of gender-responsive needs. J Contemp Crim Just 23(4):310–340

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on World Wide Web. International World Wide Web Conferences Steering Committee, pp 1171–1180

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: International conference on machine learning, pp 325–333

Žliobaitė I, Custers B (2016) Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artif Intell Law 24(2):183–201