CPH TECH POLICY BRIEF #5

SEPTEMBER 2023

PUBLISHED BY THE CROWN PRINCESS MARY CENTER, COPENHAGEN CENTER FOR SOCIAL DATA SCIENCE & DEPARTMENT OF POLITICAL SCIENCE / UNIVERSITY OF COPENHAGEN

ALGORITHMIC ADMISSIONS: PREDICTING COLLEGE DROPOUT RISKS WITH GRANULAR ACADEMIC DATA

BY MAGNUS LINDGAARD NIELSEN & ANDREAS BJERRE-NIELSEN

This fifth edition of the CPH Tech Policy Brief presents insights from a new research study by Magnus Lindgaard Nielsen, Jonas Skjold Raaschou-Pedersen, Emil Chrisander, Julien Grenet, Anna Rogers, David Dreyer Lassen, and Andreas Bjerre-Nielsen: "Transforming Prediction Policy: How Novel Machine Learning Methods Can Improve Higher Education Admission"[1].

OVERVIEW

The development of artificial intelligence and machine learning in the recent decade has been nothing short of meteoric. This CPH Tech Policy Brief investigates how the recent advances in AI and machine learning could be employed for algorithmic policy making. We summarize our recent work that outlines how to employ the recent advances in machine learning and that provides evidence of how such AI-based systems can improve policy decisions while preserving the same level of fairness as human decisions[1]. The context of the study is admissions to higher education in Denmark[2] where we find that the new methods outperform standard approaches and are more objective in assessing academic aptitude than current admission criteria when comparing different students' backgrounds.

RENAISSANCE IN AI

Over the past decade, the realm of artificial intelligence (AI) has witnessed a transformative renaissance. What was once confined to the realm of academic curiosity and rudimentary algorithms has swiftly blossomed into a powerhouse of innovation, affecting virtually every facet of our daily lives. This shift was largely catalyzed by the rediscovery of models known as neural networks in the guise of deep learning, a technique that has demonstrated remarkable prowess in tasks ranging from image recognition to natural language processing. Such advancements are epitomized by iconic milestones, such as Google DeepMind's AlphaGo triumphing over human champions in the intricate game of Go[3], a feat that many speculated would take decades to achieve. Likewise, the advent of sophisticated language models based on the transformer architecture[4], like OpenAI's GPT-4[5] has redefined what machines can generate and comprehend, blurring the lines between human and machine-generated content.

TRANSFORMING PREDICTION POLICY

Within this dynamic landscape, AI's potential has extended far beyond mere technological marvels and has begun to address intricate societal challenges. Higher education, a cornerstone of societal development, has not remained untouched. Traditional models of admissions and educational prediction, once heavily reliant on linear metrics like Grade Point Average (GPA), are being rigorously challenged by AI's nuanced and holistic approach. As institutions strive for both excellence and equity, AI's promise lies in its ability to glean insights from vast, complex datasets, as has been demonstrated in different aspects of human lives, such as individuals' health[6], jobs and careers[7] and learning[8]. Our study[1] goes a step further and asks whether increased predictive accuracy translates into improved algorithmic policy making. Our study delves into the policy context of higher education admissions, seeking to harness the new transformer architecture in machine learning[4] to forge a path toward a more informed and fair admissions process.

RESEARCH DESIGN AND DATA

We tailor our research design to mimic a natural scenario for prediction policy; selecting which students to admit at schools and colleges. The goal for policymakers and the algorithms is to screen students and choose the ones that have higher likelihood of completing. Such a goal is natural as educational institutions receive funding and tuition for a combination of enrollment and study completion. Once models are estimated, their predicted likelihood of completion can serve as new, synthetic admission criteria.

To emulate a policy situation, we exclusively leverage pre-admission data and use the last observed year to measure the performance of our model. Our data comes from Statistics Denmark and our sample covers 444,884 students in Danish higher education who fulfill two criteria: 1) they applied to higher education using The Coordinated Enrolment System in the period 2006-2017, and 2) primary and secondary grades were available in the national registries. Our model covers all institutions offering higher education, is trained on the cohorts in the period 2006-2016 and tested on the 2017 cohort.

TEXT-LIKE REPRESENTATIONS OF RECURRING RECORDS

Our method leverages the power of the model architecture of the transformer approach[4] to efficiently represent and analyze complex data regarding grades and application patterns. Such data is very complex, and each user has a unique set of records consisting of course enrollment and grades they have obtained. Traditional data representations often struggle to capture intricate relationships and temporal patterns. We overcome this by converting the data into sequences of events. This event sequence feeds into the transformer model, sidestepping traditional aggregations like computing means (e.g., as in GPA) and incorporating the temporal dimension. The transformer model uses a concept known as a self-attention mechanism, which allows it to effectively capture both short-term and long-term dependencies in the data.

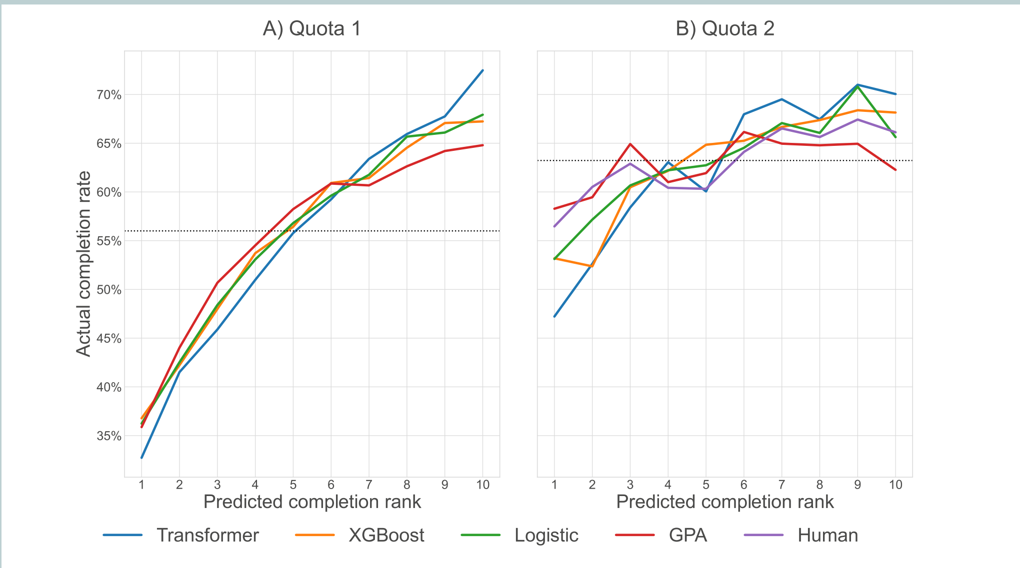
PERFORMANCE BOOST WITH OUR MODEL

Admission to higher education often uses student GPA to rank which students to prioritize. However, we find that this isn't the best way to predict future success in college, nor are models built on other linear metrics. Instead, our new model, which uses advanced computer techniques to make sense of the complex data, showed better results.

We find that our model improves prediction of students' degree completion. We measure this using a well-known metric to capture model accuracy, AUC-ROC[a]. Our new method achieved AUC scores of 67.7% when just using school data and 69.0% when using all available data. This is an improvement of 3.4% and 1.7%, respectively, compared to standard machine learning methods based on linear metrics of grades.

a This measure is short for Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC), where 100\% correspond to a perfect model and 50\% corresponds to random guessing. The ROC curve is a curve that measures the trade-off between two ways of failing to predict.

FIGURE 1 CONTRACTION CURVES FOR MODELS USING SCHOOL DATA ONLY



Note: Students' predicted likelihood of completion are binned in an ascending manner and mean completion rates within each bin is calculated. This is done for two subpopulations: Applicants for admission through Quota 1 in panel A and Quota 2 in panel B. Quota 1 students are accepted on the basis of GPA, whereas Quota 2 students are accepted on the basis of faculty assessment. Students are perfectly ranked if completion rates start at 0% and abruptly jumps to 100%. Horizontal dotted lines indicate mean completion rate.

HOW OUR MODEL COULD CHANGE ADMISSIONS

If colleges started using our model's rankings, they could change how they decide who gets in. For example, they might admit students our model predicts are more likely to graduate. We tested this by creating a hypothetical situation where colleges would reject more students than they do now. The results showed our model was the best at identifying those who might not graduate, especially for Quota 2 applicants who are ranked by faculty, with the performance seen in Figure 1. Here, students are binned into deciles, and we compute actual completion rates within each bin. Models perform better at identifying the most at-risk students if bin 1 has a lower actual completion rate.

FAIRNESS IN ADMISSIONS

A crucial aspect of prediction algorithms used for policies, but also policies in general, is that they treat those affected in a fair manner. We investigated whether our models favored or discriminated against certain groups of students in two ways.

The first approach was to measure whether our model was well-calibrated according to student groups, meaning that adding that information about group membership does not improve the predictions. Our model satisfies this fairness concept across a wide variety of sensitive attributes. The second approach is to measure whether those who completed or not differed systematically in their predicted likelihood of completing or not. Our models do not satisfy this criterion, which is known as separation. The fact that our model only is well-calibrated is not surprising as the two measures in general are mutually exclusive[9].

To compare the fairness of our models with the current admission criteria (GPA or faculty assessment), we measure predictive differences using the ABROCA measure[10 b]. We find that our models exhibit similar fairness properties compared to the current admission policies, but that GPA-based admission is the relatively fairest by a slight amount. We find that there are challenges in achieving fully fair outcomes for all students, especially for immigrants and their descendants, even with the current systems in place.

IMPLICATIONS

We find that machine learning predictions and methods based on the transformer architecture can be used for policy to improve efficiency: Our models outperform current admission policies and standard machine learning methods. While the integration of machine learning into decision-making processes has raised concerns about potential biases, our findings indicate that fairness is not necessarily compromised. However, it's crucial to note that the preservation of fairness is contingent upon the design and implementation of the admissions system. When correctly utilized, machine learning can offer insights without introducing additional biases.

DILEMMAS

The usage of machine learning models for algorithmic policy making in the context of higher education and more raises a number of dilemmas:

- Should more advanced and better performing models be used at the cost of transparency? (e.g., if adopting algorithmic admission criteria applicants will have a harder time figuring out where they can be admitted).
- To what extent should information that is not task-specific about individuals, such as gender and immigration status, be used? Using this information can paradoxically make models fairer if there was discrimination in the selected sample[11] (e.g., to predict the graduation of those admitted).
- Is an algorithmic fairness approach sufficient to quantify and evaluate the fairness of the proposed systems, or are more fundamental notions required?
- The adoption of algorithmic policy making is likely to have consequences for other actors (e.g., students rejected by new admission criteria will apply to other programs).

POLICY RECOMMENDATIONS

There are obvious potential efficiency gains to be had by using machine learning models, and based on this we have the following recommendations:

- *Deployment of prediction model in finding and offering increased guidance for at risk students.* Using the prediction model to identify at-risk students before admission allows guidance counselors to intervene proactively, enhancing student retention and success rates from early stages of enrollment.
- *Reexamine the merit of Quota 2, which underperforms in both prediction and fairness by our measures.* Our evaluation indicates that faculty assessment underperforms both in terms of ranking students and in fairness, necessitating a reexamination to assess whether it aids in creating a fair and effective admissions system or should be replaced.
- *Investigate scope for deployment for advanced machine learning systems beyond higher education.* Owing to the centralized data architecture in Denmark, the applicability of machine learning systems in various scenarios beyond higher education in Denmark merits further exploration.

b The Absolute Between-ROC Area (ABROCA) is a measure of how far apart the ROC curves for two different subpopulations are, defined by a binary sensitive attribute. As the ROC curve is a measure of performance at different thresholds, larger differences indicate a more unfair model.

REFERENCES

1 MAGNUS LINDGAARD NIELSEN, JONAS SKJOLD RAASCHOU-PEDERSEN, EMIL CHRISANDER, JULIEN GRENET, ANNA ROGERS, DAVID DREYER LASSEN, AND ANDREAS BJERRE-NIELSEN. “TRANSFORMING PREDICTION POLICY: HOW NOVEL MACHINE LEARNING METHODS CAN IMPROVE HIGHER EDUCATION ADMISSION”. IN: UNPUBLISHED WORKING PAPER. 2023.

2 ANDREAS BJERRE-NIELSEN AND EMIL CHRISANDER. VOLUNTARY INFORMATION DISCLOSURE IN CENTRALIZED MATCHING: EFFICIENCY GAINS AND STRATEGIC PROPERTIES. ARXIV:2206.15096 [ECON, Q-FIN]. JUNE 2022.

3 ELIZABETH GIBNEY. “WHAT GOOGLE’S WINNING GO ALGORITHM WILL DO NEXT”. EN. IN: NATURE 531.7594 (MAR. 2016). NUMBER: 7594 PUBLISHER: NATURE PUBLISHING GROUP, PP. 284–285.

4 ASHISH VASWANI, NOAM SHAZEER, NIKI PARMAR, JAKOB USZKOREIT, LLION JONES, AIDAN N. GOMEZ, LUKASZ KAISER, AND ILLIA POLOSUKHIN. “ATTENTION IS ALL YOU NEED”. IN: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 2017-DECEMBER (JUNE 2017). ARXIV: 1706.03762 PUBLISHER: NEURAL INFORMATION PROCESSING SYSTEMS FOUNDATION, PP. 5999–6009.

5 OPENAI. GPT-4 TECHNICAL REPORT. ARXIV:2303.08774 [CS]. MAR. 2023.

6 LAILA RASMY, YANG XIANG, ZIQIAN XIE, CUI TAO, AND DEGUI ZHI. “MED-BERT: PRETRAINED CONTEXTUALIZED EMBEDDINGS ON LARGE-SCALE STRUCTURED ELECTRONIC HEALTH RECORDS FOR DISEASE PREDICTION”. EN. IN: NPJ DIGITAL MEDICINE 4.1 (MAY 2021). NUMBER: 1 PUBLISHER: NATURE PUBLISHING GROUP, PP. 1–13.

7 KEYON VAFA, EMIL PALIKOT, TIANYU DU, AYUSH KANODIA, SUSAN ATHEY, AND DAVID BLEI. “CAREER: ECONOMIC PREDICTION OF LABOR SEQUENCE DATA UNDER DISTRIBUTION SHIFT”. EN. IN: NOV. 2022.

8 DONGMIN SHIN, YUGEUN SHIM, HANGYEOL YU, SEEWOO LEE, BYUNGSOO KIM, AND YOUNG-DUCK CHOI. “SAINT+: INTEGRATING TEMPORAL FEATURES FOR EDNET CORRECTNESS PREDICTION”. IN: LAK21: 11TH INTERNATIONAL LEARNING ANALYTICS AND KNOWLEDGE CONFERENCE. LAK21. NEW YORK, NY, USA: ASSOCIATION FOR COMPUTING MACHINERY, APR. 2021, PP. 490–496.

9 JON KLEINBERG, SENDHIL MULLAINATHAN, AND MANISH RAGHAVAN. “INHERENT TRADE-OFFS IN THE FAIR DETERMINATION OF RISK SCORES”. IN: 8TH INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE (ITCS 2017). ED. BY CHRISTOS H. PAPADIMITRIOU. VOL. 67. LEIBNIZ INTERNATIONAL PROCEEDINGS IN INFORMATICS (LIPICS). ISSN: 1868-8969. DAGSTUHL, GERMANY: SCHLOSS DAGSTUHL–LEIBNIZ-ZENTRUM FUER INFORMATIK, 2017, 43:1–43:23.

10 JOSH GARDNER, CHRISTOPHER BROOKS, AND RYAN BAKER. “EVALUATING THE FAIRNESS OF PREDICTIVE STUDENT MODELS THROUGH SLICING ANALYSIS”. EN. IN: PROCEEDINGS OF THE 9TH INTERNATIONAL CONFERENCE ON LEARNING ANALYTICS & KNOWLEDGE. TEMPE AZ USA: ACM, MAR. 2019, PP. 225–234.

11 ASHESH RAMBACHAN AND JONATHAN ROTH. “BIAS IN, BIAS OUT? EVALUATING THE FOLKWISDOM”. IN: 1ST SYMPOSIUM ON FOUNDATIONS OF RESPONSIBLE COMPUTING (FORC 2020). ED. BY AARON ROTH. VOL. 156. LEIBNIZ INTERNATIONAL PROCEEDINGS IN INFORMATICS (LIPICS). ISSN: 1868-8969. DAGSTUHL, GERMANY: SCHLOSS DAGSTUHL–LEIBNIZ-ZENTRUM FÜR INFORMATIK, 2020, 6:1–6:15.