Academic performance prediction in a gender-imbalanced environment

Piotr Sapiezynski
Northeastern University

Valentin Kassarnig
Graz University of Technology

Christo Wilson
Northeastern University

Sune Lehmann
Technical University of Denmark

Alan Mislove
Northeastern University

ABSTRACT

Individual characteristics and informal social processes are among the factors that contribute to a student's performance in an academic context. Universities can leverage this knowledge to limit drop-out rates and increase performance through interventions targeting at-risk students. Data-driven recommendation systems have been proposed to identify such students for early interventions. However, as we show in this paper, it is possible to identify certain groups of students whose performance is best predicted using indicators that differ from those predictive for the majority. Naïve approaches that do not account for this fact might favor the majority class and lead to disparate mistreatment in the case of minorities. In this paper we investigate the low academic performance predictors of female and male participants of the Copenhagen Networks Study. We find that social indicators (e.g. mean grade point average of peers or fraction of low-performing peers) predict low-performance of male participants more accurately than they do for female participants, and that this situation is reversed for individual behaviors. Because of the gender imbalance among the participants, optimal gender-oblivious models detect low-performing male students with higher accuracy than low-performing female students. We review the existing approaches to addressing the disparate mistreatment problem and propose our own method that outperforms the alternatives on the dataset in question.

ACM Reference format:
Piotr Sapiezynski, Valentin Kassarnig, Christo Wilson, Sune Lehmann, and Alan Mislove. 2017. Academic performance prediction in a gender-imbalanced environment. In *Proceedings of FATREC Workshop on Responsible Recommendation at ACM RecSys, Como, Italy, August 2017 (FATREC'17),* 4 pages.
https://doi.org/10.18122/B20Q5R

1 INTRODUCTION

One of the central driving forces behind the adoption of algorithmic decision-making is the goal of eliminating biases from the decision process. However, it has recently been shown that these algorithms can have the opposite effect, possibly as a consequence of how the data is mined [2]. Algorithmic biases have been demonstrated in the systems that make decisions (or aid the human decision making process) in areas as diverse as loans [10], parole [12], hiring [10], and policing [15].

A growing body of fairness research emphasizes a range of problems with black box algorithms. There exist multiple definitions of fairness, some of which have been shown to be mutually exclusive [9]. The discussion is especially heated around *disparate mistreatment*: a situation in which error rates in a decision making process are not balanced between representatives of a particular characteristic (e.g. gender or race). Angwin et. al. [12] argued that the system judges use as an assistant in their parole decisions is more likely to wrongly imprison blacks than whites. The article provoked a series of responses, which argued that the system was indeed fair, but according to a different definition of fairness [6, 8]. The notion of disparate mistreatment was formalized by Zafar et al. in a recent article which also introduces an approach of solving the problem through constrained training of the classifier [23].

Independently of the research on fairness, there is increasing interest in data-driven predictions of academic performance and intervention recommendations. For example Balfanz, et al. [1] proposed a system based on school records that recommends targeted interventions to activate students at high risk of dropping out from high school. More recently, Wang et al. [21] showed that the academic performance can also be predicted from behavioral data collected using smartphones. In a student population we studied recently, social indicators proved to be more predictive of academic performance than the behavior or characteristics of the individual [14]. These social factors (including mean grade point average of peers and the fraction of low-performing peers) were more highly correlated with an individual performance than, for example, class attendance. In this paper, we ask whether these findings hold equally for men and women in the dataset. Further, we ask whether a model built on these features works equally well for the two sexes. Finally, we review the existing methods of avoiding disparate mistreatment and propose a novel approach, based on constrained forward feature selection. Instead of optimizing the classifier for best overall performance, we constrain the training process by progressively adding features so that the model maintains comparable performance for all groups of the protected feature (i.e. for men and women). While this simple approach might not work on datasets where balanced features are absent, it does outperform other methods on our dataset. Of course, while our method can accurately identify low-performing male and female students, recommending particular interventions lies beyond the scope of this study.

FATREC'17, August 2017, Como, Italy

DOI: 10.18122/B20Q5R

Table 1: Summary statistics of the dataset. There is no statistically significant difference between performance among men and women in the study ($p_{val} = 0.65$ in Kolomogorov-Smirnov test)

	Performance			
	Low	Medium	High	**Total**
Male	142	141	137	**420**
Female	38	39	43	**120**
Total	**180**	**180**	**180**	**540**

2 METHODS

2.1 Data

The data used in this paper was collected as part of the Copenhagen Networks Study (CNS), a large scale computational social science study designed to measure human interactions and mobility with high resolution [20]. The approximately 800 participants of the study were freshmen and sophomores at the Technical University of Denmark. After responding to an online questionnaire on psychological and health indicators, they were equipped with an instrumented smartphone (Google Nexus 4) that—with their consent—tracked their location, proximity to other participants, and communication instances (metadata of short messages and calls, without the content). Finally, the vast majority of the participants (717 out of 839) opted in to share their Facebook data as well, which was acquired using Facebook API. The data collection campaign lasted two years. In this study we focus on participants who interacted with at least three other subjects through phone calls, short messages, face to face, and on Facebook. There are 420 men and 120 women in the dataset, and this gender imbalance corresponds to the imbalance in the overall student population. We divide the students into three equally-sized groups based on their GPA after two years. Table 1 presents summary statistics.

We derive a number of variables in the following feature categories:

Individual behaviors. *Class attendance* is computed from location data combined with class schedule using the method we previously described [13]; it corresponds to the fraction of lectures and exercises a student attended within the courses they signed up for. *Facebook activity* score is defined as the mean number of status updates a student posted in a week during the duration of the observation.

Individual characteristics. This dataset was obtained through an online questionnaire and includes: The Big Five [11] (*neuroticism*, *openness*, *conscientiousness*, *extraversion*, *agreeableness*), Rotter's *Locus of Control* [18], *stress* [4], *self-esteem* [17], *satisfaction with life* [5], PANAS (*positive* and *negative*) [22], *loneliness* [19], *depression* [3], and narcissism (*rivalry*, *admiration*, *overall*) [7].

Network characteristics. *Degree centrality* measures, one for each of four interactions networks: in physical space (person-to-person proximity measured using Bluetooth), calls and short message exchanges, and Facebook interactions.

Peer performance. Knowing the underlying social networks (proximity, phone communication, and Facebook) as well as the grades of each participant, we computed the *mean GPA* of each persons' peers, as well as *fraction of low/high-performers* (two features for each interaction network).

2.2 Classifier training

In each problem, we train a common classifier, oblivious to gender. We use k-fold cross-validation with $k = 3$ (due to the low number of female samples in the dataset we maintained a small k to avoid folds with no women). In each test fold, we calculate the performance on (a) all test samples, (b) only male samples, and (c) only female samples, and report these in figures. As we showed in our previous work [14], Linear Discriminant Analysis (LDA) is the machine learning approach that achieves the highest results with the dataset (compared against logistic regression, random forest, and SVC). We tune hyper-parameters through grid search cross-validation separately for each feature-set.

3 RESULTS

3.1 Detecting low-performing students

We divide students into three equally sized groups based on their grade point average (GPA): low-, mid-, and high-performing students. In this article we focus on identifying low performing students. Hence, we rephrase the problem as a binary classification task, where the target class are the low-performers, consisted with identifying students to intervene. We then use four fine-tuned LDA models to predict student performance each based on a different feature-set: individual characteristics, individual behaviors, network centrality, and peer performance. We then combine first two categories and train the 'individual' model; we combine the third and fourth sets and train the 'network' model. We then combine all features into a 'combined' model.

As shown in Figure 1, peer-performance is a good predictor of low performance amongst men, but the signal is weaker for female students. Combining the individual and network features into a common model results in a gap in predictive performance between men and women (*AUC ROC* = 0.84 and 0.67, respectively). To better illustrate this effect, we investigate example cumulative distributions of social and individual features among the genders with respect to performance, see Figure 2.

3.2 Fair predictions through feature selection

Now we build a model which maximizes a prediction performance metric in the low-performers' detection problem, while constraining the difference of performance between genders. We adapt a forward feature selection strategy: we start by selecting the feature that has the highest predictive power for the entire population while satisfying the requirement given in Eq. 1:

$$\frac{|P_m - P_w|}{P_{total}} \leq \epsilon, \tag{1}$$

where ϵ is a parameter controlling how much inter-gender difference we are willing to allow, and P is the selected performance metric, for example area under receiver characteristic curve (*AUC ROC*), or Matthew's Correlation Coefficient (*MCC*). We then add more features, one by one, in a way that the new model has increasing P score and satisfies the requirement from Eq. 1.

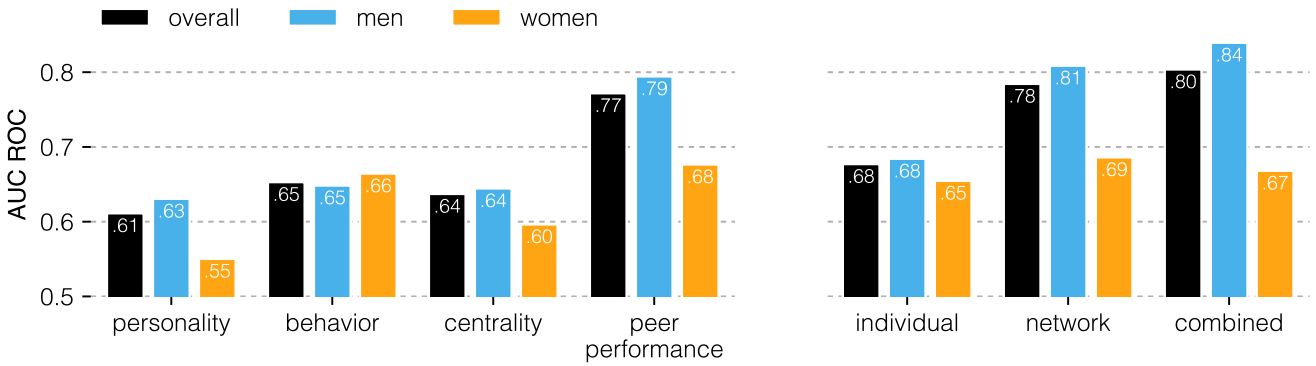

Figure 1: Low-performers' detection. Peer-performance is an efficient predictor of low performance amongst men, but the signal is much weaker for female students. Note, that the *AUC ROC* of a random classifier would be equal to 0.5, so all feature categories provide signal related to low academic performance.

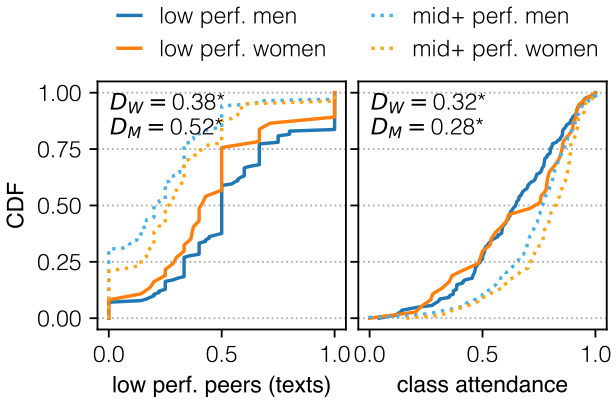

Figure 2: We use the Kolmogorov-Smirnov test on cumulative distribution functions (CDF) of two features (*fraction of low-performing peers* in the text network, and *class attendance*) to measure how dissimilar low-performing students of each gender are from the high performers. We find larger differences for men than women in the peer performance feature. However, the difference is larger for women in the individual behavior feature. Annotated are the results of K-S test, marked with the (*) symbol wherever significant with $p_{val} < 0.05$.

Figure 3 shows the results of training such fair classifiers. It emphasizes the trade-off between overall performance and fairness: the bigger the allowed difference between genders, the higher the overall performance. Typically, in binary classification tasks *AUC ROC* is used to measure the performance of the classifier. In this case, however, using *AUC ROC* might be misleading: it summarizes the performance of a classifier at all thresholds, but a classifier put to use would have to operate at a chosen threshold. Even if *AUC ROC* scores are balanced, the classifier at a particular threshold might still suffer from the disparate mistreatment problem. Therefore, we perform the constrained forward feature selection using Matthew's correlation coefficient [16]. It quantifies the performance at a threshold and—contrary to the popularly used F_1 score—penalizes the classifier for classifying all samples as the target class (such a classifier on this dataset has $MCC = 0$ and $F_1 = 0.5$). We define *MCC* in Eq 2.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \quad (2)$$

3.3 Alternative approaches

Figure 4 compares the results achieved through constrained forward feature selection (CFFS), the method proposed by Zafar et al. [23], re-balancing the dataset, as well as training separate models for men and women. Because of too few female subjects in the data, training separate models results in severe penalty on performance of the female-only model. Re-balancing the dataset as well as the approach proposed by Zafar et al. [23] achieve better results. Constrained forward feature selection achieves high and nearly equal *MCC* for both genders.

4 DISCUSSION

In this work we showed that empirical data can be more predictive for a one group of subjects than other groups, and the problem might go unnoticed unless specifically investigated. The situation we described is not simply the case of imbalance, as re-balancing the data does not solve the issue. Instead, we found that fair learning can be achieved by only learning on selected features. The solution is not generalizable to all datasets–depending on the problem, there might be no features that perform similarly well for representants of all classes among the protected feature. We tested our approach on other datasets. It fails, for example, to solve the disparate mistreatment problem in the COMPAS dataset [12], where all predictive features achieve higher performance for one of the races. Therefore, rather than recommending our approach for use in all scenarios, we limit our conclusion to emphasizing the need for considering the diversity of users in machine learning systems.

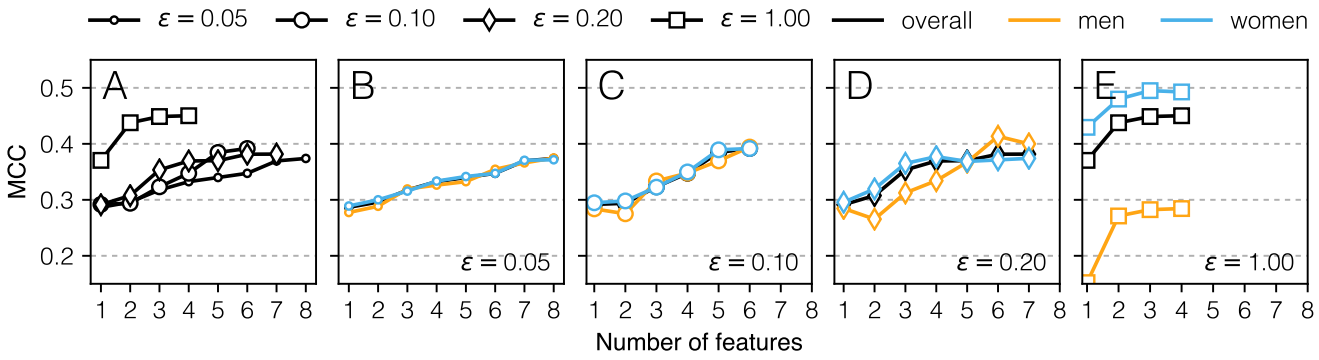

Figure 3: Learning fair classifiers. In each step we extend the model with a feature to maximize the overall performance of the classifier while maintaining the maximum disparity ϵ between genders. $\epsilon = 1$ means there is no constraint on parity. Note, that a constrained classifier has a higher performance for the underrepresented class than the unconstrained classifier. Note that for a random classifier $MCC = 0$. The selection process stops when no more features can be added to improve performance while maintaining performance parity, hence a possible difference in the number of features used depending on ϵ.

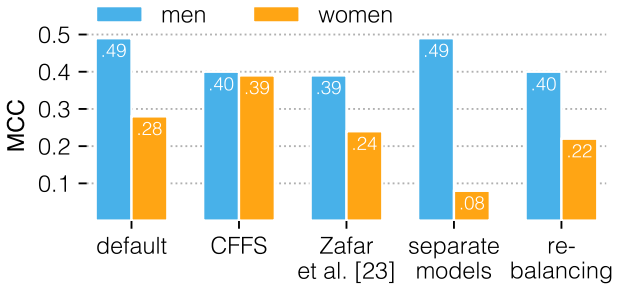

Figure 4: Alternative approaches to learning fair classifiers. On the dataset in question, the constrained forward feature selection (CFFS) method outperforms other approaches.

REFERENCES

[1] Robert Balfanz, Liza Herzog, and Douglas J. Mac Iver. 2007. Preventing Student Disengagement and Keeping Students on the Graduation Path in Urban Middle-Grades Schools: Early Identification and Effective Interventions. *Educational Psychologist* 42, 4 (2007), 223–235. https://doi.org/10.1080/00461520701621079 arXiv:http://dx.doi.org/10.1080/00461520701621079

[2] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671–732. https://ssrn.com/abstract=2477899

[3] P Bech, N-A Rasmussen, L Raabæk Olsen, V Noerholm, and W Abildgaard. 2001. The sensitivity and specificity of the Major Depression Inventory, using the Present State Examination as the index of diagnostic validity. *Journal of affective disorders* 66, 2 (2001), 159–164.

[4] Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior* (1983), 385–396.

[5] Ed Diener, Robert A. Emmons, Randy J. Larsen, and Sharon Griffin. 1985. The satisfaction with life scale. *Journal of Personality Assessment* 49, 1 (1985), 71–75.

[6] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. *COMPAS risk scales: Demonstrating accuracy equity and predictive parity.* Technical Report. Technical report, Northpointe, July 2016. http://www. northpointeinc. com/northpointe-analysis.

[7] Robert A Emmons. 1984. Factor analysis and construct validity of the narcissistic personality inventory. *Journal of personality assessment* 48, 3 (1984), 291–300.

[8] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks. *Fed. Probation* 80 (2016), 38.

[9] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *CoRR* abs/1609.07236 (2016). http://arxiv.org/abs/1609.07236

[10] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr.. In *CSCW*. 1914–1933.

[11] Oliver P John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 1999 (1999), 102–138.

[12] Angwin Julia, Larson Jeff, Mattu Surya, and Lauren Kirchner. 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-in-criminal-sentencing

[13] Valentin Kassarnig, Andreas Bjerre-Nielsen, Enys Mones, Sune Lehmann, and David Dreyer Lassen. 2017. Class attendance, peer similarity, and academic performance in a large field study. *CoRR* abs/1702.01262 (2017). http://arxiv.org/abs/1702.01262

[14] Valentin Kassarnig, Andreas Bjerre-Nielsen, Enys Mones, Piotr Sapiezynski, Sune Lehmann, and David Dreyer Lassen. 2017. Academic Performance and Behavioral Patterns. (2017). Under review.

[15] Kristian Lum and William Isaac. 2016. To predict and serve? *Significance* 13, 5 (2016), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

[16] Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405, 2 (1975), 442–451.

[17] Morris Rosenberg. 1965. *Society and the adolescent self-image.* Princeton university press Princeton, NJ.

[18] Julian B Rotter. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied* 80, 1 (1966), 1.

[19] Daniel W Russell. 1996. UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment* 66, 1 (1996), 20–40.

[20] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. 2014. Measuring large-scale social networks with high resolution. *PloS one* 9, 4 (2014), e95978.

[21] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14.

[22] David Watson, Lee A Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[23] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 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, 1171–1180.