EVALUATING A LEARNED ADMISSION-PREDICTION MODEL AS A REPLACEMENT FOR STANDARDIZED TESTS IN COLLEGE ADMISSIONS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Hansol Lee

December 2022

ABSTRACT

A growing number of college applications has presented an annual challenge for college admissions in the US. In response to this challenge, admission offices have often relied on standardized test scores to parse their large applicant pools into viable subsets. However, this approach may be subject to bias in test scores and fails to work in test-optional admissions. In this work, we explore a machine learning-based approach to replace the role of standardized tests in subset generation while taking into account a wide range of factors extracted from student applications to support a more holistic review. We evaluate the approach on data from an undergraduate admissions office at a selective US institution and discuss how machine learning can be leveraged to support human decision-making in college admissions.

BIOGRAPHICAL SKETCH

Hansol completed her undergraduate studies in Computer Science with a minor in Psychology at Cornell University in 2019. She continued her studies in Computer Science by enrolling in the Master of Science program at Cornell University in the same year, advised by professors René Kizilcec and Thorsten Joachims. Hansol joined the Ph.D. program in Education Data Science at the Stanford Graduate School of Education in Fall 2021.

Dedicated to my parents.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisors Thorsten Joachims and René Kizilcec for their constant support and helpful guidance. Their mentorship has been invaluable both in completing my Master's research as well as my growth as a scholar. I am grateful to Wendy Ju, Andrea Cuadra, and Alexander Ruch, for showing me how rewarding and exciting research could be, and to Lillian Lee and Kilian Weinberger for being incredible instructors as well as mentors. I would also like to thank Ashudeep Singh for being an amazing Ph.D. mentor and Scott Campbell for generously sharing his expertise and resources for the admission project.

I am also grateful to the wonderful friends and family who made my time at Cornell so special and memorable. Thanks to Ejin, Juyoung, So Youn, Soyoung, and Seyun for all of the laughs, all-nighters, and sweet memories. My special thanks to Jason for being my biggest supporter in everything that I do. Most importantly, I would like to thank my parents for providing me with unconditional love and support throughout my years of study.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Colleges and universities across the United States receive an increasing number of student applications for admission to their incoming class each year [24]. The Common Application[1], the primary tool used by students to apply to colleges in the United States, has received over 6.6 million first-year applications during the 2021-22 admission cycle. This is a 9.1% increase in applications from the previous year, and a 21.3% increase from the 2019-20 cycle [13]. The growing number of applications received by selective colleges and universities has presented an annual challenge for college admissions that adhere to a holistic review process, which aims to assess each applicant as a whole by considering a wide range of factors presented in a student's application [11, 4, 30].

The sheer volume of applications received by colleges makes it challenging for human reviewers to perform a thoughtful and equitable review of individual applications given the limited admission timeline, which typically spans only a few months. Hiring more human reviewers is a potential solution to this problem, with the advantage of spreading out the reviewing load by assigning fewer applications to each human reviewer. However, this approach may not be sustainable as hiring more admission professionals adds organizational complexity that makes it hard to scale with the growing number of applications. More importantly, it raises the problem of maintaining consistency among multiple reviewers in their admission decisions. A lack of consistency among reviewers could further complicate the task of selecting a coherent class of first-year students [19, 22].

[1] https://www.commonapp.org/

Absent alternatives to reduce the reviewing load, admission officers may find it necessary to lean on quantitative measures such as standardized test scores to complete admission decisions in time. Many selective colleges in the United States have used heuristics that are based on standardized test scores such as the ACT or SAT to "triage" their large applicant pool—that is, to organize their large applicant pool to better allocate the limited resources available for application review [25].

However, there are several limitations to this standardized testing-based approach (which we will refer to as an "SAT-based" approach in this thesis). First, there are many unresolved concerns about gender, racial and socioeconomic biases in standardized test scores which undermine the fairness of the SAT-based approach for organizing the applicant pool [12, 33, 10, 28, 29, 26]. Second, this SAT-based approach is dependent on requiring all applicants to submit their test scores, which may impose a significant financial burden on many applicants. Moreover, many institutions began test-optional admission in response to testing site closures during the coronavirus disease (COVID-19) pandemic, which makes the traditional SAT-based heuristics incomplete, impractical, and potentially biased. This change in the admission policy requires an alternative method for triaging large applicant pools in the absence of standardized test scores.

To overcome the issue of bias and costliness of standardized tests, their increasing unavailability in test-optional admission, and their non-holistic nature as a basis for organizing the admission process, we explore a machine learning-based approach that takes into account a wide range of factors extracted from student applications to enable a more holistic review in the absence of standard-

ized test scores. In particular, we focus on answering the following research question: How well can an admission prediction model trained on past admission data replace and improve on the traditional SAT-based heuristic to organize the applicant pool for review? We examine this question in the context of first-year undergraduate admissions at a selective US institution and find that the prediction model is better aligned with existing admission practices at the case institution compared with an SAT-based heuristic.

CHAPTER 2

RELATED WORK

Our work builds on the idea of human-machine collaboration, which advocates for the design and use of machine learning systems with the intention of augmenting, not replacing, human contributions [20, 32]. Autor et al. argued that machines may replace humans in performing routine tasks while complementing humans in performing nonroutine cognitive tasks [3]. Jarrahi suggested that machines may extend humans' cognition by equipping human decision-makers with comprehensive data analytics, whereas humans may offer a more holistic and intuitive approach to decision-making [16]. In our work, we aim to leverage the complementary strengths of machines and humans in the admission process; a machine-learned admission prediction model can be used to organize a large applicant pool in a way consistent with past institutional decision-making, allowing admission officers to use the freed-up resources to engage in the process of holistic review in a more meaningful way.

Approaches based on machine learning are increasingly studied to support various aspects of college admissions. For example, Basu et al. used machine learning algorithms to predict whether a student who is offered admission would accept the offer, helping institutions to better estimate the sizes of their entering class [5]. From the applicants' perspective, Gupta et al. and Kiaghadi et al. used machine learning techniques to evaluate applicants' chances of admission to help students make informed decisions about where to apply to college [14, 17]. In the context of college essays, Alvero et al. explored the use of computational text analysis to assist human readers in their evaluation of college application essays [1, 2].

In particular, our work focuses on the use of machine learning to predict admission to support the holistic review process of admission officers. As early as the 1990s, Bruggink et al. and Moore et al. utilized domain knowledge to build statistical models to predict undergraduate admissions [7, 21]. More recently, Lux et al. used multi-layer perceptron and support vector machine algorithms to predict admission decisions at a small private liberal arts college [18]. Rees and Ryder evaluated the usefulness of the random forest algorithm in assisting in the process of an internal medicine residency program in northern New England [27]. Neda and Gago-Masague [23] compared classification performances of various machine learning algorithms trained on applications submitted to the Computer Science department at the University of California, Irvine.

Most of the prior work only primarily evaluated the admission prediction model using metrics such as overall accuracy and Area Under the Receiver Operating Characteristic Curve (AUROC) that do not adequately provide an understanding of the model behavior trained on imbalanced datasets such as admission datasets. Our work utilizes a similar modeling approach to estimate admission probability for each applicant but provides additional analyses of the uncertainty of model predictions as well as specific recommendations for using the predicted admission probabilities in admission practice in order to to avoid potential misuse of the prediction model that is superficial in nature.

In addition to evaluating the overall predictive accuracy, Waters and Miikkulainen examined a contentious use of an admission prediction model: to save time in the admission process [31, 8]. In their 2014 study, Waters and Miikkulainen used logistic regression to predict graduate admissions in the computer science department at the University of Texas at Austin. Their work focused

on improving the efficiency of review by cutting the time spent on application reading and the number of applications to review. We emphasize that our work focuses on improving the process of holistic review by providing a tool that better supports the organization of the applicant pool than the SAT-based method previously used in admission, not by providing a shortlisting tool to make the admission process cost-effective.

CHAPTER 3

METHODS

In this section, we describe the first-year undergraduate admissions process at the case university and explain how we develop and evaluate the admissions prediction model using past admissions data at the case university. We then explore how to construct new applicant pools based on the admission prediction model for admissions officers to use in the admissions process.

3.1 Context

The case university is one of many selective institutions in the United States where standardized test scores such as the SAT have been used in the admissions process to prioritize the review order of applications. Like many other colleges, the case institution began test-optional undergraduate admissions during the COVID-19 pandemic which provided an emergent need for an alternative to the SAT-based heuristic.

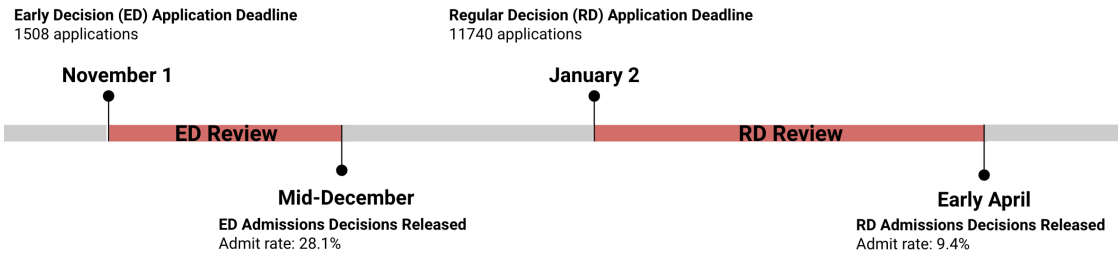

Figure 3.1: The 2019-2020 admissions timeline for early and regular decisions at the case institution.

Admissions at the case institution begin with students submitting their applications for either the Early Decision (ED) or the Regular Decision (RD) admissions cycle (see Fig 3.1). In the 2019-20 cycle, 1,508 ED applications were received by November 1, and admissions decisions were released in mid-December. During the RD admissions cycle, 11,740 applications were submitted by January 2, and decisions were sent out by early April. The admit rates for ED and RD were 28.1% and 9.4%, respectively.[1] Application review was performed by professional admissions staff in the case institution's undergraduate admissions office. Each admissions officer was responsible for reviewing the applications from certain geographical regions, and a larger team of external reviewers supported their initial review of the applications. The review process spanned about one month for ED and about three months for RD.

The institution requires all first-year applicants to apply through the Common Application portal. Each completed application via Common Application includes a variety of information such as a personal essay, descriptions of extracurricular activities, honors and awards, AP and/or IB scores, SAT or ACT scores, SAT subject scores, TOEFL/IELTS scores, and a college-specific essay. In addition, recommendation letters from two teachers, secondary school reports that include a recommendation letter from the guidance counselor, school profile and the official transcript of the applicant, and the current mid-year grade report are submitted alongside the Common Application.

Like many other selective institutions in the US, the case institution has relied on standardized test scores such as the SAT and the ACT in order to organize the applicant pool into three subsets, the "Top", "Middle", and "Bottom"

[1] As an ED applicant, students apply to only one institution and must enroll at that school if admitted. Because of this binding nature, ED applicant pools usually have a higher rate of admission relative to RD applicant pools.

pool. Once these pools are defined using the SAT-based model, all student applications were sent to admissions staff for review. Every application was reviewed by human reviewers regardless of their initial pool assignment and admissions staff reviewed applications in an iterative process to fill the annual admit target (i.e., the number of spaces available for matriculation). Grouping similar applicants together is necessary for equitable review, since many qualified applicants apply to selective colleges like the case institution, but the human resources and the size of the first-year class are limited.

Indeed, this SAT-based segmentation into pools shows a correlation with the admission outcome: 83.1% of the final admitted class for the 2019-2020 admissions cycle was identified from the Top pool alone, which only consisted of 56.6% of the entire applicant pool. However, it is clear that the SAT-based model did not replace holistic human review: only 28.5% and 8.8% of applicants in the Top pool were identified as female and underrepresented minority (URM) students, respectively, whereas 51.4% were female and 30.0% were URM students in the final admitted class. In this work, we explore an admission prediction model that could replace and improve on the SAT-based heuristic that had been used at the case institution for triaging the applicant pool, aligning the formation of applicant pools with the holistic human review that follows.

3.2 Dataset description

The dataset we used to build the prediction model comprises the student application data submitted to the case institution during the 2019-20 admission cycle as well as their final admission outcomes. First-year applicants apply through

the Common Application which contains a fixed set of data fields including SAT and ACT scores, SAT subjects, Advanced Placement (AP) International Baccalaureate (IB), English proficiency test scores (TOEFL/IELTS), high school GPA, class rank, high school information, intended major, legacy status, career interests, languages spoken, personal essays, application information (early or regular, application fee waivers, etc.), extracurricular activities and time commitment, courses taken in the current year, high school disciplinary records, honors and awards, and several demographic information such as gender, ethnic background, citizenship, age, first-generation status, and parental levels of education. We consider all information presented in the Common Application except for personally identifiable information such as names, addresses, contact information, and names of high schools.

In the 2019-20 admission cycle, all applicants were still required to report their SAT or ACT scores, and international students from non-English speaking countries were also required to submit their TOEFL or IELTS scores. In building the admission prediction model, however, we choose to remove SAT/ACT and TOEFL/IELTS scores from our feature set in order to simulate the test-optional admission policy. Importantly, we note that we did not have access to other application materials apart from what is presented in the Common Application; as a result, several crucial pieces to application review such as college-specific essays, teacher and guidance counselor recommendation letters, high school reports, and transcripts are omitted in the feature set.

We filter out a small number of duplicate applications. We also remove student-athletes and Reserve Officers' Training Corps (ROTC) applications who were recruited to the university prior to application submission. We impute

missing values with a unique placeholder value and add an indicator variable for missingness. Categorical features are one-hot encoded, while some categorical values with fewer than 1% frequency are merged together as "RARE". The final processed data has 13,248 rows and 1,435 columns where each row corresponds to a single completed application submitted to the case institution for either the ED or RD admission cycle.

3.3 Modeling from past admissions

We frame college admissions prediction as a probabilistic binary classification problem and focus on the following two admissions outcomes: admitted (including admitted and conditionally admitted applicants) and denied (including denied, wait-listed, and withdrawn applicants). We leave out a randomly sampled 20% of the dataset for testing (n = 2650), and train the model using the remaining 80% of the dataset. We note that 11.5% of applicants in the training data and 11.7% in the testing data were granted final admission from the case institution in the 2019-20 admission cycle. We fit a Gradient Boosting Decision Trees model using sklearn's XGBoost library using its default parameter settings [9].

We assess the potential of the prediction model to replace and improve the institution's SAT-based triaging process by using the following evaluation strategy. In the testing data, 1,509 applicants (57.0% of the testing data) were placed in the Top pool by the SAT-based model, and 82.5% of the final admitted class was identified from the Top pool. Similarly, we identify the top 1,509 applications based on the predicted probability of admission from the admission pre-

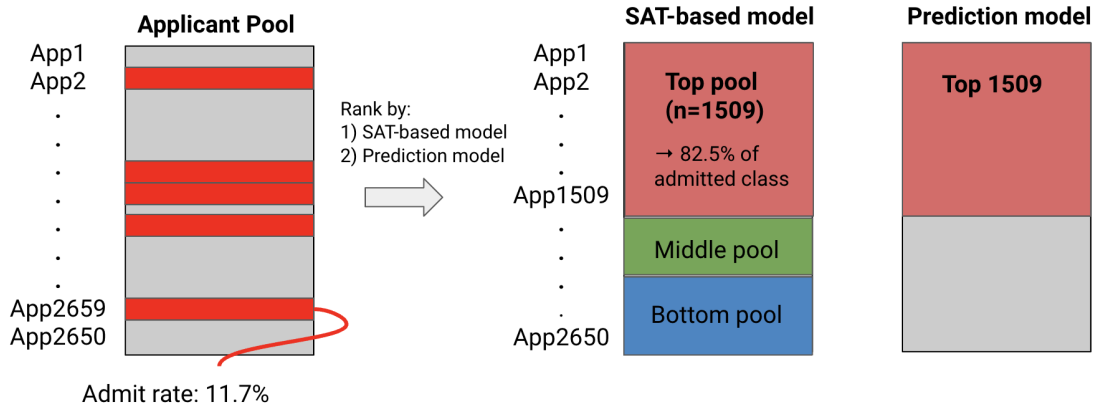

Figure 3.2: Illustration of the SAT-based and the prediction-based "Top pool" in the testing data (n = 2650). The initial applicant pool is shown on the left; the red rows represent the admitted applicants. The two figures on the right represent the applicant pools organized by the SAT-based model (middle) and the prediction model (right). In the SAT-based model, 1,509 applicants are grouped into the Top pool, which identifies 82.5% of the admitted class. In the prediction model, the top 1,509 applicants as sorted by their predicted probability of admission are grouped into the prediction model's notion of the "Top pool".

diction model in order to simulate the model's version of the Top pool. We then compare the proportion of the eventually admitted class identified in the Top pools between the SAT-based model and the prediction model. We also compare the proportion of URM applicants and of female applicants represented in the SAT-based and prediction-based Top pools to the proportion in the final admitted class (see Fig 3.2 for a visual illustration).

3.4 Construction of new applicant pools

While the SAT-based model used by the case institution defined only three pools (Top, Middle, and Bottom), the prediction model offers a straightforward way

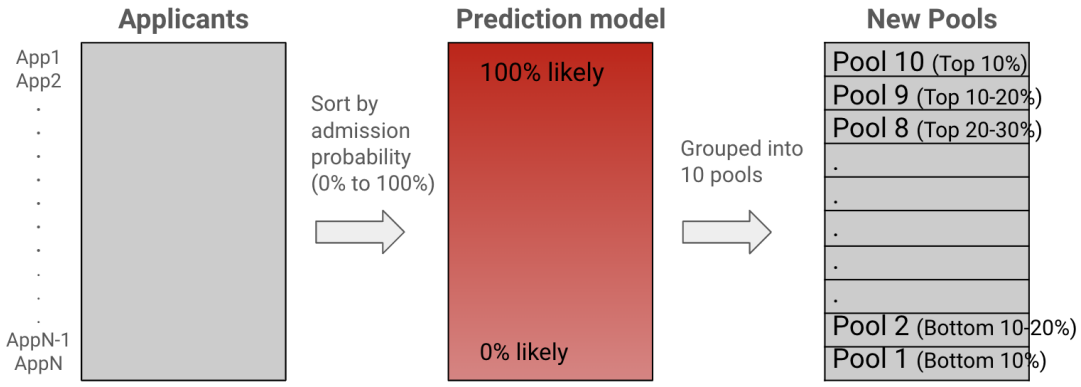

Figure 3.3: Construction of the new pools based on the admission prediction model. All applicants (left) are sorted by the predicted probability of admission (middle). The applicants with the top 10% admission probability are placed in Pool 10, the next 10% in Pool 9, and so on (right), to produce 10 different applicant pools with varying predicted acceptance rates.

to create more fine-grained pools and convey their probabilistic semantics. One option is to sort applicants by their predicted probabilities of admission and aggregate them into 10 different pools as follows: applicants with the top 10% admission probability are placed in Pool 10, the next top 10% is placed in Pool 9, and so on (see Fig 3.3 for a visual illustration). Pool 10 is then predicted to have the highest number of admitted applicants, Pool 9 is predicted to have the second-highest number of admitted applicants, and so on.

We assess the calibration of the new pools by checking whether or not the predicted admission rate matches the actual admission rate in each pool. In addition to analyzing the prediction uncertainty in the aggregate pools of students, we further explore the uncertainty of the prediction model by examining the distribution of the individual predicted probabilities for denied and admitted applicants.

CHAPTER 4

RESULTS

We evaluate the potential of the admission prediction model as a replacement for and an improvement over the SAT-based triaging process previously used at the case institution. We find that the prediction model outperforms the SAT-based model by placing more admits in the Top pool and fewer admits in the Bottom pool. The Top pool of the prediction model also more closely matches the final admitted class in terms of the female and URM composition than the SAT-based Top pool. We analyze both the aggregate and individual predicted probabilities of the admission prediction model and find that the new applicant pools constructed from the prediction model are well-calibrated by comparing the predicted and the actual admit rates in each pool.

4.1 Comparison to the SAT-based model

As shown in Fig 4.1, we find that the prediction model outperforms the SAT-based model in having more admits in the Top pool and fewer admits in the Bottom pool. Specifically, the Top pool of the prediction model identifies 91.9% of the final admitted class, compared to the 82.5% in the SAT-based Top pool ($N = 309$, $\chi^2(1) = 12.206$, $p < 0.001$). Conversely, the Bottom pool of the prediction model consists of 4.5% of the admitted class, while the SAT-based Bottom pool consists of 9.7% of the admitted class ($N = 309$, $\chi^2(1) = 6.2642$, $p < 0.05$). These results show that the prediction model is significantly better aligned with the admission criteria of the case institution than the SAT-based model.

In addition, Fig 4.2 shows that the Top pool of the prediction model more

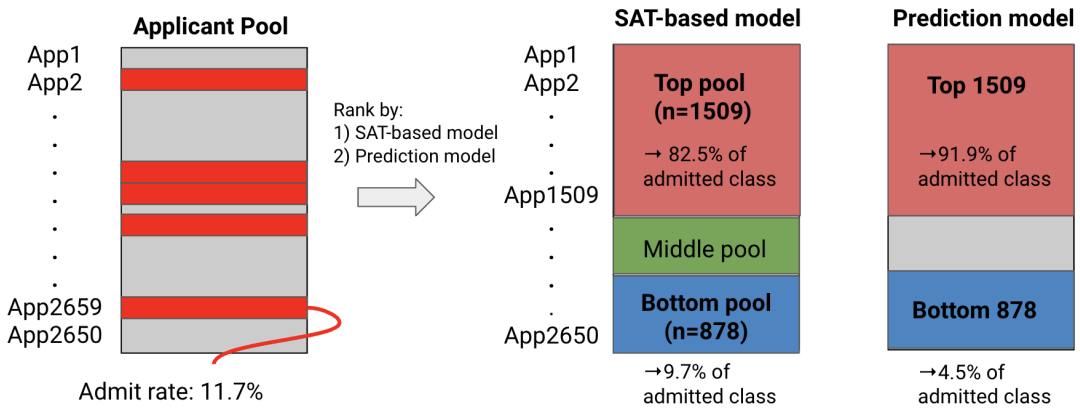

Figure 4.1: Illustration of the proportion of admitted class identified by the SAT-based model (middle) and by the prediction model (right).

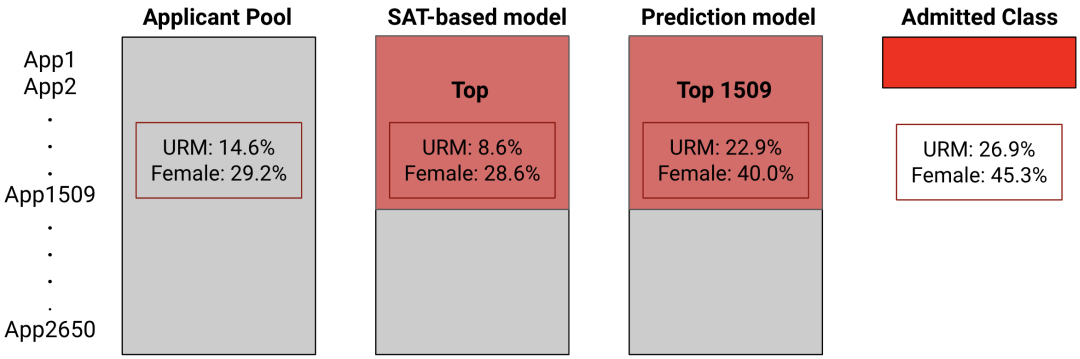

Figure 4.2: Proportion of URM and female students in each of the following pools (from left to right): 1) the entire applicant pool (n = 2650), 2) the Top pool of the SAT-based model (n = 1509), 3) the Top pool of the prediction model (n = 1509), and 4) the final admitted class (n = 309).

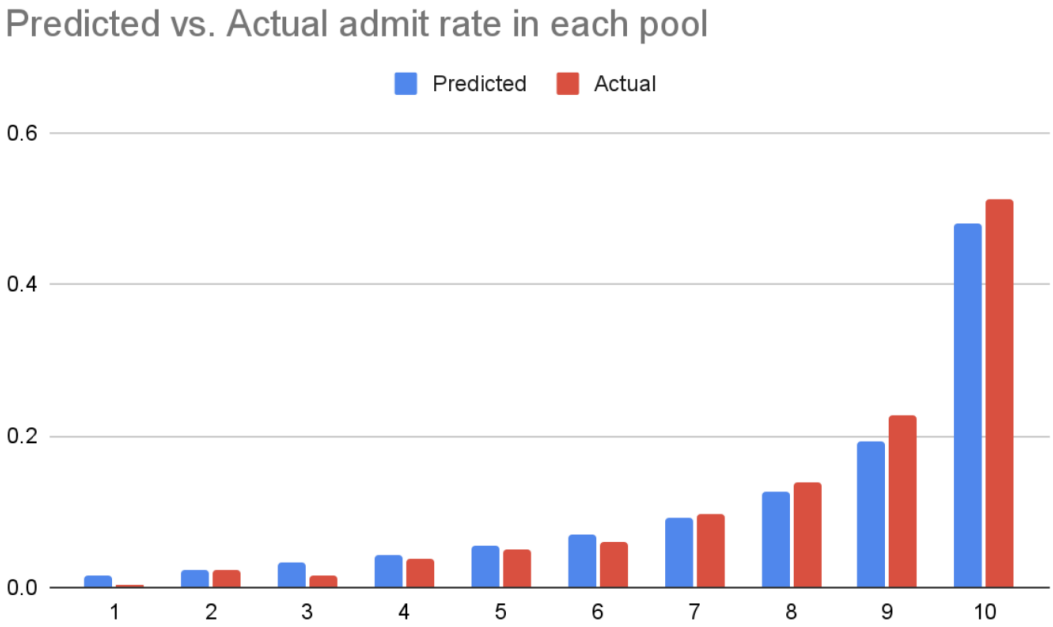

Figure 4.3: Predicted admit rate (blue) and actual admit rate (red) for each of the 10 pools constructed from the prediction model.

closely matches the final admitted class in terms of the female and URM distributions than the SAT-based Top pool. The URM and female student proportion in the overall applicant pool is 14.6% and 29.2%, respectively. We see that 8.6% and 28.6% of the SAT-based Top pool are URM and female students, while 22.9% and 40.0% of the prediction-based Top pool are URM and female students. The prediction model produces a Top pool that more closely matches the demographic makeup of the final admitted pool, where 26.9% of admits are URM students and 45.3% are female students.

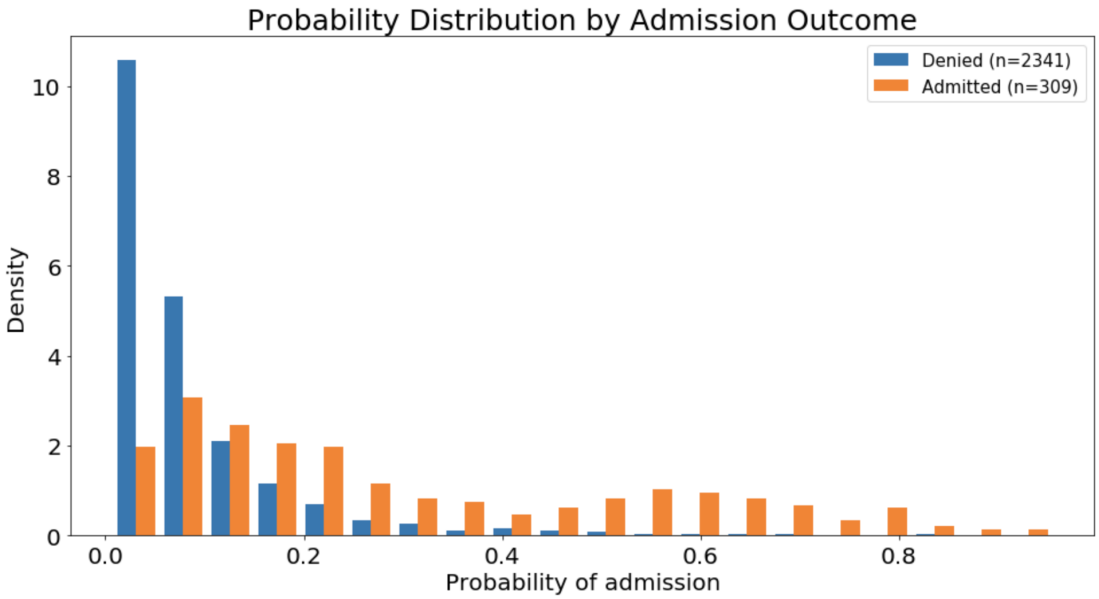

Figure 4.4: Density plot of estimated admission probability scores from 0.0 to 1.0 for denied (blue) and admitted (orange) applicants in the testing set.

4.2 Evaluating uncertainty of the prediction model

We find that new applicant pools constructed with the predicted probabilities are reasonably well-calibrated at the level of the defined pools. We compute the predicted admit rate by taking the average of the predicted probabilities in each pool, and the actual admit rate by taking the proportion of actual admits in each pool. As presented in Fig 4.3, we see that the predicted admission rate closely follows the actual admission rate in each pool ($r(10) = 0.998$, $p < 0.001$).

While the model is well-calibrated across the ten pools, an applicant's placement in a given pool does not indicate whether they should be admitted or not. For example, a student in Pool 10 is not necessarily a stronger applicant than a student in Pool 9. We highlight this issue by examining the distribution of

predicted probabilities for individual students. Consider the following distribution of predicted probabilities for denied and admitted students shown in Fig 4.4. We see that most of the predicted probabilities are concentrated on the lower end for the denied applicants, suggesting that the prediction model is able to accurately discern most of the denied applicants. On the other hand, the predicted probabilities for the admitted applicants are more widely spread across the entire range of predicted probabilities, implying that the prediction model has limited knowledge about who gets admitted. This suggests that the final admission decisions may depend largely on human evaluation of other parts of the applications such as teacher recommendations, personal essays, and transcripts that are missing from the prediction model; in other words, the prediction model exhibits significant epistemic uncertainty [15]. The aggregate pools are meant to provide a coarse yet meaningful organizational structure of the application pool for effective reading for admission officers, and as a safety net to minimize the chance that any qualified applicant gets overlooked.

CHAPTER 5

DISCUSSION

In this work, we set out to answer the following question: How well can a prediction model trained on past admission data replace and improve on the traditional SAT-based heuristic to organize the applicant pool for review? We found three potential ways in which the admission prediction model can replace and improve the SAT-based triaging heuristic previously used at the case institution. First, we found that the prediction model outperforms the SAT-based model by placing more admits in the Top pool and fewer admits in the Bottom pool. Second, the Top pool of the prediction model also more closely matches the final admitted class in terms of the female and URM composition than the SAT-based Top pool. Finally, we showed that the new applicant pools constructed from the prediction model are well-calibrated by comparing the predicted and the actual acceptance rates in each pool, allowing for direct interpretation of the prediction model's applicant pools in admission.

5.1 Implications

As described earlier, many institutions began test-optional admission in response to testing site closures during the COVID-19 pandemic, which made the traditional SAT-based heuristic impractical. In our development and evaluation of the admission prediction model, we found that the prediction model may serve as a practical alternative to the SAT-based model for organizing the applicant pool; it can be trained using other already available student information in the Common Application excluding the standardized test scores (SAT

and ACT) and English proficiency scores (TOEFL and IELTS). Because the prediction model represents a larger set of information provided in the student application, it can also be seen as more holistic compared to the SAT-based model which is only based on standardized test scores and a few demographic markers.

Comparative analysis of the prediction-based and the SAT-based Top pools suggests that the prediction model may not only replace but also improve the triaging process in important ways. We found that in the past application cycle, the prediction model would have identified 9.4% more admitted students in the Top pool and 5.2% fewer admitted students in the Bottom pool compared to the SAT-based model. In other words, the prediction model outperforms the SAT-based model in terms of both the true positive rate (i.e. more admits in the Top pool) and the false positive rate (i.e. fewer admits in the Bottom pool). The prediction model further improves on the traditional SAT-based model as the prediction-based Top pool contains a pool of students with a similar female and URM composition to the actual admitted class compared to the SAT-based Top pool. These results suggest that the prediction model is able to organize the applicant pool in a way that better reflects the institution's admission goals, which is in line with the fact the prediction model is indeed trained on the institution's past admission decisions.

The upside of using the prediction model instead of the SAT-based model is that the admission practices are no longer tied to any biases in the standardized test scores that do not align with the institution's values. However, because the admission prediction model is directly informed by past admission decisions, it is important to ensure that the past admission data is appropriate for develop-

ing and using the resulting prediction model. So, any use of the model should be accompanied by a governance process that ensures that the dynamics of the admission process continue to reflect the potentially changing values of the institution. A prediction model is a tool, where thoughtful use could untether admission practices from biases in external scores to iteratively improve admission practices, but thoughtless use can also cement existing inequities.

One new affordance of the prediction model is that the new applicant pools are well-calibrated in that their predicted acceptance rates seem to match the actual acceptance rates. In contrast, the SAT-based model did not yield pools that are calibrated to actual acceptance rates. This suggests that the prediction model may further improve the admission process by enabling a direct interpretation of the applicant pools in terms of their predicted acceptance rates. For example, the institution may set itself the goal to double the acceptance rates in the lower pools, encouraging admission staff to find creative new ways of identifying qualified students that do not fit the typical profile. In this way, the prediction model may serve as a way to target a subset of applicants to examine more closely and highlight new applicants who might have been overlooked.

We note that such use of machine-predicted probabilities to guide decision-making in admission is not making judgments about individuals; rather, it is creating pools of students with different admission rates. In particular, we conjecture that much of the uncertainty captured by the prediction model is epistemic in nature, not aleatoric [15]. This means that the uncertainty comes from the model's limitation to truly understand an applicant's qualification, not from some external randomness. In the extreme, for a pool with a 90% acceptance rate, 9 out of 10 students in that pool may individually have a 100% probability

of getting admitted, while one student has a 0% probability. The probability therefore only makes sense if we think about pools of students instead of individual students. We thus stress that it is important to present results to human admission staff in terms of pools, not in terms of predicted probabilities of admission for individuals, to avoid any misconception of what the model is capable of.

5.2 Limitations and Future Work

There are several limitations to building the admission prediction model. First, admission decisions are not independent as they are about creating a class, but they were modeled as if they are independent when building the prediction model. Next, the application data used to train the admission prediction model did not include the full information that the human reviewers use to make their decisions. This work only included the data available in the Common Application, but there are many other aspects of student applications that heavily influence admission decisions (e.g. student essays, and letters of recommendation).

Another potential direction for improving the prediction model is to account for covariate shifts. For example, the admission prediction model was developed on admission data where SAT/ACT scores were mandatory, which may have a different distribution from future years where tests are optional or no longer factor in the decision at all. Even though the SAT and ACT scores are excluded entirely from the feature set for building the admission prediction model, there may potentially be a large covariate shift between the year in which

the model is trained and tested and the year in which the trained model is to be deployed. The model evaluation results presented in this work are based on the testing set from the same year that the model is trained on, and this may not be an accurate estimation of the model performance in another year because of the potential shift in data distributions. Future work may attempt to account for the potential covariate shift across datasets from different years by training a classifier for identifying whether a given data point belongs to the training or the testing distribution, and using it to assign more weights to training instances that are closer to a testing distribution [6].

Given the interaction of the prediction model with the human decision-making process, the full impact of the prediction model on the admission process and final decisions requires evaluation in a controlled randomized trial. In addition to assessing the impact of including the prediction model in the admission practice, future work could explore ways to provide explanations to admission staff as to how the pools are computed and why a given applicant is placed in a particular pool. Comparing the admission decisions between the prediction model and admission staff may help human decision-makers reflect on their decisions and see where they may have blind spots. It may also help improve the prediction model by identifying where the model may have blind spots; admission staff may be able to offer insight into where additional data collection may be helpful to make the predictions better.

BIBLIOGRAPHY

[1] AJ Alvero, Noah Arthurs, Anthony Lising Antonio, Benjamin W Domingue, Ben Gebre-Medhin, Sonia Giebel, and Mitchell L Stevens. Ai and holistic review: informing human reading in college admissions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 200–206, 2020.

[2] AJ Alvero, Jasmine Pal, and Katelyn M Moussavian. Linguistic, cultural, and narrative capital: computational and human readings of transfer admissions essays. *J Comput Soc Sci*, 5(2):1709–1734, September 2022.

[3] David H. Autor, Frank Levy, and Richard J. Murnane. The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333, 2003.

[4] Michael Bastedo, Nicholas Bowman, Kristen Glasener, and Jandi Kelly. What are we talking about when we talk about holistic review? selective college admissions and its effects on low-ses students. *The Journal of Higher Education*, 89:1–24, 04 2018.

[5] Kanadpriya Basu, Treena Basu, Ron Buckmire, and Nishu Lal. Predictive models of student college commitment decisions using machine learning. *Data*, 4(2), 2019.

[6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *J. Mach. Learn. Res.*, 10:2137–2155, December 2009.

[7] Thomas H. Bruggink and Vivek Gambhir. Statistical models for college admission and enrollment: A case study for a selective liberal arts college. *Research in Higher Education*, 37(2):221–240, 1996.

[8] Lilah Burke. The death and life of an admissions algorithm. *Inside Higher Ed*, December 2020.

[9] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Ezekiel J. Dixon-RomÁN, Howard T. Everson, and John J. Mcardle. Race, poverty and sat scores: Modeling the influences of family income on black

and white high school students' sat performance. *Teachers College Record*, 115(4):1–33, 2013.

[11] Lorelle L. Espinosa, G. Orfield, and M. Gaertner. Race, class, and college access: Achieving diversity in a shifting legal landscape. *UCLA: The Civil Rights Project / Proyecto Derechos Civiles*, 2015.

[12] Roy O. Freedle. Correcting the sat's ethnic and social-class bias: A method for reestimating sat scores. *Harvard Educational Review*, 73(1):1–43, 2003.

[13] Mark Freeman, Brian Heseung Kim, Preston Magouirk, and Trent Kajikawa. Deadline update: first-year application trends through march 15. *Common App*, March 2022.

[14] Narender Gupta, Aman Sawhney, and Dan Roth. Will i get in? modeling the graduate admission process for american universities. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 631–638, 2016.

[15] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021.

[16] Mohammad Hossein Jarrahi. Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586, 2018.

[17] M. Kiaghadi and P. Hoseinpour. University admission process: a prescriptive analytics approach. *Artif Intell Rev*, 2022.

[18] Thomas Lux, Randall Pittman, Maya Shende, and Anil Shende. Applications of supervised learning techniques on undergraduate admissions data. In *Proceedings of the ACM International Conference on Computing Frontiers*, CF '16, page 412–417, New York, NY, USA, 2016. Association for Computing Machinery.

[19] Patrícia Martinková, Dan Goldhaber, and Elena Erosheva. Disparities in ratings of internal and external applicants: A case for model-based inter-rater reliability. *PLOS ONE*, 13(10):1–17, 10 2018.

[20] Steven M. Miller. Ai: Augmentation, more so than automation. *Asian Management Insights.*, 5(1):1–20, 2018.

[21] James S Moore. An expert system approach to graduate school admission decisions and academic performance prediction. *Omega*, 26(5):659–670, 1998.

[22] Rüdiger Mutz, Lutz Bornmann, and Hans-Dieter Daniel. Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLOS ONE*, 7(10):1–10, 10 2012.

[23] Barbara Martinez Neda and Sergio Gago-Masague. Feasibility of machine learning support for holistic review of undergraduate applications. In *2022 International Conference on Applied Artificial Intelligence (ICAPAI)*, pages 1–6, 2022.

[24] Michael Nietzel. "new data: Applications surge at larger, selective colleges". *Forbes*, Mar 2021.

[25] United States Department of Education. Title vi: Admissions: Princeton university (02086002). *United States Department of Education Office for Civil Rights, Region II*, September 2015.

[26] Sean F. Reardon, Demetra Kalogrides, Erin M. Fahle, Anne Podolsky, and Rosalía C. Zárate. The relationship between test item format and gender achievement gaps on math and ela tests in fourth and eighth grades. *Educational Researcher*, 47(5):284–294, 2018.

[27] Christiaan A. Rees and Hilary F. Ryder. Machine learning for the prediction of ranked applicants and matriculants to an internal medicine residency program. *Teaching and Learning in Medicine*, 0(0):1–10, 2022. PMID: 35591808.

[28] Phyllis Rosser. *The SAT Gender Gap: Identifying the Causes.* Washington, DC: Center for Women Policy Studies., 1989.

[29] Art Sawyer. How the new sat has disadvantaged female testers. *Compass Education Group*, October 2017.

[30] Mitchell L. Stevens. *Creating a Class*. Harvard University Press, 2020.

[31] Austin Waters and Risto Miikkulainen. Grade: Machine learning support for graduate admissions. In *Proceedings of the 25th Conference on Innovative Applications of Artificial Intelligence*, 2013.

[32] H James Wilson and Paul R Daugherty. Collaborative intelligence: Humans and ai are joining forces. *Harvard Business Review*, 96(4):114–123, 2018.

[33] Rebecca Zwick. Is the sat a 'wealth test'? *Phi Delta Kappan*, 84(4):307–311, 2002.