ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-699

Human versus Machine: Do college advisors outperform a machine-learning algorithm in predicting student enrollment?

Suchitra Akmanchi
University of Virginia

Kelli A. Bird
University of Virginia

Benjamin L. Castleman
University of Virginia

Prediction algorithms are used across public policy domains to aid in the identification of at-risk individuals and guide service provision or resource allocation. While growing research has investigated concerns of algorithmic bias, much less research has compared algorithmically-driven targeting to the counterfactual: human prediction. We compare algorithmic and human predictions in the context of a national college advising program, focusing in particular on predicting high-achieving, lower-income students' college enrollment quality. College advisors slightly outperform a prediction algorithm; however, greater advisor accuracy is concentrated among students with whom advisors had more interactions. The algorithm achieved similar accuracy among students lower in the distribution of interactions, despite advisors having substantially more information. We find no evidence that the advisors or algorithm exhibit bias against vulnerable populations. Our results suggest that, especially at scale, algorithms have the potential to provide efficient, accurate, and unbiased predictions to target scarce social services and resources.

VERSION: January 2023

Suggested citation: Akmanchi, Suchitra, Kelli A. Bird, and Benjamin L. Castleman. (2023). Human versus Machine: Do college advisors outperform a machine-learning algorithm in predicting student enrollment?. (EdWorkingPaper: 23-699). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/gadf-ey53

Human versus Machine: Do college advisors outperform a machine-learning algorithm in predicting student enrollment?

Suchitra Akmanchi, University of Virginia
Kelli A. Bird, University of Virginia
Benjamin L. Castleman, University of Virginia
*Authors listed in alphabetical order

Abstract

Prediction algorithms are used across public policy domains to aid in the identification of at-risk individuals and guide service provision or resource allocation. While growing research has investigated concerns of algorithmic bias, much less research has compared algorithmically-driven targeting to the counterfactual: human prediction. We compare algorithmic and human predictions in the context of a national college advising program, focusing in particular on predicting high-achieving, lower-income students' college enrollment quality. College advisors slightly outperform a prediction algorithm; however, greater advisor accuracy is concentrated among students with whom advisors had more interactions. The algorithm achieved similar accuracy among students lower in the distribution of interactions, despite advisors having substantially more information. We find no evidence that the advisors or algorithm exhibit bias against vulnerable populations. Our results suggest that, especially at scale, algorithms have the potential to provide efficient, accurate, and unbiased predictions to target scarce social services and resources.

Acknowledgements

We are grateful to program leadership and advisors from CollegePoint, College Possible, and College Advising Corps for their engagement on this project. We are grateful to Bloomberg Philanthropies for their financial support for this research. Any errors or omissions are our own.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B200005 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Introduction

Prediction algorithms are used across numerous public policy domains to aid in the identification of at-risk individuals and guide service provision or resource allocation. Prediction models inform which patients receive costly medical treatments, the felony sentences judges assign to defendants, and which college students receive additional academic support (Bird et al., 2021; Obermeyer et al., 2019; Stevenson and Doleac, 2019). In parallel to the rising role of prediction algorithms are growing concerns about algorithm accuracy and fairness, and whether increased reliance on algorithms will result in worse outcomes for vulnerable populations (Bird, Castlman, and Song, in progress; Lee et al., 2019; Osoba and Welser, 2017; Slaughter, Kopec, and Batal, 2021). For instance, Obermeyer et al. (2019) demonstrate that an algorithm used to guide health decisions assigns similar risk levels to Black patients who are in fact sicker than their White counterparts, and that this algorithmic bias reduces the number of Black patients recommended for additional treatments.

Alongside concerns about algorithmic bias accuracy and fairness, it is also important to consider the counterfactual of how inaccuracies and biases in human-generated predictions may also negatively affect program and service allocation. Extensive research demonstrates, for instance, that people working across a variety of professions, from police officers to physicians, exhibit implicit biases towards others based on their identity and attributes, and that these implicit biases impact consequential decisions, such as medical treatments and employment offers (e.g. Bertrand and Mullainathan, 2004; Eagly and Karau, 2002; Eberhardt et al., 2004; Green et al., 2007). Some evidence suggests that algorithms may in fact help mitigate the effect of human bias, e.g. in employment settings (Cowgill, 2020).

In some public domains, such as criminal justice, studies have investigated the comparative accuracy and fairness of algorithms and humans in making important predictions. Lin et al. (2020) demonstrate that a commercial algorithm used to predict recidivism in criminal justice outperformed humans when humans did not receive immediate feedback on the accuracy of their predictions or when researchers provided the algorithm with additional predictors. Kleinberg et al. (2018) show that using algorithms to determine bail decisions could substantially reduce subsequent criminal activity relative to bail decisions determined by judges.

Investigations of model versus human accuracy are much more nascent in other domains, however, including in education, despite rapidly growing use of algorithms to predict student outcomes (Baker et al., 2020; Bird et al., 2021; Ekowo and Palmer, 2016; Ositelu and Acosta, 2021).[1] On the one hand, we might expect teachers, counselors, and other educators to dominate algorithms in prediction accuracy given their direct interactions with students and awareness of contextual factors that might be difficult to incorporate into an algorithm (e.g. family or household factors) but which affect student performance. On the other hand, substantial evidence demonstrates that teachers exhibit racial biases and adjust their expectations and evaluation of students based on their identity (Dee, 2005; Grissom and Redding, 2016; Starck et al., 2020). Yet research to date exploring human versus model accuracy in education has been limited to contexts outside of the U.S. or to very specialized settings, e.g. medical school enrollment at one institution or a single ten-day online course (Burkhardt et al., 2018; Kanter and Veeramachanemi, 2015; Ye, 2021).

We contribute new evidence by comparing algorithm versus human accuracy in the context of a national remote college advising program, CollegePoint, that has served tens of thousands of high school seniors over the last decade (for more information on the CollegePoint program, see Sullivan et al., 2021). CollegePoint pairs low- and moderate-income, high-achieving high school seniors with professional college advisors or peer mentors who provide them with individualized assistance with college and financial aid applications. The primary goal of CollegePoint is to increase the share of lower-income, high-achieving students that matriculate at selective colleges and universities with high graduation rates -- the program refers to these as “CollegePoint schools”. Historically, advising has primarily taken place from the start of senior year, when students are deciding which colleges to apply to, through the end of senior year, when students decide which college they want to attend.

While CollegePoint has historically concluded advising at the end of students’ senior year in high school, during the COVID-19 pandemic the program extended advising into the summer

[1] For instance, at least 1,400 colleges and universities now use predictive analytics, collectively spending hundreds of millions of dollars each year on prediction software (Barshay and Aslanian, 2019).

after high school. Due to resource limitations CollegePoint was not able to provide summer advising to all students. To guide which students advisors focused their time on during the summer after high school, we trained a logistic regression algorithm (using data from several CollegePoint cohorts) to predict whether students would enroll at a CollegePoint school. The predictors included measures related to academic performance, student and high school demographics; the number and selectivity of college applications submitted by students; and college cost. We provide further detail on the algorithm and associated data in the Data and Methods section of the paper. We applied the algorithm to students from the CollegePoint Class of 2021 cohort to generate predictions for enrollment in a CollegePoint school during the Fall 2021 semester.

In parallel, we worked with program leadership to have CollegePoint advisors complete a guided assessment of whether students in their Class of 2021 caseload would attend a CollegePoint school.[2] Thirty-three advisors participated in the activity, providing assessments for 856 students drawn from across the country. Advisors had access to all of the student-level data we used as algorithm predictors. In addition, advisors had detailed notes from their interactions with students as well as student self-reported data on where students had been accepted to college and where they planned to enroll as of the end of high school that was not available when we trained the algorithm. Advisors thus had substantially more information to inform their predictions than did the algorithm, at least for students with whom they had more frequent interactions.

Our primary analysis focuses on assessing the accuracy with which the algorithm and advisors predicted students' enrollment at CollegePoint schools, for the 856 students for whom we observe both algorithm predictions and advisor assessments. The analysis yields several key results. First, both the algorithm and advisors accurately predict attendance at a CollegePoint school for the majority of students (66-69 percent). Second, advisors, on average, slightly outperform the algorithm in predicting student enrollment. However, the modestly greater advisor accuracy is entirely concentrated among students with whom advisors had more frequent interactions. Specifically, advisors outperform the algorithm for students with whom advisors

[2] We provide the guided assessment advisors used in the Appendix and discuss the assessment process in more detail in the Data and Methods section.

had above the 60th percentile of interactions for the analytic sample (10 or more interactions). By comparison, the algorithm achieved similar accuracy among students lower in the distribution of interactions. This is despite advisors having substantially more information than the algorithm, both from their qualitative notes from interacting with students and from additional milestone completion data (e.g. where students were accepted to college) to which advisors had access but which we were not able to incorporate into the algorithm. Third, we find no evidence that the algorithm's or advisors' predictions are differentially accurate across student groups, suggesting that neither the algorithm nor advisors exhibit clear bias in their assessments of students' probability of enrolling at a CollegePoint school.

Data and Methods

Sample Overview

CollegePoint is a remote college advising program funded by Bloomberg Philanthropies that has been serving high-achieving, low- and moderate-income students since 2014. The primary objective of CollegePoint is to increase enrollment at high-quality colleges and universities for this population of students. Specifically, CollegePoint focuses on institutions with graduation rates above 70 percent, which we refer to as "CollegePoint schools". CollegePoint recruits high school students who score above the 90th percentile on their PSAT, SAT or ACT and have a high school GPA of 3.5 or higher, and also have annual family income below $80,000. Once the student signs up for CollegePoint, they are connected to a remote advisor from one of four participating advising organizations. Sullivan et al (2021) provides additional detail about CollegePoint organizational practices.

In this paper, we compare advisor-generated and algorithm-generated predictions of the enrollment outcomes for high school students in the CollegePoint Class of 2021; we trained the algorithm on historical data from the CollegePoint Classes of 2017 through 2020. Table 1 presents baseline characteristics for three samples: (1) the historical training sample; (2) the full Class of 2021; and (3) the subset of the Class of 2021 for whom we have advisor predictions (described below). Overall, approximately half of the students are first generation college students, and all of the students in the Class of 2021 are low- or moderate-income. The majority of students identify as Asian, White, or Hispanic/Latino- with these three groups making up approximately 60 percent of the students served. The subset of students with advisor predictions are very similar to the full Class of 2021.[3]

Data Overview

Through CollegePoint, its partner advising organizations, and external data sources, we observe student demographics, test scores, high school characteristics, college application information, and college enrollment. College Board and the ACT provided student-level demographics (race/ethnicity, gender, parent education); SAT, ACT, or PSAT score; high school attended and GPA, and family income category. Over the course of the advising period (which generally spans the students' senior year of high school), advisors made a record of each interaction they have with a student and the topics discussed. Advisors also recorded the list of colleges to which a student applied; we supplement this list with information about SAT score sends from the College Board.

We linked each student's high school to the National Center for Education Statistics (NCES) database and the list of colleges to which they apply to the Integrated Postsecondary Education Data System (IPEDS) to obtain institutional characteristics. Among these are the racial composition of the student's high school, the net price of college, the distance of the student to the colleges to which they applied, and the average SAT score percentile for the incoming college class among the application schools. We observe students' eventual college enrollment outcomes through National Student Clearinghouse (NSC) matches.[4]

Advisor Prediction

Our data on advisor predictions of whether students would enroll at a CollegePoint school come from a guided reflection CollegePoint administered to a subset of CollegePoint advisors in May 2021. The guided reflection prompted advisors to gather and consider data, notes, and contextual

[3] For all CollegePoint Classes, our main source of demographic data is the College Board. For the Classes of 2017 to 2020, we supplemented this with demographic information from both Common App and the ACT. We did not have access to these supplemental data sources for the Class of 2021, which accounts for the differences in missingness on baseline measures for the Class of 2021.

[4] The NSC includes student-by-term enrollment and graduation records for nearly all colleges and universities in the United States. The NSC currently covers 97.3 percent of all higher education enrollments; the coverage is 99.6 percent and 96.4 percent for public 4-year and not-for-profit 4-year institutions, respectively.

information they had on specified students from their advising caseload, and for each student, to assess how likely they would be to matriculate at a CollegePoint school. The reflection then prompted advisors to translate this assessment into a 1 to 10 rating, for each student on their caseload, of the likelihood of enrollment at a CollegePoint school. The reflection also prompted advisors to consider and identify the extent to which various barriers could potentially impede students from successfully matriculating at a CollegePoint school. We include the full guided reflection advisors were asked to complete in the Appendix.

CollegePoint invited advisors from two out of four organizations, College Advising Corps and College Possible, to participate in this project. 68.8 percent of College Advising Corps advisors (16 advisors in total) and 81.4 percent of College Possible advisors (27 advisors in total) who advised the Class of 2021 agreed to participate. CollegePoint prompted advisors to complete the reflection for a randomly selected one-third of their advising caseload, yielding predictions of enrollment at CollegePoint schools and associated potential barriers to enrollment for 856 students.

Algorithm Prediction

To generate our predictions for the Class of 2021, we use a logistic regression algorithm trained on data from previous CollegePoint classes (2017-2020).[5] We used logistic regression as our prediction algorithm because of its widespread use in predictive modeling, its suitability for binary outcomes, and relatively low cost to implement (in terms of computing power and technical skills required). In a systematic comparison of prediction models in postsecondary education applications, we moreover found little difference in performance between logistic regression and more advanced prediction methods (e.g. random forest models) (Bird et al., 2021).

[5] We originally developed this prediction algorithm to predict CollegePoint school enrollment for the Class of 2020, using historical training data from the Classes of 2017 to 2019. The initial goal for creating this algorithm was to allow CollegePoint to target resources toward at-risk students. Specifically, during the COVID-19 pandemic, CollegePoint decided to extend advising through the summer after high school to support students after high school graduation with difficulties they might encounter matriculating to their intended college or university. CollegePoint had limited summer advising to extend to students, however, and was interested in using an algorithm to inform which students advisors targeted for more intensive support.

Our primary outcome of interest is whether students enrolled at a CollegePoint school, which we define based on student-level college enrollment records from the National Student Clearinghouse. We included a total of 98 student-level predictors in our algorithm. These predictors include demographic predictors (student's gender, race/ethnicity, income category, parental education, state of residence, whether state of residence has a CollegePoint school, and percent of high school population that is non-white), predictors describing student-advisor interactions (the number of times the student and advisor interacted, whether the student and advisor discussed one of the several common topics)[6], predictors describing student test scores (students' first SAT/ACT score, best SAT/ACT score, and best pre-CollegePoint SAT/ACT score)[7], predictors describing program implementation (what month the student began receiving CollegePoint, which of organization the student was assigned to), and predictors describing the colleges within a student's application set (whether college is a Barrons 1 or 2, total in-state cost of college, student's distance from college, average SAT percentile score of incoming college class, net price of college based for family income categories <$30k, $30k-48k, and $48-75k). For this last category of predictors describing college application behavior, if a student applied to multiple colleges, then we aggregated or averaged each predictor across the set of colleges to which the student applied (e.g. number of Barron's 1 or 2 schools applied to; average distance in miles from student's home to each college applied to).[8] For student observations missing predictor values, we converted the missing values to zero and include an indicator to capture whether or not the predictor is missing; these missing indicators are included in the predictor counts in the above paragraph.[9]

[6] The top four topics are applying for financial aid, applying to college, college list, and college decision.

[7] We convert the composite SAT and ACT scores to percentile scores, and combine these percentile scores into one measure.

[8] Barron's index is commonly used as a high-level measure of institutional selectivity. Barron's 1 institutions are classified as "most competitive" and Barron's 2 are classified as "highly competitive". Examples of Barron's 1 institutions include Harvard University, Williams College, and University of Virginia; examples of Barron's 2 institutions include Boston University, Trinity College, and Ohio State University.

[9] Data are missing for three reasons. First, demographic data is missing if the student is not in the College Board data (i.e. they did not take the SAT). For the Classes of 2017-2020, we supplemented demographic data with information from the Common App and the ACT, but we did not receive this data for the Class of 2021, resulting in higher rates of demographic missingness for the Class of 2021. Second, students' high school characteristics are missing if we do not have a valid high school code for the student. For the Classes of 2017-2020, CollegePoint recorded students' high school's numerical code via their intake form, but they stopped this process for the Class of 2021. Consequently, we must rely fully on the high school code provided by College Board for the Class of 2021, and roughly 25 percent of the Class of 2021 cannot be linked to high school characteristics (see Table 1). Finally, college application data are missing if the advisor did not have a record of the student applying to any college as of the time of data collection (or we do not observe this application, due to lack of advisor record or SAT score send), or if the student applied to a college that is not in the IPEDS data (e.g. a university outside the United States).

Following standard predictive modeling practices, we randomly divided the historical cohorts into a training set (80 percent) and a test set (20 percent); this division allowed us to evaluate the performance of the algorithm on out-of-sample data. The training set established the algorithm parameters (i.e. the coefficients of the logistic model), and used those parameters to assign each student in the test set a predicted score ranging from zero to one. To convert these predicted scores to a binary prediction (i.e. will or will not enroll at a CollegePoint school), we followed a commonly used procedure to set the threshold to maximize the F1-score.[10] The resulting threshold was 0.40, indicating that any student with a predicted score at or above 0.40 is predicted to enroll at a CollegePoint school, while all other students are predicted not to enroll at a CollegePoint School.

We report three standard metrics of algorithm accuracy: c-statistic, precision, and recall. The c-statistic is a goodness of fit measure, and is equal to the probability that a randomly chosen true positive is assigned a higher predicted score by the algorithm than a randomly chosen true negative. Our algorithm has a c-statistic of 0.76, indicating moderately good performance.[11] The algorithm's c-statistic of 0.76 means that for any two selected students -- one who enrolled at a CollegePoint school and one who didn't -- the algorithm assigns a higher predicted score to the enrolled student 76 percent of the time. Our algorithm's precision is 0.66, indicating that 66 percent of students that the algorithm predicted to enroll in a CollegePoint school actually did so. Our algorithm's recall is 0.87, indicating that 87 percent of the students who actually did enroll at a CollegePoint school were accurately predicted by the algorithm.

Comparing Advisor and Algorithm Prediction

Our main comparison of interest is between the advisor- and algorithm-generated predictions of enrollment in a CollegePoint school. Since a student's enrollment outcome is binary, we convert the algorithm's and advisor's predictions to also be binary. As mentioned above, the algorithm internally generates a binary enrollment prediction based on a threshold value (0.40). We map

[10] The F1-score is the harmonic mean of precision and recall. We use 10-fold cross-validation to find the threshold to maximize the F1-score.

[11] According to Hosmer, Lemeshow, and Sturdivant (2013), a c-statistic between 0.7 and 0.8 is considered acceptable performance. Similarly, Bowers and Zhou (2019) characterize a c-statistic between 0.75 and 0.85 as "partially convincing".

advisor ratings to a binary variable where a rating of 1-4 (inclusive) is a prediction of no enrollment in a CollegePoint schools and a rating of 5-10 (inclusive) is a prediction of enrollment. It is then straightforward to compare the binary predictions of the algorithm and advisors to students' actual (binary) enrollment outcome.[12]

Results

Advisor versus Algorithm Accuracy

We begin by comparing the overlap between algorithm predicted scores and advisor ratings. Figure 1 displays the correlation between algorithm and advisor predictions; the size of the circles is directly proportional to the number of students each represents.[13] There is a moderate positive relationship (correlation=0.44) between the algorithm and advisors' predictions, with fairly common "disagreement" between the algorithm and advisors on their assessment of the likelihood a student would enroll at a CollegePoint school. For instance, among students the advisors identify as most likely to enroll at a CollegePoint school ("10" out of the 1-10 rating), there is a nontrivial mass of students (19.14 percent) for whom the algorithm generated a predicted score of 0.40 or lower. When we convert the advisor ratings and algorithm predicted scores to the binary predictions of "will enroll at a CollegePoint school" (rating >= 5; predicted score >= 0.4) or "will not enroll at a CollegePoint schools", we find that the advisors and algorithm predict different outcomes for 31.3 percent of students. This level of "disagreement" is perhaps not surprising, given that advisors had access to more, and arguably richer, information and data about students.

The magnitude of "disagreement" between the algorithm and advisors did not translate into substantially greater accuracy of predictions for the advisors relative to the algorithm. Specifically, the overall accuracy for advisors is slightly higher compared to the algorithm, though this difference in accuracy is not statistically significant (68.46 percent versus 65.5 percent; p =0.11). As we indicate in the lower left of Figure 2, the algorithm and advisor

[12] When we repeat our analysis treating the rating of "5" as either a prediction students would not enroll at a CollegePoint school or as an inconclusive prediction, our results are very similar.
[13] Since advisor predictions of likelihood of enrollment were on a 1-10 scale, and algorithm-predicted scores were from 0-1, we rescaled the algorithm scores to be from 1-10 such that (0.0, 0.1) = 1; [0.1,0.2) = 2; ... [0.9, 1) = 10.

provided the same accurate predictions for just under half of students (49.9 percent). When the algorithm and advisors disagreed, the advisors had accurate predictions for slightly more students (18.8 versus 15.6 percent). Neither were correct in predicting outcomes for the remaining 15.7 percent of the sample. The algorithm has a lower false positive rate -- i.e. incorrectly predicting a student will enroll when they will not -- compared with the advisors (22.5 percent versus 27.4 percent; $p < 0.01$). However, the algorithm also has a significantly higher false negative rate -- i.e. incorrectly predicting a student will not enroll when they do -- compared with the advisors (12.0 percent versus 3.9 percent; $p < 0.01$). In this sense, the algorithm is slightly more "pessimistic" about students' enrollment outcomes than the advisors.

We also present in Figure 2 the accuracy rates for the algorithm and advisor across the prediction distribution. The x-axis corresponds to the algorithm predicted scores (blue circle) and advisor ratings (red triangle), with the size of the shapes being directly proportional to the number of students each represents. The y-axis corresponds to the accuracy of the algorithm and advisor predictions (i.e. whether students predicted to enroll at a CollegePoint school actually did so). We define accuracy as whether the advisor or algorithm made a correct binary prediction, i.e. whether or not the student would enroll at a CollegePoint school.[14] Several notable features emerge from this plot. First, both the algorithm and advisors are generally more accurate closer to the tails of the distribution.[15] This pattern makes intuitive sense, since the profiles of students most or least likely to enroll at a CollegePoint school may be most apparent to advisors and have data attributes most strongly associated with CollegePoint enrollment, facilitating accurate prediction by the algorithm. By comparison, accuracy rates are lowest in the middle of the distribution, where students' outcomes may have been more ambiguous. Second, at most points in the distribution the algorithm and advisors had fairly similar accuracy rates. When we focus on predicted scores and ratings of 2, 3, and 4, we see that advisors were more accurate than the algorithm, which perhaps indicates the advisors had access to contextual information suggesting students faced barriers to enrolling at a CollegePoint school. The algorithm was modestly more accurate for predicted scores or ratings of 5, 6, and 7. This may reflect that, for students for

[14] If a student did enroll at a CollegePoint school, then an accurate advisor prediction would be a rating >= 5; an accurate algorithm prediction would be a predicted score >= 0.40.

[15] The algorithm accuracy for students in the bin of 0.9 - 1.0 is quite low, but this represents very few students (0.1%).

whom the advisors had greater uncertainty about their enrollment outcomes, they defaulted to assigning students a rating of 5. As we see above in Figure 1, this seems to be reflected in the distribution of advisor ratings, with advisors assigning the ratings of 1, 5, and 10 with the greatest frequency.[16]

Relationship between Prediction Accuracy and Frequency of Interactions with Advisor

In the CollegePoint initiative, students had the opportunity to interact with their advisor at numerous points throughout the college and financial aid application process and while they were making their choice about which college to attend. The average student interacted with their advisor 10.1 times, though the full distribution ranges from 1 to 61 with an interquartile range of 3 to 13. Over the course of these interactions, students may have shared meaningful information about their personal circumstances (e.g. a parent losing a job, making paying for college more difficult), which advisors captured in qualitative notes and could factor into their predictions. By comparison, the algorithm was fully reliant on the comparatively easy-to-quantify measures we describe above. Table 2 compares advisor versus algorithm accuracy based on the quintile of number of interactions a student had with their advisor. Across the first three quintiles (students with 9 or fewer interactions) the algorithm has slightly better accuracy compared to advisors (67.6 percent versus 63.4 percent). However, advisors are significantly more accurate than the algorithm for students in the top two quintiles (10 or more interactions, 76.5 percent accurate compared to 61.6 percent for the algorithm). This result is intuitive -- more interactions with students mean advisors have more information on which to base their ratings.

When we explore the relationship between interactions and advisor ratings, we find that advisors are more likely to rate students as very low (1 or 2) or very high (8, 9, or 10) if they had more interactions with students, while advisors were more likely to assign a middle rating for students

[16] We chose to have advisor ratings of 5 corresponding as the threshold for assigning a binary advisor prediction because of its closeness to the algorithm's predicted score threshold of 0.40. We acknowledge that a rating of 5 could instead be interpreted as "unsure either way". However, because approximately half of students with a rating of 5 do enroll at a high-quality college, the advisor accuracy is virtually unchanged if we treat ratings of 5 differently. Specifically, if we treat a rating of 5 as predicting a negative outcome, the overall advisor accuracy is 70.1 percent. If we treat a rating of 5 as 50 percent predicting a negative outcome and 50 percent predicting a positive outcome, then the overall advisor accuracy is 69.5 percent.

with whom they did not frequently interact. Again, this pattern suggests that advisors likely used a rating of 5 as a default "I don't know" option for students they knew little about.

We next explore whether the timing of interactions is related to advisor accuracy. If an advisor only interacted with a student during the Fall of their senior year, then that advisor would only know that student's intended or actual application behavior (information that is also incorporated as predictors into the algorithm). However, if an advisor continued to interact with the student during the Spring, the advisor may learn to which schools the student was admitted, and eventually to which school the student accepted the admission offer. In Figure 3, we show advisor and algorithm accuracy separately for the subgroups of students who did or did not interact with their advisor after key dates in the admissions process: January 31st (typical date for colleges to send Early Decision admissions decisions); April 1st (typical date for colleges to send regular admissions decisions); and May 1st (typical deadline for students to accept admissions offer). In Plot A, we see that advisors have worse accuracy than the algorithm for students who did not interact on or after January 31st (58 versus 66 percent, respectively, $p = 0.04$), but significantly better accuracy for students they did interact with on or after January 31st (73 versus 65 percent, respectively, $p < 0.01$). We find similar differences in advisor/algorithm accuracy for students who interacted on or after April 1st (11.9 percent better advisor accuracy, $p = 0.01$) and on or after May 1st (19.4 percent better advisor accuracy, $p = 0.02$). These results suggest that the modestly higher level of overall accuracy from the advisors compared to the algorithm is driven, at least in part, by advisors having additional knowledge of admissions and acceptance decisions.

Investigating potential accuracy bias in algorithm and advisor predictions

Finally, we compare the advisors' and algorithm accuracy across student demographic and socio-economic subgroups. As seen in Table 3, we find that advisors are directionally more accurate than the algorithm for every subgroup (or equally accurate, in the case of the middle-income subgroup), with more precisely-measured differences for the non-URM[17] ($p =0.03$), low-income ($p =0.01$), and female ($p =0.08$) subgroups.

[17] URM is an abbreviation for underrepresented minoritized, which includes non-White and non-Asian students.

When we compare advisors' accuracy across subgroups, we find that advisors have similar levels of accuracy for more disadvantaged subgroups and less disadvantaged subgroups. In fact, advisors are slightly more accurate for low-income students compared with middle-income students, though this difference is not statistically significant (p = 0.43). When we compare the algorithm's accuracy across subgroups, we find more variation in accuracy across subgroups; however, the algorithm is generally more accurate for more disadvantaged subgroups compared to less disadvantaged subgroups. Specifically, the algorithm is more accurate for URM students compared with non-URM students (p = 0.07).

The results that both advisors and the algorithm are similarly or more accurate for more disadvantaged populations run contrary to previous work that both humans and algorithms are worse at predicting outcomes for marginalized people (Bertrand and Mullainathan, 2004; Koch, D'Mello, and Sackett, 2014; Gershenson, Holt, and Papageorge, 2016), and therefore may allay concerns that making predictions (either by humans or algorithms) could lead to bias against vulnerable populations, at least in the context of this CollegePoint algorithm.

Discussion

Prediction algorithms have the potential to increase the accuracy and efficiency of service provision and resource allocation across a variety of policy domains. The importance of efficient and effective targeting has only increased in the wake of the COVID-19 pandemic, with growing shares of the population potentially benefiting from limited social services, e.g. to address learning loss in schools or mental health counseling needs (Centers for Disease Control, 2022; Kuhfeld et al., 2022). While growing research has investigated concerns that algorithmic bias could result in fewer resources being allocated to more vulnerable populations, much less research (especially outside criminal justice) has empirically compared service and resource targeting through algorithms to the counterfactual: human judgment and prediction.

Our results indicate that humans are slightly more accurate in their predictions than algorithms, at least as it pertains to college enrollment quality, when they are able to consider a broader and more qualitatively-rich range of data. Specific to our context, advisor accuracy was highest among students with whom advisors had interacted frequently and thus may have had insight into relevant contextual factors affecting whether and where students would enroll in college. Our results also indicate that humans are less confident and less accurate in their predictions when the outcome is more ambiguous. Whereas advisor predictions were more accurate among students least or most likely to enroll at a CollegePoint school, they appeared to default to a neutral rating of “5” for students whose enrollment outcomes were less clear, and for these students advisor predictions were less accurate than the algorithm.

Especially at scale, people tasked with allocating scarce services and resources may often find themselves in the position of having had fewer interactions with the individuals they are serving or facing uncertainty about individuals’ potential outcomes. Our results suggest that, in these scenarios, algorithms may equal or exceed humans in the accuracy with which they predict individual outcomes. Algorithms may also allow for more cost-efficient prediction at scale. While the upfront costs to train algorithms are non-trivial, primarily in terms of skilled analytic capacity, once operational the variable costs to generate predictions from algorithms are likely to be far lower than having humans manually assess individuals’ probability of achieving different outcomes.

Despite concerns about algorithm accuracy and fairness in education, we do not find evidence that algorithmic or human predictions are less accurate for historically vulnerable populations pursuing postsecondary education. If anything, our results indicate that both the algorithm and advisors’ predictions are *more* accurate for disadvantaged populations, which could result in better-targeted services and resources for these students.

Given the nascence of research on algorithmic and human prediction accuracy, it will remain important for researchers to investigate algorithm accuracy and fairness in other policy domains and at other educational margins where these algorithms are applied. But our results suggest that algorithms have the potential to provide efficient, accurate, and unbiased predictions to target scarce social services and resources.

References

Baker, Ryan S., Andrew W. Berning, Sujith M. Gowda, Shizhu Zhang & Aaron Hawn (2020) Predicting K-12 Dropout, Journal of Education for Students Placed at Risk (JESPAR), 25:1, 28-54, DOI: 10.1080/10824669.2019.1670065

Bertrand, Marianne and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, Vol 94, No 4, September 2004 (pp 991-1013). DOI: 10.1257/0002828042002561

Bird, K. A., Castleman, B. L., Mabel, Z., & Song, Y. (2021). Bringing Transparency to Predictive Analytics: A Systematic Comparison of Predictive Modeling Methods in Higher Education. *AERA Open*, *7*. https://doi.org/10.1177/23328584211037630

Bowers, A.J., Zhou, X. (2019) Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. Journal of Education for Students Placed At Risk, 24(1) 20-46. https://doi.org/10.1080/10824669.2018.1523734

Burkhardt, J., DesJardins, S. L., Teener, C. A., Gay, S. E., et al. Predicting Medical School Enrollment Behavior: Comparing an Enrollment Management Model to Expert Human Judgement. *Academic Medicine*, 93 (11S) pp. S68-73. https://doi.org/10.1097/ACM.0000000000002374

Centers for Disease Control and Prevention. "New CDC data illuminate youth mental health threats during the COVID-19 pandemic." Press release, March 31, 2022. https://www.cdc.gov/media/releases/2022/p0331-youth-mental-health-covid-19.html

Cowgill, Bo. Bias and Productivity in Humans and Machines (July 30, 2019). Columbia Business School Research Paper Forthcoming, Available at SSRN: https://ssrn.com/abstract=3584916 or http://dx.doi.org/10.2139/ssrn.3584916

Dee, Thomas, S. 2005. "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review*, 95 (2): 158-165.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573

Ekowo, Manuela and Iris Palmer. "The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis." New America: Policy Paper, Oct. 24, 2016. https://www.newamerica.org/education-policy/policy-papers/promise-and-peril-predictive-analytics-higher-education/#:~:text=In%20a%20new%20paper%2C%20The,well%20in%2C%20and%20provide%20digital

Gershenson, Steh, Stephen B. Holt, Nicholas W. Papageorge. Who believes in me? The effect of student–teacher demographic match on teacher expectations. *Economics of Education Review*, Volume 52, 2016, Pages 209-224, ISSN 0272-7757, https://doi.org/10.1016/j.econedurev.2016.03.002.

Green, A.R., Carney, D.R., Pallin, D.J., Ngo, L.H., Raymond, K.L., Iezzoni, L.I., and Banaji, M.R. Implicit Bias among Physicians and its Prediction of Thrombolysis Decisions for Black and White Patients. *J GEN INTERN MED* 22, 1231–1238 (2007). https://doi.org/10.1007/s11606-007-0258-5

Grisson, Jason A. and Christopher Redding. Discretion and Disproportionality: Explaining the Underrepresentation of HIgh-Achieving Students of Color in Gifted Programs. *AERA Open*, 2016. https://doi.org/10.1177/2332858415622175.

Hosmer, David W. Jr., Stanley Lemeshow, and Rodney X. Sturdivant (2013). Applied Logistic Regression, Third Edition. John Wiley & Sons, Inc., Hoboken New Jersey. ISBN 978-0-470-58247-3.

Kanter, J. M. and K. Veeramachaneni, "Deep feature synthesis: Towards automating data science endeavors," *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2015, pp. 1-10, doi: 10.1109/DSAA.2015.7344858.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan, Human Decisions and Machine Predictions, *The Quarterly Journal of Economics*, Volume 133, Issue 1, February 2018, Pages 237–293, https://doi.org/10.1093/qje/qjx032

Kuhfeld, Megan, James Solan, and Karyn Lewis (2022). “Test Score Patterns Across Three COVID-19-impacted School Years.” EdWorking Paper No. 22-51. Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ga82-6v47.

Koch AJ, D'Mello SD, Sackett PR. A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. J Appl Psychol. 2015 Jan;100(1):128-61. doi: 10.1037/a0036734. Epub 2014 May 26. PMID: 24865576.

Lee, Nicole Turner, Paul Resnick, and Genie Barton. “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms.” Brookings Report, May 22, 2019. https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Lin, Zhiyuan & Jung, Jongbin & Goel, Sharad & Skeem, Jennifer. (2020). The limits of human predictions of recidivism. Science Advances. 6. eaaz0652. 10.1126/sciadv.aaz0652.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 25 Oct 2019, Vol 366, Issue 6464, pp. 447-453. DOI: 10.1126/science.aax234

Ositelu, Monique O. and Alejandra Acosta. “Equity and Predictive Analytics in Enrollment Management.” New America: Collection. Oct. 6, 2021 https://www.newamerica.org/education-policy/collections/equity-and-predictive-analytics-in-enrollment-management/

Osoba, Osonde A. and William Welser IV, An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence. Santa Monica, CA: RAND Corporation, 2017. https://www.rand.org/pubs/research_reports/RR1744.html.

Slaughter, Rebecca Kelly, Janice Kopec, and Mohamad Batal. "Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission. Yale Low School ISP Digit Future Whitepaper & YJoLT Special Publication, August 2021. https://law.yale.edu/sites/default/files/area/center/isp/documents/algorithms_and_economic_justice_master_final.pdf

Starck, Jordan G., Travis Riddle, Stacey Sinclair, and Natasha Warikoo. Teachers Are People Too: Examining the Racial Bias of Teachers COmpared to Other American Adults. *Educational Researcher*, Volume 49, Issue 4, April 2020.

Stevenson, Megan T. and Jennifer L. Doleac. "Algorithmic Risk Assessment in the Hands of Humans." IZA DP no. 12853. December 2019

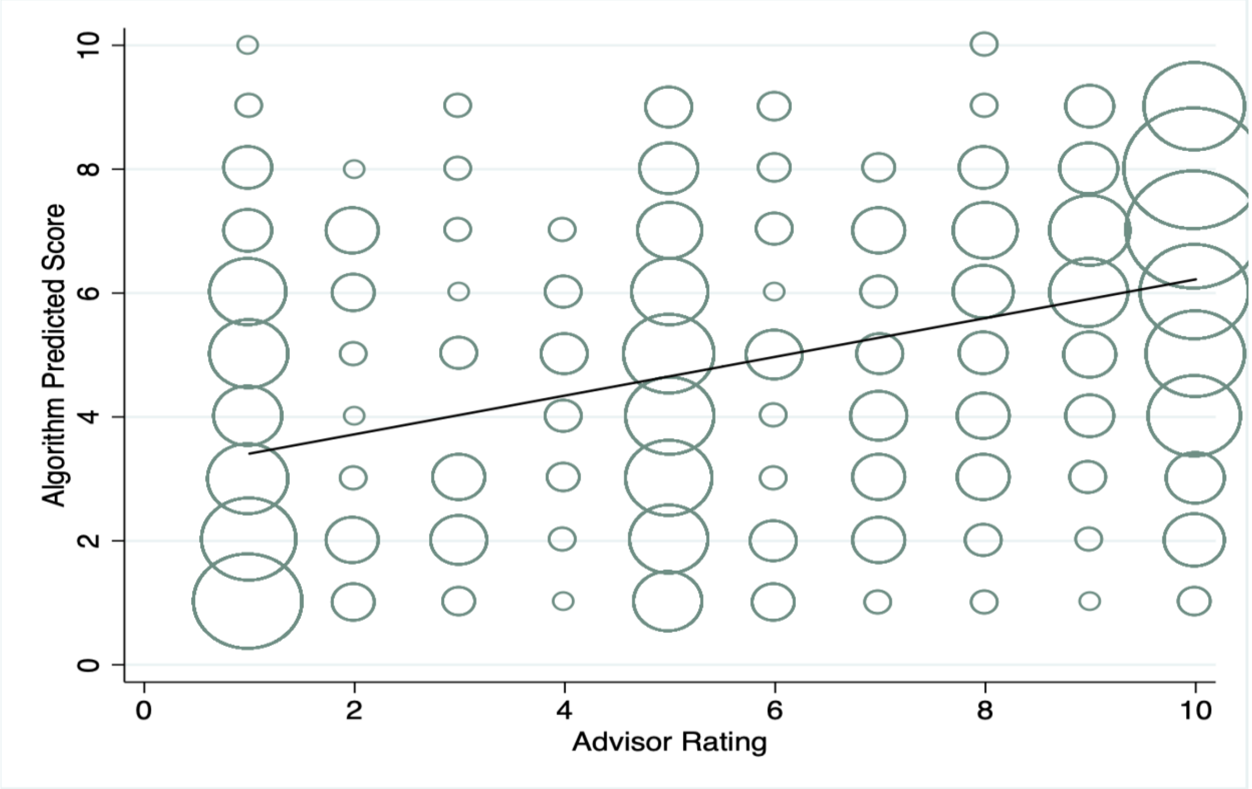
Sullivan, Zach, Ben Castleman, Gabrielle Lohner, and Eric Bettinger. (2021). College Advising at a National Scale: Experimental Evidence from the CollegePoint initiative. (EdWorkingPaper: 19-123). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/s323-5g64

Ye, Xiaoyang. Personalized advising for college match: Experimental evidence on the use of human expertise and machine learning to improve college choice. Working paper, 2022

Table 1: Student characteristics of relevant samples

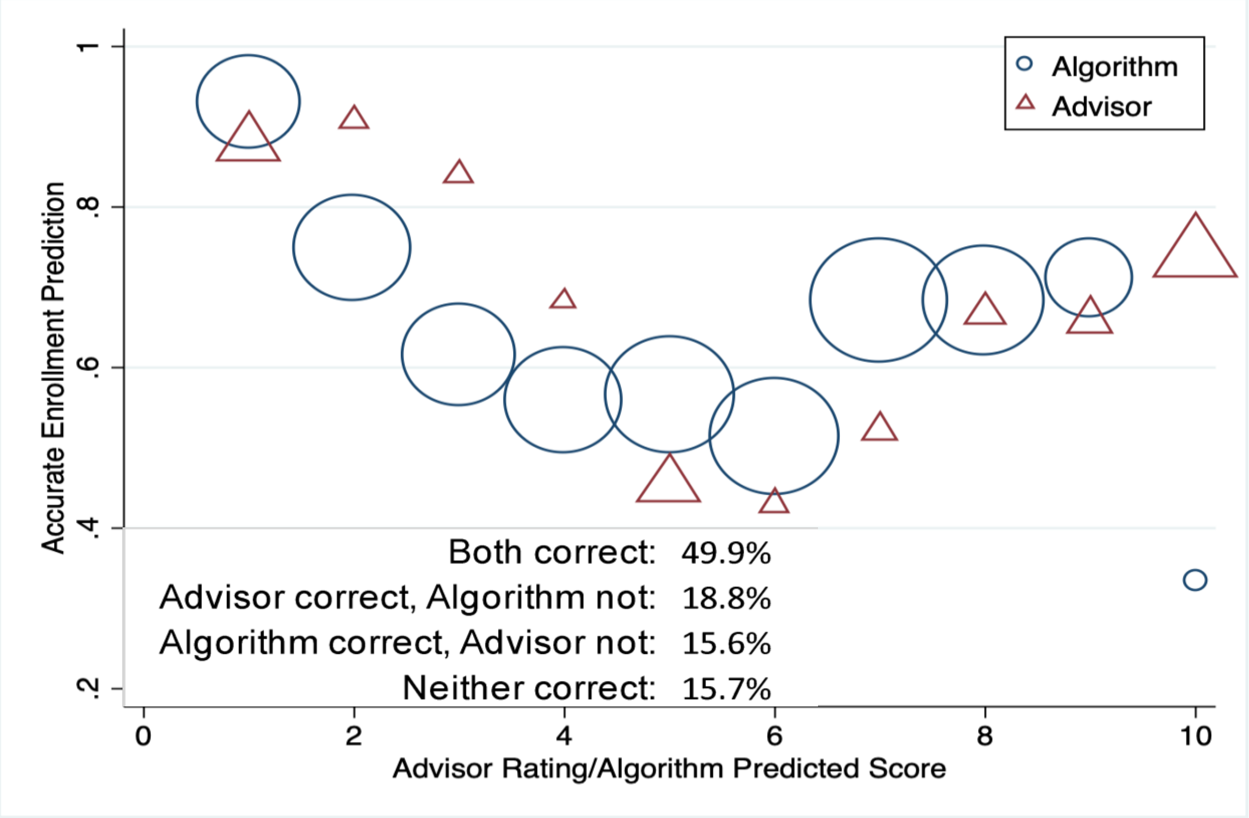
	Classes of 2017-2020	Full Class of 2021	Class of 2021 with advisor predictions
	(1)	(2)	(3)
Female	45.6%	43.7%	45.2%
Male	38.3%	32.6%	29.1%
Missing Gender	16.1%	23.6%	25.6%
Asian	24.1%	23.6%	19.3%
Black	6.3%	8.6%	7.9%
Hispanic/Latino	18.5%	22.1%	22.9%
White	28.9%	17.8%	18.8%
Other Race	4.8%	3.7%	4.6%
Missing Race	17.4%	24.3%	26.5%
First Gen	42.7%	54.4%	52.5%
Not First Gen	48.4%	20.6%	20.8%
Missing Parent Ed	9.0%	25.0%	26.8%
Low Income	37.1%	45.8%	43.8%
Middle Income	58.2%	54.2%	56.2%
High Income	4.7%	0.0%	0.0%
Maximum SAT or ACT Score (Percentile)	95.5	92.0	90.5
Number of Advisor Interactions	7.43	10.33	10.07
N	27,599	5,775	856

Figure 1: Relationship between Advisor ratings and Algorithm predicted scores



Notes: the x-axis denotes an advisor's assessment of the likelihood a student will enroll at a CollegePoint school, with 1 being the lowest and 10 being the highest. The y-axis denotes the logistic model's predicted score, rescaled from 0-1 to 1-10 such that (0.0, 0.1) = 1; [0.1,0.2) = 2; ... [0.9, 1) = 10. The size of each circle is directly proportional to the number of students within that rating-by-predicted score cell.

Figure 2: Comparing accuracy of advisor ratings and algorithm's predicted scores



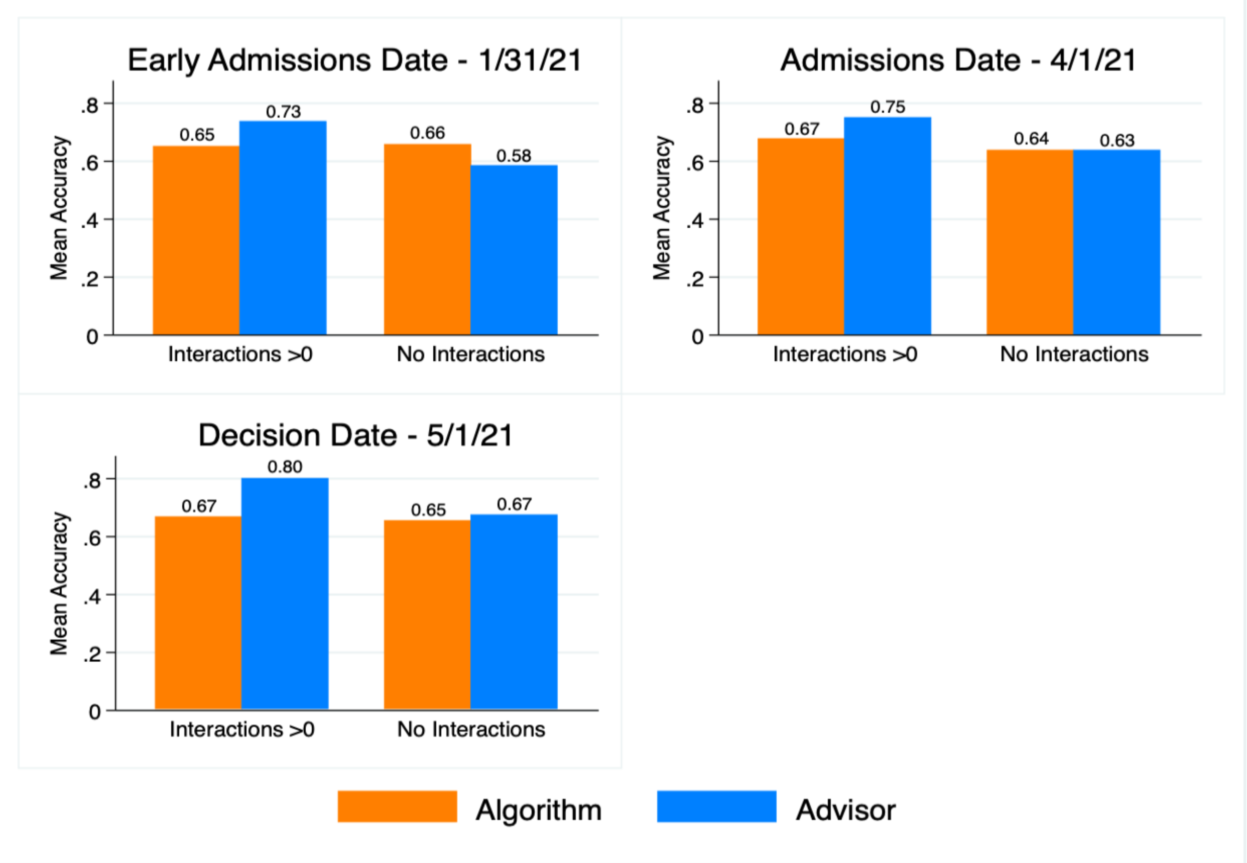
Notes: the x-axis denotes either an advisor's assessment of the likelihood a student will enroll at a CollegePoint school (with 1 being the lowest and 10 being the highest) or the algorithm's predicted score, rescaled from 0-1 to 1-10 such that (0.0, 0.1) = 1; [0.1,0.2) = 2; ... [0.9, 1) = 10. The y-axis denotes the share of students with a particular rating or scaled predicted score for whom the advisor or algorithm were correct in their prediction. The size of each circle (algorithm) or triangle (advisor) is directly proportional to the number of students within that rating or predicted score cell.

Table 2: Comparing Advisor and Algorithm Accuracy, by Frequency of Advisor-Student Interactions

	Advisor Accuracy	Algorithm Accuracy	P-value (1) = (2)
	(1)	(2)	(3)
Overall	0.688	0.655	0.11
Quintile of interactions			
1st	0.598	0.685	0.04
2nd	0.685	0.623	0.28
3rd	0.653	0.699	0.286
4th	0.767	0.599	<0.01
5th	0.764	0.634	<0.01
Bottom 3 quintiles	0.635	0.676	0.12
Top 2 quintiles	0.765	0.616	<0.01

Notes: Accuracy is equal to the share of students for whom the advisor or algorithm made a create prediction. The 1st (bottom) quintile corresponds to the 20% of students with the fewest interactions with their advisors. The p-values are from a paired t-test that advisor accuracy is equal to algorithm accuracy within each grouping.

Figure 3: Comparing advisor and algorithm accuracy, by timing of last advisor-student interaction



Notes: Accuracy is equal to the share of students for whom the advisor or algorithm made a create prediction. Within each plot, the "Interactions >0" bars include students who had at least one interaction with their advisor on or after the relevant date (e.g. January 31st), while the "No Interactions" bars include students who did not have any interactions with their advisor on or after that date.

Table 3: Comparing advisor and algorithm accuracy by student demographic characteristics

	Advisor Accuracy	Algorithm Accuracy	P-value: Advisor = Algorithm
	(1)	(2)	(3)
Overall	0.688	0.653	*0.11*
URM	0.687	0.68	*0.85*
Non-URM	0.678	0.61	*0.03*
P-value: URM = non-URM	*0.82*	*0.07*	
Low Income	0.699	0.627	*0.01*
Middle Income	0.674	0.674	*1*
P-value: Low income = Middle Income	*0.43*	*0.15*	
First Gen	0.684	0.657	*0.34*
Not First Gen	0.691	0.601	*0.2*
P-value: First Gen = Not First Gen	*0.96*	*0.8*	
Female	0.685	0.633	*0.08*
Male	0.684	0.67	*0.58*
P-value: Female = Male	*0.99*	*0.27*	

Notes: Accuracy is equal to the share of students for whom the advisor or algorithm made a create prediction. The p-values in column (3) are the result of a paried t-test of algorithm accuracy and advisor accuracy within each subgroup. The p-values presented in separate rows (e.g. P-value: URM = non-URM) represent a paired-test of advisor accuracy across demographic subgroups (column 1) or of algorithm accuracy across demographic subgroups (column 2)

CO2021 College Enrollment Predictions

* Required

Please fill out the following information separately for each of the students on your spreadsheet. We recognize that for some students you may have had limited substantive interactions. We'd still appreciate if you could provide your best assessment for the questions that follow.

1. EASE ID *

2. Student's first name *

3. Student's last name *

4. On a 1-10 scale, with 10 being most likely to attend a CollegePoint school, what is your best estimate for how likely this student is to attend a CollegePoint school this fall? *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Least Likely	○	○	○	○	○	○	○	○	○	○	Most Likely

Please indicate from 1-10, with 10 being the most influential, whether each of the following barriers might contribute to the student not enrolling at a CollegePoint school.

5. They did not apply to a broad enough range of CollegePoint schools. *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Least Influential	○	○	○	○	○	○	○	○	○	○	Most Influential

6. Their academic background may not have been competitive enough for the CollegePoint schools they applied to. *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Least Influential	○	○	○	○	○	○	○	○	○	○	Most Influential

7. The CollegePoint schools they were accepted to were too far from home. *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Least Influential	○	○	○	○	○	○	○	○	○	○	Most Influential

8. The student had concerns about whether they would “belong” or “fit in” at the CollegePoint schools they were accepted to. *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Least Influential	○	○	○	○	○	○	○	○	○	○	Most Influential

9. The student did not have enough support from her family to attend one of the CollegePoint schools she was accepted to. *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Least Influential	○	○	○	○	○	○	○	○	○	○	Most Influential

10. The student could not afford to attend one of the CollegePoints he/she was accepted to. *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Least Influential	○	○	○	○	○	○	○	○	○	○	Most Influential

11. Please list other concrete barriers that may prevent this student from attending a CollegePoint school. *

This content is neither created nor endorsed by Google.

Google Forms