ADDRESSING MACHINE LEARNING BIAS IN LAW SCHOOL ADMISSIONS

By Jada Mosier

Introduction

This paper will focus on the applications of ML algorithms to aid in decision making for law school admittance with the goal of developing an algorithm that evaluates individual features such as LSAT score to predict the likelihood of passage of the bar exam while minimizing bias.

The implementation of ML in this domain can be challenging as admission decisions are deeply impactful on both the prospects and careers of students as well as to the reputations of colleges and universities. Yet, ML presents new opportunities for the improvement of education as algorithms may be better equipped to predict the success of future law school students and have the potential to reduce human biases. Nevertheless, it is privy to unethical bias that could have problematic consequences.

The provided dataset from Kaggle contains variables such as LSAT scores, undergraduate GPA, and bar exam outcomes, as well as several more nuanced variables such as gender, race, and income. With the data provided, our objective is to audit an algorithm to uncover any racial and gender bias through an intersectional lens, attempt to mitigate bias, and to provide a critical analysis of existing uses and methods of model training to benefit the future development of ML in education.

Characterization of Use

Within education, machine learning is increasingly being used as a tool for various administrative and learning processes. Currently, ML is being integrated into educational systems in the form of intelligent tutoring platforms that assess learner performance to offer targeted feedback and guidance (Johnson, 2023). The data currently being used for developing ML models catering to education typically consists of testing scores, demographics, and GPA. The data provided is then used to predict targets such as academic performance, graduation probability, and career success, and drive decisions of admission for students.

Another prominent use, and one we will be focusing on for the purposes of our research is the use of predictive models to forecast student success. This is usually done for the purpose of admission decisions into programs or universities, as traditional admission methods are both time-consuming and prone to human bias. Such models use historical data to predict future outcomes such as grades, graduation probabilities, and post-graduation success, but are susceptible to bias as factors such as family income, parental occupation, and parental education levels had a significant effect on college retention and graduation. (Hutt et al., 2019)

Ethical Considerations

Sampling Bias

Sampling bias poses a significant risk in the education domain, especially for predictive models assessing student success or aiding in admission decisions. This bias occurs when the training data fails to accurately represent the target population, leading to skewed predictions and furthering existing inequities. In law school admissions and bar exam performance modeling, sampling bias could be present through:

- Underrepresentation of demographic groups: Lack of diverse data across races, ethnicities, or socioeconomic statuses affects predictions for underrepresented groups (e.g., lower-income or non-traditional students).
- Selection bias: Datasets from self-selected applicants or specific institution types (e.g., highly selective schools) may lack diversity.
- Historical bias: Data reflecting past discrimination or systemic inequalities in education or the legal profession can perpetuate biases in the model.

Sampling bias in this context can lead to unfair denials of admission or inaccurate exam predictions, exacerbating educational disparities. To address this, it's crucial to include data from underrepresented groups and regions. Rigorous data collection and sampling practices should be implemented to ensure a representative dataset. If necessary, reweighting and resampling techniques can be used during data preprocessing. Additionally, examining historical context and potential biases in the data is essential to identify and address systemic inequalities or discriminatory practices.

Differential Subgroup Validity

When considering the use of ML in the domain of education, particularly in the prediction of bar exam performance, differential subgroup validity can arise due to discrepancies in the data the model learns from, leading to differences in accuracy across varying demographic groups. For example, variables commonly used in law school admissions, such as LSAT scores and GPA, largely depend on access to adequate resources and opportunities.

Use of a model biased against demographic groups whose data is underrepresented in the training data or who have historically been disadvantaged in the education system can limit their admission possibilities and scholarship opportunities, thus perpetuating existing disparities. The groups impacted by this include racial minorities such as Black applicants who have, on average, one of the two lowest LSAT scores of all racial groups (Lauth & Sweeney, n.d.) and who are significantly less likely to be accepted to law school, with nearly half of Black law school applicants (49%) being rejected on every law school application. “That share is larger than that of any other racial or ethnic group” (Jaschik, 2019). Black applicants are also less likely to apply to law school. In 2024, Black applicants made up only 12% of all applications as compared to White applicants who made up 51% (*Law school admission council: Volume summary,* 2024). Additionally, law school applicants who identify as female, those from lower socioeconomic backgrounds or from non-traditional educational backgrounds may also be negatively impacted as these groups have been historically disadvantaged by or are underrepresented in traditional legal education.

To mitigate the risk of differential subgroup validity presenting itself in ML algorithms used to influence admission decisions and to ensure that the data used to train such algorithms is properly representative of all potential applicants, ML engineers can resample or reweight the data. They can also implement changes to the model itself such as using decoupled classifiers or regularizing loss. Moreover, it is vital to consider the impacts on applicants who represent intersectional identities, as the predictive power for people falling into these subgroups may be lower unless they are specifically considered.

Accuracy/Non-accuracy Affecting Injustices

The accuracy and non-accuracy affecting injustices that are produced by a model in the education domain can have broad implications on justice and fairness. In the context of law school admissions and bar exam performance predictions, the consequences of an accuracy affecting injustice can affect

individual opportunities and perpetuate systemic inequalities. For example, qualified individuals could be denied admission or opportunities due to the model's errors or biases. This could disproportionately impact underrepresented or marginalized groups, increasing existing educational disparities and limiting access to legal careers. Inaccurate predictions of bar exam performance could also result in misallocated resources or support systems. Students who are inaccurately predicted to struggle may receive additional resources or interventions, while those who are inaccurately predicted to succeed may not receive the support they need, potentially hindering their chances of success.

To address this issue and its impact on justice, it is crucial to evaluate and monitor the model's performance across different subgroups and scenarios. Additionally, making sure the predictive model is interpretable and understandable can aid in making sure we are inspecting the model for potential biases or unfair treatment of certain groups, and take appropriate actions to mitigate such issues. By prioritizing accuracy and fairness, predictive models in the education domain can promote justice and equitable opportunities for all individuals, regardless of their background or demographic characteristics.

Algorithmic Bias Assessment

We chose to develop a ML model around the task of predicting bar passage with the goal of informing law school admission decisions. The dataset we used included data on 22,407 law school students and information on if they eventually passed the bar exam. The features we used to train the model include an applicant's undergraduate GPA, LSAT score, birth year, whether they were a full time or part time student, whether they graduated or dropped out, their family income, age, gender, and race. For simplicity in our initial analysis, we chose to focus on black vs. non-black applicants (denoted "OtherRaces" in our model) for the sensitive feature within our model. We split the dataset into training and testing subsets, and applied the data to a logistic regression model as we had too few features to achieve the results we wanted with a random forest classifier and found that it underfit our data.

The overall metrics of the model showed more inaccuracies for Black applicants than for other races. The accuracy for Black applicants was 32.37%, indicating a bias, while other races had a much higher observed accuracy of 77.3%. There were also substantial differences in the false negative rate (FNR) and false positive rate (FPR) between subgroups, with Black applicants having a very high FNR of 84.96% as compared to other races which had a combined FNR of 22.01%, indicating that the model inaccurately classified a very large portion of Black applicants as unqualified to pass the bar. Interestingly, the FPR for other races was higher at 40.31% as compared to Black applicants that had a FPR of 3.03%, telling us that the model expected non-Black applicants to pass the bar at much higher rates than was accurate. The precision metric was high for both groups at 94.87% for Black applicants and 98.01% for others. Recall was good for other races at 77.99% and very low for Black applicants at a meager 15.04%, indicating that the model incorrectly mislabeled the vast majority of Black applicants as unable to pass the bar exam.

Algorithmic Bias Metrics

To perform our audit, we focused on race as a sensitive feature, specifically classifying applicants as Black vs non-Black. Additionally, we chose to focus on race with the additional feature of binary gender to perform a more thorough and intersectional analysis of the model's treatment of Black women in predicting educational outcomes. Thus, we created four subgroups to focus on: Black women, Black men, Other women, and Other men. This choice was driven by our research into the disparities in law school admission and LSAT scores, as Black applicants and female applicants were less likely to take the

LSAT and consequently, apply to law school (Lauth & Sweeney, n.d.). This led us to be interested in analyzing the unique features of intersectionality and the predictive metrics for Black female applicants, for whom we predicted the model to hold the least predictive power.

Assessing bias in our algorithm involved looking at multiple fairness metrics including equal opportunity difference and disparate impact. Using other men as the reference group (due to them having the highest recall), the equal opportunity difference for Black men was 0.594 and 0.661 for Black women as compared to 0.011 for other women. Black men and women show a high disparity in recall, indicating that the model is much less likely to correctly predict for Black women and men to pass than other men. Furthermore, to evaluate the difference in fairness between Black and non-Black women, by using other women as the reference group, we can see that the equal opportunity difference for Black women is still substantial at 0.65. Additionally, while small, there is an equal opportunity difference of 0.67 between Black men and women when Black men are used as the reference group. Thus, the model is significantly biased against both Black men and women, even more so against Black women, as they showed the highest equal opportunity difference in comparison to every other subgroup.

Moreover, we looked at disparate impact with a threshold of 0.8 to measure fairness. The disparate impact ratio for Black men is 0.242 and 0.157 for Black women as opposed to 0.986 for other women with other men used as the reference group. This indicates unequal treatment of certain subgroups by the model, as the ratios for Black men and women both fall below the fairness threshold. While other women are above the threshold, the model is less fair to this group than to other men, however, the impact can still be deemed acceptable and wouldn't classify as disparate impact based on our threshold. The fairness measures used are important ways to ensure fairness in algorithms used to predict education outcomes. Fairness in such algorithms is critical as access to education can influence one's intragenerational mobility and the overall equality of educational spaces.

Regarding our audit results, it is clear that there are significant disparities in how the model performs for Black vs non-Black applicants, and furthermore for Black women vs non-Black women and non-Black men, indicating the occurrence of differential subgroup validity and the presence of weaker predictive power for specific intersectional identities (in our case, Black female applicants). As the model incorrectly mislabeled most Black women as unable to pass the bar, this intersectionality marginalized group faced magnified challenges from underrepresentation.

Since the model is less effective at correctly identifying positive outcomes for Black applicants than others, if implemented, could lead to Black applicants being rejected at higher rates from law schools or being less likely to receive merit-based scholarship funding, limiting their opportunities and contributing to systemic inequalities. This disparity can be attributed to sampling bias, as there are fewer female and Black law school applicants, and Black applicants have lower average LSAT scores as a result of disadvantage. Other sources of injustice include historical bias and disadvantage embedded in the data, cultural biases in standardized testing and non-cognitive factors influencing academic success. Additionally, our model exhibited a high false positive rate for non-Black applicants, meaning it expected them to pass the bar at much higher rates than was accurate. Such inaccuracies could lead to misallocation of resources and opportunities, disproportionately affecting underrepresented groups and racial minorities.

Conclusion

The implications of these findings are extremely meaningful for educational institutions. Failing to address these ethical concerns can reinforce systemic inequalities, historical injustices, and limited access to educational and career opportunities. To address these ethical concerns we recommend ensuring

representative and inclusive datasets including data from underrepresented groups, regions, and institutions. We also recommend a critical examination of historical biases present in the training data, and to consider gathering different features that are more holistic.

References

Application gallery. Clever. (2024, March 27). https://www.clever.com/app-gallery/schoolinks

Hutt, S., Gardner, M., Duckworth, A. L., & D'Mello, S. K. (2019). Evaluating fairness and generalizability in models predicting ... https://files.eric.ed.gov/fulltext/ED599210.pdf

Jaschik, S. (2019, April 14). *Study argues that law schools limit black enrollment through the LSAT*. Inside Higher Ed | Higher Education News, Events and Jobs. https://www.insidehighered.com/admissions/article/2019/04/15/study-argues-law-schools-limit-black-enrollment-through-lsat#:~:text=In%20the%202016%2D17%20admission,Asian%20test%20takers%20is%20153

Johnson, R. (2023, October 23). *The role of Artificial Intelligence in Elearning: Integrating AI tech into education*. eLearning Industry. https://elearningindustry.com/role-of-artificial-intelligence-in-elearning-integrating-ai-tech-into-education#:~:text=AI%2Dpowered%20intelligent%20tutoring%20systems,offer%20targeted%20feedback%20and%20remediation.

Lauth, L. A., & Sweeney, A. T. (n.d.). *LSAT performance with regional, gender, and racial and ethnic breakdowns: 2011–2012 through 2017–2018 testing years (TR 22-01)*. LSAC. https://www.lsac.org/data-research/research/lsat-performance-regional-gender-and-racial-and-ethnic-breakdowns-2011-2018#:~:text=Average%20LSAT%20scores%20were%20highest,the%20lowest%20average%20LSAT%20scores.

Law school admission council: Volume summary. Law School Admission Council | Volume Summary. (2024, April 20). https://report.lsac.org/VolumeSummary.aspx

Muhie, Y. A., & Wolde, A. B. (2023b, April). (PDF) integration of Artificial Intelligence Technologies in teaching and learning in higher education. https://www.researchgate.net/publication/370230833_Integration_of_Artificial_Intelligence_Technologies_in_Teaching_and_Learning_in_Higher_Education