Fairness Machine:
A Web Application for Studying Machine Learning Fairness Scenarios

JIANQIU BAI, University of California, Santa Cruz
ADVISOR: YANG LIU*, University of California, Santa Cruz
ADVISOR: JAMES DAVIS, University of California, Santa Cruz
ADVISOR: NARGES NOROUZI, University of California, Santa Cruz

Machine learning has been utilized in various decision-making situations, however, the fairness of such machine learning algorithms can be hardly defined. Though current machine learning models have high accuracy on training and testing dataset, the results still have biases. For instance, a well-known recidivism risk evaluation software, COMPAS, has been found out that white defendants are stated to have low risk more often than black defendants in general [1].

Previously, (Liu and et. al. [3]) have investigated on whether the predictions of trained models are acceptable to the public, on the scenario of loan allocation. This report proposes a crowdsourcing platform for investigating whether a machine learning algorithm is fair for the scenario of **bail judgement**. Three machine learning models are pre-trained to predict the recidivism risk percentage according to the COMAPAS dataset. A web application is then build as an experimental tool to collect clients' perspectives on, given a certain designed scenario, which model of three they think has the most valid prediction. In terms of such responses, it could be used for evaluating whether our pre-trained models have biases or not.

Additional Key Words and Phrases: Machine Learning Fairness

ACM Reference Format:
Jianqiu Bai, Advisor: Yang Liu, Advisor: James Davis, and Advisor: Narges Norouzi. 2020. Fairness Machine: A Web Application for Studying Machine Learning Fairness Scenarios. 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the desire of auto service, machine learning has been utilized to make decision for human in various areas. Though machine learning algorithms provide high accuracy in most training cases, biases are still found from such algorithms used in practical cases, which is caused by irrelevant attributes taken into account and relevant attributes ignored in machine learning training phase [3]. Therefore, in a context of decision-making using machine learning, a perfect model needs to restrict any bias based on inherent and acquired characteristics of the target, which is known as the study of machine learning fairness [2].

This report introduces a framework that investigates the fairness of algorithmic decisions, especially for bail judgement scenario. The following section provides related work that helps readers to understand current issues on predicting future criminals: trained models are biased. Section 3 proposes our approach on evaluating fairness

Authors' addresses: Jianqiu Bai, jbai14@ucsc.edu, University of California, Santa Cruz; Advisor: Yang Liu, yangliu@ucsc.edu, University of California, Santa Cruz; Advisor: James Davis, davis@soe.ucsc.edu, University of California, Santa Cruz; Advisor: Narges Norouzi, nanorouz@ucsc.edu, University of California, Santa Cruz.

XXXX-XXXX/2020/6-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

of the trained models, by building a web survey application. The website interface is then introduced in Section 4. Future work are concluded in Section 5.

2 RELATED WORK

The software used for predicting future criminals is biased on race

(Angwin and et. al. [1]) stated that there are significant racial disparities for future crime assessment using COMPAS. In terms of ProPublica's study of risk scores, black defendants are most likely to be labelled wrongly as future criminals as at almost twice the rate as white defendants. To further verify their conclusion, they have proposed four show cases: two petty theft arrests, two drug possession arrests, two DUI arrests, two shoplifting arrests. For each case, there is a pair of black and white defendants who are judged for the same criminal. However, the algorithm always provides higher risk scores for black defendants than white defendants against the fact. Further statistical analysis is also supported to emphasize that such instances are not case-special.

3 METHOD

Based on Section 2, the question is addressed on how can we determine a recidivism risk evaluation model is fair.

3.1 Pre-trained Models on COMPAS Dataset

We first trained three models using different algorithms according to raw COMPAS dataset as shown in Figure **??**. To achieve relevant fair models, not all attributes are utilized but we have filtered with:

- sex
- age
- age cat: age range (<25, 25-45, >45)
- decile score: recidivism risk score
- text score: character describing level of decile score (low, media, high)
- priors count: previous criminal count
- two year recid: a boolean value indicating where this person is committed with future crimes
- in jail time
- out jail time

id	name	first	last	compas_scre	sex	dob	age	age_cat	race	juv_fel_coun	decile_score	juv_misd_co	juv_other_co	priors_coun
8902	samuel louis	samuel	louis	1/22/13	Male	7/5/92	23	Less than 25	Other	0	3	0	0	[illegible]
8908	leverdieu jas	leverdieu	jasmin	12/1/13	Male	7/11/77	38	25 - 45	African-Ame	0	4	0	0	[illegible]
8945	ronald smith	ronald	smith	1/15/13	Male	11/3/94	21	Less than 25	African-Ame	0	7	0	0	[illegible]
8948	mikemsonn	mikemsonn	michaud	4/26/13	Male	8/3/89	26	25 - 45	Other	0	2	0	0	[illegible]
8949	sunil jagpal	sunil	jagpal	1/27/13	Female	1/20/77	39	25 - 45	African-Ame	0	6	0	0	[illegible]
8960	dominic zack	dominic	zackery	3/15/13	Male	9/12/79	36	25 - 45	African-Ame	0	4	0	0	[illegible]
8972	shanaice pin	shanaice	pinkney	2/5/14	Female	9/5/88	27	25 - 45	African-Ame	0	6	0	0	[illegible]
8997	john kemp	john	kemp	9/10/13	Male	2/10/61	55	Greater than 45	African-Ame	0	1	0	0	[illegible]
9020	gabriel magr	gabriel	magnone	7/1/14	Male	3/20/86	30	25 - 45	Caucasian	0	5	1	0	[illegible]
9022	randail spen	randail	spencer	11/20/13	Male	2/7/84	32	25 - 45	African-Ame	0	4	0	0	[illegible]
9045	corey smagli	corey	smaglik	12/10/13	Male	8/19/69	46	Greater than 45	Caucasian	0	5	0	0	[illegible]
9056	leroy farley	leroy	farley	3/4/14	Male	7/26/52	63	Greater than 45	African-Ame	0	6	0	0	[illegible]
9079	jerko delacru	jerko	delacruz	8/10/13	Male	10/13/93	22	Less than 25	Hispanic	0	4	0	0	[illegible]
9080	juan guevara	juan	guevara	4/22/13	Male	10/16/78	37	25 - 45	Caucasian	0	1	0	0	[illegible]
9089	michael pres	michael	preston	8/16/13	Male	7/8/78	37	25 - 45	Caucasian	0	2	0	0	[illegible]
9165	eric mckenzi	eric	mckenzie	8/26/13	Male	3/15/90	26	25 - 45	African-Ame	0	9	0	0	[illegible]
9169	jean georges	jean	georges	5/21/13	Male	5/9/82	33	25 - 45	Other	0	1	0	0	[illegible]
9172	lorenzo mcki	lorenzo	mckinney	10/29/13	Male	6/17/94	21	Less than 25	African-Ame	0	7	0	0	[illegible]
9178	aleshea king	aleshea	king	2/9/13	Female	4/5/83	33	25 - 45	African-Ame	0	4	0	0	[illegible]
9226	gerard spelm	gerard	spelman	8/21/13	Male	4/5/87	29	25 - 45	Caucasian	0	1	0	0	[illegible]
9227	jeremy peleh	jeremy	pelehowski	3/28/13	Male	3/24/85	31	25 - 45	Caucasian	0	2	0	0	[illegible]
9239	steven strob	steven	strobridge	2/14/14	Male	8/21/92	23	Less than 25	African-Ame	0	4	0	0	[illegible]
9245	angelo armb	angelo	armbrister	4/29/13	Male	10/6/89	26	25 - 45	African-Ame	0	3	0	0	[illegible]

Fig. 1. Raw COMPAS Dataset Samples.

The evaluation results are shown as:

	A	B	C	D	E	F	G	H	I
1	id	Y	A	predictor1	predictor2	predictor3	predictor1_prob	predictor2_prob	predictor3_prob
2	8902	1	5	0	0	0	0.321120209	0	0.478827185
3	8908	0	0	1	1	1	0.578750921	1	0.883751766
4	8945	0	0	1	1	1	0.575520949	1	0.883751766
5	8948	0	5	0	0	0	0.332877291	0.112615601	0.478827185
6	8949	0	0	1	1	0	0.542530261	0.781180833	0
7	8960	1	0	1	0	0	0.522382734	0.246005913	0
8	8972	1	0	1	0	0	0.519356862	0.246005913	0
9	8997	0	0	1	1	0	0.549296048	1	0.420018323
10	9020	1	2	0	0	0	0.457215197	0	0.579981677
11	9022	1	0	1	1	1	0.577894165	1	0.883751766
12	9045	0	2	0	0	0	0.416835679	0	0.116248234
13	9056	1	0	1	1	0	0.535375886	0.781180833	0
14	9079	1	3	0	0	1	0.426026631	0.218819167	0.941191139
15	9080	1	2	0	0	0	0.476090658	0	0.579981677
16	9089	0	2	0	0	0	0.428150915	0	0.579981677
17	9165	0	0	1	1	0	0.582190532	1	0.883751766
18	9169	1	5	0	0	0	0.330114024	0.112615601	0.478827185
19	9172	1	0	1	1	1	0.575520949	1	0.883751766
20	9178	1	0	1	1	0	0.552661062	0.781180833	0
21	9226	1	2	0	0	0	0.42364859	0	0.579981677
22	9227	0	2	0	0	0	0.406940001	0	0.116248234

Fig. 2. Evaluation Samples.

3.2 Survey Design

To learn about which model of three is more fair, a web application is build, to post test dataset and evaluation results in a form of survey. For each survey, ten test pieces are selected randomly, two of three model evaluation results are selected randomly as well to make survey robust. Users can vote a model from their own perspectives by given information, and responses will be saved in the database. The following section shows the website interface.

4 WEB UI

The web platform is build using Django framework as it is python based. SQL database is used for store and transfer data.

4.1 Home Page

The home page is shown as Figure 3, 4 and 5. The theme is black as we want to present a clear and neat website. When users scroll down, we provide background information about COMPAS. In addition, contact information are provided in common at the bottom of home page.

Fig. 3. Home Page Screenshot (Part 1).

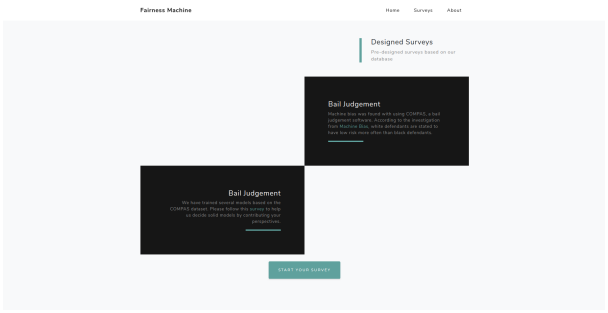

Fig. 4. Home Page Screenshot (Part 2).

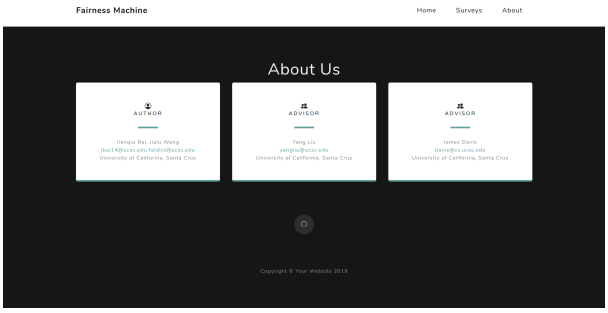

Fig. 5. Home Page Screenshot (Part 3).

4.2 Survey Page

We designed clear "Start Survey" button navigating users to start a survey. Before starting the survey, we provide a modal to present survey guide to let users know how to do the survey as shown in Figure 6. We present some basic information about our pre-trained models as well.

According to Figure 7, we provide information section divided into sensitive information and relative information. Users can vote a model by clicking the button. We also implement hover function to provide users more detail information.

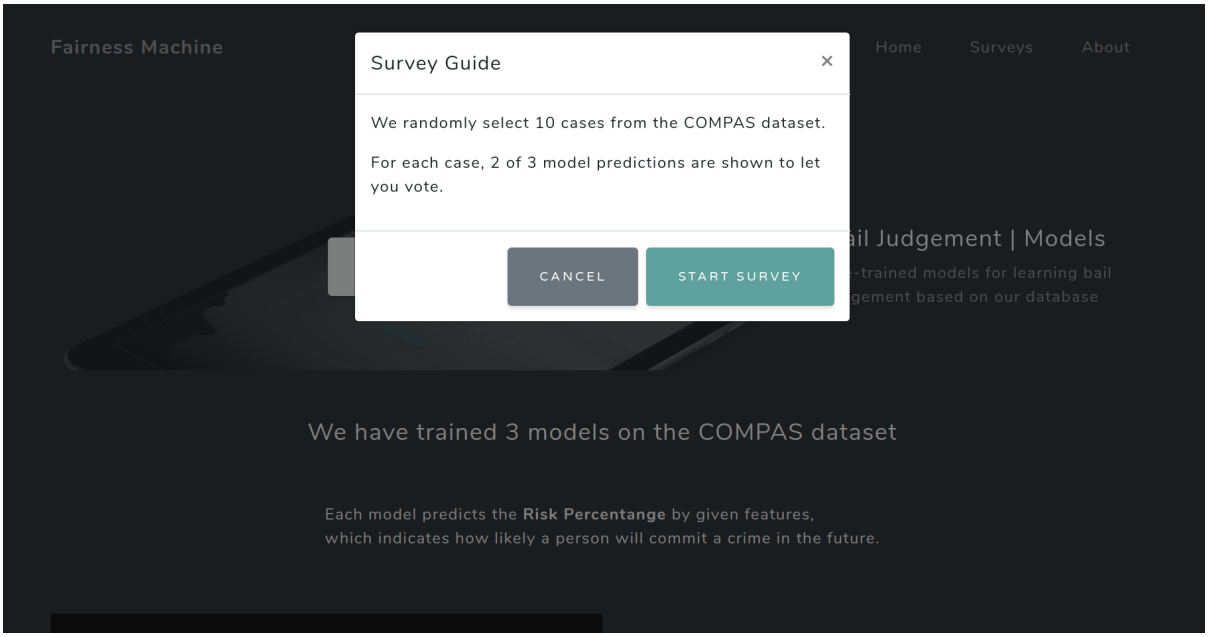

Fig. 6. Model Page Screenshot.

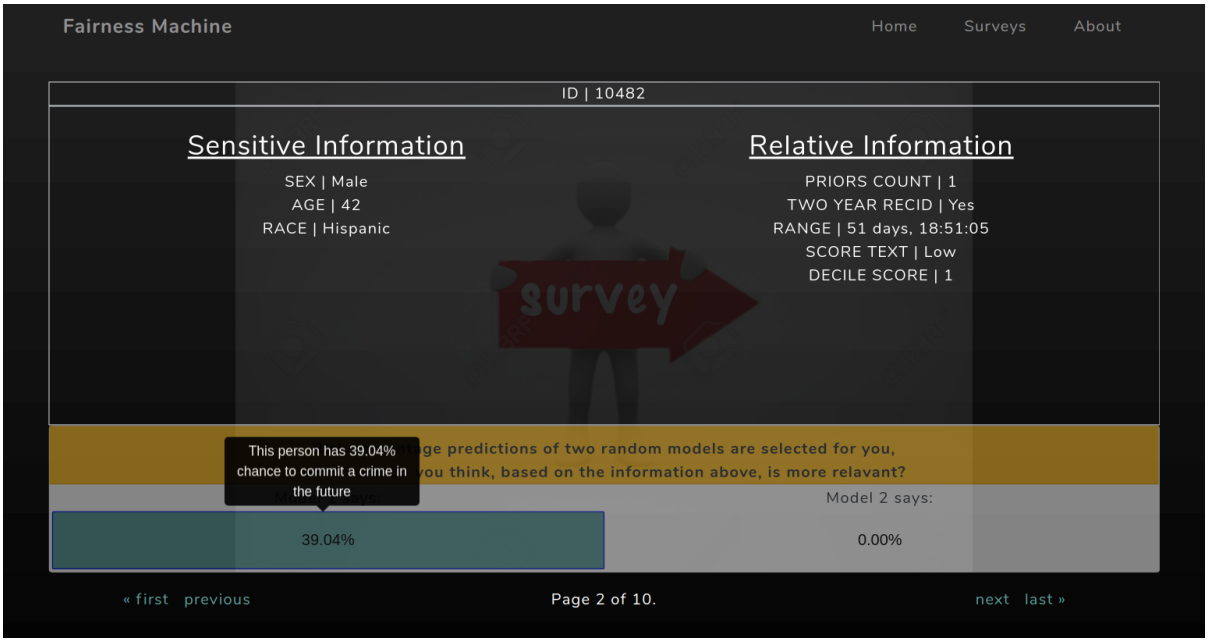

Fig. 7. Survey Page Screenshot.

4.3 Data Page

Once a user finishes a survey set, users can review their responses from the client side as shown in Figure 8. Once the response is saved, the web url will redirect to the data page, where clients can view original data and responsive data as shown in Figure 9 and 10.

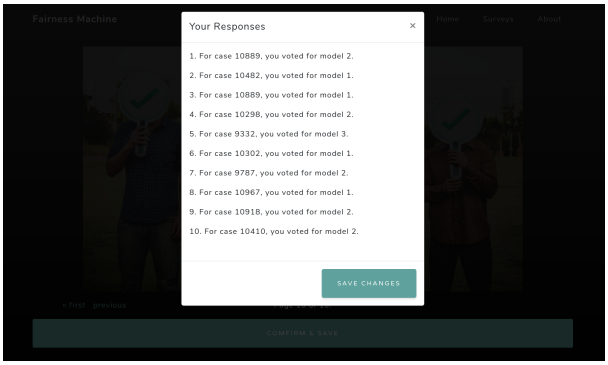

Fig. 8. Response Review Screenshot.

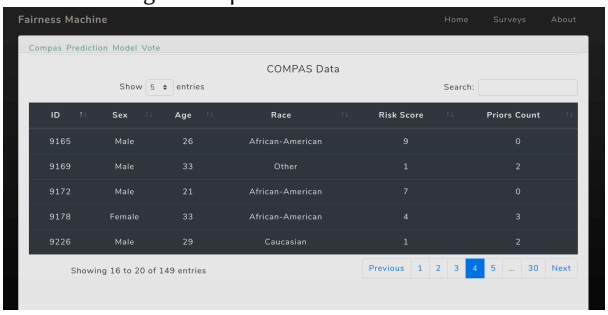

Fig. 9. COMPAS Data.

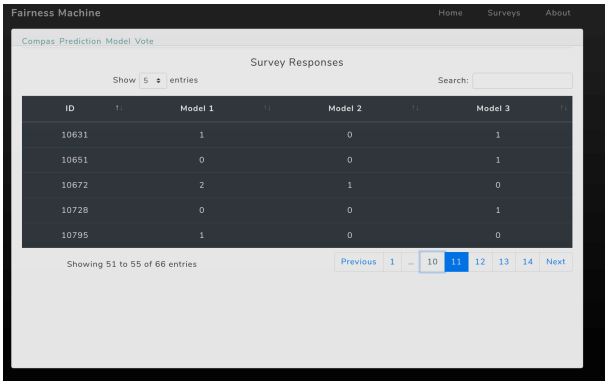

Fig. 10. Response Data.

4.4 Admin Page

Django has build-in admin page for superusers to manage the database. Once we have collect sufficient responses from client side, we could export users' voting information into a local csv file as shown in figure 11, which could be utilized to evaluate fairness of our pre-trained models.

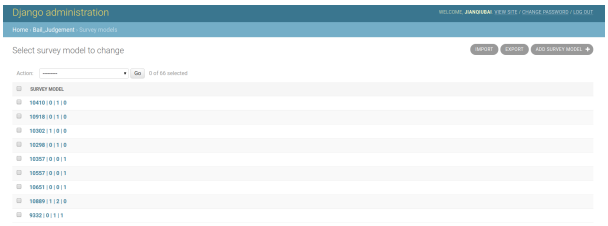

Fig. 11. Admin Page Screenshot.

5 FUTURE WORK

This website is currently served using AWS on http://54.71.194.47/. The next step is to deploy the website to Google Cloud service in order to access free database storage. More information and source code can be found from my github repository for development and deployment. Once the pipeline is fully tested, it is possible to import larger dataset and design more models in other areas such as university admission and etc.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, and ProPublica. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Propublica* (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data5* 2 (2016). https://arxiv.org/pdf/1610.07524.pdf

[3] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* 99-106 (January 2019). https://doi.org/10.1145/3306618.3314248