Auditing the COMPAS Recidivism Risk Assessment Tool: Predictive Modelling and Algorithmic Fairness in CS1

Claire S. Lee* Princeton University Jeremy Du^{*} Princeton University Michael Guerzhoy Princeton University, University of Toronto, Li Ka Shing Knowledge Institute

ABSTRACT

We present an assignment in which students apply predictive modelling to build a model that predicts re-arrest of criminal defendants using real data. Students assess the algorithmic fairness of a realworld criminal risk assessment tool (RAT), and reproduce results from an impactful story in ProPublica and a 2018 *Science Advances* paper. Students explore different measures of algorithmic fairness, and adjust the model they build to satisfy the false positive parity measure.

Our target audience is students in Introduction to Data Science courses that do not require previous computing experience, as well as students in standard CS1 courses. We advocate for teaching predictive modelling in CS1. To facilitate the teaching of predictive modelling in CS1, we provide tutorials on predictive modelling and algorithmic fairness, in both Python and Java; we also provide a simplified "Learning Machine" API in those languages.

Our approach enables teaching algorithmic fairness and predictive modelling more generally very early in the students' computing career. A companion website with all our teaching materials is available at https://PredictiveModellingEarly.github.io/.

KEYWORDS

cs1, algorithmic fairness, data science, predictive modelling

ACM Reference Format:

Claire S. Lee, Jeremy Du, and Michael Guerzhoy. 2020. Auditing the COM-PAS Recidivism Risk Assessment Tool: Predictive Modelling and Algorithmic Fairness in CS1. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '20), June* 15–19, 2020, Trondheim, Norway. ACM, New York, NY, USA, 2 pages. https: //doi.org/10.1145/3341525.3393998

1 INTRODUCTION

Applications of machine learning, and predictive modelling in particular, are now everywhere: machine learning models diagnose our illnesses, decide whether to extend credit to us, and decide whether to grant us bail. Considerations of algorithmic fairness are coming to the forefront. It is important that the predictive models that have

ITiCSE '20, June 15-19, 2020, Trondheim, Norway

© 2020 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-6874-2/20/06.

https://doi.org/10.1145/3341525.3393998



Figure 1: The predictive modelling framework, as presented to students in our tutorials.

an impact on so many people's lives are not biased with respect to protected characteristics such as race or sex.

In the U.S. criminal justice system, risk assessment tools (RATs) are increasingly being used to assess a criminal defendant's probability of re-offending. RATs use information such as the number of priors as well as questionnaire data from defendants. In 2016, the non-profit journalism organization ProPublica analyzed COMPAS, a RAT made by Northpointe, Inc., to assess whether it was biased against African-American defendants. ProPublica found that COM-PAS incorrectly labeled innocent African-American defendants as likely to reoffend twice as often as innocent white defendants [1]. In technical terms, this means the false positive parity measure of algorithmic fairness was not satisfied by the COMPAS system.

In a follow-up to ProPublica's investigation, Julia Dressel and Hany Farid showed that a score almost equivalent to the COM-PAS score can be obtained by using only the defendant's sex, age, and number of priors [4]. Researchers in the algorithmic fairness community pointed out that different observational measures of algorithmic fairness cannot all be simultaneously satisfied [2].

We present an assignment suitable for Introduction to Data Science and Introduction to Programming courses in which students replicate the findings from ProPublica and *Science Advances*, and investigate a way to adjust the models they build so that innocent African-American defendants are not mislabeled at a higher rate than innocent white defendants (with only a marginal impact on accuracy). We successfully used the assignment in a class that has no programming or statistics prerequisites.

We have two goals in mind: to reinforce students' understanding of predictive modelling and to teach them important CS1 concepts.

^{*}Both authors contributed equally to this work.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

It turns out that computing various measures of fairness for a predictive model and adjusting a model's thresholds to achieve a specified criterion of fairness involve programming tasks that are just at the right level of difficulty for mid-to-late CS1.

Predictive modelling is usually not taught in introductory courses. In part, this is because it is traditionally taught in courses that require knowledge of calculus. Another reason is the complexity of working with machine learning libraries, most of which require extensive computing experience.

We believe that predictive modelling and algorithmic fairness can be made accessible to all students. We teach predictive modelling with logistic regression early by treating logistic regression as a black box. For students who are learning Python and Java, we address the complexity of the commonly used machine learning libraries by creating a small number of wrapper functions that the students can use as black boxes. We supply drafts of tutorials covering the usage of those black boxes.

2 ASSIGNMENT OUTLINE

Students use a real dataset of defendants in Broward Country, Florida, which was obtained by ProPublica via a Freedom of Information Act request. ¹ Both the COMPAS scores and data about re-arrests is available for the defendants in the dataset. Students visualize the data and assess the COMPAS scores using several algorithmic fairness criteria. Students then reproduce the results from Dressel and Farid [4] by fitting a logistic regression to the data. Finally, students adjust the logistic regression classifier they built in order to obtain a classifier that is fair with respect to False Positive Parity and assess the consequences of their adjustment for the model.

3 A BRIEF INTRO TO OBSERVATIONAL ALGORITHMIC FAIRNESS MEASURES

Knowledge of the algorithmic fairness literature is not required to complete this assignment. We give a quick summary of what the students need to know.

3.1 Measures of Classifier Performance

We are considering a dataset where the output of interest is "positive" / "yes" (1) or "negative" / "no" (0). The outputs of our classifier are stored in the vector pred, and the correct outputs (i.e., the ground truth) are stored in the vector y. We can compute the following measures.

- Correct Classification Rate (CCR): for what proportion of the inputs does the correct output y[i] match the classifier output pred[i]?
- False Positive Rate (FPR): for what proportion of the inputs for which the correct output y[i] is negative is the classifier output pred[i] positive?

3.2 Observational Measures of Algorithmic Fairness

Algorithmic fairness can be assessed with respect to an input characteristic. Typically, algorithmic fairness would be assessed with respect to characteristics such as race or sex. We can compute the following measures:

- False positive parity with respect to characteristic C is satisfied if the false positive rate for inputs with C = 0 is the same as the false positive rate for inputs with C = 1. ProPublica found that false positive parity was not satisfied by classifiers based on the COMPAS score with respect to race. The false positive rate for African-American defendants (i.e., the percentage of innocent African-American defendants classified as likely to re-offend) was higher than for white defendants.
- Calibration with respect to characteristic C is satisfied if an individual who was labeled "positive" has the same probability of actually being positive, regardless of the value of C, and if an individual who was labeled "negative" has the same probability of actually being negative regardless of the value of C. The makers of the COMPAS tool claimed that COMPAS satisfies calibration [3].

It is fairly easy to show that in general, only one measure of fairness at a time can be satisfied for any particular classifier. We provide further resources in our Learning Machine API tutorials.

3.3 The Learning Machine API

We provide versions of the assignment for R, Python, and Java. For the Python and Java versions of the assignment, we encapsulate logistic regression in a learningmachine module/LearningMachine class. The idea is to let CS1 students concentrate on programming and to quickly grasp the idea of a black box that takes in a training set and spits out a model that can make predictions for new data. We hope that avoiding the need to spend several lectures on properly teaching subsets of scikit-learn or Weka will enable CS1 instructors to teach predictive modelling and algorithmic fairness early on in their course.

3.4 Predictive Modelling Tutorials

We provide tutorials on predictive modelling in Python and in Java. We focus on getting students to understand the predictive modelling framework (Fig. 1) without dwelling on the mathematics of, for example, logistic regression. We find that it is possible to get students up to speed in a few hours.

REFERENCES

- Julia Angwin and Jeff Larson. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-incriminal-sentencing
- [2] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.
- [3] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018).
- [4] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. Science Advances 4, 1 (2018).

 $^{^{1}} https://github.com/propublica/compas-analysis/blob/master/compas-scores-raw.csv$