Evaluating the Fairness of Predictive Student Models Through Slicing Analysis

Josh Gardner
Paul G. Allen School of Computer Science & Engineering
The University of Washington
jpgard@cs.washington.edu

Christopher Brooks
School of Information
The University of Michigan
brooksch@umich.edu

Ryan Baker
Graduate School of Education
The University of Pennsylvania
rybaker@upenn.edu

ABSTRACT

Predictive modeling has been a core area of learning analytics research over the past decade, with such models currently deployed in a variety of educational contexts from MOOCs to K-12. However, analyses of the differential effectiveness of these models across demographic, identity, or other groups has been scarce. In this paper, we present a method for evaluating unfairness in predictive student models. We define this in terms of differential accuracy between subgroups, and measure it using a new metric we term the Absolute Between-ROC Area (ABROCA). We demonstrate the proposed method through a gender-based "slicing analysis" using five different models replicated from other works and a dataset of 44 unique MOOCs and over four million learners. Our results demonstrate (1) significant differences in model fairness according to (a) statistical algorithm and (b) feature set used; (2) that the gender imbalance ratio, curricular area, and specific course used for a model all display significant association with the value of the ABROCA statistic; and (3) that there is *not* evidence of a strict tradeoff between performance and fairness. This work provides a framework for quantifying and understanding how predictive models might inadvertently privilege, or disparately impact, different student subgroups. Furthermore, our results suggest that learning analytics researchers and practitioners can use slicing analysis to improve model fairness without necessarily sacrificing performance.[1]

CCS CONCEPTS

• **General and reference** → **Metrics**; • **Applied computing** → *Computer-assisted instruction*;

KEYWORDS

Fairness, machine learning, MOOCs

ACM Reference Format:
Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *The 9th International Learning Analytics & Knowledge Conference (LAK19), March 4–8, 2019, Tempe, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3303772.3303791

[1]Code to fully replicate or extend this analysis is open-source and available at https://github.com/educational-technology-collective/slicing-analysis

LAK19, March 4–8, 2019, Tempe, AZ, USA

ACM ISBN 978-1-4503-6256-6/19/03…$15.00
https://doi.org/10.1145/3303772.3303791

1 INTRODUCTION

Following the rapid expansion of predictive machine learning models across many societal contexts over the past decade, recent work has begun to assess the ethical impact of these models when deployed for real-world decision-making [36]. Generally, this work has supported the conclusion that models which ignore the differential impact of their predictions on different groups of individuals – e.g. those of different ethnicities, genders, or national identities – can yield undesirable properties which can produce and reinforce unanticipated inequalities across groups.

While work to measure, understand, and correct unfairness produced by predictive models has gained increasing traction even in the past year [e.g. 10, 13, 20, 23, 24], there has not yet been sufficient attention to this issue in education, despite extensive prior research on prediction in education, widespread deployment of predictive models in "live" educational environments, and the high stakes of educational outcomes for increasing social equity. In this work, we develop a methodology for the measurement of unfairness in predictive models using "slicing analysis" [30], in which model performance is evaluated across different dimensions or categories of the data. We specifically apply this approach to examine gender-based differences of MOOC dropout models as a case study. Further, we propose a novel method for measuring and evaluating the differential performance of predictive models across groups using a metric we term the Absolute Between-ROC Area (ABROCA). This method is particularly suited for evaluating predictive models in education, as detailed in Section 4. In order to demonstrate the use of the ABROCA method for slicing analysis and to provide empirical benchmarks of existing models, we conduct a large-scale replication study of [11, 14] to evaluate five previously-proposed models in Section 5. Our results demonstrate (1) significant variation in model unfairness across the statistical algorithms and feature sets evaluated; (2) significant associations between unfairness and course gender imbalance, course curricular area, and specific courses. Furthermore, (3) we do *not* observe a strict tradeoff between unfairness and performance, suggesting that slicing analysis may be used to improve model fairness without sacrificing predictive performance in practice.

To date, the authors are not aware of any work which has answered the original call for slicing analysis in [30]. The current work thus provides a methodological foundation and an empirical benchmark for future work both within learning analytics and

across domains which apply predictive models to diverse individuals in practice. This work addresses the growing concern in the field of learning analytics regarding the lack of understanding of the differential impact of predictive models [27, 28, 31].

This analysis uses the MOOC Replication Framework (MORF) [17], a framework for replication of previous findings and reproducible research. While this analysis uses data from Massive Open Online Courses (MOOCs) as a case study, the need for slicing analysis (and the method proposed here) applies to any educational context in which predictive models are used.

1.1 Notation

In this work, $\mathcal{M}$ refers to a specific parameterization of a statistical model, trained on a dataset $\{X, A, Y\}_{i=1}^{n}$. X is a vector of features for each of n observations, and the aim of $\mathcal{M}$ is to minimize the loss of its predictions, $\hat{Y}$, with respect to the true labels $Y \in \{0, 1\}$ (in Section 5, this label represents MOOC dropout). $\mathcal{M}$ assigns predicted class labels by first predicting the probability that a given observation will be a positive case, $\hat{p}(x_i) = \mathbb{P}(Y = 1|x_i)$, and comparing this predicted probability $\hat{p}$ to some threshold t. $\mathcal{M}$ predicts $\hat{y}_i = 1$ if and only if $\hat{p}(x_i) > t$, and $\hat{y}_i = 0$ otherwise. It is important to note that t must be selected and is not given by $\mathcal{M}$; t is often selected based on a variety of factors, including optimal model performance and the acceptable levels of false positives vs. true positives.

Each observation also has a membership in some subgroup of a "protected attribute" A, which is a dimension (such as race, gender, or nationality) along which we desire the model to be fair. We use the labels "baseline" and "comparison", denoted by $\{A_b, A_c\}$, respectively, to denote the two subgroups of A. Typically, the "baseline" group would be a privileged or majority group to which all other groups are compared to ensure non-discrimination. Our analysis does not assume that A is explicitly provided to the model (indeed, none of the models replicated here do so). The size of A can be arbitrarily large, but in this work, we focus on the two-class case of A; the proposed method can be extended to such cases as long as there is a clear notion of a baseline subgroup to which all other subgroups might be compared.

2 RELATED WORK

2.1 Predictive Modeling in Learning Analytics

There is an extensive research base on the use of predictive models in learning analytics, and in MOOCs in particular. A comprehensive survey of such work is beyond the scope of this paper, and we refer the reader to [15, 26] for such reviews. However, we provide a brief overview of the diversity of approaches taken to the task of student success prediction in MOOCs in order to motivate the need for a more comprehensive evaluation of these methods, beyond their predictive accuracy, by using slicing analysis.

Research on student success prediction in MOOCs began almost immediately after the launch of the initial "big three" MOOC platforms (Coursera, edX, Udacity) in 2012. These platforms provide massive, granular, and diverse data on student activity and understanding which serve as the foundation for complex and varied student success models. The data used in predictive MOOC models include counts of various student activity and navigation actions [22]; natural language processing [8] and social network metrics [35] derived from course discussion forums; measures of student performance extracted from assignment [32] or in-video quiz submissions [4]. The statistical algorithms used for modeling and prediction across such research also vary, spanning from parametric models such as logistic regression e.g. [33], to nonparametric tree-based models [14], to modern neural network models [11]. In addition to being used for research, such models are increasingly used in practice, for example, to support student interventions [2] and data collection [34]. The active use of predictive models increases the urgency of understanding whether such models are equally effective for all student groups. Rigorous comparisons of the relative efficacy of these models in MOOCs have been rare [14], and existing comparisons almost exclusively evaluate models' predictive performance.

2.2 Fairness in Machine Learning Algorithms

Recently, the machine learning research community and even the popular press have begun to engage with and investigate issues of fairness, bias, and the disparate impact of machine learned models [7, 36]. Borrowing from several disciplines, recent work has made efforts at measuring, understanding, and correcting for the presence of such biases in machine learned models. We present an overview of selected work relevant to the proposed method here.

Defining and *measuring* predictive fairness has been a central task of prior work, with many definitions of fairness previously proposed [13]. As we will argue below, however, many of these definitions are inadequate for learning analytics research.

One of the simplest conceptions of fairness is the notion of *demographic parity*. Demographic parity requires that, for all groups of the protected attribute A (e.g. gender), the overall probability of a positive prediction of a given outcome should be the same – the protected attribute should be independent of the prediction (this condition has also been referred to as "disparate impact" [23]). However, as [9, 19] both argue, demographic parity is not an ideal fairness criterion. First, it fails to ensure individual fairness, requiring that we make "matching" predictions across demographic groups so that, in aggregate, the probability of a positive prediction across groups is equal, even when there are legitimate reasons for these differences. Second, demographic parity can *prevent* models from providing their intended utility in cases where the protected attribute is correlated with the outcome. In education, for example, such a situation might arise when membership in a disadvantaged group is in fact correlated with a higher rate of dropout (this situation has been demonstrated, for example, among female students in data science MOOCs [2, 5]). A demographic parity predictor would not be able to predict higher rates of dropout for female students relative to males in this case, despite the fact that this is both correct and necessary for the model to support effective student intervention. For further critiques of the demographic parity criterion see [9, 19].

In [9] it is suggested that a measure of *individual* fairness seems more appropriate than the group fairness measured by demographic parity. Specifically, [9] argues that *similar individuals should be classified similarly*, proposing a conception of fairness based on a hypothetical distance metric between individuals. In the presence of such a method, the problem of creating a "fair" classifier can be

reduced to a tractable optimization problem. However, this shifts the challenge of defining fairness to one of finding a suitable similarity metric between individuals – a task no less difficult, and which raises many of the same questions as defining fairness itself.

Seminal work by Hardt et al. proposes a definition of fairness based on *equalized odds* [19].[2] A predictor satisfies equalized odds if and only if the false positive rate and the true positive rate are equal among the baseline and comparison classes A_b and A_c. That is, a predictor satisfies equalized odds if and only if:

$$Pr\{\hat{Y} = 1 | A = A_c, Y = 0\} = Pr\{\hat{Y} = 1 | A = A_b, Y = 0\} \quad (1)$$

$$Pr\{\hat{Y} = 1 | A = A_c, Y = 1\} = Pr\{\hat{Y} = 1 | A = A_b, Y = 1\} \quad (2)$$

Equation (1) represents the condition that the predicted false positive rate must be equivalent across the two groups of the protected attribute; (2) represents the same for the true positive rate.

There are several advantages to the equalized odds method over previous conceptions of fairness. First, it avoids the group fairness measures which, as noted above, can still lead to unfair outcomes for individuals and also diminish the quality of the resulting classifier. Second, this method can, in principle, be implemented directly from a dataset and classifier – [19] demonstrates that achieving equalized odds amounts to solving a linear optimization program. This can be applied in a simple post-processing step without modifying a model's training process – a clear practical advantage. Third, the equalized odds method does not rely on an additional metric, such the hypothetical similarity measure required in [9] – it is, in this sense, a *complete* solution. While [19] constitutes important progress toward defining fairness in a specific classification task, in Section 4 we discuss characteristics of this method which limit its usefulness in learning analytics and in a variety of other contexts where predictions might be used for a range of decisions or interventions, not just a single intervention where the decision threshold is predetermined.

While we do not directly consider work which is rooted in legal and political notions of fairness, we note that there is a rich literature on applying legal concepts, such as disparate impact, to prediction fairness [e.g. 3, 12, 29]. While learning analytics must, at minimum, satisfy such concepts, we believe that learning systems should aspire to an even higher standard of fairness.

2.3 Replication in Learning Analytics

We demonstrate our approach to measuring fairness in educational predictive models through the replication of previous work, for two reasons. First, there has been only limited replication research in the field of learning analytics to date, with the replication rate (the percentage of studies which are replications) estimated at approximately 0.13% in the field of education [25]. The limited previous efforts at replication in the field of learning analytics, e.g. [1, 16], have demonstrated the additional insights that replication of existing findings across large, representative datasets can bring, particularly because of the restricted nature of most educational data.

[2]The authors propose a related, but more relaxed, criterion known as *equal opportunity*, which we do not address, but to which our criticism still applies.

Second, our methodology applied to previously-researched predictive models allows for a deeper understanding of the previous work which has already already had its predictive performance thoroughly documented. Replication thus builds on this knowledge by contributing a new perspective on this work, viewing it through the lens of fairness. Such analysis stands to be more informative than merely proposing novel classification approaches without fully evaluating previous feasible options for student performance prediction. Additionally, a failure to conduct any slicing analysis of prior work would fail to provide empirical benchmarks for future work evaluated via the slicing analysis method proposed below: a knowledge of how discriminatory (or not) previous work may be is highly relevant to future research into non-discriminatory student success models. Only through replication can we even determine whether the field may have a problem that needs solving.

3 SLICING ANALYSIS: A NEW MODE OF MODEL UNDERSTANDING

3.1 Motivation and Definition

The concept of slicing analysis was originally proposed in [30] as a methodology necessary to correct a predictive modeling culture focused on "winning" above all else – that is, to encourage deeper evaluation of predictive models beyond their predictive performance. Sculley et al. [30] argue that this "winner's curse" weakens the overall impact of machine learning research which is pursued with a singleminded focus on predictive performance ("wins"), sacrificing deeper evaluation of model fairness and impact as a result. The field of learning analytics has also matured to the point where such deeper evaluation is required, particularly because its goals often involve serving a diverse array of students [28].

The current work aims to answer the call for such increased empirical rigor, transparency, and evaluation now that predictive modeling research in learning analytics has reached a critical mass, and to provide a demonstration and toolkit to support such analyses. While standardized methods exist for evaluating a variety of high-stakes products, from the brakes on an automobile to the autopilot system on an airplane, no such standardized method currently exists for predictive modeling and AI systems [20]. As [30] notes:

> Performance measures such as accuracy or AUC on a full test set may mask important effects, such as quality improving in one area but degrading in another. Breaking down performance measures by different dimensions or categories of the data is a critical piece of full empirical analysis.

This captures the inspiration for slicing analysis: to evaluate a predictive model's performance by "slicing," the results of that model across different dimensions or categories in the test set. In doing so, we uncover the relative performance or fairness of a model across subgroups which comprise a given dataset. The degree to which both performance *and* fairness vary over all groups is of critical importance for learning analytics research and practice.

Like the model performance analysis which is common to both learning analytics research and machine learning research, slicing analysis is an exploratory methodology that can be used to understand models or, optionally, to inform model selection. As such,

we do not require that slicing analyses lead to *corrections* of model unfairness (although they may do so). The current work does not explicitly offer a method for correcting for the results of a slicing analysis post-hoc; however, such methods have been proposed [e.g. 19], and our method is indeed compatible with such a correction (see Section 4). We see the first step of slicing analysis as identifying, understanding, and quantifying unfairness within models.

3.2 Why Slice? The Importance of Model Fairness in Learning Analytics

While we believe that most readers accept *a priori* that fairness is a necessary component for educational predictive models, we present several considerations in support of this intuition.

First, as we noted above, fairness (or lack of discrimination) is a legal requirement of many systems [3]. For example, [23] notes that in the United States, Tile VII of the Civil Rights Act of 1964 bans even facially neutral practices that might nevertheless have an "unjustified adverse impact on members of a protected class" (p. 2). In most developed nations, equal educational opportunity is a specific legal requirement. Second, one of the primary aims of education is to increase opportunity and knowledge for all learners. This is particularly relevant for MOOCs, which have been highlighted for their ability to reach diverse, underserved learners across the globe [6]. Predictive models which systematically disadvantage these groups fail to deliver on this promise. Third, models which unfairly discriminate against minority groups are detrimental in a practical sense: by providing predictions which may be unfairly biased by attributes not relevant to learning, such models fail to support the goal of providing the optimal learning environment for all students. Fourth, on another practical note, discriminatory models risk creating negative feedback loops, in which discriminatory predictions become self-reinforcing [31]. Fifth, as it stands, the field of learning analytics has little to no evidence about whether its models display predictive bias, and identifying these biases without a formal method such as slicing analysis is likely to only become more challenging as state-of-the-art models used for prediction increase in sophistication. Slicing analysis allows us to observe the impacts of the models on sensitive groups even if we cannot explain the underlying models.

Collectively, these considerations provide a clear motivation for conducting slicing analysis in learning analytics. This motivation is particularly strong given the current state of the field, which has seen increasing deployment of "live" predictive modeling tools in both digital and in-person learning environments [2, 28].

4 A METHOD FOR SLICING ANALYSIS

This section introduces the proposed method for slicing analysis, which we apply to a large-scale predictive modeling replication experiment in Section 5. The goal of slicing analysis is to measure the unfairness[3] of the predictions of a model $\mathcal{M}$.

We propose a method for strictly *measuring* fairness, and not correcting it, for several reasons. First, measurement is a necessary condition for correcting any detected unfairness. Second, satisfying any individual definition of fairness is not a proof of fairness: following [19], "we envision our framework as providing a reasonable way of discovering and measuring potential concerns that require further scrutiny" [19, pp. 21]. Finally, the procedure of correcting a fairness metric is at least marginally more complex than the measurement itself, and is beyond the scope of this paper. However, as a demonstration of a naïve solution, note that the ABROCA statistic can be reduced to zero (at a cost to model performance) by introducing randomization into the predictions of one or more groups.

A primary goal of predictive modeling in learning analytics is to produce accurate predictions to support personalized learner support or adaptive learner pathways [15]. The concept of fairness implies that subgroups should be treated equally by such models, and, in particular, that subgroups should benefit equally from their predictions. Thus, the current work presents a conception of fairness rooted in *equal model performance across subgroups of A.*

4.1 Proposed Method: Absolute Between-ROC Area (ABROCA)

Measuring the extent to which a model $\mathcal{M}$ meets a definition of fairness based on equivalent performance requires a robust measurement of model performance across subgroups. We propose a method based on the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of the false positive rate and true positive rate of a model's predictions across all possible thresholds $t \in [0, 1]$, where t determines at what predicted probability a model predicts $\hat{Y} = 1$. The area under the ROC curve, or AUC, is a common metric used to evaluate predictive models. AUC values range between 0 and 1, where an AUC value of 0.5 is equivalent to random guessing and a higher AUC indicates better performance. This measure has been noted for its robustness to imbalanced data [21] and its straightforward interpretation as the probability that a randomly-selected positive observation will be (correctly) assigned a higher predicted probability than a randomly-selected negative observation [18]. Figure 1 shows two ROC curves (in red and blue).

AUC is one of the most commonly-used model performance metrics in predictive modeling research in MOOCs [15, 26]. At least one reason for this widespread adoption is that predictive models are often used for a wide range of interventions, and for different interventions, different thresholds are often appropriate; AUC measures performance across all possible thresholds.

The analysis above may seem to suggest that we directly compare AUC values across various demographic groups in order to compare performance across different subgroups, perhaps by taking the difference of the areas, given by

$$\text{AUC}_b - \text{AUC}_c \tag{3}$$

However, in cases where the ROC curves cross, (3) allows differences in performance to "cancel out" across thresholds:"positive" differences in one area are mitigated by "negative" differences where the other ROC curve is greater. Such a case is shown in Figure 1, and our case study below shows that such cases are commonly encountered in practice. (3) can substantially underestimate the degree of unfairness in such cases. (3) is, simply put, an unreliable method

[3]The terms bias, unfairness, and discrimination are used interchangeably in this work to refer to inequitable prediction across identity groups. We seek to avoid confusion with the statistical definition of bias, with which the current work is not concerned.

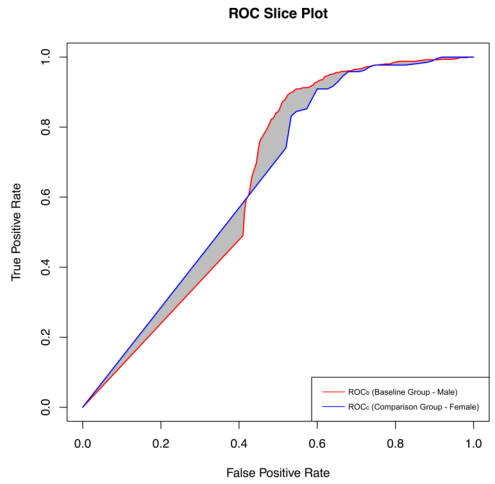

Figure 1: A "slice plot," which shows the Receiver Operating Characteristic (ROC) curves for a model across two groups (male and female learners). The shaded region between the two curves represents the ABROCA statistic, which measures differences in model performance across groups.

for measuring how unfairness affects these subgroups across the full range of potential thresholds for a learning analytics model.

This motivates the need for a measure which captures the differences in performance for each subgroup of the protected attribute, across all possible thresholds, without allowing for the possibility of differences which cancel out, or negate each other.

We propose the Absolute Between-ROC Area (ABROCA) to measure fairness. ABROCA measures the *absolute value* of the area between the baseline group ROC curve ROC_b and those of one or more comparison groups ROC_c. ABROCA thus captures the divergence between the baseline and comparison group curves across all possible thresholds, and aggregates this divergence without respect to which subgroup's model may achieve better performance at any particular threshold (by taking the absolute value). Not only does this prevent the ABROCA metric from allowing positive and negative difference in performance at various thresholds to "cancel out"; it also captures the unfairness present when minority-group models outperform majority-group models (because, as described above, equal performance across all groups is necessary to meet our proposed conception of fairness).

The ABROCA statistic is the total difference, across all possible thresholds, in the probability of $\mathcal{M}$ correctly classifying a randomly-selected data point from the majority vs. the non-majority group. Formally, ABROCA is defined as:

$$\int_0^1 |ROC_b(t) - ROC_c(t)|\, dt \tag{4}$$

Visually and geometrically, the ABROCA statistic has a straightforward interpretation: it is the area between the two ROC curves, shown in Figure 1. ABROCA can vary between 0 and 1, but in most cases its range is practically restricted to the interval $[0.5, 1]$, as an AUC of 0.5 is achievable by random guessing.

4.2 Advantages Over Existing Fairness Metrics

The ABROCA metric builds on the conceptions of fairness proposed in previous works, but overcomes several limitations of the specific metrics discussed in Section 2.

First, ABROCA is not threshold-dependent. ABROCA accounts for difference in model performance between subgroups *across the entire range of possible thresholds* t. In contrast, many of the most widely-used existing methods, such as the equalized odds definition of [19], apply only for a *specific* threshold. As [19] states, equalized odds can only be achieved at points where the ROC curves cross, or by using different thresholds (or some form of a randomized threshold) for the different demographic groups. This is problematic because (a) it fails to account for or evaluate any difference in model performance at any other threshold, when a fixed threshold is used; (b) it forces an evaluator to choose a specific threshold at which the equalized odds condition is to be evaluated, when in practice, the actual threshold used may depend on the intervention or the specific context of model deployment, which is often not known at the time of model evaluation; or (c) it forces the modeler to vary thresholds for different subgroups of A in order to enforce equalized odds. Such *treatment disparity* is also commonly rejected as discriminatory and undesirable for predictive models [23]. Figure 2 demonstrates a case where equalized odds is achieved at some thresholds, while at others there are substantial gaps between the odds in each group. We also note that for some use cases in learning analytics, one may wish to compare a restricted range of potential thresholds (e.g. using $t > 0.5$ to develop high-recall models). Computing ABROCA across a restricted range of thresholds in such cases is straightforward, and amounts to changing the limits of integration in (4).

Second, in contrast to many existing conceptions of fairness which largely center on decision-making systems (such as credit loan applications) where a "positive" prediction is considered a disadvantageous outcome for the individual and only the probability of positive prediction is used to evaluate fairness [e.g. 9], ABROCA evaluates the overall accuracy of the models without strictly fixating on the positive case. This is particularly relevant to learning analytics contexts, which are different from the contexts in which many existing systems have been evaluated (e.g. loan applications [19], parole decisions, graduate admissions [23]). For example, in a MOOC modeling scenario, a *prediction* that a given student may drop out is neither "good" nor "bad"; it is simply a prediction that is used to inform future action used to support that student. In contrast, in e.g. the credit modeling scenarios used to justify metrics such as demographic parity which evaluate only the positive case, a prediction that the same individual is likely to default on a loan will directly lead to a rejection of their loan application – clearly a disadvantageous and more undesirable outcome.

Third, this method is "complete": the ABROCA statistic can be computed directly from the results of a predictive modeling experiment, with minimal computation, and no additional data collection or selection of additional metrics required: the ABROCA statistic is computed from the predicted probabilities $\hat{p}$ and the labels Y, and is simply an integral over the ROC curve (for which there exist several open-source implementations and a robust literature on constructing). ABROCA does not assume the existence of another

metric such as a hypothetical distance measure, relying only on a simple function of the ROC.

Fourth, there are practical considerations which make ABROCA an appealing method for slicing analysis in learning analytics, and in similar disciplines where predictive models are used to support a variety of interventions with different and varying thresholding requirements. ABROCA has a simple interpretation mathematically, geometrically, and visually. Additionally, because ABROCA is a function of the ROC curve which has well-known and straightforward statistical properties [18], several useful statistical properties of the ABROCA can also be derived. This allows us to compute standard deviations, confidence intervals, and significance testing for slicing analyses performed with ABROCA (an explication of these properties is beyond the scope of the current work).

5 SLICING ANALYSIS: MOOC DROPOUT MODELS

In this section, we conduct a slicing analysis in a large-scale case study of MOOC data. We apply the ABROCA method to measure the relative discrimination of five different models replicated from two prior predictive modeling studies [11, 14]. In addition to serving as a case study of a slicing analysis in learning analytics generally, and of the ABROCA method in particular, this analysis demonstrates that model unfairness is affected by the modeling algorithm and possibly the feature set used for this model; that course gender imbalance, curricular area, and specific courses are all related to unfairness; and that unfairness does *not* bear a direct relation to performance in the results of our experiment.

5.1 Experiment

We conduct a dropout modeling experiment, where the goal is to predict a binary dropout label indicating whether the student will persist to the final week of a course. Models extract features from the raw MOOC data exports, including clickstream files, course assignment and video activity, and natural language and social network processing of discussion forum posts.

We replicate five total predictive models using MORF, a platform for large-scale replication and original research on a large repository of MOOC data [17]. We extract the exact feature sets to replicate [11, 14] from the raw MOOC data for every course with at least two sessions (one for testing, and all earlier sessions for training), a total of 44 MOOCs. We train the predictive models using all sessions except the final session of a course, predicting on the final, held-out session. A total of 3,080,230 students were in the training dataset, with 1,111,591 students in the testing dataset. We replicate three models from [11], and two from [14]. In each case, we replicate the authors' preferred or "best" model according to the original experiment, as well as additional model(s) which serve to demonstrate a common statistical algorithm used in MOOC dropout prediction. We also replicate the underlying feature set used for each model. This experiment allows us to explore two very different feature sets, along with five commonly-used statistical algorithms for MOOC dropout modeling [15, 26]. An overview of the models and features replicated for this experiment is shown in Table 1. Collectively, these models represent five of the most common algorithms for MOOC dropout modeling [15].

Cite	Feature Types (N)	Modeling Algorithms
[11]	Activity Counts (7)	Long Short-Term Memory Network (LSTM)* Logistic Regression (LR) Support Vector Machine (SVM)
[14]	Activity Counts (15) Assignment Metrics (123) Forum & NLP Metrics (51)	Classification Tree (CART)* Naive Bayes (NB)

Table 1: Summary of models replicated in experiment. For details on the exact features used, see the original publications or the open-source code for this experiment. * indicates highest-performing model in original experiment.

Each model is constructed using the first three weeks of data for each course. We use three weeks of data, rather than a smaller, earlier window for model training, in order to allow the models to achieve the best possible fit to the data, as some evidence in the original experiments demonstrated that model performance improved over time [11]. For each course, the trained model is used to predict learner dropout in the most recent offering of the course, again using the first three weeks of data. These predictions are compared to the true labels in order to conduct the slicing analysis.

The ABROCA statistic and AUC are computed for each of the five models on the held-out session of each course. We perform both aggregated statistical analysis, as well as detailed analysis of individual courses by inspecting informative graphical plots which we term "slice plots."

The gender data used in this experiment was inferred from the names of learners using the `gender-guesser` library[4] which has been used in other analyses of MOOCs [5]. This analysis only includes users for which high-confidence inferences were possible (users with e.g. gender-neutral names were excluded). While an imperfect measure, it is useful for demonstrating differences between subpopulations.

5.2 Results

The results of the experiment are summarized in Table 2, and shown in Figures 3 and 4. We provide a detailed slicing analysis on a single course in Section 5.2.1; we provide a more comprehensive overall analysis of our results in Sections 5.2.2 and 5.3. These results demonstrate several novel findings, which we evaluate using non-parametric statistical tests in order to make minimal assumptions about the distribution of the ABROCA statistic.

5.2.1 Detailed Case Study. Figure 2 shows the results of a single model (here, the LSTM model of [11]) applied to one of the 44 courses in our experiment, a business MOOC. Figure 2 shows the ROC curves for male students (ROC_b) in red, and female students (ROC_c) in blue. The ABROCA region is shaded (recall that the ABROCA statistic is the area of this shaded region). This same plot, along with the slice plots for all other models on the same course for comparison, is shown in Figure 3.

Figure 2 illustrates several features of slicing analysis using ABROCA. First, the plot demonstrates how much broader a slicing

[4] See https://pypi.org/project/gender-guesser/

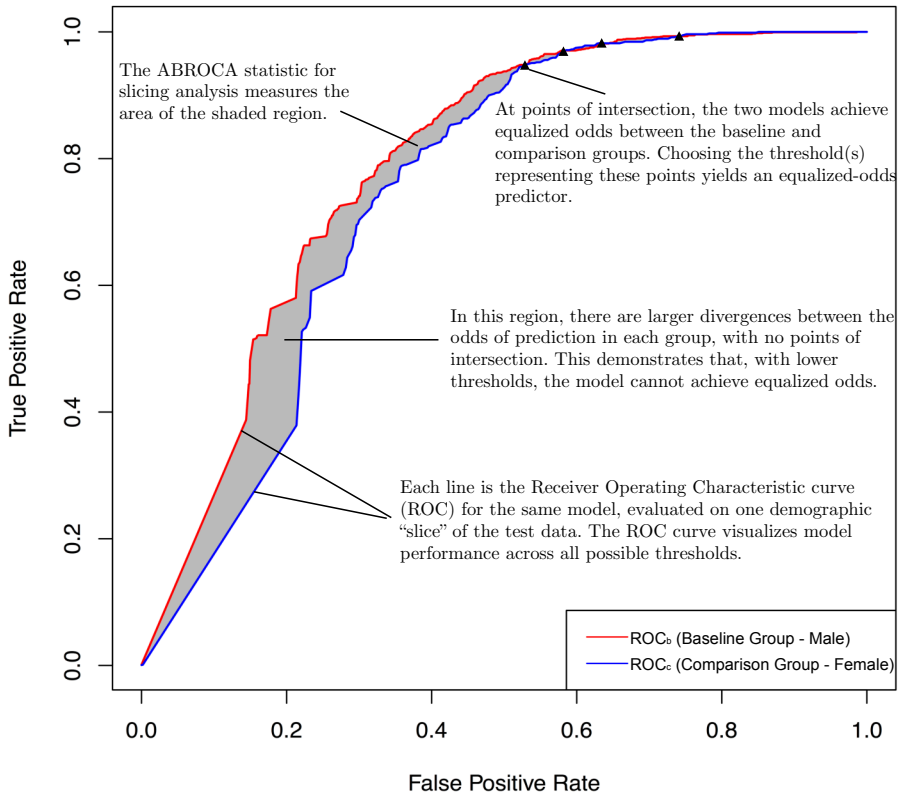

Figure 2: Annotated slice plot for a single course (a business course) in the case study, using gender as the protected attribute. Results from LSTM model shown. For comparison to results from other algorithms in the same course, see Figure 3.

analysis via ABROCA is than analysis using an equalized odds criterion. In particular, Figure 2 shows that there are only four thresholds (shown as four points annotated with ▲) where the model achieves equalized odds. Note that the equalized odds method would require choosing a specific threshold, and computing the odds of positive prediction for each of the two groups (comparing only the two specific points, one on each ROC curve, representing the odds for each group at that specific threshold). If the threshold representing any of those four points were chosen in this case, the model would achieve equalized odds. However, viewing the entire slice plot and using ABROCA demonstrates a more nuanced picture. The plot reveals that, while there are individual points where equalized odds are achieved, for the majority of the space of potential thresholds, the models cannot achieve even close to equalized odds, as demonstrated by the large vertical distance between the A-conditional ROC curves. This demonstrates that while the model might achieve equalized odds for higher thresholds, for lower thresholds, it does not. This is particularly relevant for models which may be used in learning analytics, where interventions may require choosing a very high or low threshold (such as a resource-intensive intervention which requires a low false-positive rate, and therefore a lower threshold). Unlike an equalized odds slicing analysis, ABROCA accounts for these differences across thresholds and can support decision making about the full range of thresholds.

Second, Figure 3 demonstrates an example of the variability between the models considered in this experiment. There are visible differences between the five models when applied to this business MOOC, both in terms of the size of the ABROCA statistic and the difference in the shape of the ROC curves. For example, the CART and NB models replicated from [14] achieve the lowest ABROCA statistics, both approximately 0.013. Both models' comparatively higher fairness across all thresholds is indicated by the relative closeness of the A-conditional ROC curves across their respective slice plots in Figure 3. The CART model achieves considerably better predictive performance, as indicated by the larger area underneath the A-conditional ROC curves (Table 1 also shows that the CART model achieved better average performance than the NB model, measured by AUC, across all courses in this experiment). These differences are likely due to a combination of factors, including the features extracted from the raw data, as well as the statistical algorithms and hyperparameters used in this implementation. Third, Figures 2 and 3 demonstrate, practically, how an educational modeler or practitioner might approach the process of model selection using slicing analysis. By conducting a visual comparison of the ABROCA regions, and combining this analysis with knowledge of which region(s) of the space of potential thresholds might be used for interventions the model will support, a user can choose a model for deployment, or determine which additional models to explore.

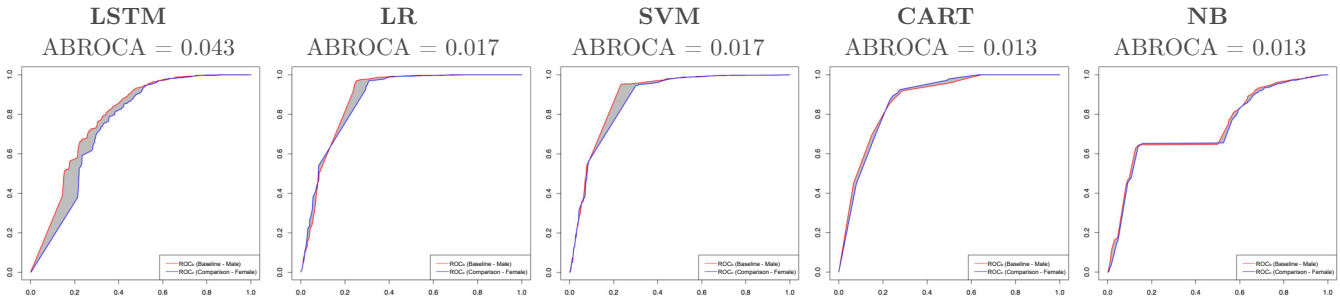

Figure 3: Slice plots for all models applied to the business MOOC used in Figure 2.

	ABROCA			AUC
	Mean	SD	95% CI	Mean
LSTM [11]	0.0351	0.006	[0.0221, 0.0480]	0.653
LR [11]	0.0353	0.005	[0.0261, 0.0444]	0.669
SVM [11]	0.0311	0.004	[0.0226, 0.0395]	0.693
CART [14]	0.0280	0.005	[0.0174, 0.0387]	0.715
NB [14]	0.0177	0.0037	[0.0103, 0.0251]	0.558

Table 2: Slicing analysis results for each model replicated across the 44 courses in MORF.

5.2.2 Overall Results. The complete results from the replication of the five predictive models across the MORF dataset is shown in Table 2, and ABROCA plots representing the performance of each model across a single business MOOC are shown in Figure 3. The results demonstrate several initial findings.

First and most generally, the results demonstrate that there is moderate variability in the ABROCA statistic across the models evaluated. We note, for example, that the mean ABROCA value of the classification tree model replicated from [14] is nearly 25% lower than the mean ABROCA value of either the LSTM or logistic regression model replicated from [11], while achieving the highest average AUC over all 44 courses. A Kruskal-Wallace test (a nonparametric statistical test of the null hypothesis that the mean ABROCA values are the same in each group, and of which the Wilcoxon rank sum test is a special case for two samples) of differences by model rejects the hypothesis that there is no difference in mean ABROCA between models, with $p = 8.106 \times 10^{-5}$. These results demonstrate non-trivial variability in the fairness of the models considered, providing evidence that it is important to consider fairness in learning analytics models, and to make slicing analysis, not simply performance analysis, a critical component of model evaluation for both research and deployment.

Second, these results provide evidence that the underlying *feature set* used as input may be equally relevant to the fairness of the resulting model. Note that the LSTM, LR, and SVM models use the same underlying set of 7 base features from [11], while the CART and NB models use a disjoint set of 191 features from [14] (see Table 1). Despite major differences in the way the algorithms statistically represent input features, our results show clear differences in fairness by feature set: Kruskal-Wallace test of differences by feature set (all models from [11] vs. all models from [14]) rejects the hypothesis that there is no difference in mean ABROCA between the groups with $p = 1.99 \times 10^{-5}$. This coheres with the initial findings in [14], which demonstrated similar results regarding the relative contributions of features in comparison to statistical algorithms or hyperparameter settings with respect to model performance. These results suggest that exploring different *feature sets* might be a particularly fruitful avenue for future slicing analyses, and that some characteristics of the feature set in [14] might lead to better fairness than the features in [11].

5.3 Exploratory Analysis

In this section we conduct an exploratory analysis of course factors and their association with the observed differences in model performance. In particular, we explore whether course size, gender imbalance, and subject area are associated with varying levels of model unfairness measured by ABROCA, and whether our results demonstrate a tradeoff between fairness and performance. Note that the exploration in this section does not permit causal inference about the observed associations, but serves as an exploratory evaluation of where such associations may exist.

Figure 4a shows the ABROCA of each model measured on the test data (the most recent session of each course in MORF).

First, the results in Figure 4a demonstrate a clear but complex relationship between the gender imbalance in a course and the ABROCA. Statistical testing confirms that the quadratic curve for each model visible in Figure 4a represents a statistically significant quadratic relationship between gender imbalance (measured by percentage of male students in the dataset). An F-test was used to compare a simple linear regression model and a quadratic regression which included intercept terms for each algorithm, and the F-test indicated (a) a statistically significant increase in the proportion of variance explained by the quadratic model ($p = 1.533 \times 10^{-5}$), and (b) a highly significant model fit from the resulting quadratic regression model, with $p = 2.277 \times 10^{-8}$. Collectively, these results suggest that very high, or moderately low (< 0.45) proportion of male students is associated with a significantly larger ABROCA statistic. We can see from Figure 4a that the shape of this relationship is remarkably consistent across all five models evaluated. This means that, in particular, models trained with highly-imbalanced training data may be particularly susceptible to differential performance across demographic groups, and that algorithm selection cannot overcome the limitations of biased training data.

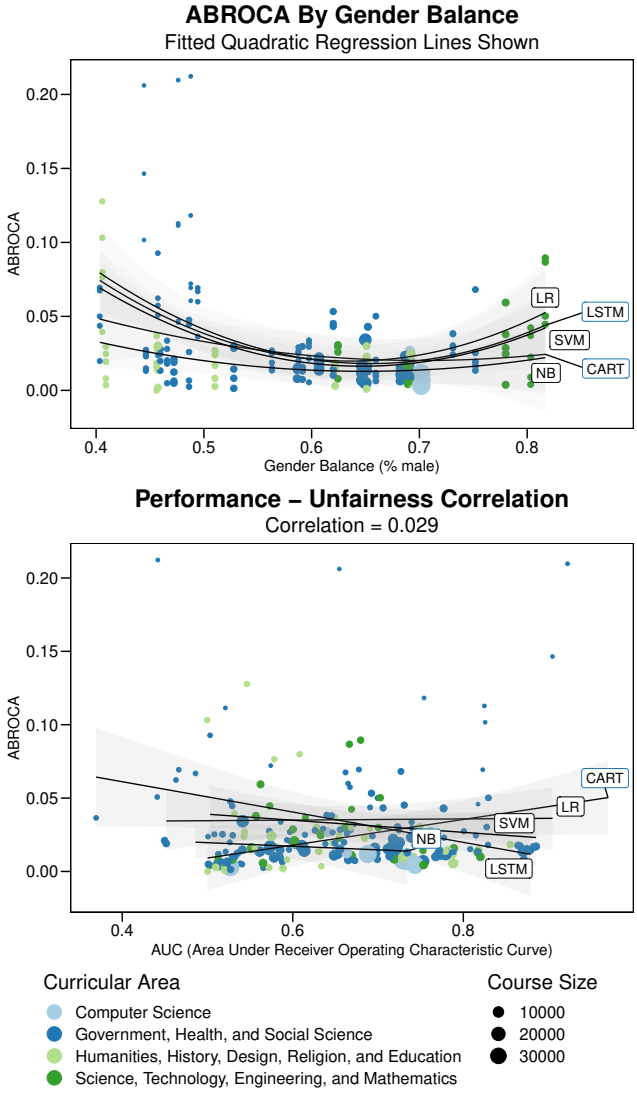

Figure 4: Top (a): Relationship between gender balance in course and ABROCA, with quadratic fitted lines by model. Below (b): Relationship between performance (measured by AUC) and unfairness (measured by ABROCA) with fitted regression lines by model. 95% confidence regions shaded.

Second, Figure 4 demonstrates findings related to both curricular area (coded according to the four-category grouping from [6]) and specific courses. With respect to curricular area, the results demonstrate an association between curricular area and ABROCA statistic: a Kruskal-Wallis rank sum test shows $p = 0.03692$, suggesting evidence of an association between curricular area and ABROCA statistic across the courses and models surveyed. An analysis by *course*, where each individual course is treated as a group (note that there are five observations of ABROCA for a given course, one per model in our experiment), shows $p = 1.195 \times 10^{-8}$. This indicates that differences between courses are quite strong: different courses in the dataset were associated with very different ABROCA statistics. This is visible in Figure 4 by noting that the same course (indicated by same-sized points, aligned vertically) often occupies a similar relative position for each model. These results suggest that modeling which accounts for course topic (measured by curricular area), and course-specific modeling, are promising avenues for future research on constructing unbiased MOOC models.

Third, and perhaps most critically, we present Figure 4b. Figure 4b shows the observed relationship between model performance (measured by overall AUC) and model fairness (measured by ABROCA) across each individual course-model pair applied to the test data, with fitted regression lines. Figure 4b demonstrates that we observe almost no relationship between performance and fairness across any of the five models evaluated. The overall correlation between AUC and ABROCA is 0.029, with $p = 0.6692$ using a Pearson test of whether the correlation $\rho = 0$. This means that we fail to reject a null hypothesis that there is no correlation between performance (AUC) and fairness (ABROCA) in the observed data. For every individual model, Figure 4b shows that a flat line is within the 95% linear regression band. This demonstrates strong initial evidence that we can have models that are both *fair* and *accurate*: if there is no correlation between the two, there is likely no strict tradeoff, or the bound at which a theoretical tradeoff may limit performance has not been reached by state-of-the-art MOOC modeling methods), providing further impetus for the use of slicing analysis to support learning analytics models.

6 CONCLUSION AND FUTURE RESEARCH

6.1 Discussion

This paper demonstrates a method for evaluating the fairness of predictive models in learning analytics through slicing analysis, and the specific insights that such an analysis can achieve. While we must be careful not to overstate the generalizability of the results of this specific experiment, our results reveal that slicing analysis in general, and the proposed ABROCA statistic in particular, can provide a perspective on model performance beyond mere predictive accuracy. ABROCA can provide insights into model discrimination across the entire range of possible thresholds. In particular, our analysis revealed that there were differences between various statistical algorithms and feature sets. Furthermore, our results show that model discrimination is related to the course gender imbalance, course curricular area, and the individual course itself. Finally, these results show that there does *not* appear to be a strict tradeoff between model performance and discrimination encountered by state-of-the-art MOOC dropout models. Some of these results may confirm readers' intuitions – e.g., that models trained on courses most skewed toward a subgroup also produce the greatest bias toward these students; model performance varies considerably by course and some courses may simply be "harder" to predict fairly on than others. The results demonstrating lower (although statistically significant) variability across algorithms, in particular, may also be counterintuitive to many learning analytics researchers – largely, prior research has primarily focused on applying and evaluating different feature engineering methods and algorithms, not evaluating the demographic balance of the training data or specific course characteristics [15]. This analysis suggests that better

transfer methods, or models which incorporate course-level characteristics, may produce considerable advances in both fairness and, potentially, predictive performance. Finally, our demonstration that these results do *not* show a strict tradeoff between fairness and performance suggests that slicing analysis can support the improvement of biased models *without* necessarily detracting from performance in practice.

6.2 Future Research

This work suggests several promising lines of future research, related to theory, methods, and applications extending its analysis. While the ABROCA statistic itself represents a useful measure for conducting slicing analysis, future work detailing the asymptotic properties of this statistic, which is a function of the ROC curve, which itself has well-established statistical properties [18], would aid in developing procedures for inference (for example, providing estimates of standard errors and significance for ABROCA). With respect to methods, the slicing analysis methodology can be added as an additional informative section to results reporting for applied predictive modeling research both within and outside learning analytics. Finally, the case study presented here represents only an initial demonstrative application of the slicing analysis method via ABROCA. Extensive future work detailing further predictive models, both replicated and novel, and slicing them along several demographic characteristics, would contribute to the learning analytics community's understanding of the potential impact of its predictive models across the many complex dimensions of student identity.

7 ACKNOWLEDGEMENTS

This work was supported in part under the Holistic Modeling of Education (HOME) project funded by the Michigan Institute for Data Science (MIDAS).

REFERENCES

[1] JML Andres, RS Baker, G Siemens, D Gašević, and S Crossley. 2018. Studying MOOC Completion at Scale Using the MOOC Replication Framework. In *Proc. LAK*. ACM, New York, 71–78.
[2] V Bakthavachalam and A Hickey. 2018. Using Technology to Increase Equality of Opportunity. https://digital.hbs.edu/data-and-analysis/using-technology-to-increase-equality-of-opportunity/. (2018). Accessed: 2018-9-17.
[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. Law Rev.* 104 (2016), 671.
[4] C G Brinton and M Chiang. 2015. MOOC performance prediction via clickstream data and social learning networks. In *IEEE INFOCOM*. IEEE, New York, 2299–2307.
[5] Christopher Brooks, Joshua Gardner, and Kaifeng Chen. 2018. How Gender Cues in Educational Video Impact Participation and Retention. In *Proc. ICLS*, J Kay and R Luckin (Eds.). ISLS, New York, 1835–1842.
[6] Isaac Chuang and Andrew Dean Ho. 2016. *HarvardX and MITx: Four Years of Open Online Courses – Fall 2012-Summer 2016.* Technical Report. Harvard/MIT.
[7] Kate Crawford and Ryan Calo. 2016. There is a blind spot in AI research. *Nature News* 538, 7625 (Oct. 2016), 311.
[8] S Crossley, DS McNamara, R Baker, Y Wang, L Paquette, T Barnes, and Y Bergner. 2015. Language to Completion: Success in an Educational Data Mining Massive Open Online Class. In *Proc. EDM*. 388–391.
[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proc. ITCS*. ACM, 214–226.
[10] C Dwork, N Immorlica, AT Kalai, and MDM Leiserson. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *Proc. FAT**, SA Friedler and C Wilson (Eds.), Vol. 81. PMLR, New York, NY, USA, 119–133.
[11] M Fei and D Y Yeung. 2015. Temporal Models for Predicting Student Dropout in Massive Open Online Courses. In *ICDMW)*. IEEE, New York, 256–263.
[12] M Feldman, SA Friedler, J Moeller, C Scheidegger, and S Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proc. KDD*. ACM, 259–268.
[13] SA Friedler, C Scheidegger, S Venkatasubramanian, S Choudhary, EP Hamilton, and D Roth. 2018. A comparative study of fairness-enhancing interventions in machine learning. (2018). arXiv:stat.ML/1802.04422
[14] J Gardner and C Brooks. 2018. Evaluating Predictive Models of Student Success: Closing the Methodological Gap. *Journal of Learning Analytics* 5, 2 (2018), 105–125.
[15] Josh Gardner and Christopher Brooks. 2018. Student Success Prediction in MOOCs. *User Modeling and User-Adapted Interaction* 28, 2 (2018), 127–203.
[16] J Gardner, C Brooks, JML Andres, and R Baker. 2018. Replicating MOOC Predictive Models at Scale. In *Proc. Learning@Scale*. ACM, New York, 1–10.
[17] Josh Gardner, Christopher Brooks, Juan Miguel L Andres, and Ryan Baker. 2018. MORF: A Framework for Predictive Modeling and Replication At Scale With Privacy-Restricted MOOC Data. (2018). arXiv:1801.05236
[18] J A Hanley and B J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 1 (1982), 29–36.
[19] Moritz Hardt, Eric Price, Nati Srebro, and Others. 2016. Equality of opportunity in supervised learning. In *NIPS*. 3315–3323.
[20] M Hind, S Mehta, A Mojsilovic, R Nair, K Natesan Ramamurthy, A Olteanu, and KR Varshney. 2018. Increasing Trust in AI Services through Supplier's Declarations of Conformity. (Aug. 2018). arXiv:cs.CY/1808.07261
[21] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. 2013. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. *Int Conf Affect Comput Intell Interact Workshops* 2013 (2013), 245–251.
[22] Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. 2014. Predicting MOOC dropout over weeks using machine learning methods. In *Proc. EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*. ACL, 60–65.
[23] ZC Lipton, A Chouldechova, and J McAuley. 2018. Does mitigating ML's impact disparity require treatment disparity? (2018). arXiv:1711.07076
[24] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. (March 2018). arXiv:cs.LG/1803.04383
[25] Matthew C Makel and Jonathan A Plucker. 2014. Facts are more important than novelty: Replication in the education sciences. *Educ. Res.* 43, 6 (2014), 304–316.
[26] P M Moreno-Marcos, C Alario-Hoyos, P J Muñoz-Merino, and C Delgado Kloos. 2018. Prediction in MOOCs: A review and future research directions. *IEEE Trans. Learn. Technol.* NA, 99 (2018), 1–1.
[27] P Prinsloo and S Slade. 2014. Educational triage in open distance learning: Walking a moral tightrope. *The Intl. Rev. of Res. in Open and Distributed Learning* 15, 4 (Aug. 2014).
[28] Lynne D Roberts, Vanessa Chang, and David Gibson. 2017. Ethical Considerations in Adopting a University- and System-Wide Approach to Data and Learning Analytics. In *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*, Ben Kei Daniel (Ed.). Springer, Cham, 89–108.
[29] Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.* 29, 5 (Nov. 2014), 582–638.
[30] D Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's Curse? On Pace, Progress, and Empirical Rigor. In *ICLR Workshops*. Vancouver, CA.
[31] Sharon Slade and Paul Prinsloo. 2013. Learning Analytics: Ethical Issues and Dilemmas. *American Behavioral Scientist* 57, 10 (March 2013), 1510–1529.
[32] C Taylor, K Veeramachaneni, and U O'Reilly. 2014. Likely to stop? Predicting Stopout in Massive Open Online Courses. (2014). arXiv:cs.CY/1408.3382
[33] J Whitehill, K Mohan, D Seaton, Y Rosen, and D Tingley. 2017. Delving Deeper into MOOC Student Dropout Prediction. (Feb. 2017). arXiv:cs.AI/1702.06404
[34] J Whitehill, J Williams, G Lopez, C Coleman, and J Reich. 2015. Beyond prediction: Toward automatic intervention to reduce MOOC student stopout. In *Proc. EDM*. 171–178.
[35] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. 2013. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proc. 2013 NIPS Data-driven education workshop*, Vol. 11. 14.
[36] James Zou and Londa Schiebinger. 2018. AI can be sexist and racist — it's time to make it fair. *Nature* 559, 7714 (July 2018), 324.